STATISTICS 665
ASYMPTOTIC METHODS
IN STATISTICAL INFERENCE

Doug Wiens*

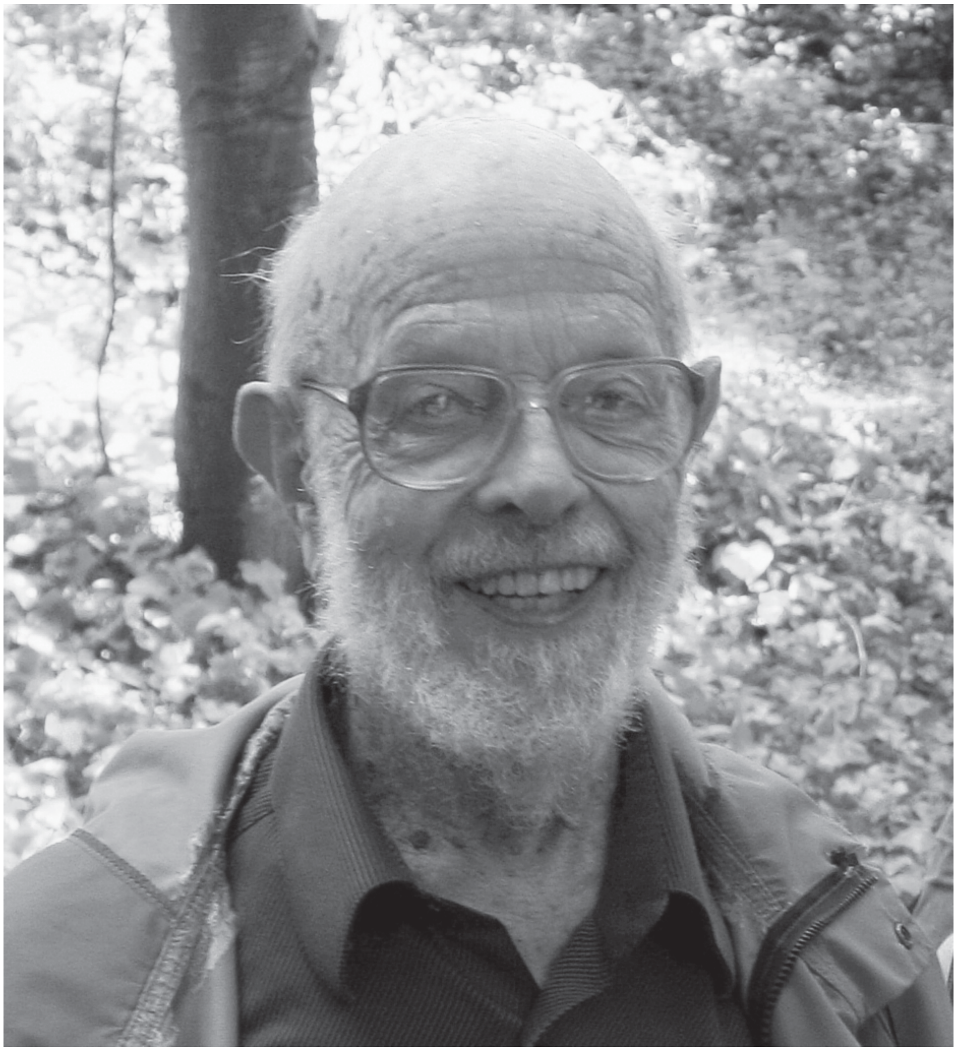
April 12, 2018

Contents

III MULTIVARIATE EXTENSIONS 111

IV NONPARAMETRIC ESTIMATION 129

V LIKELIHOOD METHODS 170



Erich Lehmann, 1917-2009. See the obituary on the course web site.

Part I

PROBABILISTIC PRELIMINARIES

1. Convergence concepts

- Large sample procedures are typically used to give tractable, approximate solutions to otherwise intractable problems. The example at the end of this lecture illustrates the fact that the 'improvement' realized by an exact solution, over an approximate one, is typically of a smaller order (asymptotically) than the underlying sampling variation.

- **Example**: We commonly base inferences about a population mean μ on the 't-statistic'
$$T_n = \frac{\sqrt{n}\,(\bar{x} - \mu)}{S}.$$
If the – independent, identically distributed ('i.i.d.') – observations are non-normal then the exact distribution of T_n is, in general, very intractable. But
$$P\left(T_n \leq x\right) \rightarrow \Phi(x) \text{ as } n \rightarrow \infty.$$

The 'rate' of convergence refers to the fact that the error in the approximation is typically of order $1/\sqrt{n}$, and a correction of the form

$$P\left(T_n \leq x\right) = \Phi(x) + \frac{1}{\sqrt{n}} A(x) + \text{smaller order terms} \tag{1.1}$$

is available (Edgeworth expansion). Question – what if the observations are not even independent?

- Sequences $\{a_n\}, \{b_n\}$.
  - They are asymptotically equivalent ($a_n \sim b_n$) as $n \to \infty$ if $a_n/b_n \to 1$
  - a_n is of a smaller order than $b_n(a_n = o\left(b_n\right))$ if $a_n/b_n \to 0$
  - a_n is of a smaller, or the same, order as $b_n(a_n = O\left(b_n\right))$ if $|a_n/b_n|$ is bounded
  - a_n and b_n are of the same order ($a_n \asymp b_n$) if $a_n = O\left(b_n\right)$ and $b_n = O\left(a_n\right)$

- e.g. $a_n = 1/n + 2/n^2 + 3/n^3$, $b_n = 1/n$, $c_n = 2/n^2 + 3/n^3$. Then:
  - $b_n = o(1)$
  - $c_n = o\left(b_n\right)$
  - $a_n/b_n = 1 + \frac{c_n}{b_n} = 1 + o(1) \rightarrow 1$; in particular $a_n = O\left(b_n\right)$
  - Also $b_n = O\left(a_n\right)$, so $a_n \asymp b_n$. In fact $a_n \sim b_n$, which is stronger.

- **Embedding sequences**: Given a statistic T_N, we can view this as a member of a possible sequence $\{T_n\}$. Of course we observe only the one member of this sequence, and so an appropriate approximation will depend on the manner in which we anticipate the limit being approached. For example suppose we observe $X_n \sim bin(n, p_n)$. If X_n is the number of electors, out of n sampled, who favour a particular political party, then we might imagine n 'large' but p remaining fixed; in this case the CLT gives

$$\frac{X_n - np}{\sqrt{np(1-p)}} \xrightarrow{L} N(0,1) \text{ as } \ n \to \infty.$$

This results in a normal approximation to the distribution of X_n and in the use of $\hat{p}_n = X_n/n$ as a consistent estimate of p (i.e. $\hat{p}_n \xrightarrow{pr} p$). On the other hand, if X_n is the frequency of a rare type of X-ray out of a large number of emissions, we might imagine 'n large and p small with the mean number of emissions constant'; a way to formalize this is that $np_n \to \lambda > 0$ as $n \to \infty$. In this case $X_n \xrightarrow{L} \mathbb{P}(\lambda)$, the Poisson distribution with mean λ. Thus we get two possible but very different limit distributions, depending on the sequence within which X_n is embedded.

- **Convergence in law**: Let $\{Z_n\}$ be a sequence of r.v.s. We say that $Z_n \xrightarrow{L} F$, or $Z_n \xrightarrow{L} Z$, where Z has d.f. F, if

  $P(Z_n \leq z) \rightarrow F(z)$ at every continuity point of F.

  An equivalent and often more convenient formulation is

  $$Z_n \xrightarrow{L} Z \Leftrightarrow E\left[g\left(Z_n\right)\right] \rightarrow E\left[g\left(Z\right)\right]$$
  whenever g is bounded and continuous.

  A consequence of this, obtained by taking $g(z) = \cos(tz)$ for fixed t, then $g(z) = \sin(tz)$, and then using $\exp(itz) = \cos(tz) + i\sin(tz)$, is the ' $\Rightarrow$' in

  $Z_n \xrightarrow{L} Z \Leftrightarrow E\left[e^{itZ_n}\right] \rightarrow E\left[e^{itZ}\right]$ for all (real) t.

- For the $bin(n,p_n)$ distribution,

$$\begin{aligned} E\left[e^{itX_n}\right] &= \sum e^{itx}\binom{n}{x}p_n^x(1-p_n)^{n-x} \\ &= \left(1+p_n\left(e^{it}-1\right)\right)^n. \end{aligned}$$

  If $np_n \to \lambda$ as $n \to \infty$ then $p_n(e^{it}-1) = \frac{\lambda(e^{it}-1)}{n} + o\left(\frac{1}{n}\right)$ and this is

$$\begin{aligned} E\left[e^{itX_n}\right] &= \left(1+\frac{\lambda(e^{it}-1)}{n}+o\left(\frac{1}{n}\right)\right)^n \\ &\to \text{exp}\left\{\lambda(e^{it}-1)\right\} \end{aligned}$$

  by Exercise 4.8 (assigned). This is the $\mathbb{P}(\lambda)$ c.f. So $P\left(X_n \le x\right) \to P\left(X \le x\right)$, where $X \sim \mathbb{P}(\lambda)$, at *continuity points* x, i.e. x *not an integer*. What if $x = m$, an integer? (These after all are the only points that matter, in this application!)

- The approach to normality for fixed p can be obtained in the same way, but it is easier to do this as a consequence of the CLT.

- **Convergence in probability:** A sequence $\{Y_n\}$ of r.v.s tends to a constant c *in probability* ($Y_n \overset{pr}{\to} c$) if

$$P\left(|Y_n - c| \geq \varepsilon\right) \to 0 \text{ for any } \varepsilon > 0.$$

If Y is a r.v., then $Y_n \overset{pr}{\to} Y$ means that $Y_n - Y \overset{pr}{\to} 0$.

  - **Chebyshev's Inequality** is a basic tool here. It states that for any c and any $a \geq 0$,

$$E\left[(Y-c)^2\right] \geq a^2 P(|Y-c| \geq a).$$

Proof: Put $Z = Y - c$. Define

$$I(A) = \begin{cases} 1, & \text{if } A \text{ occurs,} \\ 0, & \text{otherwise,} \end{cases}$$

and note that $P(A) = E\left(I(A)\right)$. Then

$$Z^2 \geq Z^2 I(|Z| \geq a) \geq a^2 I(|Z| \geq a),$$

so that

$$E[Z^2] \geq a^2 E[I(|Z| \geq a)] = a^2 P(|Z| \geq a).$$

- **Corollary 1**: If $E\left[(Y_n - c)^2\right] \to 0$ we say $Y_n \overset{q.m.}{\to} c$ (convergence in quadratic mean) and then $Y_n \overset{pr}{\to} c$.

- **Corollary 2**: **WLLN**. If $Y_n = \frac{1}{n}\sum_{i=1}^n X_i$, where the X_i are i.i.d. with mean ξ and variance $\sigma^2 < \infty$, then $E\left[(Y_n - \xi)^2\right] = \sigma^2/n \to 0$, so $Y_n \overset{pr}{\to} \xi$.

- Note $Y_n \overset{q.m.}{\to} c \Rightarrow E[Y_n] \to c$ (assigned) but $Y_n \overset{pr}{\to} c \not\Rightarrow E[Y_n] \to c$ (you should find counterexamples), hence $Y_n \overset{pr}{\to} c \not\Rightarrow Y_n \overset{q.m.}{\to} c$.

- Does $Y_n \overset{L}{\to} Y$ imply $Y_n \overset{pr}{\to} Y$? Why or why not? What if $Y = c$, a constant?

- It is worth noting that the errors introduced by using an approximate (but easy) solution rather than an exact (but difficult or impossible) one are typically of a smaller order than $n^{-1/2}$, which is the order of the usual sampling variation (e.g. the standard error of an estimate). For instance consider the (absolute) difference in widths between an exact confidence interval on a possibly non-normal mean: $CI_1 = \bar{x} \pm t_{\alpha/2} s/\sqrt{n}$ (here $t_{\alpha/2}$ is the exact quantile, which may be impossibly hard to compute) and an approximation $CI_2 = \bar{x} \pm z_{\alpha/2} s/\sqrt{n}$. This has expectation

$$\begin{aligned} & E\left[\left|width_1 - width_2\right|\right] \\ = \ & 2\left|t_{\alpha/2} - z_{\alpha/2}\right| \frac{E[S]}{\sqrt{n}} \\ = \ & 2\left\{\frac{a}{\sqrt{n}} + o\left(n^{-1/2}\right)\right\}\left\{\frac{\sigma}{\sqrt{n}} + o\left(n^{-1/2}\right)\right\} \\ = \ & O\left(n^{-1}\right); \end{aligned}$$

here a is a constant which can be computed by applying (1.1).

2. Slutsky's Theorem; consequences

- Some 'arithmetic' re convergence in probability:

  – If $A_n, B_n \overset{pr}{\rightarrow} a, b$ resp., then $A_n \pm B_n \overset{pr}{\rightarrow} a \pm b$, $A_n B_n \overset{pr}{\rightarrow} ab$, $A_n/B_n \overset{pr}{\rightarrow} a/b$ if $b \neq 0$.
  Proof of the first (you should do the others):

$$\begin{aligned} & P\left(\left|(A_n \pm B_n) - (a \pm b)\right| > \varepsilon\right) \\ = \; & P\left(\left|(A_n - a) \pm (B_n - b)\right| > \varepsilon\right) \\ \leq \; & P\left(|A_n - a| > \varepsilon/2\right) + P\left(|B_n - b| > \varepsilon/2\right) \\ \rightarrow \; & 0. \end{aligned}$$

  A consequence is that a rational function of $A_n, B_n \overset{pr}{\rightarrow}$ the corresponding rational function of a, b. (A rational function is a ratio of multinomials.)

  – If f is continuous at a, and $A_n \overset{pr}{\rightarrow} a$, then $f(A_n) \overset{pr}{\rightarrow} f(a)$ (proof is straightforward, or see Stat 512 Lecture 7 notes).

- **Order in probability**.

  - If $A_n/B_n \overset{pr}{\to} 0$ then we say that $A_n = o_P(B_n)$.

  - $A_n = O_P(B_n)$ if A_n/B_n is 'bounded in probability': For any ε there is an $M = M(\varepsilon)$ and an $N = N(\varepsilon)$ such that $n > N$ implies $P(|A_n/B_n| \leq M) \geq 1 - \varepsilon$. Equivalently, $\lim_{M,n\to\infty} P\left(|A_n/B_n| \leq M\right) = 1$. Note that $\lim_{n\to\infty} \lim_{M\to\infty} P\left(|A_n/B_n| \leq M\right) = 1$ is not enough – $A_n/B_n = Y_n \sim N(n, 1)$ satisfies this but is not $O_P(1)$. Why not?

- In particular, if $X_n \sim bin(n, p)$, then:

  - $X_n/n = n^{-1}\sum_{i=1}^n Z_i$, where $Z_i = I($'i^{th} experiment is a success'$) \sim bin(1, p)$, and so $\hat{p}_n = X_n/n \overset{pr}{\to} E[Z_1] = p$ by the WLLN. (Bernoulli's Law of Large Numbers). Thus $\hat{p}_n - p = \left(X_n - np\right)/n \overset{pr}{\to} 0$, hence is $o_P(1)$.

– By the CLT,

$$\frac{X_n - np}{\sqrt{np(1-p)}}$$

has a limit distribution, hence is $O_P(1)$ (proven below).

- You should show: $Y_n \overset{pr}{\to} c \Rightarrow Y_n = O_P(1)$.

- **Proof** that convergence in law implies bounded in probability: Suppose $Z_n \overset{L}{\to} F$. We are to find M such that $P(|Z_n| \leq M) \geq 1-\varepsilon$ for all sufficiently large n. Let F_n be the d.f. of Z_n. Suppose we can choose numbers M, n_0 such that:

  1. $\pm M$ are continuity points of F;

  2. $n > n_0 \Rightarrow |F_n(\pm M) - F(\pm M)| < \varepsilon/4$ (possible since $F_n(\pm M) \to F(\pm M)$);

  3. $F(M) - F(-M) \geq 1 - \frac{\varepsilon}{2}$.

Then for $n > n_0$,

$$\begin{aligned} & P(|Z_n| \leq M) \geq P(-M < Z_n \leq M) \\ = \; & F_n(M) - F_n(-M) \\ = \; & [F_n(M) - F(M)] + [F(M) - F(-M)] \\ & + [F(-M) - F_n(-M)] \\ \geq \; & -\frac{\varepsilon}{4} + \left(1 - \frac{\varepsilon}{2}\right) - \frac{\varepsilon}{4} = 1 - \varepsilon. \end{aligned}$$

Are these choices possible? Yes, if F has enough continuity points; in particular if they can be arbitrarily large in absolute value. But the discontinuity points of F are $\cup_{j=1}^{\infty}\{x|F(x) - F(x-) \geq \frac{1}{j}\}$; this is a countable union of finite sets (why?), hence is countable. Thus every interval contains infinitely many continuity points; in particular they may be chosen arbitrarily large or small.

- **Slutsky's Theorem**: If $Y_n \xrightarrow{L} Y$ and $A_n, B_n \xrightarrow{pr} a, b$ then $A_nY_n + B_n \xrightarrow{L} aY + b$.
  A proof, valid for multivariate r.v.s, will be outlined in Lecture 13. It relies on the special case of this univariate version with $A_n, a = 1$; this is assigned. See also the classic (1946) text *Mathematical Methods in Statistics*, by Harald Cramér.

- **Corollary 1**: If $Y_n \overset{pr}{\to} Y$ then $Y_n \overset{L}{\to} Y$. **Proof**: $Y_n = (Y_n - Y) + Y = B_n + Y$, where $B_n \overset{pr}{\to} 0$.

- **Corollary 2**: If $Y_n \overset{L}{\to} Y \sim F$, and $X_n \overset{pr}{\to} x$, where x (finite) is a continuity point of F, then $P(Y_n \leq X_n) \to F(x)$.
  Proof: $Y_n - X_n + x \overset{L}{\to} Y$, so $P(Y_n \leq X_n) = P(Y_n - X_n + x \leq x) \to F(x)$.

- **Corollary 3** (A special case of Corollary 2):
  If $Y_n \overset{L}{\to} Y$ with d.f.s F_n, F, and $x_n \to x$, where x (finite) is a continuity point of F, then $F_n(x_n) \to F(x)$.
  Question: What if $x = \pm\infty$ in Corollary 3?

- **Central Limit Theorem**: Let $\{X_j\}_{j=1}^n$ be a sample (i.e. i.i.d., but *not necessarily normal*), with mean μ_X and finite standard deviation $\sigma_X > 0$. Put

$$\begin{aligned} S_n &= \sum_{j=1}^n X_j, \\ T_n &= \frac{S_n - E\left[S_n\right]}{\sqrt{VAR\left[S_n\right]}} = \frac{\sqrt{n}\left(\bar{X} - \mu_X\right)}{\sigma_X}. \end{aligned}$$

The CLT states that $T_n \xrightarrow{L} \Phi$.

- You should be familiar with the derivation of this CLT which is based on an expansion of the moment generating function or characteristic function – if not, see e.g. Stat 512 Lecture 15. We will prove a refined version of this CLT in the next class.

- First an application. Let S^2 be the sample variance. Note
$$S^2 = \frac{n}{n-1}\left[\frac{\sum X_j^2}{n} - \bar{X}^2\right],$$
which by the WLLN and the 'arithmetic' above $\xrightarrow{pr} E[X^2] - E[X]^2 = \sigma_X^2$. Thus the 't-statistic' used to make inferences about μ without requiring knowledge of σ is
$$t_n = \frac{\sqrt{n}\left(\bar{X} - \mu\right)}{S} = T_n \cdot \frac{\sigma}{S} \xrightarrow{L} \Phi,$$
by Slutsky's Theorem, since $\frac{\sigma}{S} = \sqrt{\frac{\sigma_X^2}{S^2}} \xrightarrow{pr} 1$. (Why?)

3. Berry-Esséen Theorem; Edgeworth expansions

- The previously stated version of the CLT was for i.i.d. r.v.s., and asserts that the d.f. of the normalized average of such r.v.s converges in law to the Normal d.f. A strengthening is the **Berry-Esséen theorem**, which applies to 'triangular arrays which are i.i.d. within rows':

$$\begin{array}{lllll} X_{1,1} & & & & \sim F_1 \\ X_{1,2} & X_{2,2} & & & \sim F_2 \\ & & \ddots & & \vdots \\ X_{1,n} & X_{2,n} & \cdots & X_{n,n} & \sim F_n \\ & & \vdots & & \vdots \end{array}$$

This theorem asserts that if the r.v.s $X_{j,n}$ are i.i.d. for $j = 1, ..., n$, with means ξ_n, variances σ_n^2 and normalized third absolute moments

$$\gamma_n \overset{def}{=} E\left[\left|\frac{X_{j,n} - \xi_n}{\sigma_n}\right|^3\right],$$

and if

$$Z_n = \frac{S_n - E[S_n]}{\sqrt{VAR[S_n]}} = \frac{\sum_{j=1}^n \left(X_{j,n} - \xi_n\right)}{\sqrt{n}\sigma_n},$$

with d.f. $G_n(z)$, then there is a universal constant C such that

$$\sup_z |G_n(z) - \Phi(z)| \leq \frac{C}{\sqrt{n}}\gamma_n.$$

(They gave $C = 3$; it is now (since 2011) known that $C \in [.4097, .4784]$.)
Thus if $\gamma_n = o(\sqrt{n})$, we have $Z_n \xrightarrow{L} \Phi$.
(In fact $G_n(z) \rightrightarrows \Phi(z)$.)

- An example is if each $X_{j,n}$ is $bin(1, p_n)$, so that their sum $S_n = \sum_{j=1}^{n} X_{j,n}$ is $bin(n, p_n)$. Then $\xi_n = p_n$, $\sigma_n = \sqrt{p_n(1-p_n)}$ and $Z_n = \frac{S_n - np_n}{\sqrt{np_n(1-p_n)}}$. Suppose that $p_n \to p \in (0,1)$. We have

$$\begin{aligned} \gamma_n &= \left|\frac{1-\xi_n}{\sigma_n}\right|^3 p_n + \left|\frac{0-\xi_n}{\sigma_n}\right|^3 (1-p_n) \\ &= O(1) = o(\sqrt{n}), \end{aligned}$$

so $Z_n \xrightarrow{L} \Phi$.

- **Related example**: Let $\{X_j\}_{j=1}^n$ be a sample from a d.f. F with density f. We wish to estimate the p^{th} quantile $\xi_p = F^{-1}(p)$ $(p \neq 0, 1)$. Assume $f(\xi_p) > 0$. In particular F is continuous at ξ_p and so $F(\xi_p) = p$. If $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ are the order statistics, then a possible estimate of ξ_p is $X_{(k_n)}$, where $X_{(k_n)}$ has $\approx np$ observations smaller than or equal to it. Formally, we require
$$\frac{k_n}{n} = p + o(n^{-1/2}).$$
This is satisfied if $k_n = [np]$, since then $|k_n - np| \leq 1$. Put
$$T_n = \sqrt{n}\left(X_{(k_n)} - \xi_p\right).$$
We have
$$P\left(T_n \leq z\right) = P\left(X_{(k_n)} \leq \xi_p + \frac{z}{\sqrt{n}}\right)$$
$$\begin{aligned} &= P\begin{pmatrix} \text{the number of obsn's} \\ > \xi_p + \frac{z}{\sqrt{n}} \text{ is } \leq n - k_n \end{pmatrix} \\ &= P\left(S_n \leq n - k_n\right), \end{aligned}$$
where $S_n \sim bin\left(n, q_n = 1 - F\left(\xi_p + \frac{z}{\sqrt{n}}\right)\right)$. Note that, by the Mean Value Theorem (details in the

next lecture),

$$\begin{aligned} q_n &= 1 - \left[F(\xi_p) + f(\xi_p)\frac{z}{\sqrt{n}} + o(n^{-1/2})\right] \\ &= 1 - p - f(\xi_p)\frac{z}{\sqrt{n}} + o(n^{-1/2}). \end{aligned}$$

In particular $q_n \to 1 - p \neq 0, 1$, so the Berry-Esséen theorem holds. Then

$$\begin{aligned} P\left(T_n \leq z\right) &= P\left(\frac{S_n - nq_n}{\sqrt{nq_n(1-q_n)}} \leq \frac{n - k_n - nq_n}{\sqrt{nq_n(1-q_n)}}\right) \\ &= G_n\left(z_n = \frac{n - k_n - nq_n}{\sqrt{nq_n(1-q_n)}}\right) \\ &\to \Phi\left(\lim z_n\right). \text{ (Why?)} \end{aligned}$$

Now

$$\begin{aligned} z_n &= \frac{\sqrt{n}\left(1-\frac{k_n}{n}-q_n\right)}{\sqrt{q_n(1-q_n)}} \\ &= \frac{\sqrt{n}\left(\begin{array}{c} 1-p+o(n^{-1/2}) \\ -\left(1-p-f(\xi_p)\frac{z}{\sqrt{n}}+o(n^{-1/2})\right) \end{array}\right)}{\sqrt{q_n(1-q_n)}} \\ &\to \frac{f(\xi_p)z}{\sqrt{(1-p)p}} = \frac{z}{\sigma},\ \text{say.} \end{aligned}$$

Thus

$$T_n \xrightarrow{L} N\left(0, \sigma^2 = \frac{p(1-p)}{f^2(\xi_p)}\right).$$

In particular, for the median $\left(m = \xi_{1/2}\right)$ one has

$$\sqrt{n}\,(\hat{m}-m) \to N\left(0, \frac{1}{4f^2(m)}\right),$$

where $\hat{m}$ is any of the usual order statistics used to estimate m (or their averages – why?).

- **CLT via the Edgeworth expansion**. Let X be a r.v. with c.f. $\psi(t) = E\left[e^{itX}\right]$. (If X is $N(0,1)$ we will write $\xi(t) = e^{-t^2/2}$ for the c.f.) The *cumulants* β_r are defined by

$$\begin{aligned} \log\psi(t) &= \sum_{r=1}^{\infty} \frac{\beta_r}{r!}(it)^r; \text{ equivalently} \\ \beta_r &= (-i)^r \frac{d^r}{dt^r} \log\psi(t)_{|t=0}. \end{aligned}$$

In particular

$$\begin{aligned} \beta_1 &= E[X], \\ \beta_2 &= VAR[X], \\ \beta_3 &= E[(X-\mu_X)^3] \text{ (`coefficient of skewness')}, \\ \beta_4 &= E[(X-\mu_X)^4] - 3\sigma_X^4 \text{ (`coefficient of excess')}. \end{aligned}$$

For the $N(0,1)$ d.f., $\log\xi(t) = -t^2/2$, so the second cumulant $= 1$ and all others vanish. We then have

$$\begin{aligned} \log\frac{\psi(t)}{\xi(t)} &= \log\psi(t) - \log\xi(t) \\ &= \sum_{r=1}^{\infty} \frac{[\beta_r - I(r=2)]}{r!}(it)^r, \end{aligned}$$

hence

$$\psi(t) = \tag{3.1}$$
$$\left[\exp\left\{\sum_{r=1}^{\infty} \frac{[\beta_r - I(r=2)]}{r!}(it)^r\right\}\right] \cdot \xi(t).$$

Now let $\{X_j\}_{j=1}^n$ be a sample (i.e., i.i.d.), and put

$$Y_j = \frac{X_j - \mu_X}{\sigma_X},\ T_n = \frac{1}{\sqrt{n}}\sum_{j=1}^n Y_j.$$

The cumulants of Y are

$$\lambda_1 = 0,\ \lambda_2 = 1,\ \lambda_3 = E[Y^3], \dots\ .$$

Now the c.f. of T_n is

$$\psi_{T_n}(t) = E\left[e^{it\frac{1}{\sqrt{n}}\sum_{j=1}^n Y_j}\right] = \psi_Y^n\left(\frac{t}{\sqrt{n}}\right),$$

(why?) with ‘cumulant generating function’ (c.g.f.)

$$\begin{aligned}\log \psi_{T_n}(t) &= n \log \psi_Y \left(\frac{t}{\sqrt{n}}\right) \\ &= n \sum_{r=1}^{\infty} \frac{\lambda_r}{r!} \left(i\frac{t}{\sqrt{n}}\right)^r \\ &= \sum_{r=1}^{\infty} \frac{\lambda_r n^{-\left(\frac{r}{2}-1\right)}}{r!}(it)^r.\end{aligned}$$

Thus T_n has cumulants $\beta_r = \lambda_r n^{-\left(\frac{r}{2}-1\right)}$ ($= 0$ if $r = 1$ and $= 1$ if $r = 2$). Then (3.1) becomes

$$\psi_{T_n}(t) = \left[\exp\left\{\sum_{r=3}^{\infty} \frac{\lambda_r n^{-\left(\frac{r}{2}-1\right)}}{r!}(it)^r\right\}\right] \cdot \xi(t)$$

$$= \left[\begin{array}{c} 1 + \frac{\lambda_3}{6\sqrt{n}}(it)^3 + \frac{\lambda_4}{24n}(it)^4 \\ + \frac{\lambda_3^2}{72n}(it)^6 + o\left(n^{-1}\right) \end{array}\right] \cdot \xi(t). \qquad (3.2)$$

This gives the standard CLT: $\psi_{T_n}(t) \rightarrow \xi(t)$, so $T_n \xrightarrow{L} \Phi$.

4. Delta method; Liapounov's Theorem

- **Hermite polynomials**: Let $\phi = \Phi'$ be the $N(0,1)$ density, note $\phi'(z) = -z\phi(z)$; continuing we obtain

$$\phi^{(k)}(z) = (-1)^k H_k(z)\phi(z), \tag{4.1}$$

where $H_k(z)$ is the Hermite polynomial:

$H_1(z) = z,\ H_2(z) = z^2 - 1,\ H_3(z) = z^3 - 3z, \ldots$.

Differentiating both sides of (4.1) gives

$$H_{k+1}(z) = zH_k(z) - H_k'(z).$$

Note that (integrating by parts repeatedly)

$$\begin{aligned}
& \int_{-\infty}^{\infty} e^{itz}(-1)^k H_k(z)\phi(z)dz \\
= \ & \int_{-\infty}^{\infty} e^{itz}\phi^{(k)}(z)dz = (-it)\int_{-\infty}^{\infty} e^{itz}\phi^{(k-1)}(z)dz \\
= \ & \ldots = (-it)^k \int_{-\infty}^{\infty} e^{itz}\phi(z)dz \\
= \ & (-it)^k \xi(t).
\end{aligned}$$

Thus in (3.2), $(it)^k\xi(t)$ is the c.f. of $H_k(z)\phi(z)$, i.e.

$$\psi_{T_n}(t)=\int_{-\infty}^{\infty}e^{itz}\left[\begin{array}{c}1+\frac{\lambda_3}{6\sqrt{n}}H_3(z)\\ +\frac{3\lambda_4H_4(z)+\lambda_3^2H_6(z)}{72n}\end{array}\right]\phi(z)dz+o\left(n^{-1}\right).$$

- By uniqueness of c.f.s, an expansion for the density of T_n is

$$\begin{aligned} f_n(z) &= \phi(z)+\frac{\lambda_3}{6\sqrt{n}}H_3(z)\phi(z) \\ &\quad +\frac{3\lambda_4H_4(z)+\lambda_3^2H_6(z)}{72n}\phi(z)+o\left(n^{-1}\right) \end{aligned}$$

  and an expansion for the d.f. is, since $H_k(z)\phi(z)=\left[(-1)^k\,\phi^{(k-1)}(z)\right]'=\left[-H_{k-1}(z)\phi(z)\right]'$,

$$\begin{aligned} F_n(z) &= \Phi(z)-\frac{\lambda_3}{6\sqrt{n}}H_2(z)\phi(z) \\ &\quad -\frac{3\lambda_4H_3(z)+\lambda_3^2H_5(z)}{72n}\phi(z)+o\left(n^{-1}\right). \end{aligned}$$

  The first term gives the classical CLT for normalized averages of i.i.d.s: $T_n\overset{L}{\to}\Phi$. The error is $O\left(n^{-1/2}\right)$; it is $O\left(n^{-1}\right)$ if the X_j are symmetrically distributed ($\lambda_3=0$).

- The above is for continuous r.v.s; it can be shown to hold (to the order $n^{-1/2}$) for integer-valued r.v.s as well, with the continuity correction – see the discussion following Theorem 2.4.3 in the text.

- **Taylor's theorem**. You should read in text; we typically need only the special case of the Mean Value Theorem: If f is differentiable at x, then $f(x+c) = f(x) + f'(x)c + o(c)$ as $c \to 0$.
  (**Proof**: ...)
  Typically $c = O(n^{-1/2})$ or $O(n^{-1})$.
  We also have

  $$f(x+c) = f(x) + f'(\xi)c,$$

  for some point $\xi = \xi(c, x)$ between x and $x + c$ (see Stat 312 Lecture 15 for a proof).

- **Delta method**. Suppose that X_n is $AN(\theta, \frac{\sigma^2}{n})$, i.e. $\sqrt{n}\,(X_n - \theta) \overset{L}{\rightarrow} Z \sim N(0, \sigma^2)$, and that $f'(\theta)$ exists and is $\neq 0$. Define R_n by

$$f\,(X_n) - f(\theta) = (X_n - \theta)\, f'(\theta) + R_n.$$

We claim that $R_n = o_P(X_n - \theta)$ (if the X_n are constants then this is the MVT), so $\sqrt{n}R_n = o_P(\sqrt{n}\,(X_n - \theta)) = o_P(O_P(1)) = o_P(1)$ (assigned). This gives

$$\begin{aligned}\sqrt{n}\,(f\,(X_n) - f(\theta)) &= \sqrt{n}\,(X_n - \theta)\, f'(\theta) + o_P(1) \\ &\overset{L}{\rightarrow} N(0, \left[\sigma f'(\theta)\right]^2),\end{aligned}$$

by Slutsky.

Proof of claim: $\frac{R_n}{X_n - \theta} = \frac{f(X_n) - f(\theta)}{X_n - \theta} - f'(\theta) = h(X_n)$, say. Define $h(\theta) = 0$, so that h is continuous at θ. Now $X_n \overset{pr}{\rightarrow} \theta$ (why?), so $h(X_n) \overset{pr}{\rightarrow} h(\theta) = 0$. □

By Slutsky's theorem, we also have

$$\frac{\sqrt{n}\,(f\,(X_n) - f(\theta))}{s_n f'(\hat{\theta}_n)} \overset{L}{\rightarrow} N(0, 1)$$

as long as $s_n \overset{pr}{\rightarrow} \sigma$, $\hat{\theta}_n \overset{pr}{\rightarrow} \theta$, and f' is continuous at θ. (We could use $\hat{\theta}_n = X_n$.)

- **Example**: Let S_n be $bin(n,p)$, so that $X_n = S_n/n = \hat{p}$ is $AN(p, p(1-p)/n)$. This is inconvenient for making inferences about p. To get confidence intervals on p we can instead use the fact (which you should now be able to show) that

$$\frac{\sqrt{n}\,(X_n - p)}{\sqrt{\hat{p}(1-\hat{p})}} \xrightarrow{L} N(0,1),$$

leading to CIs '$X_n \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$'. A more accurate method is to use a **variance stabilizing transformation**. We choose $f(\cdot)$ so that '$\sigma f'(\theta)$'$=\sqrt{p(1-p)}f'(p)$ is independent of p:

$$f'(p) \propto \frac{1}{\sqrt{p(1-p)}} \Rightarrow f(p) \propto \text{arcsin}\,\sqrt{p}.$$

Since $\left(\text{arcsin}\,\sqrt{p}\right)' = \frac{1}{2\sqrt{p(1-p)}}$ we have

$$\text{arcsin}\,\sqrt{\hat{p}} \sim AN(\text{arcsin}\,\sqrt{p}, \frac{1}{4n}).$$

From this we get CIs on arcsin $\sqrt{p}$, and transform them to get CIs on p which are typically more accurate than those above.

- Uniformity – read §2.6; note in particular Polya's Theorem.

- **CLT for non i.i.d. r.v.s**.
  There are important applications requiring a CLT for r.v.s which are independent, but not identically distributed. Regression is an example – we end up working with terms like $\sum x_i Y_i$, where the Y_i may be equally varied but the $x_i Y_i$ are not. Suppose then that $\{X_{j,n}\}_{j=1}^{n}$ are independent r.v.s with d.f.s $F_{j,n}$, means $\xi_{j,n}$ and variances $\sigma_{j,n}^2$. (Triangular array; independence within rows.) Put $S_n = \sum_{j=1}^{n} X_{j,n}$, with mean $\xi_n = \sum_{j=1}^{n} \xi_{j,n}$ and variance $\sigma_n^2 = \sum_{j=1}^{n} \sigma_{j,n}^2$. Then **Liapounov's theorem** states that

  $$\frac{S_n - \xi_n}{\sigma_n} \xrightarrow{L} \Phi,$$

  provided

  $$\frac{1}{\sigma_n^3} \sum_{j=1}^{n} E\left[\left|X_{j,n} - \xi_{j,n}\right|^3\right] \to 0.$$

  (Check that this becomes '$\gamma_n/\sqrt{n} \to 0$' if the $X_{j,n}$ are i.i.d., as in the Berry-Esséen Theorem.)

- **Lemma**: $\{Y_j\}_{j=1}^n$ independent, with zero mean, variance σ^2, common $E\left[\left|Y^3\right|\right]$. Consider a linear combination $S_n = \sum w_{j,n} Y_j$ with $\sum w_{j,n}^2 = 1$. Then Liapounov's Theorem applies, and yields $\frac{S_n}{\sigma} \xrightarrow{L} \Phi$, if

$$\sum_{j=1}^n |w_{j,n}|^3 \to 0. \tag{4.2}$$

Equivalently,

$$w_n \overset{def}{=} \max_{1 \le j \le n} |w_{j,n}| \to 0. \tag{4.3}$$

Proof: (This is essentially Theorems 2.7.3, 2.7.4 of the text.) In the notation above $X_{j,n} = w_{j,n} Y_j$,

$$\sigma_n^2 = \sum_{j=1}^n \sigma_{j,n}^2 = \sum_{j=1}^n w_{j,n}^2 \sigma^2 = \sigma^2,$$

and so $S_n/\sigma \xrightarrow{L} \Phi$ as long as

$$\frac{1}{\sigma_n^3} \sum_{j=1}^n E\left[\left|X_{j,n}\right|^3\right] = \frac{E\left[|Y|^3\right]}{\sigma^3} \sum_{j=1}^n |w_{j,n}|^3 \to 0.$$

To see that (4.2) and (4.3) are equivalent, note that

$$\sum_{j=1}^{n} |w_{j,n}|^3 = \sum_{j=1}^{n} |w_{j,n}|^2 |w_{j,n}| \ \leq w_n \sum_{j=1}^{n} |w_{j,n}|^2 = w_n,$$

and that

$$w_n^3 = \left(\max_{1 \leq j \leq n} |w_{j,n}| \right)^3 = \left(\max_{1 \leq j \leq n} |w_{j,n}|^3 \right) \leq \sum_{j=1}^{n} |w_{j,n}|^3,$$

so that

$$0 \leq w_n^3 \leq \sum_{j=1}^{n} |w_{j,n}|^3 \leq w_n;$$

thus one → 0 iff the other one does. □

- We apply this to simple linear regression: $Y_i = \alpha + \beta x_i + \varepsilon_i$ with the usual assumptions (but NOT assuming normal errors), so Y_i has mean $\alpha + \beta x_i$ and variance σ^2. The LSEs are $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$ and

$$\begin{aligned}\hat{\beta} &= \frac{\sum (x_i - \bar{x}) Y_i}{\sum (x_i - \bar{x})^2} \\ &= \frac{1}{\sqrt{\sum (x_i - \bar{x})^2}} \sum w_{i,n} Y_i \\ &= \frac{1}{\sqrt{\sum (x_i - \bar{x})^2}} \sum w_{i,n} \left[\alpha + \beta x_i + \varepsilon_i\right],\end{aligned}$$

with $w_{i,n} = (x_i - \bar{x}) / \sqrt{\sum (x_i - \bar{x})^2}$. Since $\sum w_{i,n} = 0$ and $\sum w_{i,n} x_i = \sum w_{i,n} (x_i - \bar{x}) = \sqrt{\sum (x_i - \bar{x})^2}$, we have that

$$\hat{\beta} = \beta + \frac{\sum w_{i,n} \varepsilon_i}{\sqrt{\sum (x_i - \bar{x})^2}},$$

hence

$$\sqrt{\sum (x_i - \bar{x})^2} \left(\hat{\beta} - \beta\right) = \sum w_{i,n} \varepsilon_i \xrightarrow{L} N(0, \sigma^2)$$

as long as $\max_{1 \le i \le n} |x_i - \bar{x}|^2 = o\left(\sum (x_i - \bar{x})^2\right)$.

- Under this condition $\hat{\beta}$ is $AN(\beta, \sigma^2/S_{XX})$ for $S_{XX} = \sum (x_i - \bar{x})^2$. For Normal errors, $\hat{\beta}$ is $N(\beta, \sigma^2/S_{XX})$ exactly.

- For simple linear regression, in terms of the 'hat' matrix
$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$$
we have that $h_{ii} = n^{-1} + w_{i,n}^2$. In general linear regression models, asymptotic normality of the estimates holds if $\max h_{ii} \to 0$ (Yohai & Maronna).

- Read §2.8 on CLT for dependent r.v.s, in particular Theorems 2.8.1 and 2.8.2.

Part II

LARGE-SAMPLE INFERENCE

5. Introduction to asymptotic tests

- **General testing framework**: We observe $\mathbf{X} = (X_1, ..., X_n)$ where the distribution of $\mathbf{X}$ depends on a (univariate) parameter θ. We test $H : \theta = \theta_0$ vs. $K : \theta > \theta_0$ by rejecting H if a 'test statistic' $T_n = T(\mathbf{X})$ is too large, say $T_n > C_n$, the 'critical value' defining the 'rejection region'. For a specified 'level'

$$\alpha = P(\text{Type I error}) = P(\text{reject } H|H \text{ true})$$

we wish to attain α asymptotically:

$$P_{\theta_0}(T_n > C_n) = \alpha + o(1).$$

- Suppose that, <u>under H</u>, T_n is an *asymptotically normal* (AN) estimate of θ_0:

$$\sqrt{n}\left(T_n - \theta_0\right) \xrightarrow{L} N\left(0, \tau^2\left(\theta_0\right)\right).$$

(In general, T_n is $AN\left(\mu_n, \tau_n^2\right)$ if $\frac{T_n - \mu_n}{\tau_n} \xrightarrow{L} \Phi$.) Then

$$P_{\theta_0}(T_n > C_n) \to 1 - \Phi\left(\lim \frac{\sqrt{n}\left(C_n - \theta_0\right)}{\tau\left(\theta_0\right)}\right).$$

(Why?) This asymptotic level $= \alpha$ for

$$\lim \frac{\sqrt{n}\left(C_n - \theta_0\right)}{\tau\left(\theta_0\right)} = \Phi^{-1}\left(1 - \alpha\right) \overset{def}{=} u_\alpha.$$

Thus we require

$$C_n = \theta_0 + \frac{\tau\left(\theta_0\right) u_\alpha}{\sqrt{n}} + o\left(n^{-1/2}\right).$$

- We often work instead with the observed value $t_{obs} = T(\mathbf{x})$ of T_n and then calculate the **p-value**:

$$\begin{aligned}\hat{\alpha}(t_{obs}) &= P_{\theta_0}(T_n > t_{obs}) \\ &= 1 - \Phi\left(\frac{\sqrt{n}\left(t_{obs} - \theta_0\right)}{\tau\left(\theta_0\right)}\right) + o(1),\end{aligned}$$

  if T_n is $AN\left(\theta_0, \tau^2\left(\theta_0\right)\right)$. [The error is *uniformly* (in t_{obs}) $o(1)$ by Theorem 2.6.1 - how?]

- **Studentization**: The above derivation holds with $\tau\left(\theta_0\right)$ replaced by any consistent estimate. More generally, suppose that

$$\sqrt{n}\left(T_n - \theta_0\right) \xrightarrow{L} N\left(0, \tau^2\left(\theta_0, \phi\right)\right)$$

  for a (possibly vector-valued) 'nuisance parameter' ϕ. E.g. $\sqrt{n}\left(\bar{X} - \mu_0\right) \xrightarrow{L} N\left(0, \sigma^2\right)$ and σ^2 is a nuisance parameter if interest is on testing for μ. Suppose that, <u>when H is true</u>, $\hat{\tau}_n \xrightarrow{pr} \tau\left(\theta_0, \phi\right)$. Then Slutsky's Theorem yields the critical point

$$C_n = \theta_0 + \frac{\hat{\tau}_n u_\alpha}{\sqrt{n}} + o_P\left(n^{-1/2}\right)$$

(Corollary 2 p. 21) and the p-value

$$\hat{\alpha}(t_{obs}) = 1 - \Phi\left(\frac{\sqrt{n}\,(t_{obs} - \theta_0)}{\hat{\tau}_n}\right) + o_P(1).$$

In the problem of testing for μ, $\hat{\tau}_n = S$ yields the 't' statistic.

- The same approach holds for non-normal limits. **Example**: $X_1, ..., X_n \overset{i.i.d.}{\sim} U(0,\theta)$, the uniform distribution on $[0,\theta]$. The natural estimate of θ is (a multiple of) $X_{(n)}$, and we reject H, at asymptotic level α, for $X_{(n)} > C_n$ with C_n determined by

$$P_{\theta_0}\left(X_{(n)} > C_n\right) = 1 - P_{\theta_0}(\text{ all } X_i \leq C_n) = 1 - \left(\frac{C_n}{\theta_0}\right)^n.$$

Represent C_n as $\theta_0\left(1 - \frac{t}{n}\right)$ for some t, then

$$1 - \left(\frac{C_n}{\theta_0}\right)^n = 1 - \left(1 - \frac{t}{n}\right)^n \to 1 - e^{-t},$$

and so the limiting level is α if $t = -\log(1-\alpha)$ and $C_n = \theta_0\left(1 + \frac{\log(1-\alpha)}{n}\right)$. (In this problem $C_n = \theta_0 (1-\alpha)^{1/n}$ gives α exactly.)

- **Two-sample problems**. Suppose $\mathbf{X} = (X_1, ..., X_m)$ and $\mathbf{Y} = (Y_1, ..., Y_n)$ are independent, that $U_m = U(\mathbf{X})$ and $V_n = V(\mathbf{Y})$ are AN (marginally and hence jointly, using their independence – exercise):

$$\sqrt{m}\,(U_m - \xi) \xrightarrow{L} N\left(0, \sigma^2\right),$$
$$\sqrt{n}\,(V_n - \eta) \xrightarrow{L} N\left(0, \tau^2\right),$$

and that with $N = m + n$,

$$\frac{m}{N} \to \rho > 0,\ \ \frac{n}{N} \to 1 - \rho > 0.$$

By Slutsky + independence,

$$\sqrt{N}\left(U_m-\xi\right)\xrightarrow{L}N\left(0,\sigma^2/\rho\right) \text{ and}$$

$$\sqrt{N}\left(V_n-\eta\right)\xrightarrow{L}N\left(0,\tau^2/(1-\rho)\right), \text{ hence}$$

$$\sqrt{N}\left((U_m-V_n)-(\xi-\eta)\right)\xrightarrow{L}N\left(0,\frac{\sigma^2}{\rho}+\frac{\tau^2}{1-\rho}\right)$$

$$\text{and } \frac{(U_m-V_n)-(\xi-\eta)}{\sqrt{\frac{\sigma^2}{N\rho}+\frac{\tau^2}{N(1-\rho)}}}\xrightarrow{L}N\left(0,1\right);$$

$$\text{also } \frac{(U_m-V_n)-(\xi-\eta)}{\sqrt{\frac{\sigma^2}{m}+\frac{\tau^2}{n}}}\xrightarrow{L}N\left(0,1\right).$$

Thus we can test $H:\xi-\eta=\Delta$ (specified) vs. $K:\xi-\eta>\Delta$ at level α by rejecting if

$$\frac{(U_m-V_n)-\Delta}{\sqrt{\frac{\sigma^2}{m}+\frac{\tau^2}{n}}}>u_\alpha. \qquad (5.1)$$

Again by Slutsky we can replace σ^2 and τ^2 by consistent (under H) estimates.

- **Example 1**: Two Normal means; Behrens-Fisher problem if $\sigma^2\neq\tau^2$; not a problem asymptotically.

- **Example 2**: $X_1, ..., X_m \overset{i.i.d.}{\sim} \mathbb{P}(\lambda)$, $Y_1, ..., Y_n \overset{i.i.d.}{\sim} \mathbb{P}(\mu)$, $U_m = \bar{X}$, $V_n = \bar{Y}$, $\sigma^2 = \lambda$, $\tau^2 = \mu$. Test with $\Delta = 0$.

  (To get Poisson moments, note that all cumulants $= \lambda$ since the $\mathbb{P}(\lambda)$ c.g.f. is

  $$\begin{aligned} \log E[e^{itX}] &= \log e^{\lambda\left(e^{it}-1\right)} \\ = \ & \lambda\left(e^{it}-1\right) = \sum_{r=1}^{\infty} \frac{\lambda}{r!}(it)^r \end{aligned}$$

  with all $\beta_r = \lambda$. Thus in particular the mean, variance and third central moment equal λ.)

  Then (5.1) becomes

  $$\frac{\left(\bar{X} - \bar{Y}\right)}{\sqrt{\frac{\lambda}{m} + \frac{\mu}{n}}} > u_\alpha;$$

  we could also use

  $$\frac{\left(\bar{X} - \bar{Y}\right)}{\sqrt{\frac{\bar{X}}{m} + \frac{\bar{Y}}{n}}} > u_\alpha.$$

- Asymptotic tests are generally not unique, in that they have 'equivalent' modifications. Two sequences of tests, with rejection regions R_n and R_n' and test statistics U_n and V_n are **asymptotically equivalent** if, under H, the probability of their leading to the same conclusion tends to 1:

$$\begin{aligned} &P_H(U_n \in R_n \text{ and } V_n \notin R_n') + \\ &P_H(U_n \notin R_n \text{ and } V_n \in R_n') \to 0. \quad (5.2) \end{aligned}$$

- Now consider AN test statistics with differing estimates of scale. We test $H : \theta = \theta_0$ vs. $K : \theta > \theta_0$ using

$$\begin{aligned} R_n &= \left\{ T_n \mid \ U_n = \frac{\sqrt{n}\,(T_n - \theta_0)}{\hat{\tau}_n} > u_\alpha \right\}, \\ R_n' &= \left\{ T_n \mid \ V_n = \frac{\sqrt{n}\,(T_n - \theta_0)}{\hat{\tau}_n'} > u_\alpha \right\}, \end{aligned}$$

where $\hat{\tau}_n$ and $\hat{\tau}_n'$ are consistent estimates of $\tau\,(\theta_0, \phi)$ and T_n is $AN(\theta_0, \tau^2\,(\theta_0, \phi)\,/n)$ under H. Then $U_n, V_n \xrightarrow{L} \Phi$, so that (5.2) holds, e.g. the first term is

$$P_H\left(\frac{\hat{\tau}_n}{\tau\left(\theta_0,\phi\right)}u_\alpha < \frac{\sqrt{n}\left(T_n-\theta_0\right)}{\tau\left(\theta_0,\phi\right)} \ \leq \frac{\hat{\tau}_n'}{\tau\left(\theta_0,\phi\right)}u_\alpha\right)$$
$$\rightarrow \ \ \Phi(u_\alpha)-\Phi(u_\alpha)=0.$$

- Reconsider the Poisson example above. Under H the denominator is $\sqrt{\frac{\lambda}{m}+\frac{\mu}{n}} = \sqrt{\lambda}\sqrt{\frac{1}{m}+\frac{1}{n}}$ and λ is consistently estimated by $\frac{m\bar{X}+n\bar{Y}}{m+n}$ (or any other weighted average; this one minimizes the variance); thus we can equivalently reject if

$$\frac{\left(\bar{X}-\bar{Y}\right)}{\sqrt{\frac{m\bar{X}+n\bar{Y}}{m+n}}\sqrt{\frac{1}{m}+\frac{1}{n}}} = \frac{\left(\bar{X}-\bar{Y}\right)}{\sqrt{\frac{\bar{X}}{n}+\frac{\bar{Y}}{m}}} > u_\alpha.$$

- You should browse §3.2: Comparing two treatments; in particular Examples 3.2.1, 3.2.5 (Wilcoxon tests).

6. Power; sample size; efficacy

- To carry out a test, i.e. to determine the rejection region, one needs only to calculate under H. To assess the 'power' of the test ($= P_K$(reject)) we need the behaviour under alternatives. Again, test $H : \theta = \theta_0$ vs. $K : \theta > \theta_0$ by rejecting if $T_n > C_n$. The **power** against θ is

$$\beta_n(\theta) = P_\theta\,(\text{reject}) = P_\theta\,(T_n > C_n)\,.$$

  (Thus $\beta_n(\theta_0) = \alpha + o(1)$.)

- A sequence of tests is **consistent** if $\beta_n(\theta) \to 1$ for any $\theta > \theta_0$. This is a very mild requirement. It is easily seen to hold for the AN test statistics considered in the previous lecture, assuming that they are also AN under K. (See Theorem 3.3.2. in the text for a proof.)

- More useful is to study the performance against 'contiguous' alternatives $\theta_n \to \theta_0$ at rate $1/\sqrt{n}$:
$$\theta_n = \theta_0 + \Delta/\sqrt{n} + o(n^{-1/2}).$$

- Suppose that, <u>under the sequence of alternatives</u>,
$$\frac{\sqrt{n}\,(T_n - \theta_n)}{\tau\,(\theta_0, \phi)} \overset{L}{\to} N\,(0,1) \tag{6.1}$$
where ϕ is a nuisance parameter, and that $\hat{\tau}_n$ is a consistent estimator of $\tau\,(\theta_0, \phi)$ for each ϕ. Then the rejection region is
$$\frac{\sqrt{n}\,(T_n - \theta_0)}{\hat{\tau}_n} > u_\alpha.$$
In checking (6.1), if $\tau\,(\theta, \phi)$ is continuous at θ_0 for each ϕ we may replace it by $\tau\,(\theta_n, \phi)$, or we may replace it by $\hat{\tau}_n$.

The power against θ_n is then

$$\begin{aligned} & \beta_n(\theta_n, \phi) \\ = \; & P_{\theta_n}\left(\frac{\sqrt{n}\,(T_n - \theta_0)}{\hat{\tau}_n} > u_\alpha\right) \\ = \; & P_{\theta_n}\left(\frac{\sqrt{n}\,(T_n - \theta_n)}{\hat{\tau}_n} > u_\alpha - \frac{\sqrt{n}\,(\theta_n - \theta_0)}{\hat{\tau}_n}\right) \quad (6.2) \\ = \; & 1 - P_{\theta_n}\left(\frac{\sqrt{n}\,(T_n - \theta_n)}{\hat{\tau}_n} \leq u_\alpha - \frac{\sqrt{n}\,(\theta_n - \theta_0)}{\hat{\tau}_n}\right). \end{aligned}$$

Under K, $\sqrt{n}\,(T_n - \theta_n)\,/\hat{\tau}_n$ is AN(0,1) so that by Corollary 2 in Lecture 2,

$$\begin{aligned} & \beta_n(\theta_n, \phi) \to \\ & 1 - \Phi\left(\textit{plim}\left\{u_\alpha - \frac{\sqrt{n}\,(\theta_n - \theta_0)}{\hat{\tau}_n}\right\}\right) \\ = \; & 1 - \Phi\left(u_\alpha - \frac{\Delta}{\tau\,(\theta_0, \phi)}\right) = \Phi\left(\frac{\Delta}{\tau\,(\theta_0, \phi)} - u_\alpha\right). \end{aligned}$$

- **Example**: t-test of a mean. $X_1, ..., X_n$ are i.i.d. with mean ξ, variance σ^2 and bounded (in ξ) third absolute central moment. Test $\xi = \xi_0$ vs. $\xi > \xi_0$. Consider alternatives

$$\xi_n = \xi_0 + \Delta/\sqrt{n} + o(n^{-1/2}).$$

Under this sequence of alternatives the Berry-Eséen theorem gives that

$$\frac{\sqrt{n}\left(\bar{X} - \xi_n\right)}{\sigma} \xrightarrow{L} N\left(0, 1\right),$$

since

$$\gamma_n = E_{\xi_n}\left[\left|\frac{X_1 - \xi_n}{\sigma}\right|^3\right] = O(1) = o\left(\sqrt{n}\right).$$

Replacing σ by the std. dev. S_n gives the t-test, with

$$\begin{aligned} \beta_n(\xi_n, \sigma) &\to \Phi\left(\frac{\Delta}{\sigma} - u_\alpha\right) \\ &= \Phi\left(\frac{\sqrt{n}\left(\xi_n - \xi_0\right)}{\sigma} - u_\alpha\right) + o(1). \end{aligned}$$

- These considerations are often used to determine an appropriate **sample size**. Suppose that, in the preceding example, we wish to attain an asymptotic power of β against alternatives which are $k\sigma$ away from ξ_0. Thus we require

  $$\beta = \Phi\left(\frac{\sqrt{n}\left(\xi_n - \xi_0\right)}{\sigma} - u_\alpha\right) = \Phi\left(\sqrt{n}k - u_\alpha\right),$$

  leading to

  $$n \approx \left(\frac{u_\alpha - u_\beta}{k}\right)^2.$$

  E.g. for a level $\alpha = .05$ test, attaining a power of $\beta = .9$ against $k = .5$ requires

  $$n \geq \left(\frac{1.645 + 1.282}{.5}\right)^2 = 34.27,$$

  i.e. $n \geq 35$.

- **Efficacy**. Test $H : \theta = \theta_0$ vs. $K : \theta > \theta_0$ by rejecting if
$$\frac{\sqrt{n}\left(T_n - \mu\left(\theta_0\right)\right)}{\hat{\tau}_n} > u_\alpha. \qquad (6.3)$$
Here we assume that, under H, T_n is $AN\left(\mu\left(\theta_0\right), \tau^2\left(\theta_0, \phi\right)/n\right)$ and that $\hat{\tau}_n$ is consistent for $\tau\left(\theta_0, \phi\right)$. Then the asymptotic level is α.

- Suppose as well that
$$\frac{\sqrt{n}\left(T_n - \mu\left(\theta_n\right)\right)}{\tau\left(\theta_0, \phi\right)} \xrightarrow{L} N(0,1) \qquad (6.4)$$
under alternatives $\theta_n = \theta_0 + \Delta/\sqrt{n} + o(n^{-1/2})$, and that $\mu'\left(\theta_0\right)$ exists and is > 0. The positivity is a natural requirement if we reject for *large* T_n.

- As at p. 54, but replacing θ by $\mu\left(\theta\right)$, we obtain
$$\beta_n(\theta_n, \phi) \rightarrow \Phi\left(\textit{plim}\ \frac{\sqrt{n}\left(\mu\left(\theta_n\right) - \mu\left(\theta_0\right)\right)}{\hat{\tau}_n} - u_\alpha\right).$$

Since

$$\begin{aligned} \mu(\theta_n) - \mu(\theta_0) &= \mu'(\theta_0)(\theta_n - \theta_0) + o(\theta_n - \theta_0) \\ &= \mu'(\theta_0)\,\Delta/\sqrt{n} + o(n^{-1/2}), \end{aligned}$$

we obtain

$$\beta_n(\theta_n, \phi) \to \Phi\left(\frac{\mu'(\theta_0)\,\Delta}{\tau(\theta_0, \phi)} - u_\alpha\right).$$

Here the 'efficacy' $\frac{\mu'(\theta_0)}{\tau(\theta_0,\phi)}$ depends only on the chosen test and not on the level or alternative. A test with greater efficacy has greater asymptotic power for all Δ, at all levels. Note that the efficacy depends only on the asymptotic mean and variance, at or near θ_0.

- **Example**: Matched subjects (e.g. brothers and sisters) each receive one of treatments A and B (e.g. a remedial reading course or not) with random assignments within the pairs; the data are

$$\begin{aligned} (X_i = \ & \text{response to A, } Y_i = \text{response to B)} \\ & \text{in the i}^{\text{th}} \text{ pair } (i = 1, ..., N). \end{aligned}$$

Put $Z_i = X_i - Y_i$ and test for treatment differences. Assume that $X - \theta \overset{def}{=} Y' \sim Y$ for a 'shift' parameter θ, with $\theta > 0$ indicating that treatment A is superior to treatment B. Then

$$\begin{aligned} & P_\theta(Z \leq z) = P_\theta(X - Y \leq z) \\ = \ & P_\theta(Y' + \theta - Y \leq z) = F(z - \theta), \end{aligned}$$

where F is the d.f. of $Y' - Y$. Since $Y' - Y$ and $Y - Y'$ are distributed in the same way (by virtue of the random assignments within pairs) we have that F is symmetric: $F(-z) = 1 - F(z)$. Here we have assumed that F is continuous, and will also assume it to be differentiable at 0 with derivative $f(0) > 0$.

- Consider the **sign test**. Put $N_+ =$ number of positive Z_i and $T_n = N_+/n$. For any θ we have

  $$N_+ \sim bin(n, P_\theta(Z > 0) = 1 - F(-\theta) = F(\theta)).$$

  Under $K : \theta = \theta_n = \Delta/\sqrt{n} + o(n^{-1/2})$ the test statistic T_n has mean $F(\theta_n)$ and variance $F(\theta_n)\left(1 - F(\theta_n)\right)/n$.

- For contiguous alternatives we have $\theta_n \to \theta_0 = 0$ and hence $F(\theta_n) \to F(\theta_0) = 1/2$ so that (as in an earlier application of Berry-Esséen)

$$\frac{\sqrt{n}\,(T_n - F(\theta_n))}{\sqrt{F(\theta_n)\,(1 - F(\theta_n))}} \xrightarrow{L} N\,(0,1)\,.$$

  Here $\mu\,(\theta) = F\,(\theta)$ and $\tau\,(\theta,\phi) = \sqrt{F(\theta)\,(1 - F(\theta))}$, with $\mu\,(\theta_0) = 1/2$, $\mu'\,(\theta_0) = f\,(0) > 0$ and $\tau\,(\theta_0,\phi) = 1/2$. Thus (6.4) holds (upon invoking the continuity of $\tau\,(\theta,\phi)$ at θ_0).

- Applying (6.3) and (6.4) with $\Delta = 0$ gives the rejection region

$$\frac{\sqrt{n}\left(T_n - \frac{1}{2}\right)}{1/2} = \sqrt{n}\,(2T_n - 1) > u_\alpha.$$

  As above,

$$\beta(\theta_n,\phi) \to \Phi\,(c\Delta - u_\alpha)$$

  with efficacy

$$c = \frac{\mu'\,(\theta_0)}{\tau\,(\theta_0,\phi)} = 2f(0).$$

7. Relative efficiency

- **Relative efficiency**. We compare tests by requiring that, asymptotically, they attain the same power against the same sequence of alternatives. The 'asymptotic relative efficiency' (ARE) is the limiting ratio of (sample sizes)$^{-1}$ required for this. Formally, consider sequences

  $$\left\{T_k^{(1)}\right\}_{k=1}^{\infty}, \left\{T_k^{(2)}\right\}_{k=1}^{\infty}$$

  of test statistics for testing θ_0 against alternatives $\theta > \theta_0$. Let $N_k^{(i)}, i = 1, 2$ be the sample sizes ($\to \infty$ as $k \to \infty$) and

  $$\theta_k^{(i)} = \theta_0 + \frac{\Delta_i}{\left(N_k^{(i)}\right)^{\gamma/2}}$$

  the alternatives, satisfying $\theta_k^{(1)} - \theta_0 \sim \theta_k^{(2)} - \theta_0$. Suppose the power functions satisfy

  $$\beta_i\left(\theta_k^{(i)}\right) \to H_i(c_i\Delta_i - u_\alpha),$$

  for some d.f.s H_i with $H_i\left(-u_\alpha\right) = \alpha$ and 'efficacies' c_i.

- **Example (of the most common case)**: If for one value of i we have $N_k = k\ (= n)$ (i.e. sequence of tests is indexed by the sample size), $\gamma = 1$:

  $$\theta_n^{(i)} = \theta_0 + \frac{\Delta}{\sqrt{n}}$$

  and $T_n \sim AN\left(\mu\left(\theta_n\right), \frac{\tau^2(\theta_0,\phi)}{n}\right)$ under K, then $H_i = \Phi$ and we derived $c_i = \mu'\left(\theta_0\right)/\tau\left(\theta_0, \phi\right)$.

- The ARE (of $T^{(2)}$ relative to $T^{(1)}$) is

  $$e_{2,1} = \lim_{k\to\infty} \frac{N_k^{(1)}}{N_k^{(2)}},$$

  with the $N_k^{(i)}$ constrained as above – asymptotically equal powers against asymptotically equal alternatives. Then if, say, $e_{2,1} = 3$, the test based on $T^{(1)}$ requires about 3 times as many observations as one based on $T^{(2)}$, in order to attain the same power asymptotically.

- The constraint $\theta_k^{(1)} - \theta_0 \sim \theta_k^{(2)} - \theta_0$ can be rewritten as $\frac{\Delta_1}{\left(N_k^{(1)}\right)^{\gamma/2}} \sim \frac{\Delta_2}{\left(N_k^{(2)}\right)^{\gamma/2}}$, yielding $\frac{N_k^{(1)}}{N_k^{(2)}} \rightarrow \left(\frac{\Delta_1}{\Delta_2}\right)^{2/\gamma}$; thus $e_{2,1} = \left(\frac{\Delta_1}{\Delta_2}\right)^{2/\gamma}$.

- If $H_1 = H_2$ (e.g. both $= \Phi$, the most common case) then the constraint of equal asymptotic power implies $c_1\Delta_1 = c_2\Delta_2$; together with the above this gives the alternate expression $e_{2,1} = \left(\frac{c_2}{c_1}\right)^{2/\gamma}$ (so greater efficacy implies greater ARE).

- If $H_1 \neq H_2$ then one analyzes Δ_1/Δ_2 directly, under the constraint of equal powers – see Example 3.

- **Example 1**. As in matched pairs example, let

$$Z_1, ..., Z_n \overset{i.i.d.}{\sim} F(z - \theta)$$

for a symmetric d.f. F, so that θ is a ‘centre of symmetry’. We can test $H : \theta = 0$ using

(i) the t-test ($T_n = \bar{X}, \mu(\theta) = \theta$), for which $\beta(\theta_n) \rightarrow \Phi(c\Delta - u_\alpha)$ with $c_t = 1/\sigma$,
(ii) the sign test, with $c_s = 2f(0)$. The ARE of the sign test to the t-test is ($\gamma = 1$)

$$e_{s,t} = \left(\frac{c_s}{c_t}\right)^{2/\gamma} = (2\sigma f(0))^2 .$$

This can be arbitrarily large or small; for $Z_i \sim N(0, \sigma^2)$ it is $e_{s,t} = 2/\pi \approx .637$. For the Laplace ($f(x) = .5 \exp(-|x|)$, $\sigma^2 = 2$) it is 2.
(iii) An alternative procedure is the one-sample Wilcoxon test: Rank the $|Z_i|$ and sum the ranks (rather than merely the signs) of the positive Z_i. Let V_n be this sum and define $T_n = V_n/\binom{n}{2}$; large values support K. For alternatives $\theta_n = \Delta/\sqrt{n}$ it can be shown (and will be in Lecture 16) that

$$\frac{\sqrt{n}\,(T_n - \mu(\theta_n))}{\sqrt{1/3}} \xrightarrow{L} N(0,1),$$

where $\mu(\theta) = P_\theta(Z_1 + Z_2 > 0)$ and Z_1, Z_2 are distributed independently, and symmetrically around θ, with d.f. $P_\theta(Z \leq z) = F(z - \theta)$. Thus $\mu(0) = 1/2$ – defining the rejection region – and

$\gamma = 1$. The efficacy is $c_W = \mu'(0)/\sqrt{1/3}$, so that the ARE (relative to the t-test) is

$$e_{W,t} = \left(\frac{\mu'(0)/\sqrt{1/3}}{1/\sigma}\right)^2 = 3\sigma^2\left[\mu'(0)\right]^2.$$

Now $P_\theta(Z - \theta \le z) = F(z)$, so $\mu(\theta)$ is

$$\begin{aligned} &= P_\theta((Z_1-\theta)+(Z_2-\theta) > -2\theta) \\ &= \int P_\theta((Z_1-\theta) > -2\theta - t | Z_2 - \theta = t) f(t)dt \\ &= \int F(2\theta + t) f(t)dt, \end{aligned}$$

with

$$\mu'(0) = 2\int f^2(t)dt;$$

hence

$$e_{W,t} = 12\sigma^2\left[\int f^2(t)dt\right]^2.$$

At the normal this $= 3/\pi \approx .955$, at the Laplace it $= 3/2$; it can be arbitrarily large. It can (and will) be shown that for any symmetric, square integrable density f with variance σ^2, $e_{W,t} \ge .864$.

- **Example 2**. Observe N matched pairs (X_i, Y_i), where the means are ξ_X and ξ_Y, common variances $\sigma^2/2$ and
$$CORR\left[X_i, Y_i\right] = \rho.$$
Then $Z_i = Y_i - X_i$ has mean $\xi_Z = \xi_Y - \xi_X$ and variance $\sigma^2(1-\rho)$ and we can test $\xi_Z = 0$ vs. $\xi_Z > 0$ by rejecting for large
$$t = \sqrt{N}\bar{Z}/S_Z.$$
With $\gamma = 1$ and $N_k = N$ the power function $\beta(\xi_N) \to \Phi\left(\frac{\Delta}{\sigma\sqrt{1-\rho}} - u_\alpha\right)$, with efficacy
$$c = \frac{1}{\sigma\sqrt{1-\rho}}.$$
How should the pairs be formed, i.e. what method of matching results in better ARE? In this case – T_n approaching normality and the alternatives approaching the null, at rate $1/\sqrt{n}$ – the ARE is proportional to the square of the efficacy, and so we should aim to form the pairs in such a way as to maximize the correlation between X_i and Y_i. (Clearly!)

- **Example 3**. In this example $H_1 \neq H_2$ (and $\gamma = 2$). Suppose $X_1, ..., X_n \overset{i.i.d.}{\sim} U(0, \theta)$ and we test $\theta = \theta_0$ vs. $\theta > \theta_0$. Put $G(x) = 1 - e^{-x}$, $\bar{G} = 1 - G$. A test considered previously is based on $T^{(1)} = n\left(1 - X_{(n)}/\theta_0\right)$, which $\xrightarrow{L}$ G, and rejects if

  $$T^{(1)} < G^{-1}(\alpha) = \bar{G}^{-1}(1-\alpha) = -\log(1-\alpha) \overset{def}{=} l_\alpha.$$

  The power function, for alternatives $\theta_n = \theta_0 + \Delta_1/n$ (with $\gamma = 2$) is

  $$P_{\theta_n}\left(T^{(1)} < l_\alpha\right) = \text{ (... you fill this in...)}$$

  $$\begin{aligned} &= P_{\theta_n}\left(\begin{array}{c} n\left(1 - \frac{X_{(n)}}{\theta_n}\right) < \\ \frac{n(\theta_n - \theta_0) + \theta_0 l_\alpha}{\theta_n} = \frac{\Delta_1 + \theta_0 l_\alpha}{\theta_n} \end{array} \right) \\ &\to G\left(\frac{\Delta_1}{\theta_0} + l_\alpha\right). \end{aligned}$$

  For robustness, we might instead reject if

  $$T^{(2)} = n\left(1 - \frac{X_{(n-1)}}{\theta_0}\right) < w_\alpha.$$

  As with $T^{(1)}$, it can be shown (check!) that

  $$P_{\theta_0}\left(T^{(2)} < w_\alpha\right) \to 1-(1+w_\alpha)e^{-w_\alpha} \overset{def}{=} F(w_\alpha),$$

so that w_α is obtained from $\bar{F}(w_\alpha) = 1-\alpha$. Then as above, against alternatives $\theta_n = \theta_0 + \Delta_2/n$,

$$P_{\theta_n}\left(T^{(2)} < w_\alpha\right) \to F\left(\frac{\Delta_2}{\theta_0} + w_\alpha\right).$$

The ARE of $T^{(2)}$ to $T^{(1)}$ is

$$e_{2,1} = \Delta_1/\Delta_2 = \left(\Delta_1/\theta_0\right)/\left(\Delta_2/\theta_0\right),$$

so we can take $\theta_0 = 1$; then for a (common) asymptotic power of β, Δ_1 and Δ_2 satisfy

$$1 - \beta = \bar{G}\left(\Delta_1 + l_\alpha\right) = \bar{F}\left(\Delta_2 + w_\alpha\right).$$

The first of these equalities gives

$$\Delta_1 = \bar{G}^{-1}\left(1-\beta\right) - l_\alpha = l_\beta - l_\alpha.$$

Similarly, the second gives

$$\Delta_2 = \bar{F}^{-1}\left(1-\beta\right) - w_\alpha = w_\beta - w_\alpha;$$

thus

$$e_{2,1} = \frac{l_\beta - l_\alpha}{w_\beta - w_\alpha}\ (<1).$$

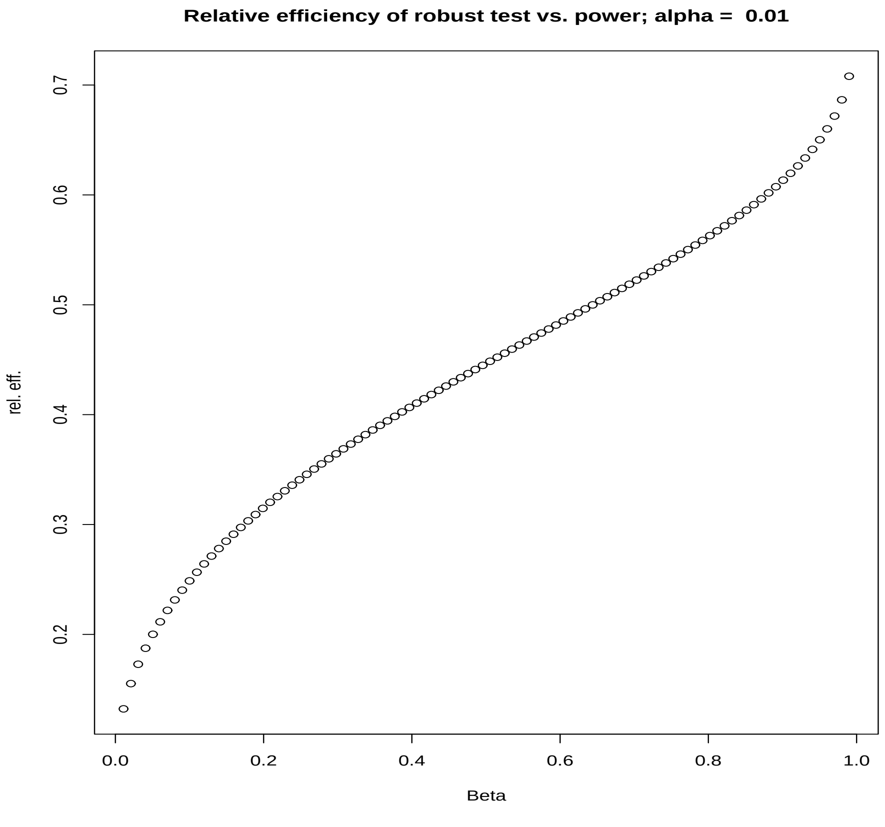
Proof of '< 1': From $1-\alpha$ $(=\bar{F}(w_\alpha)=\bar{G}(l_\alpha))=(1+w_\alpha)e^{-w_\alpha}=e^{-l_\alpha}$ we get

$$\begin{aligned} l_\beta - l_\alpha &= \log\left(\frac{1-\alpha}{1-\beta}\right)=\log\left(\frac{(1+w_\alpha)e^{-w_\alpha}}{(1+w_\beta)e^{-w_\beta}}\right) \\ &= w_\beta - w_\alpha - \left[\log\left(1+w_\beta\right)-\log\left(1+w_\alpha\right)\right], \end{aligned}$$

so that

$$e_{2,1}=1-\frac{\log\left(1+w_\beta\right)-\log\left(1+w_\alpha\right)}{w_\beta - w_\alpha}<1,$$

since $\log(\cdot)$ is increasing.



8. Robustness of test level

- **Robustness**. How do these tests perform if the assumptions underlying their derivation are violated? e.g. in the t-test we might assume that $X_1, .., X_n \overset{i.i.d.}{\sim} N\left(\xi, \sigma^2\right)$. How does the test perform if
  (i) the distribution (of $(X - \xi)/\sigma$) is non-normal: perhaps because of contamination it is $F = (1 - \varepsilon)\Phi + \varepsilon G$ for some (unknown) G, reflecting a proportion ε of erroneous sample values;
  (ii) the observations are not independent: perhaps the index i is 'time', and previous observations affect the current one, as might happen if repeated measurements are made on the same individual.

- We will consider only 'level-robustness'. Suppose we base the construction of the critical region on the assumption
$$\frac{\sqrt{n}\left(T_n - \mu\left(\theta_0\right)\right)}{\tau\left(\theta_0\right)} \xrightarrow{L} N(0,1),$$
but in fact, when F is the *true* distribution of the data,
$$\frac{\sqrt{n}\left(T_n - \mu\left(\theta_0\right)\right)}{\tau_F\left(\theta_0\right)} \xrightarrow{L} N(0,1).$$
(Here $\tau\left(\theta_0\right)$ and $\tau_F\left(\theta_0\right)$ may depend on nuisance parameters.) We reject if
$$\frac{\sqrt{n}\left(T_n - \mu\left(\theta_0\right)\right)}{\hat{\tau}_n} > u_\alpha,$$
where $\hat{\tau}_n$ is a consistent (under H) estimator of $\tau\left(\theta_0\right)$ (perhaps $\tau\left(\theta_0\right)$ itself, if there are no nuisance parameters). Take $\alpha < 1/2$, so that $u_\alpha > 0$. Let $\alpha_n(F)$ be the level of the test, with limit $\alpha(F)$; then

$$\begin{aligned}
\alpha_n(F) &= P_H\left(\frac{\sqrt{n}\left(T_n-\mu\left(\theta_0\right)\right)}{\hat{\tau}_n}>u_\alpha\right) \\
&= P_H\left(\frac{\sqrt{n}\left(T_n-\mu\left(\theta_0\right)\right)}{\tau_F\left(\theta_0\right)}>u_\alpha\frac{\hat{\tau}_n}{\tau_F\left(\theta_0\right)}\right) \\
&\rightarrow 1-\Phi\left(u_\alpha\frac{\tau\left(\theta_0\right)}{\tau_F\left(\theta_0\right)}\right) \\
&= 1-\Phi\left(u_\alpha\right)+\left[\Phi\left(u_\alpha\right)-\Phi\left(u_\alpha\frac{\tau\left(\theta_0\right)}{\tau_F\left(\theta_0\right)}\right)\right].
\end{aligned}$$

Thus

$$\begin{aligned}
\alpha(F) &= \alpha+\left[\Phi\left(u_\alpha\right)-\Phi\left(u_\alpha\frac{\tau\left(\theta_0\right)}{\tau_F\left(\theta_0\right)}\right)\right] \\
&\quad \text{is} \begin{cases} <\alpha, & \text{if } \frac{\tau(\theta_0)}{\tau_F(\theta_0)}>1, \\ =\alpha, & \text{if } \frac{\tau(\theta_0)}{\tau_F(\theta_0)}=1, \\ >\alpha, & \text{if } \frac{\tau(\theta_0)}{\tau_F(\theta_0)}<1. \end{cases}
\end{aligned}$$

If $\alpha(F)$ is $\leq \alpha$ for all F in a class $\mathbb{F}$ of distributions of the data, we say the test is 'conservative' in level. Similarly '$\geq \alpha$ for all F' is 'liberal' and '$= \alpha$ for all F' is 'robust'.

- **Example 1**. Suppose we (mistakenly) believe that a sample $X_1, ..., X_n$ arises from a $N(\xi, \sigma^2)$ population. We test $H : \sigma = \sigma_0$ vs. $K : \sigma > \sigma_0$ by using the fact that if $VAR[X] = \sigma_0^2$ then (you should show)

  $$\sqrt{n}\left(S_n^2 - \sigma_0^2\right) \Bigg/ \sqrt{VAR\left[(X-\xi)^2\right]} \xrightarrow{L} N(0,1).$$

  This does not require normality of the sample. If the data *are* normal, then under H,

  $$VAR\left[(X-\xi)^2\right] = 2\sigma_0^4 = \tau^2\left(\sigma_0\right),$$

  so we can take $\hat{\tau}_n = \sqrt{2}\sigma_0^2$. Suppose that in fact the sample arises from another distribution F with $VAR[X] = \sigma_0^2$ (so H is true) but $VAR_F\left[(X-\xi)^2\right] = \tau_F^2\left(\sigma_0\right)$. Then

  $$\frac{\tau\left(\sigma_0\right)}{\tau_F\left(\sigma_0\right)} = \sqrt{2}\frac{\sigma_0^2}{\tau_F\left(\sigma_0\right)},$$

  which may take on any positive value. Thus any $\alpha(F) \in (0, 1/2)$ is attainable and the test is very non-robust in the class $\mathbb{F}$ of distributions with finite fourth moment (i.e., non-robust against non-normality).

- **Example 2**. In the same situation as the previous example, test $\xi = \xi_0$. If the X_i are non-normal but are instead $\overset{i.i.d.}{\sim} F$ for F in the class $\mathbb{F}$ of d.f.s with mean ξ_0 and finite variance then the t-statistic ($\hat{\tau}_n = S$) still tends in law to Φ. Thus $\alpha(F) = \alpha$ and the t-test is robust in its level, in $\mathbb{F}$. Simulation studies indicate that the approach of $\alpha_n(F)$ to α is quite fast if F is symmetric, but can be quite slow if F is skewed. Note that this example does not contradict the theory above, since here $\hat{\tau}_n$ is consistent not only for $\tau(\xi_0) = \sigma_\Phi$ but for $\tau_F(\xi_0) = \sigma_F$, when F is the true distribution.

- **Example 3**. The result of the previous example (t-test) is that $\alpha_n(F) \to \alpha$ for each *fixed* $F \in \mathbb{F}$. A stronger (and more appealing) form of robustness requires uniformity (in F) of this convergence. This fails drastically; if $\mathbb{F}$ is the class of all distributions with mean ξ_0, we have that *for each* n,

$$\inf_F \alpha_n(F) = 0 \text{ and } \sup_F \alpha_n(F) = 1.$$

To see this take $\xi_0 = 0$ for simplicity. Let G be the $N(\mu_1, 1)$ d.f. and H the $N(\mu_2, 1)$ d.f. Let $F = (1 - \varepsilon)G + \varepsilon H$ for $\varepsilon \in [0, 1]$; require ε, μ_1 and μ_2 to satisfy

$$(1 - \varepsilon)\mu_1 + \varepsilon\mu_2 = 0. \tag{8.1}$$

Then $F \in \mathbb{F}$. Represent the rejection region as '$\mathbf{X} \in RR$', then the level is

$$\begin{aligned}
\alpha_n(F) &= P_F\left(\mathbf{X} \in RR\right) \\
&= \int_{RR} \prod_{i=1}^{n} \left\{(1 - \varepsilon)g(x_i) + \varepsilon h(x_i)\right\} dx_1 \cdots dx_n \\
&\geq (1 - \varepsilon)^n \int_{RR} \prod_{i=1}^{n} g(x_i) dx_1 \cdots dx_n \\
&= (1 - \varepsilon)^n P_G\left(\mathbf{X} \in RR\right) \\
&= (1 - \varepsilon)^n P_G\left(\sqrt{n}\bar{X}/S \geq u_\alpha\right).
\end{aligned}$$

For any n this may be made arbitrarily near 1 by choosing ε sufficiently small, μ_1 sufficiently large (how?), and $\mu_2 = -(1 - \epsilon)\mu_1/\epsilon$ to satisfy (8.1).

That $\inf_F \alpha_n(F) = 0$ may be shown similarly, by replacing $\alpha_n(F)$ by $1 - \alpha_n(F)$ and RR by its complement, and then proceeding as above to obtain

$1 - \alpha_n(F) \geq (1 - \varepsilon)^n P_G\left(\sqrt{n}\bar{X}/S < u_\alpha\right)$.

(Thanks to J. Sheahan for this second part.)

- **Robustness against dependence**. Suppose that $X_1, ..., X_n$ are jointly normally distributed, with mean ξ, variance σ^2. We base a test of $\xi = \xi_0$ vs. $\xi > \xi_0$ on $t_n = \sqrt{n}\left(\bar{X} - \xi_0\right)/S$. We take a very weak model of dependence, and assume that the correlations ρ_{ij} $(= \rho_{ij}^{(n)})$ satisfy (when H is true)

$$(1) \qquad \frac{1}{n}\sum_{i \neq j} \rho_{ij} \to \gamma \text{ (finite)},$$

$$(2) \qquad \frac{1}{n^2}\sum_{i \neq j} CORR\left[(X_i - \xi_0)^2, \left(X_j - \xi_0\right)^2\right] \to 0.$$

We calculate, using (1), that

$$\begin{aligned} VAR\left[\sqrt{n}\left(\bar{X} - \xi_0\right)\right] &= \sigma^2\left(1 + \frac{1}{n}\sum_{i \neq j} \rho_{ij}\right) \\ &\to \sigma^2 (1 + \gamma), \qquad (*) \end{aligned}$$

implying that under H, $\bar{X} \to \xi_0$ in $q.m.$ and hence (Corollary 1 of Lecture 1) in pr. In the same way, (2) yields that

$$\left[\frac{1}{n}\sum\left(\frac{X_i - \xi_0}{\sigma}\right)^2\right] \overset{pr}{\to} 1, \qquad (**)$$

so

$$S^2 = \frac{1}{n-1}\sum (X_i - \xi_0)^2 - \frac{n}{n-1}\left(\bar{X} - \xi_0\right)^2 \xrightarrow{pr} \sigma^2.$$

The numerator of t_n is normally distributed since the X_i are normal, hence it $\xrightarrow{L} N(0, \sigma^2 (1+\gamma))$. It follows that $t_n \xrightarrow{L} N(0, (1+\gamma))$, and that the level of the t-test (carried out *assuming independence*) is

$$\begin{aligned} P(t_n > u_\alpha) &= P\left(\frac{t_n}{\sqrt{1+\gamma}} > \frac{u_\alpha}{\sqrt{1+\gamma}}\right) \\ &\to 1 - \Phi\left(\frac{u_\alpha}{\sqrt{1+\gamma}}\right). \qquad (8.2) \end{aligned}$$

- **Example**: AR(1). Suppose that $H : \xi = 0$ is true but that, instead of being independent, the X_i follow a stationary AR(1) model:

$$X_{i+1} = \beta X_i + w_{i+1}, \ (|\beta| < 1) \qquad (8.3)$$

where $\{w_i\}$ is 'white noise', i.e. i.i.d. $N(0, \sigma_w^2)$. The t-test rejects if $t_n = \sqrt{n}\bar{X}/S > u_\alpha$. If (*)

and (**) hold then so does (8.2). To verify (*) and determine γ, note that $\sigma_X^2 = \sigma_w^2 / \left(1 - \beta^2\right)$ (calculate variances in (8.3); use stationarity). Sum (8.3) over $i = 1, ..., n-1$ to get

$$\sqrt{n}\bar{X} = \frac{U_n + \sqrt{n}\bar{w}}{1-\beta},$$

where $U_n = \left(X_1 - \beta X_n - w_1\right)/\sqrt{n}$. Since

$$\left|cov\left(U_n, \sqrt{n}\bar{w}\right)\right| \leq \sqrt{var\left[U_n\right] var\left[\sqrt{n}\bar{w}\right]}$$

and $var\left[U_n\right] \to 0$, we have $\lim var\left[\sqrt{n}\bar{X}\right] =$

$$\frac{\lim var\left[\sqrt{n}\bar{w}\right]}{(1-\beta)^2} = \frac{\sigma_w^2}{(1-\beta)^2} = \sigma_X^2 \frac{1+\beta}{1-\beta}.$$

This results in $1 + \gamma = \frac{1+\beta}{1-\beta}$, which varies over all of $(0, \infty)$ (so the asymptotic level varies over $(0, .5)$) as β varies over $(-1, 1)$. (Condition (**) can be established in the same way; this is left to you.)

The t-test is very non-robust against even very weak dependencies of this form.

9. Confidence intervals

- $\mathbf{X} = (X_1, ..., X_n)$ the data, from a distribution parameterized by θ. An interval $[\underline{\theta}(\mathbf{X}), \bar{\theta}(\mathbf{X})]$ is an asymptotic $1-\alpha$ confidence interval (CI) if

  $$P_\theta\left(\underline{\theta} \le \theta \le \bar{\theta}\right) \to 1-\alpha, \text{ for each } \theta.$$

  It is a *strong* CI if

  $$\inf_\theta P_\theta\left(\underline{\theta} \le \theta \le \bar{\theta}\right) \to 1-\alpha,$$

  entailing a form of uniformity in the convergence. (Strictly speaking, uniformity has convergence of both inf and sup.)

  – We can replace P_θ by $P_{\theta,\psi}$ in the above, where ψ is a nuisance parameter; the inf in the definition of strong CI is then taken over (θ, ψ).

  – Relationship to tests: We can reject $H : \theta = \theta_0$ in favour of $K : \theta \neq \theta_0$ iff $\theta_0 \notin [\underline{\theta}, \bar{\theta}]$, this defines a level α test from a $1-\alpha$ CI, and vice versa.

- An example of a strong CI: $X_1, ..., X_n \overset{i.i.d.}{\sim} N(\xi, \sigma^2)$. The CI derived from the 2-sided t-test is $\bar{X} \pm u_{\alpha/2} S/\sqrt{n}$, with

  $$\begin{aligned} P_{\xi,\sigma}(\xi \in CI) &= P_{\xi,\sigma}\left(\left|\frac{\bar{X}-\xi}{S/\sqrt{n}}\right| \leq u_{\alpha/2}\right) \\ &= P\left(|t_{n-1}| \leq u_{\alpha/2}\right) \rightarrow 1-\alpha; \end{aligned}$$

  since the probability does not depend on the parameters the convergence is uniform.

  – The above is typical when the CI is based on a ‘pivot’ – a function of the data and of the parameter whose distribution does not depend on the parameters.

- Example. $X_1, ..., X_n \sim \mathbb{P}(\lambda)$ $(\lambda > 0)$; test $\lambda = \lambda_0$. The large-sample test has acceptance region $\left|\sqrt{n}\left(\bar{X} - \lambda_0\right)/\sqrt{\lambda_0}\right| \leq u_{\alpha/2}$, equivalently $n\left(\bar{X} - \lambda_0\right)^2 \leq u_{\alpha/2}^2 \lambda_0$, resulting in the CI with endpoints
$$\bar{X} + \frac{u_{\alpha/2}^2}{2n} \pm \frac{u_{\alpha/2}}{\sqrt{n}}\sqrt{\bar{X} + \frac{u_{\alpha/2}^2}{4n}}. \tag{9.1}$$
Replacing $\sqrt{\lambda_0}$ by $\sqrt{\bar{X}}$ in the test statistic leads instead to the CI
$$\bar{X} \pm u_{\alpha/2}\sqrt{\frac{\bar{X}}{n}},$$
which agrees with the previous one up to terms which are $O(n^{-1})$ (i.e. if such terms are dropped). This interval is <u>not strong</u>. To see this, note that $\bar{X} - u_{\alpha/2}\sqrt{\frac{\bar{X}}{n}} \leq \lambda \leq \bar{X} + u_{\alpha/2}\sqrt{\frac{\bar{X}}{n}}$ implies that $\bar{X} > 0$, hence
$$\begin{aligned} & P_\lambda\left(\bar{X} - u_{\alpha/2}\sqrt{\frac{\bar{X}}{n}} \leq \lambda \leq \bar{X} + u_{\alpha/2}\sqrt{\frac{\bar{X}}{n}}\right) \\ \leq \; & 1 - P_\lambda\left(\bar{X} = 0\right) = 1 - e^{-n\lambda} \to 0 \text{ as } \lambda \to 0, \end{aligned}$$

hence

$$\inf_{\lambda>0} P_\lambda (\lambda \in CI) \leq \inf_{\lambda>0} \left(1 - e^{-n\lambda}\right) = 0.$$

(If λ is known to be bounded away from 0 then this example fails and the required uniformity holds – derivation in text, based on Berry-Esséen.)

- The same argument shows that (9.1), which $= \left[0, u^2_{\alpha/2}/n\right]$ if $\bar{X} = 0$, is also not strong if $\inf_{\lambda > u^2_{\alpha/2}/n} \left(1 - e^{-n\lambda}\right) = 1 - e^{-u^2_{\alpha/2}} < 1 - \alpha$ ($\alpha \gtrapprox .215$ will <u>suffice</u>).

- In (either form of) this Poisson example, note that (with AN test statistic $Z_n(\lambda)$)

$$\begin{aligned} & \inf_\lambda \lim_n P_\lambda (\lambda \in CI) \\ = \; & \inf_\lambda \lim_n P_\lambda \left(-u_{\alpha/2} \leq Z_n(\lambda) \leq u_{\alpha/2}\right) \\ = \; & \inf_\lambda P_\Phi \left(-u_{\alpha/2} \leq Z \leq u_{\alpha/2}\right) \\ = \; & 1 - \alpha \end{aligned}$$

  but this is not enough for a strong CI; we need $\lim_n \inf_\lambda P_\lambda (\lambda \in CI) = 1 - \alpha$ – the operations are interchanged.

- A strong CI on a population median. Suppose $X_1, .., X_n$ are a sample from F, with unique median $\theta = F^{-1}(.5)$. (Here F is viewed as a nuisance parameter.) Recall that the sample median $\hat{m}$ is $AN\left(\theta, 1/\left(4nf^2(\theta)\right)\right)$, but a CI based on this requires a knowledge of f. Consider instead the following procedure based on the binomial distribution. We test $\theta = \theta_0$ (2-sided) using the sign test, with test statistic

  $$S_n(\theta_0) = \# \text{ of observations which are } > \theta_0.$$

  We accept if $k_n \leq S_n(\theta_0) \leq n - k_n$. Note that

  $$S_n = n - i \Leftrightarrow i \text{ obs'ns are } \leq \theta_0 \Leftrightarrow \theta_0 \in [X_{(i)}, X_{(i+1)})$$

  so that

  $$\begin{aligned} & k_n \leq S_n(\theta_0) \leq n - k_n \\ \Leftrightarrow \ & \theta_0 \in \cup_{i=k_n}^{n-k_n} [X_{(i)}, X_{(i+1)}) = [X_{(k_n)}, X_{(n-k_n+1)}). \end{aligned}$$

  Under H, $S_n \sim bin(n, 1/2)$; thus $2\sqrt{n}\left(\frac{S_n}{n} - \frac{1}{2}\right) \xrightarrow{L} \Phi$ and we determine k_n by requiring that

  $$\begin{aligned} & P\left(\begin{array}{c} 2\sqrt{n}\left(\frac{k_n}{n} - \frac{1}{2}\right) \leq 2\sqrt{n}\left(\frac{S_n}{n} - \frac{1}{2}\right) \\ \leq 2\sqrt{n}\left(\frac{n-k_n}{n} - \frac{1}{2}\right) \end{array}\right) \\ \rightarrow \ & 1 - \alpha. \end{aligned}$$

This holds if $k_n = \frac{n}{2} - \frac{\sqrt{n} u_{\alpha/2}}{2}$ $(+o(\sqrt{n}))$, then the $LHS \rightarrow -u_{\alpha/2}$ and the $RHS \rightarrow u_{\alpha/2}$. We can take $k_n = \left[\frac{n}{2} - \frac{\sqrt{n} u_{\alpha/2}}{2}\right]$ or use any other rounding mechanism. We have shown that

$$\begin{aligned} & P_{\theta_0,F}\left(X_{(k_n)} \leq \theta_0 < X_{(n-k_n+1)}\right) \\ = \ & P\left(k_n \leq bin(n, 1/2) \leq n - k_n\right) \\ \rightarrow \ & 1 - \alpha \end{aligned}$$

uniformly in θ_0 and F, since it depends only on n. Thus a strong CI is $[X_{(k_n)}, X_{(n-k_n+1)})$.

- A more involved example of a strong CI. Suppose we have two independent samples: $X_1, ..., X_m \sim N(\xi, \sigma^2)$ and $Y_1, ..., Y_n \sim N(\eta, \tau^2)$; define $\theta = \eta - \xi$. Put $\psi = (\sigma^2, \tau^2)$ (nuisance parameter) and let $\hat{\sigma}^2$ and $\hat{\tau}^2$ be the sample variances. An asymptotic level α 2-sided test ($m, n \to \infty$; put $N = m + n$) has acceptance region

$$|Z|/\sqrt{R} = \left| \frac{\bar{Y} - \bar{X} - \theta}{\sqrt{\frac{\hat{\sigma}^2}{m} + \frac{\hat{\tau}^2}{n}}} \right| \leq u_{\alpha/2}.$$

where

$$\begin{aligned} R &= R(\psi) = \left(\frac{\hat{\sigma}^2}{m} + \frac{\hat{\tau}^2}{n} \right) \Big/ \left(\frac{\sigma^2}{m} + \frac{\tau^2}{n} \right) \quad \text{and} \\ Z &= \frac{\bar{Y} - \bar{X} - \theta}{\sqrt{\frac{\sigma^2}{m} + \frac{\tau^2}{n}}} \text{ are } \textit{independent,} \end{aligned}$$

and Z is distributed as $N(0,1)$ (free of any parameters!).

Thus, for any $\varepsilon > 0$,

$$\begin{aligned}
& P_{\theta,\psi}\left(\theta \in CI\right) - (1-\alpha) \\
= \; & P_{\theta,\psi}\left(|Z| \le u_{\alpha/2}\sqrt{R}\right) - (1-\alpha) \\
= \; & \left\{ \begin{array}{c} P_{\theta,\psi}\left(|Z| \le u_{\alpha/2}\sqrt{R} \mid |R-1| < \varepsilon\right) \cdot \\ P_{\psi}\left(|R-1| < \varepsilon\right) \end{array} \right\} - (1-\alpha) \\
& + \left\{ \begin{array}{c} P_{\theta,\psi}\left(|Z| \le u_{\alpha/2}\sqrt{R} \mid |R-1| \ge \varepsilon\right) \cdot \\ P_{\psi}\left(|R-1| \ge \varepsilon\right) \end{array} \right\},
\end{aligned}$$

so that

$$\begin{aligned}
& \left|P_{\theta,\psi}\left(\theta \in CI\right) - (1-\alpha)\right| \le \\
& \left| \begin{array}{c} P_{\theta,\psi}\left(|Z| \le u_{\alpha/2}\sqrt{R} \mid |R-1| < \varepsilon\right) \cdot \\ P_{\psi}\left(|R-1| < \varepsilon\right) - (1-\alpha) \end{array} \right| \\
& + P_{\psi}\left(|R-1| \ge \varepsilon\right) \\
& \overset{def}{=} \left|P_{\theta,\psi} \cdot P_{\psi} - (1-\alpha)\right| + \left(1 - P_{\psi}\right) \\
= \; & \left|P_{\theta,\psi} - (1-\alpha) - P_{\theta,\psi}\left(1 - P_{\psi}\right)\right| + \left(1 - P_{\psi}\right) \\
\le \; & \left|P_{\theta,\psi} - (1-\alpha)\right| + \left(1 + P_{\theta,\psi}\right)\left(1 - P_{\psi}\right) \\
\le \; & \left|P_{\theta,\psi} - (1-\alpha)\right| + 2\left(1 - P_{\psi}\right). \qquad (9.2)
\end{aligned}$$

We exhibit $a_1(\varepsilon), a_2(\varepsilon) \underset{\varepsilon\to 0}{\to} 1-\alpha$ and $b_N(\varepsilon) \underset{N\to\infty}{\to} 1$ such that

$$\begin{aligned} &1)\ a_1(\varepsilon) \le P_{\theta,\psi}\left(|Z| \le u_{\alpha/2}\sqrt{R} \mid |R-1| < \varepsilon\right) \le a_2(\varepsilon) \\ &2)\ P_\psi(|R-1| < \varepsilon) \ge b_N(\varepsilon). \end{aligned}$$

Then (9.2) is, uniformly in θ, ψ,

$$\begin{aligned} \le \quad & c_N(\varepsilon) \overset{def}{=} \max\{|a_1(\varepsilon) - (1-\alpha)|, |a_2(\varepsilon) - (1-\alpha)|\} \\ & +2(1 - b_N(\varepsilon)). \end{aligned}$$

Now $c_N(\varepsilon) \to \max\{|a_1(\varepsilon) - (1-\alpha)|, |a_2(\varepsilon) - (1-\alpha)|\}$ as $N \to \infty$, and this may be made arbitrarily small.

For 1) note that $P_{\theta,\psi}\left(|Z| \le u_{\alpha/2}\sqrt{R} \mid |R-1| < \varepsilon\right)$ is maximized by choosing R as large as possible ($= 1+\varepsilon$), yielding $a_2(\varepsilon) = 2\Phi\left(u_{\alpha/2}\sqrt{1+\varepsilon}\right) - 1$, and similarly $a_1(\varepsilon) = 2\Phi\left(u_{\alpha/2}\sqrt{1-\varepsilon}\right) - 1$. Here we use the independence of Z and R. Both these bounds $\to 1-\alpha$ as $\varepsilon \to 0$.

For 2), write

$$\begin{aligned} R &= \left(\frac{\hat{\sigma}^2}{m}+\frac{\hat{\tau}^2}{n}\right) \Big/ \left(\frac{\sigma^2}{m}+\frac{\tau^2}{n}\right) \\ &= \kappa\frac{\hat{\sigma}^2}{\sigma^2}+(1-\kappa)\frac{\hat{\tau}^2}{\tau^2} \\ &= \kappa R_\sigma+(1-\kappa)R_\tau \end{aligned}$$

for $\kappa = \frac{\sigma^2}{m} \Big/ \left(\frac{\sigma^2}{m}+\frac{\tau^2}{n}\right)$, where R_σ and R_τ are distributed independently of any parameters (in particular, independently of ψ) and $\overset{pr}{\to} 1$. Then $P_\psi\left(|R-1|<\varepsilon\right)$ is

$$\begin{aligned} &= P_\psi\left(|\kappa\left(R_\sigma-1\right)+\left(1-\kappa\right)\left(R_\tau-1\right)|<\varepsilon\right) \\ &\geq P_\psi\left(\kappa|R_\sigma-1|+(1-\kappa)|R_\tau-1|<\varepsilon\right) \\ &\geq P\left(|R_\sigma-1|<\varepsilon \text{ and } |R_\tau-1|<\varepsilon\right) \\ &\overset{def}{=} b_N\left(\varepsilon\right)\to 1, \text{ free of } \psi \text{ and hence uniformly.} \end{aligned}$$

10. Point estimation; Asymptotic relative efficiency

- **Point estimation**. Suppose that a statistic δ_n is computed as an estimate of a function $h(\theta)$, and that it is AN:
$$\sqrt{n}\left(\frac{\delta_n - h(\theta)}{\tau_n}\right) \xrightarrow{L} N(0,1).$$
We take τ_n as a measure of the accuracy of δ_n, since it is proportional to the width of an asymptotic CI.

- **Asymptotic variance vs. limiting variance**: The *asymptotic variance* is

  $\tau_0^2 = \lim \tau_n^2$ (or *plim* $\hat{\tau}_n^2$ in the case of studentizing),

  if the limit exists. Then
$$T_n \overset{def}{=} \sqrt{n}\left(\delta_n - h(\theta)\right) \xrightarrow{L} N(0,\tau_0^2).$$
It can be shown that the *limiting variance* (when it exists, otherwise use $\liminf$) exceeds the asymptotic variance:
$$\lim VAR\left(T_n\right) \geq \tau_0^2.$$

- **Example**: Let $X_n \sim bin(n,p)$ independently of $A_n \sim bin(1,a_n)$ with $a_n \rightarrow 1$. Define

  $$T_n = \frac{X_n - np}{\sqrt{n}} A_n + X_n \left(1 - A_n\right).$$

  Then $T_n \overset{L}{\rightarrow} N(0, \tau_0^2 = p(1-p))$ (why?) but

  $$\begin{aligned} VAR(T_n) &\geq VAR\left(E\left[T_n | A_n\right]\right) \\ &= VAR\left[np\left(1 - A_n\right)\right] \\ &= (np)^2 a_n \left(1 - a_n\right) \end{aligned}$$

  and this $\rightarrow \infty$ if $a_n \rightarrow 1$ slowly, say $a_n = 1 - 1/n$.

- Examples like this are troublesome since we would like to use the, more convenient, asymptotic variance as a measure of accuracy. Fortunately it is typically the case that the two quantities agree.

  Conditions under which they agree are given in the following result, in which we assume that $X_1, ..., X_n$ are i.i.d., with mean θ and variance σ^2, and that $\delta_n = h(\bar{X})$.

Recall that if $h'(\theta) \neq 0$ (assumed), then

$$\sqrt{n}\left(h(\bar{X}) - h(\theta)\right) \to N\left(0, \tau_0^2 = \left(\sigma h'(\theta)\right)^2\right).$$

Thus we aim to establish that

$$\lim VAR\left(\sqrt{n}h(\bar{X})\right) = \left(\sigma h'(\theta)\right)^2. \tag{10.1}$$

We assume that the X_i have a finite fourth moment; this implies that

$$\begin{aligned} \text{(i) } E\left[\left(\bar{X} - \theta\right)^3\right] &= O(n^{-2}), \\ \text{(ii) } E\left[\left(\bar{X} - \theta\right)^4\right] &= O(n^{-2}). \end{aligned}$$

(Proof of (i): $E\left[\left(\bar{X} - \theta\right)^3\right]$

$$\begin{aligned} &= n^{-3}E\left[\sum_{i,j,k}(X_i - \theta)\left(X_j - \theta\right)(X_k - \theta)\right] \\ &= n^{-3}E\left[\sum_i (X_i - \theta)^3\right] = O(n^{-2}), \end{aligned}$$

since the only nonzero terms are those in which all three indices agree. Assertion (ii) is proven similarly.)

Theorem: *Assume the conditions above, and that h is four times differentiable. We have:*
(i) if $h^{(4)}$ is bounded, then

$$E\left[h(\bar{X})\right] = h(\theta) + \frac{\sigma^2}{2n} h''(\theta) + O(n^{-2});$$

(ii) if both $h^{(4)}$and $(h^2)^{(4)}$ are bounded, then

$$VAR\left[h(\bar{X})\right] = \frac{\sigma^2}{n}\left(h'(\theta)\right)^2 + O(n^{-2}),$$

(so that in particular (10.1) holds).

Proof: We prove (i); for (ii) one writes $VAR\left[h(\bar{X})\right]$ as $E\left[h^2(\bar{X})\right] - \left(E\left[h(\bar{X})\right]\right)^2$ and applies (i) to each component. By Taylor's Theorem,

$$\begin{aligned} h(\bar{X}) &= h(\theta) + h'(\theta)\left(\bar{X}-\theta\right) + h''(\theta)\frac{\left(\bar{X}-\theta\right)^2}{2} \\ &\quad + h'''(\theta)\frac{\left(\bar{X}-\theta\right)^3}{3!} + h^{(4)}(\xi)\frac{\left(\bar{X}-\theta\right)^4}{4!}, \end{aligned}$$

for some ξ between $\bar{X}$ and θ. Then

$$E\left[h(\bar{X})\right] = h(\theta) + h''(\theta)\frac{\sigma^2}{2n} + R_n,$$

where (with M being a bound on $\left|h^{(4)}(x)\right|$),

$$\begin{aligned}
|R_n| &= \left|h'''(\theta)\frac{E\left[\left(\bar{X}-\theta\right)^3\right]}{3!} + E\left[h^{(4)}(\xi)\frac{\left(\bar{X}-\theta\right)^4}{4!}\right]\right| \\
&\le \left|h'''(\theta)\right|\left|\frac{E\left[\left(\bar{X}-\theta\right)^3\right]}{3!}\right| + M\cdot E\left[\frac{\left(\bar{X}-\theta\right)^4}{4!}\right] \\
&= O(n^{-2}).
\end{aligned}$$

- Why the *fourth* derivative, not merely the third?

- **Example**: The theorem applies to the case $X \sim N(\theta, 1)$ and $h(\theta) = P(X \le u) = \Phi(u-\theta)$. The limiting and asymptotic variances of $\sqrt{n}h(\bar{X}) = \sqrt{n}\Phi(u-\bar{X})$ are both $\left[h'(\theta)\right]^2 = \phi^2(u-\theta)$.

- When h is less well-behaved, *ad hoc* methods are required. Let $X_1, ..., X_n \overset{i.i.d.}{\sim} N(\theta, 1)$ and estimate $h(\theta) = e^\theta$ by $h(\bar{X})$. The derivatives of h are unbounded so the theorem above does not apply. But

$$\sqrt{n}\left(h(\bar{X}) - h(\theta)\right) \xrightarrow{L} N(0, \left[h'(\theta)\right]^2 = e^{2\theta}),$$

and

$$\sqrt{n}h(\bar{X}) = \sqrt{n}e^{\bar{X}} = \sqrt{n}e^{\theta}e^{\bar{X}-\theta}$$

has variance $e^{2\theta}\text{var}\left[\sqrt{n}e^{\bar{X}-\theta}\right]$, so we have equality of the limiting and asymptotic variances if $n\text{var}\left[e^{\bar{X}-\theta}\right] \rightarrow 1$. With $\sqrt{n}\left(\bar{X} - \theta\right) = Z \sim N(0,1)$ and $\psi(t) = E\left[e^{tZ}\right] = e^{t^2/2}$ we have that

$$\begin{aligned} & n\text{var}\left[e^{\bar{X}-\theta}\right] = n \cdot \text{var}\left[e^{Z/\sqrt{n}}\right] \\ = \; & n\left\{E\left[e^{2Z/\sqrt{n}}\right] - E^2\left[e^{Z/\sqrt{n}}\right]\right\} \\ = \; & n\left\{\psi(2/\sqrt{n}) - \psi^2(1/\sqrt{n})\right\} \\ = \; & n\left\{e^{2/n} - e^{1/n}\right\} \text{ (exactly!)} \\ = \; & n\left\{\left[1 + \frac{2}{n} + O\left(n^{-2}\right)\right] - \left[1 + \frac{1}{n} + O\left(n^{-2}\right)\right]\right\} \\ = \; & 1 + O\left(n^{-1}\right), \text{ as required.} \end{aligned}$$

- **Asymptotic Relative Efficiency**. Suppose that two estimators δ_1, δ_2 of $h(\theta)$ have

$$\begin{aligned} E_\theta\left[\delta_i\right] &= h(\theta) + \frac{a_i}{n} + O(n^{-2}), \\ VAR_\theta\left[\delta_i\right] &= \frac{\tau_i^2}{n} + O(n^{-2}). \end{aligned}$$

(In the theorem we had $\delta = h(\bar{X})$, $a = \sigma^2 h''(\theta)/2$ and $\tau^2 = \left[\sigma h'(\theta)\right]^2$.) Then $bias_\theta\left[\delta_i\right] = O(n^{-1})$ and

$$MSE_\theta\left[\delta_i\right] = bias_\theta^2 + VAR_\theta = \frac{\tau_i^2}{n} + O(n^{-2}).$$

In order that the two estimators have the same asymptotic MSE, i.e. $\frac{\tau_1^2}{n_1} \sim \frac{\tau_2^2}{n_2}$, it is required that

$$\frac{n_1}{n_2} \to \frac{\tau_1^2}{\tau_2^2} \overset{def}{=} e_{2,1},$$

the ARE of δ_2 with respect to δ_1. (We can use the asymptotic variances rather than the limiting variances; of course these typically agree.)

  – Does this agree with the previous definition in terms of efficacies? How?

- **Example**: Estimate $P(X \leq u) = F(u)$ from a sample $X_1, ..., X_n$. If F is the $N(\theta, 1)$ d.f. then $F(u) = \Phi(u - \theta) = h(\theta)$. We take $\delta_1 = h(\bar{X})$ with $\tau_1^2 = [h'(\theta)]^2 = \phi^2(u - \theta)$.

  One might instead use the 'plug-in' estimate:

$$\begin{aligned} \delta_2 &= \hat{F}_n(u) = \frac{\#\{X_i \leq u\}}{n} \sim \frac{bin(n, F(u))}{n}, \\ \tau_2^2 &= F(u)\,[1 - F(u)] = \Phi(u - \theta)\,[1 - \Phi(u - \theta)]\,. \end{aligned}$$

  Then

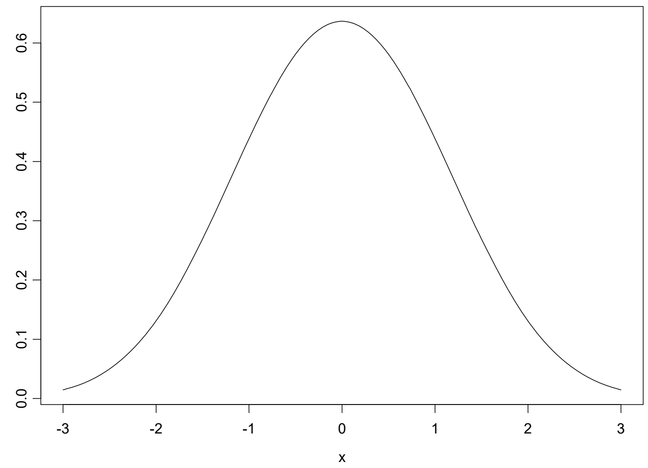
$$e_{2,1} = \frac{\phi^2(u - \theta)}{\Phi(u - \theta)\,[1 - \Phi(u - \theta)]}.$$

  Some calculus shows that this is maximized at $u - \theta = 0$, with

$$e_{2,1} \leq \frac{\phi^2(0)}{\Phi(0)\,[1 - \Phi(0)]} = 4\phi^2(0) = 2/\pi \approx .64.$$

  Thus the MLE δ_1 requires fewer than 2/3 as many observations for an asymptotic CI of the same width (<u>many</u> fewer if $|u - \theta|$ is even moderately large).

  Does this mean that δ_1 is necessarily preferred for estimating $P\,(X \leq u)$? Why or why not?



$$\frac{\phi^2(x)}{\Phi(x)[1-\Phi(x)]} \text{ vs. } x.$$

11. Comparing estimators

- **Example**: Estimating a centre of symmetry. $X_1, ..., X_n \overset{i.i.d.}{\sim} F(x-\theta)$ for a symmetric d.f. F with finite variance. Consider the competing estimators (i) Mean $\bar{X}$ with $\tau^2_{\bar{X}} = \sigma^2_X$ and (ii) Median $\tilde{X}$ with $\tau^2_{\tilde{X}} = \frac{1}{4f^2(0)}$. The ARE of the median w.r.t. the mean is $e_{\tilde{X},\bar{X}} = 4\sigma^2_X f^2(0)$. Recall that this is also the ARE of the sign test to the t-test. It may be infinitely large; at $F = \Phi$ it is $2/\pi \approx .64$.

- In the framework of the preceding example a compromise is the *trimmed mean*, i.e. the average of the middle $n-2k$ observations. Put $\alpha = k/n < .5$, then the estimate is
$$\bar{X}_\alpha = \frac{1}{n-2k}\left(X_{(k+1)} + ... + X_{(n-k)}\right).$$
Suppose that F has a density f, continuous and positive where $0 < F(x) < 1$. It can be shown

that then $\sqrt{n}\left(\bar{X}_\alpha - \theta\right) \xrightarrow{L} N(0, \sigma_\alpha^2)$ where, with $\xi_\alpha = F^{-1}(1-\alpha)$,

$$\begin{aligned}
\sigma_\alpha^2 &= E_F\left[\min\left(|X|, \xi_\alpha\right)^2\right] / (1-2\alpha)^2 \\
&= \frac{2}{(1-2\alpha)^2}\left[\int_0^{\xi_\alpha} x^2 f(x)dx + \alpha\xi_\alpha^2\right].
\end{aligned}$$

This reduces to σ_X^2 as $\alpha \to 0$ ($\xi_\alpha \to \infty$). Here we use that

$$\begin{aligned}
\sigma_X^2 &= E_F\left[|X|^2\right] \\
&\geq E_F\left[\min\left(|X|, \xi_\alpha\right)^2\right] \\
&= (1-2\alpha)^2\,\sigma_\alpha^2 \qquad (11.1) \\
&= 2\left[\int_0^{\xi_\alpha} x^2 f(x)dx + \alpha\xi_\alpha^2\right],
\end{aligned}$$

so that $\alpha\xi_\alpha^2 \to 0$ as $\alpha \to 0$. As $\alpha \to 1/2$, $\sigma_\alpha^2 \to 1/\left[4f^2(0)\right]$, by two applications of L'Hospital's rule.

- We have $e_{\bar{X}_\alpha, \bar{X}} = \sigma_X^2 / \sigma_\alpha^2$. This is > 1 for moderate α and distributions with heavier than normal tails, e.g. t on a small number of d.f. By (11.1),
$$e_{\bar{X}_\alpha, \bar{X}} \geq (1 - 2\alpha)^2$$
for any symmetric d.f. F with finite variance.

- A competing estimate of a centre of symmetry is the Hodges-Lehmann estimate $\hat{\theta}_{HL} = med\left(\frac{X_i + X_j}{2}\right)$, where the median is over all pairs $i \leq j$ or (asymptotically equivalently) $i < j$. It can be shown that
$$\sqrt{n}\left(\hat{\theta}_{HL} - \theta\right) \xrightarrow{L} N\left(0, 1/\left\{12\left[\int f^2(t)dt\right]^2\right\}\right),$$
with
$$e_{HL,\bar{X}} = 12\sigma_X^2\left[\int f^2(t)dt\right]^2.$$

– This is also the ARE of the Wilcoxon test to the t-test.

Connection: Suppose we test $H : \theta = \theta_0$ (2-sided) with $T(\theta_0)$ as test statistic, rejecting for large T. As an estimate we can take $\hat{\theta} =$ the θ_0 which minimizes $T(\theta_0)$, i.e. the value of the parameter least likely to be rejected. (For the t-test $T(\theta_0) = \sqrt{n}\left|\bar{X} - \theta_0\right|/S$ is minimized at $\theta_0 = \bar{X}$; for the sign test

$$\begin{aligned} T(\theta_0) &= \left|\{\# \text{ of obs'ns} > \theta_0\}/n - 1/2\right| \\ &= \frac{1}{2n}\left|\begin{array}{c} \{\# \text{ of obs'ns} > \theta_0\} \\ -\{\# \text{ of obs'ns} < \theta_0\} \end{array}\right| \end{aligned}$$

is minimized by what?) The HL estimate arises in the same way from the one-sample Wilcoxon test (which is based on the sum of the ranks of those $|X_i - \theta_0|$ for which $X_i - \theta_0 > 0$); this is a consequence of Problem 4 on asst. 2.

- For any symmetric, square integrable density f with variance σ^2, $e_{HL,\bar{X}} \geq .864$.

  Proof: Put $e(f) = 12\sigma_F^2 \left[\int f^2(t)dt\right]^2$ and $g(s) = \sigma_F f(\sigma_F s)$, with unit variance ($\sigma_G^2 = 1$). Then

$$e(g) = 12 \left[\int g^2(s)ds\right]^2 = e(f),$$

  so we can take $\sigma_F = 1$. We are then to

$$\text{minimize } \int f^2(t)dt \text{ subject to}$$
$$\int t^2 f(t)dt = 1, \ \int f(t)dt = 1,$$
$$\text{with } f \text{ symmetric and non-negative.}$$

  It is sufficient that $f = f(t; a, b)$ minimize

$$\int f^2(t)dt + 2a \int t^2 f(t)dt - 2b \int f(t)dt$$

  unconditionally, for constants ('Lagrange multipliers') a, b, and satisfy the side conditions. (Why? It is very instructive to write out the details of the argument.)

We minimize

$$\int \left[f^2(t) + 2at^2 f(t) - 2bf(t) \right] dt$$

by minimizing the integrand pointwise: $y^2 - (2b - 2at^2)y$ is minimized over $y \geq 0$ by $y = (b - at^2)^+$; thus

$$f(t) = (b - at^2)^+,$$

which is symmetric and non-negative. Necessarily $a > 0$ for integrability; then also $b > 0$ else $f \equiv 0$. Thus we can write f as

$$f(t) = b \left(1 - \frac{t^2}{\mu^2} \right), \ |t| \leq \mu.$$

The side condition $\int f(t)dt = 1$ gives $b = 3/(4\mu)$ and then $\int t^2 f(t)dt = 1$ gives $\mu = \sqrt{5}$; thus

$$f(t) = \frac{3}{4\sqrt{5}} \left(1 - \frac{t^2}{5} \right), \ |t| \leq \sqrt{5}$$

and we calculate that

$$e(f) = 12 \left[\int_{-\sqrt{5}}^{\sqrt{5}} f^2(t)dt \right]^2 = \frac{108}{125} = .864.$$

If T has this density f then $T^2/5 \sim Beta\,(1/2, 2)$.

12. Biased estimation; Pitman closeness

- **Biased estimation**. In general, to compare estimators δ_1, δ_2 of $h(\theta)$ with variances $\tau_i^2(\theta)$ and biases $b_i(\theta)$ we look at the ARE

  $$e_{\delta_2,\delta_1} = \text{lim}\,\frac{mse\,(\delta_1)}{mse\,(\delta_2)} = \text{lim}\,\frac{var\,(\delta_1) + bias^2\,(\delta_1)}{var\,(\delta_2) + bias^2\,(\delta_2)}.$$

  In the examples studied previously $bias^2$ was of order $O(n^{-2})$ and the ARE reduced to a comparison of the limiting, or asymptotic, variances. We look below at an example in which var and $bias^2$ are of the same order, and in which $bias^2$ plays a significant role, asymptotically. First we look at another measure of performance of an estimator.

- A related measure of accuracy is 'Pitman closeness' $P_\theta\,(|\delta - \theta| \leq a)$ for specified a; one wants this probability to be large. For an asymptotic treatment we would typically have to normalize, and consider something like lim $P_\theta\,(\sqrt{n}\,|\delta - \theta| \leq a)$.

In a broad class of cases, comparisons based on Pitman closeness of two estimators reduce to comparing the (exact or asymptotic) biases. Suppose that F is the d.f. of $(\delta - E[\delta])/\tau$, and that F is symmetric about 0 with a unimodal density f (i.e. f is a decreasing function of $|x|$). Then with (bias) $b = E[\delta] - \theta$ we have that the Pitman closeness is

$$\begin{aligned} P_\theta(|\delta - \theta| \leq a) &= F\left(\frac{a-b}{\tau}\right) - \left[1 - F\left(\frac{a+b}{\tau}\right)\right] \\ &= H_a(b), \end{aligned}$$

say. This is an even function of b whose derivative is

$$h_a(b) = \tau^{-1}\left[f\left(\frac{a+b}{\tau}\right) - f\left(\frac{a-b}{\tau}\right)\right].$$

This (odd) function of b is < 0 if $b > 0$ (since then $|a+b| > |a-b|$); thus under these conditions *the Pitman closeness is a decreasing function of* $|b|$ *for any* $a > 0$.

Thus in comparing two estimators, for each of which $(\delta - E[\delta])/\tau \sim F$ exactly or asymptotically (same F, same τ) the estimate with smaller (absolute) bias $|b|$ – hence smaller *mse* – is Pitman closer.

This conclusion (smaller *mse* $\Rightarrow$ Pitman closer) need not hold without these assumptions. The conclusion smaller *mse* $\Leftrightarrow$ smaller variance need not hold either. Suppose $X_1, ..., X_n$ are i.i.d. $U(0, \theta)$. Consider estimates $\delta_\alpha = \alpha X_{(n)}$ of θ. With $\alpha = (n+1)/n$ we have unbiasedness (hence UMVU, since $X_{(n)}$ is sufficient and the distribution is complete). We compare $\delta_{(n+1)/n}$ to δ_1 with respect to both *mse* and Pitman closeness. From $P_\theta\left(X_{(n)} \leq t\right) = (t/\theta)^n$ we calculate that $E\left[\delta_\alpha^k\right] = \frac{n}{n+k}(\alpha\theta)^k$, then obtain the bias, variance and *mse*:

$$\begin{aligned} b\left(\delta_\alpha\right) &= \left(\frac{\alpha n}{n+1} - 1\right)\theta, \\ v\left(\delta_\alpha\right) &= \alpha^2\left(\frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2\right)\theta^2, \\ mse(\delta_\alpha) &= \left[\alpha^2\frac{n}{n+2} - 2\alpha\frac{n}{n+1} + 1\right]\theta^2. \end{aligned}$$

Note *mse* is minimized by $\alpha = (n+2)/(n+1)$. Both this and $\alpha = (n+1)/n$ satisfy $n(\alpha - 1) \to 1$; we shall handle both cases by comparing δ_1 with δ_{α^*}, where $n\left(\alpha^* - 1\right) \to 1$. For arbitrary α, we rewrite the above as

$$\begin{aligned} b\left(\delta_\alpha\right) &= \left[\frac{n(\alpha-1)-1}{n+1}\right]\theta \\ v\left(\delta_\alpha\right) &= \left[\frac{\left(n\left(\alpha-1\right)\right)^2+2n\cdot n(\alpha-1)+n^2}{n\left(n+1\right)^2\left(n+2\right)}\right]\theta^2 \\ mse(\delta_\alpha) &= \frac{\left[\left(n\left(\alpha-1\right)\right)^2\frac{n+1}{n}-2n\left(\alpha-1\right)+2\right]\theta^2}{(n+1)(n+2)}. \end{aligned}$$

The ARE of the estimates considered, based on *mse*, is

$$e\left(\delta_{\alpha^*},\delta_1\right)=\lim\frac{mse(\delta_1)}{mse\left(\delta_{\alpha^*}\right)}=\lim\frac{2}{\left[\frac{n+1}{n}-2+2\right]}=2.$$

Thus any such δ_{α^*} has 1/2 the *mse*, asymptotically, even though $v\left(\delta_{\alpha^*}\right)>v\left(\delta_1\right)$ if $\alpha^*>1$. (Both bias^2 and variance of δ_1 are $O(n^{-2})$ and so reducing the bias effects a significant reduction in *mse*, even asymptotically.)

- To compare w.r.t. Pitman closeness, recall that $P_\theta\left(n\left(\theta - X_{(n)}\right) \leq a\right) \rightarrow 1 - e^{-a/\theta} = 1 - e^{-r}$, with $r = a/\theta$. Then the Pitman closeness is

$$\begin{aligned} D_\alpha &= P_\theta\left(|\delta_\alpha - \theta| \leq \frac{a}{n}\right) \\ &= P_\theta\left(-\frac{a}{n} \leq \alpha X_{(n)} - \theta \leq \frac{a}{n}\right) \\ &= P_\theta\left(\begin{array}{c} \frac{-a+n\theta(\alpha-1)}{\alpha} \leq n\left(\theta - X_{(n)}\right) \\ \leq \frac{a+n\theta(\alpha-1)}{\alpha} \end{array}\right), \end{aligned}$$

with

$$\begin{aligned} D_1 &= P_\theta\left(-a \leq n\left(\theta - X_{(n)}\right) \leq a\right) \\ &\rightarrow 1 - e^{-r}, \\ D_{\alpha^*} &\rightarrow \lim P_\theta\left(-a + \theta \leq n\left(\theta - X_{(n)}\right) \leq a + \theta\right) \\ &= \begin{cases} 1 - e^{-r-1}, & r > 1, \\ e^{r-1} - e^{-r-1}, & r \leq 1, \end{cases} \\ &= \min\left(1, e^{r-1}\right) - e^{-r-1}. \end{aligned}$$

We have $\lim_{n\rightarrow\infty}\left\{D_{\alpha_*}(r) - D_1(r)\right\} > 0$ (δ_{α^*} is asymptotically Pitman closer) if

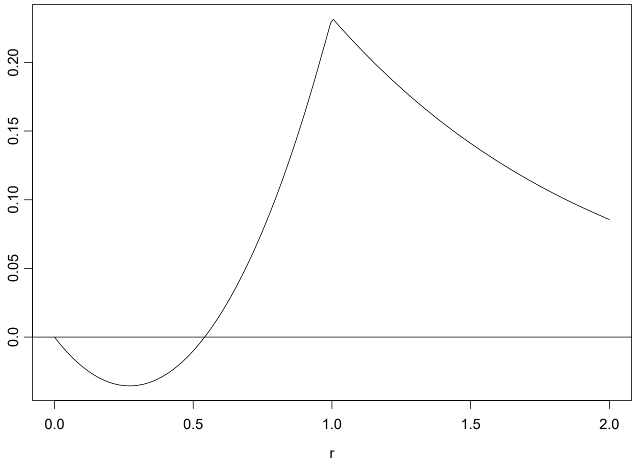
$$\min\left(1, e^{r-1}\right) - e^{-r-1} > 1 - e^{-r}.$$

This holds if
(i) $r > 1$, or if
(ii) $r \leq 1$ and $e^{r-1} - e^{-r-1} > 1 - e^{-r}$, i.e. the function

$$f(r) = e^{-1}\left(e^{r} - e^{-r}\right) + e^{-r} - 1$$

has $f(r) > 0$ $(0 < r < 1)$. Equivalently, with $t = e^{r} > 1$, we need
(i) $t > e$, or
(ii) $1 < t \leq e$ and $f(r) = \frac{(t-1)(t-(e-1))}{et} > 0$.

The second condition holds for $t \in (e-1, e]$, and so the requirement is merely $t > e - 1$. Equivalently, $r = a/\theta > \log(e-1) \approx .541$. Note that the comparison depends on the values of a and θ.



Values of $\lim_{n\to\infty}\{D_{\alpha_*}(r) - D_1(r)\}$ vs. r. The unbiased or minimum *mse* estimate δ_{α^*} is asymptotically Pitman closer than δ_1 only for $a/\theta = r > \log(e-1) \approx .541$.

Part III

MULTIVARIATE EXTENSIONS

13. Random vectors; multivariate normality

- **Convergence in probability of random vectors**. Let $\mathbf{X}^{(n)}$ be a sequence of $k \times 1$ random vectors, with elements $X_j^{(n)}$ (r.v.s) $j = 1, ..., k$. We say $\mathbf{X}^{(n)} \overset{pr}{\to} \mathbf{c}$ if $P\left(||\mathbf{X}^{(n)} - \mathbf{c}|| \geq \varepsilon\right) \to 0$ for every $\varepsilon > 0$. Equivalently (you should show) $X_j^{(n)} \overset{pr}{\to} c_j$ for each j. Exactly as in the univariate case, if $\mathbf{f} : \mathbb{R}^k \to \mathbb{R}^m$ is continuous at $\mathbf{c}$ and $\mathbf{X}^{(n)} \overset{pr}{\to} \mathbf{c}$ then $\mathbf{f}\left(\mathbf{X}^{(n)}\right) \overset{pr}{\to} \mathbf{f}\left(\mathbf{c}\right)$. In particular, $\mathbf{t}'\mathbf{X}^{(n)} \overset{pr}{\to} \mathbf{t}'\mathbf{c}$ for vectors $\mathbf{t}$.

- **Convergence in law**. Let $S \subseteq \mathbb{R}^k$. The *boundary* ∂S of S is the set of all points $\mathbf{x}$ such that any open ball around $\mathbf{x}$ intersects both S and S^c. We say $\mathbf{X}^{(n)} \overset{L}{\to} \mathbf{X} \sim H$ if $P\left(\mathbf{X}^{(n)} \in S\right) \to P_H(\mathbf{X} \in S)$ whenever S is a product $(-\infty, x_1] \times \cdots \times (-\infty, x_k]$ and $P_H\left(\partial S\right) = 0$. Note this agrees with the previous ($k = 1$) definition since the boundary of $(-\infty, c]$ in $\mathbb{R}$ is the point c, and $P_H\left(\{c\}\right) =$

0 means that c is a continuity point of H. An equivalent (and more useful) definition is that

$$\mathbf{X}^{(n)} \xrightarrow{L} \mathbf{X} \Leftrightarrow E\left[g\left(\mathbf{X}^{(n)}\right)\right] \to E_H\left[g\left(\mathbf{X}\right)\right],$$

whenever g is bounded and continuous.

- If $\mathbf{X}^{(n)} \xrightarrow{L} \mathbf{X}$ and $\mathbf{f}(\cdot)$ is continuous then $\mathbf{f}\left(\mathbf{X}^{(n)}\right) \xrightarrow{L} \mathbf{f}\left(\mathbf{X}\right)$; in particular $X_j^{(n)} \xrightarrow{L} X_j$ $(\forall j)$.

  Proof: Let g be bounded and continuous, then $E\left[g\left(\mathbf{f}\left(\mathbf{X}^{(n)}\right)\right)\right] \to E\left[g\left(\mathbf{f}\left(\mathbf{X}\right)\right)\right]$ since $g \circ \mathbf{f}$ is bounded and continuous.

- That $X_j^{(n)} \xrightarrow{L} X_j$ $(\forall j) \Rightarrow \mathbf{X}^{(n)} \xrightarrow{L} \mathbf{X}$ holds if the $X_j^{(n)}$ are independent r.v.s and the X_j are independent r.v.s (proof by c.f.s).

- $\mathbf{X}^{(n)} \xrightarrow{L} \mathbf{c}$ (constant) $\Rightarrow X_j^{(n)} \xrightarrow{L} c_j$ $(\forall j) \Rightarrow X_j^{(n)} \xrightarrow{pr} c_j$ $(\forall j) \Rightarrow \mathbf{X}^{(n)} \xrightarrow{pr} \mathbf{c}$.

- Characteristic functions: $\phi_{\mathbf{X}}(\mathbf{t}) = E\left[\text{exp}\left(i\mathbf{t}'\mathbf{X}\right)\right]$, ($t_j$ real).

  - Uniqueness: If $\phi_{\mathbf{X}}(\mathbf{t}) = \phi_{\mathbf{Y}}(\mathbf{t})$ for all $\mathbf{t}$ then $\mathbf{X} \sim \mathbf{Y}$.

  - Convergence: $\mathbf{X}^{(n)} \xrightarrow{L} \mathbf{X} \Leftrightarrow \phi_{\mathbf{X}^{(n)}}(\mathbf{t}) \to \phi_{\mathbf{X}}(\mathbf{t})$ for all $\mathbf{t}$.

- **Cramér-Wold device**:

$$\begin{aligned}
& \mathbf{X}^{(n)} \xrightarrow{L} \mathbf{X} \\
\Leftrightarrow\ & E\left[\text{exp}\left(i\mathbf{t}'\mathbf{X}^{(n)}\right)\right] \to E\left[\text{exp}\left(i\mathbf{t}'\mathbf{X}\right)\right] \text{ for all } \mathbf{t} \\
\Leftrightarrow\ & E\left[\text{exp}\left(is\mathbf{t}'\mathbf{X}^{(n)}\right)\right] \to E\left[\text{exp}\left(is\mathbf{t}'\mathbf{X}\right)\right] \text{ for all } s, \mathbf{t} \\
\Leftrightarrow\ & \mathbf{t}'\mathbf{X}^{(n)} \xrightarrow{L} \mathbf{t}'\mathbf{X} \text{ for all } \mathbf{t};
\end{aligned}$$

  i.e. we have convergence in law of $\mathbf{X}^{(n)}$ to $\mathbf{X}$ iff all linear combinations of $\mathbf{X}^{(n)}$ converge in law to those of $\mathbf{X}$.

- **Slutsky's Theorem:** If $\mathbf{X}^{(n)} \xrightarrow{L} \mathbf{X}$ and the elements of the $p \times k$ matrices $\mathbf{A}^{(n)}$ and $p \times 1$ vectors $\mathbf{b}^{(n)}$ converge in probability to the corresponding (constant) elements of $\mathbf{A}$ and $\mathbf{b}$, then
$$\mathbf{A}^{(n)}\mathbf{X}^{(n)} + \mathbf{b}^{(n)} \xrightarrow{L} \mathbf{AX} + \mathbf{b}.$$
Proof: We will prove
$$\left.\begin{array}{c} \mathbf{X}^{(n)} \xrightarrow{L} \mathbf{X}, \\ \mathbf{Y}^{(n)} \xrightarrow{pr} \mathbf{c} \text{ (constant)} \end{array}\right\} \Rightarrow \left(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}\right) \xrightarrow{L} (\mathbf{X}, \mathbf{c}). \qquad (13.1)$$
Apply this with $\mathbf{Y}^{(n)} = \left(\mathbf{A}^{(n)}, \mathbf{b}^{(n)}\right)$ and $\mathbf{f}$ the continuous function $\mathbf{f}\left(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}\right) = \mathbf{A}^{(n)}\mathbf{X}^{(n)} + \mathbf{b}^{(n)}$, which then $\xrightarrow{L} \mathbf{f}(\mathbf{X}, \mathbf{c}) = \mathbf{AX} + \mathbf{b}$.
To prove (13.1) first use the Cramér-Wold device to reduce it to the statement '$X_n \xrightarrow{L} X$, $Y_n \xrightarrow{pr} c \Rightarrow X_n + Y_n \xrightarrow{L} X + c$'; this special case of the univariate Slutsky's Theorem was proven in Assignment 1.

- A consequence of Slutsky's Theorem, established exactly as in the univariate case, is that
$$\mathbf{X}^{(n)} \xrightarrow{pr} \mathbf{X} \Rightarrow \mathbf{X}^{(n)} \xrightarrow{L} \mathbf{X}.$$

- **Multivariate normality**. We adopt a roundabout definition, to handle the case in which the density might not exist due to a singular covariance matrix. First, we say that a univariate r.v. X has the $N\left(\mu,\sigma^2\right)$ distribution if the c.f. is $E\left[e^{itX}\right] = \exp\left\{it\mu - \frac{\sigma^2 t^2}{2}\right\}$. Then

$$E\left[e^{itX}\right] \text{ is the c.f. of a r.v. with}$$
$$\begin{cases} P\left(X=\mu\right)=1, & \text{if } \sigma^2 = 0, \\ \text{p.d.f. } \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, & \text{if } \sigma^2 > 0, \end{cases}$$

  so that these are the distributions. If $\sigma^2 = 0$ then the 'density' is concentrated at a single point μ ('Dirac's delta').

- Now let $\boldsymbol{\mu}$ be a $k\times 1$ vector and $\boldsymbol{\Sigma}$ a $k\times k$ positive semidefinite matrix (i.e. $\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x} \geq 0$ for all $\mathbf{x}$). We write $\boldsymbol{\Sigma} \geq \mathbf{0}$. If $\boldsymbol{\Sigma}$ is positive definite ($\boldsymbol{\Sigma} > \mathbf{0}$), i.e. $\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$, then $\boldsymbol{\Sigma}$ is invertible.

- We say that a r.vec. $\mathbf{X}$ has the multivariate normal $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution if the c.f. is $E\left[\exp\left(i\mathbf{t}'\mathbf{X}\right)\right] = \exp\left\{i\mathbf{t}'\boldsymbol{\mu} - \frac{\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}}{2}\right\}$. Putting all but one component of $\mathbf{t}$ equal to 0 yields the consequence that then $X_j \sim N\left(\mu_j, \sigma_j^2\right)$, where $\sigma_j^2 = \boldsymbol{\Sigma}_{jj}$. Calculating

$$\begin{aligned} & \frac{\partial}{\partial t_j} E\left[\exp\left(i\mathbf{t}'\mathbf{X}\right)\right]_{|\mathbf{t}=\mathbf{0}} \\ = \; & E\left[iX_j \exp\left(i\mathbf{t}'\mathbf{X}\right)\right]_{|\mathbf{t}=\mathbf{0}} = iE[X_j] \end{aligned}$$

and

$$\frac{\partial}{\partial t_j} \exp\left\{i\mathbf{t}'\boldsymbol{\mu} - \frac{\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}}{2}\right\}_{|\mathbf{t}=\mathbf{0}} = i\mu_j$$

yields $E[X_j] = \mu_j$ and similarly $COV\left[X_j, X_l\right] = \sigma_{jl}$. We call $\boldsymbol{\Sigma}$ the covariance matrix. Note that $\boldsymbol{\Sigma} = E\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'\right]$, so that $\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t} = VAR\left[\mathbf{t}'\mathbf{X}\right]$. If $\boldsymbol{\Sigma} > \mathbf{0}$ then there is a density

$$\begin{aligned} & p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ = \; & (2\pi)^{-k/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}{2}\right\}. \end{aligned}$$

- If $\mathbf{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then an arbitrary linear combination $\mathbf{c}'\mathbf{X}$ has c.f.

$$\begin{aligned} E\left[e^{it\mathbf{c}'\mathbf{X}}\right] &= E\left[\exp\left(i\mathbf{s}'\mathbf{X}\right)\right]_{|\mathbf{s}=t\mathbf{c}} \\ = \quad \exp\left\{i\mathbf{s}'\boldsymbol{\mu} - \frac{\mathbf{s}'\boldsymbol{\Sigma}\mathbf{s}}{2}\right\}_{|\mathbf{s}=t\mathbf{c}} &= \exp\left\{it\mu - \frac{\sigma^2 t^2}{2}\right\} \end{aligned}$$

  with $\mu = \mathbf{c}'\boldsymbol{\mu}$ and $\sigma^2 = \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}$; thus $\mathbf{c}'\mathbf{X} \sim N(\mathbf{c}'\boldsymbol{\mu}, \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c})$. (If $\boldsymbol{\Sigma}$ is singular then at least one of these univariate variances is zero.)

- Conversely, suppose that *every* linear combination is normally distributed. Then $\mathbf{X}$ is multivariate normal.
  Proof: Since $Y = \mathbf{t}'\mathbf{X}$ is normal, it must be $N(\mathbf{t}'\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{t}'\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{t})$, so that

$$\begin{aligned} E\left[\exp\left(i\mathbf{t}'\mathbf{X}\right)\right] &= E\left[e^{iY}\right] = \exp\left\{i\mu_Y - \frac{\sigma_Y^2}{2}\right\} \\ &= \exp\left\{i\mathbf{t}'\boldsymbol{\mu}_{\mathbf{X}} - \frac{\mathbf{t}'\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{t}}{2}\right\}. \end{aligned}$$

  Thus $\mathbf{X}$ *is multivariate normal iff every linear combination is univariate normal*. This is the single most important property of this distribution.

- **Multivariate Central Limit Theorem**: Let $\mathbf{X}_1, ..., \mathbf{X}_n$: $k \times 1$ be i.i.d. with mean $\boldsymbol{\xi}$ and covariance matrix $\boldsymbol{\Sigma}$. Then $\sqrt{n}\left(\bar{\mathbf{X}} - \boldsymbol{\xi}\right) \xrightarrow{L} N_k\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$.
  Proof: We must show (why?) that
  $$\sqrt{n}\left(\mathbf{t}'\bar{\mathbf{X}} - \mathbf{t}'\boldsymbol{\xi}\right) \xrightarrow{L} N\left(0, \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\right)$$
  for every $\mathbf{t}$. This is the univariate CLT: put $Y_i = \mathbf{t}'\mathbf{X}_i$; these are i.i.d. with mean $\mathbf{t}'\boldsymbol{\xi}$ and variance $\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}$ and so $\sqrt{n}\left(\bar{Y} - \mathbf{t}'\boldsymbol{\xi}\right) \xrightarrow{L} N\left(0, \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\right)$. But $\bar{Y} = \mathbf{t}'\bar{\mathbf{X}}$.

- In the above if $\boldsymbol{\Sigma} > \mathbf{0}$ then the continuous function
  $$\left[\sqrt{n}\left(\bar{\mathbf{X}} - \boldsymbol{\xi}\right)\right]' \boldsymbol{\Sigma}^{-1} \left[\sqrt{n}\left(\bar{\mathbf{X}} - \boldsymbol{\xi}\right)\right] \xrightarrow{L} \mathbf{Y}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$$
  where $\mathbf{Y} \sim N_k\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$. But $\mathbf{Y}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} = \mathbf{Z}'\mathbf{Z}$ for $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}\mathbf{Y} \sim N_k\left(\mathbf{0}, \mathbf{I}\right)$. The elements Z_j are i.i.d. $N(0,1)$ and so
  $$n\left(\bar{\mathbf{X}} - \boldsymbol{\xi}\right)' \boldsymbol{\Sigma}^{-1} \left(\bar{\mathbf{X}} - \boldsymbol{\xi}\right) \xrightarrow{L} \sum_{j=1}^{k} Z_j^2 \sim \chi_k^2.$$

14. Multivariate applications

- Example of CLT: Multinomial experiments. We carry out n independent trials, each of which results in exactly one of $k+1$ mutually exclusive outcomes $O_1, ..., O_{k+1}$. On each trial, $P(O_j) = p_j$. Let Y_j $(= Y_j^{(n)})$ be the number of occurrences of O_j in the n trials. Put

$$\begin{aligned} \mathbf{Y} &= \left(Y_1^{(n)}, \cdots, Y_{k+1}^{(n)}\right)' \\ &= \sum_{i=1}^{n} \begin{pmatrix} I(O_1 \text{ occurs in i}^{\text{th}} \text{ trial}) \\ \vdots \\ I(O_{k+1} \text{ occurs in i}^{\text{th}} \text{ trial}) \end{pmatrix} \\ &= \sum_{i=1}^{n} \mathbf{I}_i, \end{aligned}$$

where $\mathbf{I}_i = \left(I_{i,1}, \cdots, I_{i,k+1}\right)^T$ and $I_{ij} = I(O_j$ occurs in the i$^{\text{th}}$ trial). We say that $\mathbf{Y} \sim multinom(n; \mathbf{p})$, where $\mathbf{p} = (p_1, ..., p_{k+1})'$.

The CLT applies: $\sqrt{n}\left(\frac{1}{n}\mathbf{Y}-E\left[\mathbf{I}_1\right]\right) \xrightarrow{L} N\left(\mathbf{0}, COV\left[\mathbf{I}_1\right]\right)$. The indicators I_{1j} are marginally $bin(1, p_j)$ with means p_j, so that $E\left[\mathbf{I}_1\right]=\mathbf{p}$. The covariances are

$$\begin{aligned}
\sigma_{jl} &= E\left[I_{1j}I_{1l}\right]-p_jp_l \\
&= P(I_{1j}=I_{1l}=1)-p_jp_l \\
&= \begin{cases} -p_jp_l, & j \neq l, \\ p_j-p_j^2, & j=l. \end{cases}
\end{aligned}$$

Thus

$$\sqrt{n}\left(\frac{1}{n}\mathbf{Y}-\mathbf{p}\right) \xrightarrow{L} N\left(\mathbf{0}, \mathbf{\Sigma}\right),$$

where Σ is given by

$$\left(\begin{array}{ccccc}
p_1-p_1^2 & \cdots & -p_1p_j & \cdots & -p_1p_{k+1} \\
\vdots & \ddots & \vdots & \vdots & \vdots \\
-p_jp_1 & \cdots & p_j-p_j^2 & \cdots & -p_jp_{k+1} \\
\vdots & & \vdots & \ddots & \vdots \\
-p_{k+1}p_1 & \cdots & -p_{k+1}p_j & \cdots & p_{k+1}-p_{k+1}^2
\end{array}\right),$$

i.e.

$$\mathbf{\Sigma}=\mathbf{D_p}-\mathbf{p}\mathbf{p}',$$

for $\mathbf{D_p}=diag(p_1,...,p_{k+1})$.

Note that $\mathbf{\Sigma}\mathbf{1}_{k+1}=\mathbf{0}$, so that $\mathbf{\Sigma}$ is singular; this reflects the fact that $\mathbf{1}'\mathbf{Y}=n$ (with 0 variance).

One can eliminate the last element of $\mathbf{Y}$ and sum the squares of the others (after subtracting their expectations); this results in the approximation forming the basis of Pearson's χ^2-test (used to test goodness of fit, independence etc. – hypotheses which can be phrased as $H: p_j = p_j^0$ for $j = 1, ..., k+1$). See §5.7.

- **Multivariate delta method**. Suppose

  $$\sqrt{n}\left(\mathbf{X}^{(n)} - \boldsymbol{\xi}\right) \xrightarrow{L} N_k\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$$

  and $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), ..., f_p(\mathbf{x}))'$ is a vector of functions differentiable at $\boldsymbol{\xi}$. Let $\mathbf{J} = [\partial \mathbf{f}/\partial \mathbf{x}]_{|\mathbf{x}=\boldsymbol{\xi}}$ be the $p \times k$ Jacobian matrix, with $(i,j)^{th}$ element $\left(\partial f_i/\partial x_j\right)_{|\mathbf{x}=\boldsymbol{\xi}}$. Then

  $$\sqrt{n}\left(\mathbf{f}\left(\mathbf{X}^{(n)}\right) - \mathbf{f}\left(\boldsymbol{\xi}\right)\right) \xrightarrow{L} N_p\left(\mathbf{0}, \mathbf{J}\boldsymbol{\Sigma}\mathbf{J}'\right).$$

  The proof is identical to that in the univariate case, but uses the multivariate MVT.

Test of association. We carry out N independent trials, each of which results in outcomes A or $\bar{A}$ (e.g. patient is a male or a female) and in outcomes B or $\bar{B}$ (e.g. patient is cured or not). Represent the frequencies and true probabilities as

	B	$\bar{B}$	TOTALS
A	N_{AB} ($p_{AB} = \pi_1$)	$N_{A\bar{B}}$ ($p_{A\bar{B}} = \pi_2$)	N_A (p_A)
$\bar{A}$	$N_{\bar{A}B}$ ($p_{\bar{A}B} = \pi_3$)	$N_{\bar{A}\bar{B}}$ ($p_{\bar{A}\bar{B}} = \pi_4$)	$N_{\bar{A}}$ ($p_{\bar{A}}$) .
	N_B (p_B)	$N_{\bar{B}}$ ($p_{\bar{B}}$)	N (1)

Define

$$\rho = \frac{P(B|A)}{P(B|\bar{A}))} \cdot \frac{P(\bar{B}|\bar{A})}{P(\bar{B}|A)} = \frac{p_{AB}}{p_{\bar{A}B}} \cdot \frac{p_{\bar{A}\bar{B}}}{p_{A\bar{B}}} = \frac{\pi_1}{\pi_3} \cdot \frac{\pi_4}{\pi_2}.$$

Note $P(B|A) > P(B|\bar{A}) \Leftrightarrow P(\bar{B}|\bar{A}) > P(\bar{B}|A)$, i.e. both ratios above are > 1 or both are < 1 or both $= 1$. Thus

$$\rho > 1 \Leftrightarrow P(B|A) > P(B|\bar{A}) \text{ and } P(\bar{B}|\bar{A}) > P(\bar{B}|A).$$

We say then that the attributes are 'positively associated'. (Negative association defined analogously.) Iff $\rho = 1$ the inequalities above are equalities and we have

$$\begin{aligned} p_B &= p_A P(B|A) + p_{\bar{A}} P(B|\bar{A}) \\ &= p_A P(B|A) + p_{\bar{A}} P(B|A) \\ &= P(B|A) \end{aligned}$$

so that $p_{AB} = p_A p_B$ and the attributes are independent. It is then of interest to test $\rho = 1$ against $\rho > 1$. The vector $\mathbf{Y} = \left(N_{AB}, N_{A\bar{B}}, N_{\bar{A}B}, N_{\bar{A}\bar{B}}\right)'$ has a $multinom(N; \boldsymbol{\pi})$ distribution, so that (with $\hat{\boldsymbol{\pi}} = \mathbf{Y}/N$)

$$\sqrt{N}\left(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\right) \xrightarrow{L} N\left(\mathbf{0}, \boldsymbol{\Sigma} = \mathbf{D}_{\boldsymbol{\pi}} - \boldsymbol{\pi}\boldsymbol{\pi}'\right).$$

Put

$$\hat{\rho} = \frac{\hat{\pi}_1 \hat{\pi}_4}{\hat{\pi}_3 \hat{\pi}_2} = \frac{\hat{p}_{AB} \hat{p}_{\bar{A}\bar{B}}}{\hat{p}_{\bar{A}B} \hat{p}_{A\bar{B}}}.$$

We obtain the asymptotic distribution of $\hat{\rho} = \exp\left(\log \hat{\rho}\right)$ in two stages. Define a function f by

$$f(\boldsymbol{\pi}) = \log \rho = \log \pi_1 - \log \pi_2 - \log \pi_3 + \log \pi_4,$$

with

$$\begin{aligned} \mathbf{J} &= \partial f / \partial \boldsymbol{\pi} = \left(\frac{1}{\pi_1}, -\frac{1}{\pi_2}, -\frac{1}{\pi_3}, \frac{1}{\pi_4}\right) \\ &= (1, -1, -1, 1)\mathbf{D}_{\boldsymbol{\pi}}^{-1} \end{aligned}$$

Then

$$\begin{aligned}&\sqrt{N}\left(\log\hat{\rho}-\log\rho\right)=\sqrt{N}\left(f(\hat{\boldsymbol{\pi}})-f(\boldsymbol{\pi})\right)\\&\xrightarrow{L} N\left(0,\tau^2=\mathbf{J}\left[\mathbf{D}_{\boldsymbol{\pi}}-\boldsymbol{\pi}\boldsymbol{\pi}'\right]\mathbf{J}'\right).\end{aligned}$$

We calculate

$$\begin{aligned}&\tau^2=\mathbf{J}\left[\mathbf{D}_{\boldsymbol{\pi}}-\boldsymbol{\pi}\boldsymbol{\pi}'\right]\mathbf{J}'\\=&\ (1,-1,-1,1)\mathbf{D}_{\boldsymbol{\pi}}^{-1}\left[\mathbf{D}_{\boldsymbol{\pi}}-\boldsymbol{\pi}\boldsymbol{\pi}'\right]\mathbf{D}_{\boldsymbol{\pi}}^{-1}(1,-1,-1,1)'\\=&\ (1,-1,-1,1)\left[\mathbf{D}_{\boldsymbol{\pi}}^{-1}-\mathbf{1}\mathbf{1}'\right](1,-1,-1,1)'\\=&\ \sum_{j=1}^{4}\pi_j^{-1}.\end{aligned}$$

Now with $g(x)=e^x$,

$$\sqrt{N}\left(g\left(\log\hat{\rho}\right)-g\left(\log\rho\right)\right)\xrightarrow{L} N\left(0,\ \tau^2\left[g'\left(\log\rho\right)\right]^2\right),$$

i.e. $\sqrt{N}\left(\hat{\rho}-\rho\right)\xrightarrow{L} N\left(0,\rho^2\tau^2\right)$.

A test of $\rho=1$ can be based on $\sqrt{N}\left(\hat{\rho}-1\right)/\rho\tau$, which $\xrightarrow{L}$ $N(0,1)$. The level is maintained if $\rho\tau$ is replaced by a consistent estimate, or if it is evaluated at $\rho = 1$. In this latter case

$$\begin{aligned}\tau^2_{|\rho=1} &= \left(\frac{1}{p_A p_B} + \frac{1}{p_A p_{\bar{B}}}\right) + \left(\frac{1}{p_{\bar{A}} p_B} + \frac{1}{p_{\bar{A}} p_{\bar{B}}}\right) \\ &= \frac{1}{p_A p_B p_{\bar{B}}} + \frac{1}{p_{\bar{A}} p_B p_{\bar{B}}} = \frac{1}{p_A p_{\bar{A}} p_B p_{\bar{B}}},\end{aligned}$$

with

$$\begin{aligned}&\frac{\sqrt{N}\,(\hat{\rho} - 1)}{\hat{\tau}} \\ &= \sqrt{N}\left(\frac{\hat{p}_{AB}\hat{p}_{\bar{A}\bar{B}} - \hat{p}_{\bar{A}B}\hat{p}_{A\bar{B}}}{\hat{p}_{\bar{A}B}\hat{p}_{A\bar{B}}}\right)\sqrt{\hat{p}_A \hat{p}_{\bar{A}} \hat{p}_B \hat{p}_{\bar{B}}}.\end{aligned}$$

The usual test, with the same asymptotic level, is obtained by noting that, under H, both $\hat{p}_{\bar{A}B}\hat{p}_{A\bar{B}}$ and $\hat{p}_A \hat{p}_{\bar{A}} \hat{p}_B \hat{p}_{\bar{B}} \xrightarrow{pr} p_A p_{\bar{A}} p_B p_{\bar{B}}$. Thus the denominator can be replaced by $\hat{p}_A \hat{p}_{\bar{A}} \hat{p}_B \hat{p}_{\bar{B}}$; the test statistic is then

$$T_N = \frac{\sqrt{N}\left(\hat{p}_{AB}\hat{p}_{\bar{A}\bar{B}} - \hat{p}_{\bar{A}B}\hat{p}_{A\bar{B}}\right)}{\sqrt{\hat{p}_A \hat{p}_{\bar{A}} \hat{p}_B \hat{p}_{\bar{B}}}},$$

which $\xrightarrow{L} N(0,1)$. For the alternative $\rho > 1$ the rejection region is $T_N > u_\alpha$; for a two-sided alternative one can equivalently reject for $T_N^2 > \chi^2_1(\alpha)$.

Goodness of fit. Test the hypothesis that the sample $X_1, ..., X_n \overset{i.i.d.}{\sim} F$ in fact arises from a distribution F_0, i.e. $H : F = F_0$. Divide the range into $k+1$ subregions A_j, and put $p_j = P_F\left(X \in A_j\right)$, $p_j^0 = P_{F_0}\left(X \in A_j\right)$. Test $H : p_j = p_j^0$, $j = 1, ..., k+1$.

For this, let $Y_j^{(n)}$ be the observed number of observations in A. Denote by $\tilde{\mathbf{Y}}$, $\tilde{\mathbf{p}}$ the vectors with the $(k+1)^{th}$ elements removed, so that $\sqrt{n}\left(\frac{1}{n}\tilde{\mathbf{Y}} - \tilde{\mathbf{p}}\right) \xrightarrow{L} N\left(\mathbf{0}, \tilde{\mathbf{\Sigma}} = \mathbf{D}_{\tilde{\mathbf{p}}} - \tilde{\mathbf{p}}\tilde{\mathbf{p}}'\right)$, and

$$\chi^2 = n\left(\frac{1}{n}\tilde{\mathbf{Y}} - \tilde{\mathbf{p}}^{(0)}\right)' \tilde{\mathbf{\Sigma}}_{(0)}^{-1}\left(\frac{1}{n}\tilde{\mathbf{Y}} - \tilde{\mathbf{p}}^{(0)}\right) \xrightarrow{L} \chi_k^2.$$

Claim: The variable defined above is

$$\begin{aligned}\chi^2 &= \sum_{j=1}^{k+1} \frac{\left(Y_j^{(n)} - np_j^0\right)^2}{np_j^0} \\ &= \sum_{\text{all cells}} \frac{(\text{observed - expected})^2}{\text{expected}}.\end{aligned}$$

Proof: Write

$$Y_{k+1} = n - \sum_{j=1}^{k} Y_j \text{ and } p_{k+1}^{(0)} = 1 - \sum_{j=1}^{k} p_j^{(0)}.$$

Then

$$\begin{aligned}
\tilde{\mathbf{\Sigma}}_{(0)}^{-1} &= \mathbf{D}_{\tilde{\mathbf{p}}^{(0)}}^{-1} + \frac{\mathbf{1}\mathbf{1}'}{p_{k+1}^{(0)}}, \\
\mathbf{1}'\left(\frac{1}{n}\tilde{\mathbf{Y}} - \tilde{\mathbf{p}}^{(0)}\right) &= -\left(\frac{Y_{k+1}^{(n)}}{n} - p_{k+1}^{(0)}\right),
\end{aligned}$$

and so

$$\begin{aligned}
\chi^2 &= n\left(\frac{1}{n}\tilde{\mathbf{Y}} - \tilde{\mathbf{p}}^{(0)}\right)' \mathbf{D}_{\tilde{\mathbf{p}}^{(0)}}^{-1} \left(\frac{1}{n}\tilde{\mathbf{Y}} - \tilde{\mathbf{p}}^{(0)}\right) \\
&\quad + n\left(\frac{1}{n}\tilde{\mathbf{Y}} - \tilde{\mathbf{p}}^{(0)}\right)' \frac{\mathbf{1}\mathbf{1}'}{p_{k+1}^{(0)}} \left(\frac{1}{n}\tilde{\mathbf{Y}} - \tilde{\mathbf{p}}^{(0)}\right) \\
&= n\left\{\sum_{j=1}^{k} \frac{\left(\frac{Y_j^{(n)}}{n} - p_j^{(0)}\right)^2}{p_j^0} + \frac{\left(\frac{Y_{k+1}^{(n)}}{n} - p_{k+1}^{(0)}\right)^2}{p_{k+1}^{(0)}}\right\}.
\end{aligned}$$

Part IV

NONPARAMETRIC ESTIMATION

15. Expectation functionals; U- and V-statistics

- Nonparametric estimation. Let $X_1, ..., X_n \overset{i.i.d.}{\sim} F$, where F is an unknown member of a class $\mathbb{F}$ of distributions. We wish to estimate a 'parameter' $\theta = h(F)$; we call h a *functional* since it maps $\mathbb{F}$ into $\mathbb{R}$. The 'plug-in' estimator is $\hat{\theta} = h(\hat{F}_n)$, where $\hat{F}_n$ is the empirical distribution function (e.d.f.):

$$\hat{F}_n(x) = \frac{\#\text{ of } X_i\text{'}s \leq x}{n} \sim \frac{bin(n, F(x))}{n} \xrightarrow{pr} F(x).$$

  - e.g. $\theta = F^{-1}(1/2)$ (population median); $\hat{\theta} = \hat{F}_n^{-1}(1/2) = \min\left\{x | \hat{F}_n(x) \geq 1/2\right\}$ is (one definition of) the sample median.

- We will investigate the asymptotic properties of such estimators.

First we consider the case in which $h(\cdot)$ is an 'expectation functional':

$$\theta = h(F) = E_F\left[\phi\left(X_1, ..., X_a\right)\right],$$

the probability-weighted average of all possible values of ϕ. Since θ is the expected value of a function of a i.i.d. observations from F, $\hat{\theta}$ is the expected value of this function based on a 'sample' from $\hat{F}_n$. The 'population' with distribution $\hat{F}_n$ consists of the data $\{x_1, .., x_n\}$, with each x_i occurring with probability n^{-1}. Thus a sample of a observations from $\hat{F}_n$ consists of an a-tuple $\left(x_{i_1}, ..., x_{i_a}\right)$ of data values, chosen with replacement. Any such a-tuple occurs with probability n^{-a}. This is then a discrete population, and so expected values are averages and the 'V-statistic' obtained as the plug-in estimate is

$$\begin{aligned} V &= h(\hat{F}_n) = E_{\hat{F}_n}\left[\phi\left(X_1, ..., X_a\right)\right] \\ &= \frac{1}{n^a} \sum_{i_1=1}^{n} \cdots \sum_{i_a=1}^{n} \phi\left(x_{i_1}, ..., x_{i_a}\right). \end{aligned}$$

- e.g. $\theta = E_F[X]$, $\hat{\theta} = E_{\hat{F}_n}[X] = \bar{X}$.

- A closely related estimate, and one with somewhat simpler asymptotic properties, is the 'U-statistic'. It differs from the V-statistic in disallowing repetitions such as $\phi(X_i, X_i)$.

- There is a symmetric functional ϕ with the same expectation θ: first note that for $\theta = E_F[\phi(X_1, ..., X_a)]$, $\phi(X_1, ..., X_a)$ can be replaced by $\phi\left(X_{i_1}, ..., X_{i_a}\right)$ for any permutation $(i_1, ..., i_a)$ of $(1, ..., a)$ without altering the expected value (since $X_1, ..., X_a$ are i.i.d.), or by

$$\phi^*(X_1, ..., X_a) = \frac{1}{a!}\sum_p \phi\left(X_{i_1}, ..., X_{i_a}\right),$$

where the sum is over all $a!$ permutations of $(1, ..., a)$. Thus also $\theta = E_F[\phi^*(X_1, ..., X_a)]$ where ϕ^* is symmetric, i.e. invariant under permutations of its arguments. e.g. if $a = 2$ then

$$\begin{aligned}\phi^*(X_1, X_2) &= \frac{1}{2}[\phi(X_1, X_2) + \phi(X_2, X_1)] \\ &= \phi^*(X_2, X_1)\end{aligned}$$

We will assume from now on that ϕ is symmetric.

- Note that an unbiased estimate of $\theta = E_F\left[\phi\left(X_1, ..., X_a\right)\right]$ is $\phi\left(X_{i_1}, ..., X_{i_a}\right)$ with distinct indices $i_1, ..., i_a$. Consider those ordered a-tuples $(i_1, ..., i_a)$ for which $1 \le i_1 < ... < i_a \le n$. There are $\binom{n}{a}$ of these. (**Reason**: Any one of them can be permuted in $a!$ ways to give a unique, unordered a-tuple $(i_1, ..., i_a)$ *with no repetitions*; there are in total
$$n(n-1)\cdots(n-a+1) = n!/(n-a)!$$
of these.)
Averaging the corresponding values of ϕ gives another unbiased estimator:
$$U = \frac{1}{\binom{n}{a}} \sum_{(i_1,...,i_a)} \phi\left(X_{i_1}, ..., X_{i_a}\right)$$
with the sum being over these *ordered* a-tuples. Any such statistic is called the *U-statistic corresponding to* ϕ.
  - e.g. $a = 1$: $\theta = E_F\left[\phi\left(X\right)\right], U = n^{-1}\sum_{i=1}^n \phi\left(X_i\right)$.

- e.g. $a = 2$; let $\theta = E_F\left[|X_1 - X_2|\right]$ with $\phi(x_1, x_2) = |x_1 - x_2|$ (note symmetry). This is a measure of scale – at the Normal it is 2σ – and the corresponding U-statistic ('Gini's mean difference') is

  $$U = \sum_{1\le i<j\le n} \left|X_i - X_j\right| / \binom{n}{2}.$$

- e.g. $a = 2$; let

  $$\theta = \sigma_F^2 = E_F\left[(X_1 - X_2)^2/2\right]$$

  with $\phi(x_1, x_2) = (x_1 - x_2)^2/2$. Then

  $$\begin{aligned} U &= \binom{n}{2}^{-1} \sum_{i<j} \phi(X_i, X_j) \\ &= \frac{1}{2}\binom{n}{2}^{-1} \sum_{\text{all } i,j} \phi(X_i, X_j) = S^2. \end{aligned}$$

- Note in this last example that U, in its re-expressed form, is clearly symmetric. You should convince yourselves that in general $U = U(X_1, ..., X_n)$ is symmetric in these n variables, i.e. $= U(X_{j_1}, ..., X_{j_n})$ for any permutation $\{j_1, ..., j_n\}$ of $\{1, ..., n\}$.

An optimality property of U-statistics. Recall that, in any family $\mathbb{F}$ of distributions, the vector $\mathbf{X}_{(n)}$ of order statistics is a *sufficient* statistic for $F \in \mathbb{F}$. Given $\mathbf{X}_{(n)} = \mathbf{x}_{(n)}$, the data must have been some re-arrangement of $\mathbf{x}_{(n)}$:

$$P_F\left(\mathbf{X} = \mathbf{x} | \mathbf{X}_{(n)} = \left(x_{(1)}, ..., x_{(n)}\right)\right) = \frac{1}{n!}$$

for each permutation $\mathbf{x} = \left(x_{i_1}, ..., x_{i_n}\right)$ of $\left(x_{(1)}, ..., x_{(n)}\right)$; this does not depend on the 'parameter' F. Note that the U-statistic corresponding to $\phi\left(x_1, ..., x_a\right)$ (symmetric!) is a function of $\mathbf{X}_{(n)}$:

$$E\left[\phi\left(X_1, ..., X_a\right) | \mathbf{X}_{(n)}\right] = \frac{1}{\binom{n}{a}} \sum_{\left(i_1, ..., i_a\right)} \phi\left(X_{i_1}, ..., X_{i_a}\right) = U,$$

since, given $\mathbf{X}_{(n)}$, all such a-tuples are equally likely.

Example: $n = 3$, $a = 2$. Given $\mathbf{X}_{(3)} = \langle 1, 5, 6 \rangle$ the sample must have been some permutation of $\{1, 5, 6\}$. For instance $X_1 = 1$, $X_2 = 6$, $X_3 = 5$. All three values $\phi(1, 5) = \phi(5, 1)$, $\phi(1, 6) = \phi(6, 1)$, $\phi(6, 5) = \phi(5, 6)$ are then equally likely, with

$$E\left[\phi\left(X_1, X_2\right) | \mathbf{X}_{(3)}\right] = \frac{\phi(1, 5) + \phi(1, 6) + \phi(6, 5)}{3}.$$

- **Rao-Blackwell Theorem**: We can always reduce the variance of an unbiased estimate by conditioning on $\mathbf{X}_{(n)}$.

  Proof: Let $S = S\left(X_1, ..., X_n\right)$ be any unbiased estimator of θ and define

  $$\tilde{S} = E\left[S|\mathbf{X}_{(n)}\right].$$

  Then (Double Expectation Theorem)

  $$E[\tilde{S}] = E_{\mathbf{X}_{(n)}}\left\{E_{S|\mathbf{X}_{(n)}}\left[S|\mathbf{X}_{(n)}\right]\right\} = E[S] = \theta.$$

  Furthermore the decomposition

  $$VAR[S] = E\left[VAR\left(S|\mathbf{X}_{(n)}\right)\right] + VAR\left[E\left(S|\mathbf{X}_{(n)}\right)\right]$$

  shows that

  $$VAR[S] \geq VAR\left[E\left(S|\mathbf{X}_{(n)}\right)\right] = VAR\left[\tilde{S}\right].$$

  This inequality is *strict* unless

  $$\begin{aligned} 0 &= E\left[VAR\left(S|\mathbf{X}_{(n)}\right)\right] \\ &= E_{\mathbf{X}_{(n)}}\left\{E_{S|\mathbf{X}_{(n)}}\left[\left(S - E\left[S|\mathbf{X}_{(n)}\right]\right)^2 |\mathbf{X}_{(n)}\right]\right\} \\ &= E\left\{E\left[\left(S - \tilde{S}\right)^2 |\mathbf{X}_{(n)}\right]\right\} = E\left[\left(S - \tilde{S}\right)^2\right], \end{aligned}$$

  which holds iff $P\left(S = \tilde{S}\right) = 1$.

- Thus $\tilde{S}$ is a strict improvement on S unless $S = \tilde{S}$ w.p. 1. If S is symmetric – as any reasonable function of i.i.d. r.v.s is – then $\tilde{S}$ is itself a U-statistic with $a = n$. In this sense any unbiased, symmetric estimator can be improved upon – in terms of variance – by a U-statistic.

- Do we get a different, and possibly better, $\tilde{S}$ if we begin with a different S? Not if $\mathbb{F}$ is rich enough that $\mathbf{X}_{(n)}$ is *complete* sufficient, i.e. if
$$E_F\left[g(\mathbf{X}_{(n)})\right] = 0 \text{ for all } F \in \mathbb{F} \ \Rightarrow g(\mathbf{x}) \equiv 0.$$
e.g. all F with densities, or all continuous F. Then at most one function of $\mathbf{X}_{(n)}$ can be unbiased for $\theta = h(F)$. If $S = S(X_1, ..., X_n)$ is unbiased so is $\tilde{S}$; i.e. the U-statistic $\tilde{S}$ is unique, unbiased and $VAR[\tilde{S}] \leq VAR[S]$, with strict inequality unless $S = \tilde{S}$. Thus *the U-statistic is the unique, minimum variance, unbiased – for all F – estimator of its expectation.*

- Here we get the variance of a U-statistic. For any $i \leq a$ define $\mathbf{X}_1 = (X_1, ..., X_i)$, $\mathbf{X}_2 = (X_{i+1}, ..., X_a)$ and define $\phi_i(x_1, ..., x_i) = \phi_i(\mathbf{x}_1)$ by

  $$\phi_i(\mathbf{x}_1) = E[\phi(\mathbf{X}_1, \mathbf{X}_2) | \mathbf{X}_1 = \mathbf{x}_1] = E[\phi(\mathbf{x}_1, \mathbf{X}_2)].$$

  (Note that in this definition we can fix *any* i of the Xs and average over the rest; this is since ϕ is symmetric.) Then by the Double Expectation Theorem,

  $$E[\phi_i(\mathbf{X}_1)] = E_F[\phi(\mathbf{X}_1, \mathbf{X}_2)] = \theta.$$

  Define also

  $$\sigma_i^2 = \text{var}[\phi_i(\mathbf{X}_1)].$$

- **Lemma**:

  (i) For fixed $a \leq \frac{n+1}{2}$, $VAR[U] = \sum_{i=1}^{a} \omega_i \sigma_i^2$ for

$$\begin{aligned}\omega_i &= \binom{a}{i}\binom{n-a}{a-i} \Big/ \binom{n}{a} &(15.1)\\ &= \binom{a}{i}^2 i! n^{-i}\left(1+O(n^{-1})\right). &(15.2)\end{aligned}$$

  (ii) If $\sigma_1^2 > 0$ and all $\sigma_i^2 < \infty$ then

$$VAR[\sqrt{n}U] = a^2\sigma_1^2 + O(n^{-1}).$$

Proof: (15.1) is assigned for $a = 2$; general case is similar. Then (15.2) uses $\binom{n}{k} = \left\{n^k/k!\right\}\left(1+O(n^{-1})\right)$:

$$\begin{aligned}\omega_i &= \binom{a}{i} \cdot \frac{\left\{(n-a)^{a-i} / (a-i)!\right\}\left(1+O(n^{-1})\right)}{\left\{n^a/a!\right\}\left(1+O(n^{-1})\right)}\\ &= \binom{a}{i}^2 i! \frac{n^{a-i}}{n^a}\left(1-\frac{a}{n}\right)^{a-i}\left(1+O(n^{-1})\right)\\ &= \binom{a}{i}^2 i! n^{-i}\left(1+O(n^{-1})\right).\end{aligned}$$

Now (ii) follows:

$$\begin{aligned}VAR[\sqrt{n}U] &= n\left(\omega_1\sigma_1^2 + O(n^{-2})\right)\\ &= a^2\sigma_1^2 + O(n^{-1}).\end{aligned}$$

16. Asymptotic normality of U-statistics

- Recall $\theta = E_F\left[\phi\left(X_1, ..., X_a\right)\right]$ is to be estimated; here $\phi\left(X_1, ..., X_a\right)$ is a symmetric function of its arguments, and $X_1, ..., X_n \overset{i.i.d.}{\sim} F \in \mathbb{F}$. A minimum variance, unbiased estimator is the U-statistic (unique, if $\mathbb{F}$ is sufficiently rich). For any $i \leq a$ define $\mathbf{X}_1 = (X_1, ..., X_i)$, $\mathbf{X}_2 = (X_{i+1}, ..., X_a)$ and define $\phi_i\left(\mathbf{x}_1\right) = E\left[\phi\left(\mathbf{x}_1, \mathbf{X}_2\right)\right]$. Then $E\left[\phi_i\left(\mathbf{X}_1\right)\right] = \theta$, and we defined σ_i^2 to be var$[\phi_i\left(\mathbf{X}_1\right)]$.

- **Theorem**: *If $\sigma_1^2 > 0$ and all $\sigma_i^2 < \infty$ then $\sqrt{n}\left(U - \theta\right) \xrightarrow{L} N\left(0, a^2\sigma_1^2\right)$.*
  Proof follows the example. Note that if ϕ is bounded, then the condition '*all* $\sigma_i^2 < \infty$' is automatically satisfied. Note also that by the preceding lemma the asymptotic variance is the limiting variance and so also
$$\frac{U - \theta}{\sqrt{VAR[U]}} \xrightarrow{L} N(0, 1).$$

Example: Wilcoxon 1-sample test. Let $X_1, ..., X_n$ be a sample with $P(X \leq x) = F(x - \xi)$, continuous and symmetric about ξ. To test $\xi = 0$ vs. $\xi > 0$ we can rank the $|X_i|$, let $S_1 < ... < S_{N_+}$ be the ranks arising from *positive* Xs, and reject if $S = \sum S_j / \binom{n}{2}$ is too large. It can be shown (asst. 2) that $S = \sum_{i \leq j} I\left(X_i + X_j > 0\right) / \binom{n}{2}$, hence

$$S = U + N_+ / \binom{n}{2} = U + O_P(n^{-1}),$$

where $U = \sum_{i<j} \phi\left(X_i, X_j\right) / \binom{n}{2}$ for $\phi\left(x_i, x_j\right) = I\left(x_i + x_j > 0\right)$ (bounded!). By the theorem,

$$\sqrt{n}(U - \theta_F) \sim AN\left(0, 4VAR\left[\phi_1\left(X_1\right)\right]\right).$$

Here

$$\begin{aligned} \phi_1\left(x_1\right) &= E_{X_2}\left[\phi\left(x_1, X_2\right)\right] = P\left(x_1 + X_2 > 0\right) \\ &= 1 - F(-x_1 - \xi) = F\left(x_1 + \xi\right), \end{aligned}$$

and so

$$\begin{aligned} \theta_F = E_F\left[\phi_1\left(X_1\right)\right] &= E\left[F\left(X + \xi\right)\right], \\ 4VAR\left[\phi_1\left(X_1\right)\right] &= 4VAR\left[F\left(X + \xi\right)\right]. \end{aligned}$$

Under H, $F(X+\xi) = F(X) \sim Unif(0,1)$ with mean 1/2 and variance 1/12. Thus under the hypothesis,

$$\sqrt{3n}(U - 1/2) \xrightarrow{L} N(0,1).$$

In any event, the theorem gives that

$$\frac{\sqrt{n}\,(U-\theta_F)}{\sqrt{4VAR\left[\phi_1(X_1)\right]}} \text{ and } \frac{U-E[U]}{\sqrt{VAR[U]}} \text{ both } \xrightarrow{L} N(0,1).$$

Here one can replace $\sqrt{n}\,(U-\theta_F)$ by $\sqrt{n}\,(S-\theta_F)$ (since $\sqrt{n}\,(S-U) = \sqrt{n}N_+/\binom{n}{2} = O_P(n^{-1/2})$) or by $\sqrt{n}\,(S-E[S])$ (since $\sqrt{n}\,(E[S]-\theta_F) = O(n^{-1/2})$).

Also one can replace $4VAR\left[\phi_1(X_1)\right] = \lim VAR\left[\sqrt{n}U\right]$ by $VAR\left[\sqrt{n}S\right]$, since (check!) $\frac{VAR[S]}{VAR[U]} \to 1$.

Thus

$$\frac{\sqrt{n}\,(S-\theta_F)}{\sqrt{4VAR\left[\phi_1(X_1)\right]}} \text{ and } \frac{S-E[S]}{\sqrt{VAR[S]}} \text{ both } \xrightarrow{L} N(0,1).$$

- **Proof of theorem**: We use the 'projection method', by which we approximate $T_n = \sqrt{n}\,(U-\theta)$ by a sum T_n^* of i.i.d.s, apply the CLT to T_n^*, then show that the approximation is good enough that the end result applies to T_n as well. Since the best (minimum mse) forecast of a r.v. from another is ... (what?), it is reasonable to conjecture that the 'best' approximation is the 'projection on the observations'

$$T_n^* = \sum_{i=1}^{n} E\left[\sqrt{n}\,(U-\theta)\,|X_i\right].$$

Claim 1: $E\left[(U-\theta)\,|X_i\right] = (a/n)\left(\phi_1\left(X_i\right)-\theta\right)$,

so that by the CLT, $T_n^* \xrightarrow{L} N\left(0, a^2\sigma_1^2\right)$. (The mean 0 and variance $a^2\sigma_1^2$ are exact for all n.)

Claim 2: $COV\left[T_n, T_n^*\right] = a^2\sigma_1^2$. Note this also $= \lim VAR\left[T_n\right] = VAR\left[T_n^*\right]$, so that

$$E\left[\left(T_n - T_n^*\right)^2\right] = VAR\left[T_n - T_n^*\right] \to 0$$

and $T_n - T_n^* \to 0$ in quadratic mean, hence in probability. Thus

$$T_n = T_n^* + \left(T_n - T_n^*\right) \xrightarrow{L} N\left(0, a^2\sigma_1^2\right)$$

by Slutsky's Theorem, completing the proof.

- *Proof of claim 1*:

$$\begin{aligned} & E\left[(U-\theta)\,|X_i\right] \\ = & \binom{n}{a}^{-1} \sum_{(i_1,\ldots,i_a)} E\left[\phi\left(X_{i_1},\ldots,X_{i_a}\right)-\theta|X_i\right]. \end{aligned}$$

If $i \notin \{i_1,\ldots,i_a\}$ the expectation is 0. Whenever $\{i_1,\ldots,i_a\}$ contains i it is $\phi_1(X_i)-\theta$. This occurs $\binom{n-1}{a-1}$ times; then since $\binom{n-1}{a-1}\Big/\binom{n}{a} = a/n$ the claim is established.

- *Proof of claim 2*: Assigned. Both this derivation and that of var$[U]$ employ the following result (or at least the technique):

$$COV\left[\phi\left(\mathbf{X}_1,\mathbf{X}_2\right),\phi\left(\mathbf{X}_1,\mathbf{X}_2'\right)\right]=\sigma_i^2, \tag{16.1}$$

where $\mathbf{X}_2' = \left(X_{i+1}',\ldots,X_a'\right)$ and

$$X_{i+1},\ldots,X_a,X_{i+1}',\ldots,X_a' \overset{i.i.d.}{\sim} F.$$

- Proof of (16.1): Given $\mathbf{X}_1$, $\phi(\mathbf{X}_1, \mathbf{X}_2)$ and $\phi\left(\mathbf{X}_1, \mathbf{X}_2'\right)$ are independent. Thus the covariance is

$$\begin{aligned}
& COV\left[\phi(\mathbf{X}_1, \mathbf{X}_2), \phi\left(\mathbf{X}_1, \mathbf{X}_2'\right)\right] \\
= \; & E_{\mathbf{X}_1}\left[E\left[\phi(\mathbf{X}_1, \mathbf{X}_2)\,\phi\left(\mathbf{X}_1, \mathbf{X}_2'\right)|\mathbf{X}_1\right]\right] - \theta^2 \\
= \; & E_{\mathbf{X}_1}\left[E\left[\phi(\mathbf{X}_1, \mathbf{X}_2|\mathbf{X}_1)\right] E\left[\phi\left(\mathbf{X}_1, \mathbf{X}_2'|\mathbf{X}_1\right)\right]\right] - \theta^2 \\
= \; & E_{\mathbf{X}_1}\left[\phi_i(\mathbf{X}_1) \cdot \phi_i(\mathbf{X}_1)\right] - \theta^2 \\
= \; & E_{\mathbf{X}_1}\left[\phi_i^2(\mathbf{X}_1)\right] - E^2_{\mathbf{X}_1}\left[\phi_i(\mathbf{X}_1)\right] \\
= \; & VAR\left[\phi_i(\mathbf{X}_1)\right].
\end{aligned}$$

- **V-statistics**. Recall that in the case of an expectation functional $\theta = h(F) = E_F\left[\phi\left(X_1, ..., X_a\right)\right]$, the plug-in estimate is
$$\begin{aligned} V &= E_{\hat{F}_n}\left[\phi\left(X_1, ..., X_a\right)\right] \\ &= \frac{1}{n^a} \sum_{i_1=1}^{n} \cdots \sum_{i_a=1}^{n} \phi\left(x_{i_1}, ..., x_{i_a}\right). \end{aligned}$$
You should read Example 6.2.5, leading to the result that if
$$VAR\left[\phi\left(X_{i_1}, ..., X_{i_a}\right)\right] < \infty$$
for all $1 \leq i_1 \leq \cdots \leq i_a \leq n$ then
$$\begin{aligned} \sqrt{n}\left(V - \theta\right) &= \sqrt{n}\left(U - \theta\right) + o_P(1) \\ &\xrightarrow{L} N\left(0, a^2\sigma_1^2\right) \end{aligned}$$
and $a^2\sigma_1^2$ is the limiting as well as the asymptotic variance of V.

- We say U and V are '$\sqrt{n}$-equivalent': $\sqrt{n}\left(U - V\right) \xrightarrow{pr} 0$. However, V typically has a bias of order n^{-1}.

- **Outline of proof** of $\sqrt{n}$-equivalence: Since ϕ is symmetric,

$$V = n^{-a}\left[a!\binom{n}{a}U + \boldsymbol{\Sigma}'\right],$$

where $\boldsymbol{\Sigma}'$ denotes the sum over indices with some repetitions, and is $O_P(n^{a-1})$. Since

$$n^{-a}a!\binom{n}{a} = 1 + O\left(n^{-1}\right)$$

we obtain $V = U + O_P(1)/n$ where the $O_P(1)$ term has a finite variance (this is where some work is required).
In particular, for $a = 2$,

$$V = U + \frac{n^{-1}\sum_i \phi\left(X_i, X_i\right) - U}{n}$$

with

$$E[V] = \theta + \frac{E\left[\phi\left(X_1, X_1\right)\right] - \theta}{n}.$$

- **Example**: $\phi(x_1, x_2) = (x_1 - x_2)^2/2$ results in $U = S^2$ (unbiased for $E\left[U\right] = \sigma_F^2$) but

$$V = n^{-1}\sum_i \left(x_i - \bar{x}\right)^2.$$

17. Influence function analysis

- An analysis of more general functionals requires notions akin to linear approximations in calculus. The preceding gave the asymptotic theory for a *linear* functional, i.e. one such as $h(F) = E_F[\phi(X)]$ for which

$$h((1-\varepsilon)F + \varepsilon G) = (1-\varepsilon)h(F) + \varepsilon h(G),$$

and we now need something like a mean value theorem for functionals, so as to treat them as *approximately linear*.

- **Definitions**: The *Kolmogorov distance* between d.f.s F and G is

$$d(F,G) = \sup_x |F(x) - G(x)| \ \ (\leq 1).$$

A functional h is *continuous* at F if

$$d(F_n, F) \to 0 \Rightarrow h(F_n) \to h(F).$$

(By Polya's Theorem – 2.6.1 in text – if F is continuous then $d(F_n, F) \to 0 \Leftrightarrow F_n \xrightarrow{L} F$.)

- **Theorem**: If $X_1, ..., X_n \overset{i.i.d.}{\sim} F$ and $h(\cdot)$ is continuous at F, then $h\left(\hat{F}_n\right)$ is consistent for $h(F)$.

  To prove this, one shows that $d(\hat{F}_n, F) \overset{pr}{\to} 0$; this is (a weak version of) the celebrated Glivenko-Cantelli Theorem. (Pointwise convergence is merely the WLLN.)

- **Example 1**. $h(F) = E_G\left[(F(X) - G(X))^2\right]$, the Cramér-von Mises distance. Here $h\left(\hat{F}_n\right)$ measures the goodness of fit between the e.d.f. and a hypothesized d.f. G. You might verify that, if G is continuous, then

$$h\left(\hat{F}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\left(G\left(X_{(i)}\right) - \frac{i - \frac{1}{2}}{n}\right)^2 + \frac{1}{12n^2}.$$

  Under $H : F = G$, we have $G\left(X_{(i)}\right) = F\left(X_{(i)}\right) \sim U_{(i)}$ with $E\left[U_{(i)}\right] = i/(n+1)$, and

$$h\left(\hat{F}_n\right) = n^{-1}\sum_i \left(U_{(i)} - E\left[U_{(i)}\right]\right)^2 + o_P(n^{-1}).$$

By using the fact that $|a^2 - b^2| \leq 2|a - b|$ if $|a|, |b| \leq 1$ we obtain

$$\begin{aligned} & |h(F_n) - h(F)| \\ = & \left| E_G\left[(F_n(X) - G(X))^2 - (F(X) - G(X))^2\right]\right| \\ \leq & E_G\left[\left|(F_n(X) - G(X))^2 - (F(X) - G(X))^2\right|\right] \\ \leq & 2E_G\left[|F_n(X) - F(X)|\right] \\ \leq & 2d(F_n, F); \end{aligned}$$

it follows that $h(\cdot)$ is continuous at F.

- **Example 2**. $h(F) = E_F[X]$. Let F and H_n be any d.f.s with finite means; put

$$F_n = (1 - \varepsilon_n)F + \varepsilon_n H_n$$

for $0 < \varepsilon_n < 1$. Let $\varepsilon_n \to 0$, then

$$d(F_n, F) = \varepsilon_n d(H_n, F) \leq \varepsilon_n \to 0$$

but $h(F_n) - h(F) = \varepsilon_n (h(H_n) - h(F))$ need not $\to 0$. For instance if $h(F) = 0$ and $h(H_n) = n/\varepsilon_n$ then $h(F_n) - h(F) \to \infty$. Thus this 'mean functional' is not continuous. A consequence is that $h\left(\hat{F}_n\right) = \bar{X}$ need not be consistent for $E_F[X]$ in the presence of outliers.

- **Influence functions**. Let $h(F)$ be a functional defined for $F \in \mathbb{F}$, a convex class of d.f.s:

$$F_0, F_1 \in \mathbb{F} \Rightarrow F_\varepsilon \overset{def}{=} (1-\varepsilon)F_0 + \varepsilon F_1 \in \mathbb{F}$$

for $0 \le \varepsilon \le 1$. Consider

$$\begin{aligned} \dot{h}(F_0; F_1) &= \lim_{\varepsilon \to 0} \frac{h\left((1-\varepsilon)F_0 + \varepsilon F_1\right) - h\left(F_0\right)}{\varepsilon} \\ &= \frac{d}{d\varepsilon} h\left(F_\varepsilon\right)_{|\varepsilon=0}. \end{aligned}$$

When $F_1 = \delta_x$ (point mass at x) this represents the limiting, normalized influence of a new observation, with value x, on the statistic $h\left(F_0\right)$. We call

$$\dot{h}(F_0; \delta_x) = IF(x) \text{ (or } IF(x; h, F_0))$$

the *Influence Function*. It can be used as a measure of the robustness of a procedure against outliers (ideally we would like it to be bounded); it can also be used to give a quick asymptotic normality proof for plug-in estimates.

– e.g. If $h(F) = E_F[X]$, then

$$\begin{aligned} h\left(\hat{F}_n\right) &= \bar{X}, \\ h\left((1-\varepsilon)F_0 + \varepsilon F_1\right) &= (1-\varepsilon)h(F_0) + \varepsilon h(F_1), \\ \dot{h}(F_0; F_1) &= \left(h(F_1) - h(F_0)\right), \end{aligned}$$

and so

$$IF(x) = \dot{h}(F_0; \delta_x) = x - E_{F_0}[X].$$

The IF is unbounded; this is further evidence of the lack of robustness of the sample average. Indeed, a single arbitrarily large outlier can push $\bar{X}$ beyond all bounds.

– In this example

$$\dot{h}(F_0; F_1) = E_{F_1}\left[IF(X)\right]$$

and

$$E_{F_0}[IF(X)] = 0;$$

these turn out to hold quite generally.

- **Asymptotic normality**. By Taylor's Theorem, expanding $h\left(F_\varepsilon\right)$ around $\varepsilon = 0$ gives

$$\begin{aligned} h\left(F_\varepsilon\right) &= h\left(F_0\right) + \dot{h}(F_0; F_1)\varepsilon + o(\varepsilon), \text{ whence} \\ h\left(F_1\right) &= h\left(F_0\right) + \dot{h}(F_0; F_1) + \text{Remainder.} \end{aligned}$$

Typically

$$\dot{h}(F_0; F_1) = E_{F_1}\left[\psi(X)\right] \tag{17.1}$$

for some function ψ. With $F_1 = \delta_x$ we obtain $IF(x) = \psi(x)$. Then with $F_1 = F_0$ we obtain

$$E_{F_0}[IF(X)] = \dot{h}(F_0; F_0) = 0.$$

Thus

$$h\left(F_1\right) = h\left(F_0\right) + E_{F_1}\left[IF(X)\right] + \text{Remainder}$$

and then, with $F_1 = \hat{F}_n$, we have (a 'Mean Value Theorem')

$$\sqrt{n}\left(h\left(\hat{F}_n\right) - h\left(F_0\right)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IF\left(X_i\right) + \sqrt{n}R_n,$$

where the $IF\left(X_i\right)$ are i.i.d. r.v.s with mean 0 and variance

$$\gamma^2\left(F_0\right) = E_{F_0}\left[IF^2\left(X\right)\right].$$

By the CLT and Slutsky's theorem, if

$$\sqrt{n}R_n \overset{pr}{\to} 0 \tag{17.2}$$

where

$$R_n = h\left(\hat{F}_n\right) - h\left(F_0\right) - \frac{1}{n}\sum_{i=1}^{n} IF\left(X_i\right),$$

we have

$$\sqrt{n}\left(h\left(\hat{F}_n\right) - h\left(F_0\right)\right) \overset{L}{\to} N\left(0, \gamma^2\left(F_0\right)\right). \tag{17.3}$$

Sometimes we just use (17.3) as a guide to what to expect, and then prove it using another technique. In other cases we can verify (17.1) and (17.2) so as to infer (17.3). The latter approach is taken in this example. Put $h(F) = E_G\left[(F(X) - G(X))^2\right]$, with

$G \neq F_0$ ($=$ the true distribution of the data). Then

$$\begin{aligned}\dot{h}(F_0; F_1) &= \frac{d}{d\varepsilon} h\left(F_\varepsilon\right)_{|\varepsilon=0} \\ &= \frac{d}{d\varepsilon} E_G\left[\left(F_\varepsilon(X) - G(X)\right)^2\right]_{|\varepsilon=0} \\ &= 2E_G\left[\left(F_\varepsilon(X) - G(X)\right)\frac{d}{d\varepsilon}F_\varepsilon(X)\right]_{|\varepsilon=0} \\ &= 2E_G\left[\left(F_0(X) - G(X)\right)\left(F_1(X) - F_0(X)\right)\right].\end{aligned}$$

With $F_1(t) = \delta_x(t) = I(t \geq x)$ we have

$$IF(x) = 2E_G\left[\left(F_0(X) - G(X)\right)\left(I(X \geq x) - F_0(X)\right)\right].$$

Note that $|IF(x)| \leq 2$ – any observation can have only bounded influence on $h(\cdot)$.

Condition (17.1): $\dot{h}(F_0; F_1) = E_{F_1}\left[\psi(X)\right]$ is easily checked (assigned). To check (17.2): $\sqrt{n}R_n \overset{pr}{\to} 0$ write

$$\begin{aligned}R_n &= \left(h\left(\hat{F}_n\right) - h\left(F_0\right)\right) - \frac{1}{n}\sum_{i=1}^{n} IF\left(X_i\right) \\ &= E_G\left[\left(\hat{F}_n(X) - G(X)\right)^2\right] - E_G\left[\left(F_0(X) - G(X)\right)^2\right] \\ &- \frac{2}{n}\sum_{i=1}^{n} E_G\left[\left(F_0(X) - G(X)\right)\left(I(X \geq X_i) - F_0(X)\right)\right].\end{aligned}$$

But

$$\begin{aligned} n^{-1}\sum_{i=1}^{n} I(X \geq X_i) &= n^{-1}(\#\text{ of times } X_i \leq X) \\ &= \hat{F}_n(X), \end{aligned}$$

so

$$\begin{aligned} \sqrt{n}R_n &= \sqrt{n}E_G\left[\left(\hat{F}_n(X) - G(X)\right)^2 - (F_0(X) - G(X))^2 \right. \\ &\quad \left. -2\,(F_0(X) - G(X))\left(\hat{F}_n(X) - F_0(X)\right)\right] \\ &= \sqrt{n}E_G\left[\left(\hat{F}_n(X) - F_0(X)\right)^2\right] \\ &\leq \left\{\sup_x \sqrt{n}\left|\hat{F}_n(x) - F_0(x)\right|\right\}^2 / \sqrt{n}. \end{aligned}$$

It can be shown that $\sup \sqrt{n}\left|\hat{F}_n(x) - F_0(x)\right| = O_P(1)$; in fact this sup has a limiting distribution: Kolmogorov showed that

$$\begin{aligned} & P_{F_0}\left(\sup_x \sqrt{n}\left|\hat{F}_n(x) - F_0(x)\right| \leq z\right) \\ \rightarrow \;& 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2z^2}; \end{aligned}$$

thus $\sqrt{n}R_n = O_P(n^{-1/2})$ and (17.3) follows.

- Note that all this assumes that $G \neq F_0$; otherwise $h(F_0) = 0$, $IF \equiv 0$ and we conclude only that
$$\sqrt{n}\left(h\left(\hat{F}_n\right) - h(F_0)\right) \overset{pr}{\to} 0.$$
If $F_0 = G$ we work instead with the second derivative $\ddot{h}(F_0; F_1)$, obtaining an n^{-1} rate of convergence to a non-normal limit. In this Cramér-von Mises example,
$$nh\left(\hat{F}_n\right) = \sum_i \left(U_{(i)} - E\left[U_{(i)}\right]\right)^2 + o_P(1)$$
has a complicated but well-studied limit distribution.

18. Bootstrapping

- Suppose that we estimate a functional $\theta = h(F)$ by $\hat{\theta}_n = h(\hat{F}_n)$ and then assess the performance through some measure $\lambda_n(F)$ which we estimate by $\lambda_n(\hat{F}_n)$. (Note that now the functional being estimated depends on n, so that the preceding theory need not apply.) Examples are

  (i) $\lambda_n(F) = P_F\left(\sqrt{n}\left(\hat{\theta}_n - h(F)\right) \leq a\right)$,

  (ii) bias $\lambda_n(F) = E_F\left[\hat{\theta}_n\right] - h(F)$,

  (iii) variance $\lambda_n(F) = E_F\left[\left(\hat{\theta}_n - E_F\left[\hat{\theta}_n\right]\right)^2\right]$.

  The plug-in estimator is obtained by replacing *every* occurrence of F by $\hat{F}_n$. In (i), F is replaced by $\hat{F}_n$ in P_F and in $h(F)$. But also $\hat{\theta}_n$ depends on F since the sample values are i.i.d. $\sim F$. We must then now sample from $\hat{F}_n$. Thus $\hat{\theta}_n$ is replaced by

  $$\theta_n^* = \hat{\theta}\left(X_1^*, ..., X_n^*\right),$$

  where the X_i^* are a random sample drawn with replacement from the data values $x_1, ..., x_n$, i.e.

independently drawn from the distribution

$$P\left(X^* = x_j\right) = n^{-1},\ j = 1, ..., n.$$

In (i) then we write $\lambda_n(\hat{F}_n)$ as

$$\begin{aligned}
\lambda_n(\hat{F}_n) &= P_{\hat{F}_n}\left(\sqrt{n}\left(\theta_n^* - \hat{\theta}_n\right) \leq a\right) \\
&= P_{\hat{F}_n}(S)\text{, where} \\
S &= \left\{(X_1^*, ..., X_n^*) \mid\ \sqrt{n}\left(\theta_n^* - \hat{\theta}_n\right) \leq a\right\}.
\end{aligned}$$

This probability can sometimes be calculated exactly – some examples are in the text. Generally it must be approximated. For this we draw a large number (B) of 'bootstrap' samples $\left(X_{b,1}^*, ..., X_{b,n}^*\right)$, $b = 1, ..., B$ from $\hat{F}_n$ and approximate $\lambda_n(\hat{F}_n)$ by the relative frequency of those in S. This requires calculating θ_n^* each time.

- In each of the examples above, we can write

$$\lambda_n(F) = E_F\left[g_n\left(\hat{\theta}_n; F\right)\right]$$

for some function $g_n(\cdot; F)$. This is to be estimated by

$$\lambda_n(\hat{F}_n) = E_{\hat{F}_n}\left[g_n\left(\theta_n^*; \hat{F}_n\right)\right],$$

which is in turn approximated by

$$\lambda_{B,n}^* = \frac{1}{B}\sum_{b=1}^{B} g_n\left(\theta_{b,n}^* \hat{F}_n\right).$$

Typically the WLLN applies:

$$\lambda_{B,n}^* \overset{pr}{\to} \lambda_n(\hat{F}_n) \text{ as } B \to \infty.$$

- We are still estimating $\lambda_n(F)$ by $\lambda_n(\hat{F}_n)$, but the latter is being approximated by $\lambda_{B,n}^*$. We use *approximate* rather than *estimate* because $\lambda_n(\hat{F}_n)$ is not an unknown parameter. Rather, it is a statistic which can in principle – but usually not in practice – be computed.

- **Example 1**. $n = 3$, $(x_1, x_2, x_3) = (8, 3, 5)$, $\theta =$ median of F, $\hat{\theta}_n = 5$. A bootstrap sample might be $\mathbf{X}^* = (8, 8, 5)$, obtained by randomly drawing three values, with replacement, from $(8, 3, 5)$. In this case $\theta_n^* = 8$. After B repetitions of the procedure we approximate $\lambda_n(\hat{F}_n)$ (still in case (i) above) by

$$\begin{aligned}\lambda_{B,n}^* &= \frac{1}{B}\sum_{b=1}^{B} I\left(\sqrt{n}\left(\theta_{b,n}^* - \hat{\theta}_n\right) \le a\right) \\ &= \frac{1}{B}\sum_{b=1}^{B} I\left(\sqrt{3}\left(\theta_{b,n}^* - 5\right) \le a\right).\end{aligned}$$

Thus $\lambda_{B,n}^* \sim \frac{1}{B} bin\left(B, \lambda_n(\hat{F}_n)\right) \underset{B\to\infty}{\overset{pr}{\to}} \lambda_n(\hat{F}_n)$.

- In this example $\lambda_n(\hat{F}_n)$ can be computed exactly – there are only $3^3 = 27$ possible resamples, with medians $\theta_n^* = (8, 3, 5)$ occurring with probabilities (under $\hat{F}_n$) of $(7, 7, 13)/27$. These are the probabilities of the three possible values of $\sqrt{3}\left(\theta_n^* - 5\right)$, and then one easily calculates the 4 possible values, as a varies, of

$$\lambda_n(\hat{F}_n) = P_{\hat{F}_n}\left[\sqrt{3}\left(\theta_n^* - 5\right) \le a\right].$$

- **Example 2**. $\lambda_n(F)$ = bias of $\hat{\theta}_n$. Write $\hat{\theta}_n = h(\hat{F}_n) = \delta\left(X_1, ..., X_n\right)$; then

  $$\begin{aligned} \lambda_n(F) &= E_F\left[\delta\left(X_1, ..., X_n\right)\right] - h(F), \\ \lambda_n(\hat{F}_n) &= E_{\hat{F}_n}\left[\delta\left(X_1^*, ..., X_n^*\right)\right] - h(\hat{F}_n). \end{aligned}$$

  To approximate the latter we draw bootstrap samples $\left(X_{b,1}^*, ..., X_{b,n}^*\right)$, compute $\delta_b^* = \delta\left(X_{b,1}^*, ..., X_{b,n}^*\right)$ each time and

  $$\lambda_{B,n}^* = \left[\frac{1}{B}\sum_{b=1}^{B}\delta_b^*\right] - \hat{\theta}_n.$$

  By the WLLN, $\lambda_{B,n}^* \overset{pr}{\to} \lambda_n(\hat{F}_n)$ as $B \to \infty$.

- There are two aspects of bootstrapping being considered here. The first is the computation of $\lambda_{B,n}^*$ to approximate $\lambda_n(\hat{F}_n)$, which we use as our finite-sample measure of the performance of $h\left(\hat{F}_n\right)$. Matters might end here if this is all we want. Or, we can ask – and will for the rest of this lecture – how good $\lambda_n(\hat{F}_n)$ is as an estimate of $\lambda_n(F)$:

  $$\lambda_{B,n}^* \xrightarrow[B\to\infty]{pr} \lambda_n(\hat{F}_n) \overset{?}{\approx} \lambda_n(F).$$

- The total error is

$$\lambda^*_{B,n} - \lambda_n(F) = \left(\lambda^*_{B,n} - \lambda_n(\hat{F}_n)\right) + E_B, \tag{18.1}$$

where

$$E_B \overset{def}{=} \lambda_n(\hat{F}_n) - \lambda_n(F)$$

is called the 'bootstrap error'. Typically $\lambda^*_{B,n} - \lambda_n(\hat{F}_n)$ can be made arbitrarily small by taking enough resamples. We say that 'the bootstrap works' if $E_B \overset{pr}{\rightarrow} 0$. This involves more than the continuity of the functional $\lambda_n(\cdot)$, since λ_n itself varies with n. If the bootstrap works, then

$$\lambda^*_{B,n} - \lambda_n(F) \overset{pr}{\rightarrow} 0 \text{ as } B, n \rightarrow \infty.$$

The order (in n) of the error (18.1) is that of E_B. If there are estimates of $\lambda_n(F)$ whose errors are of the same or smaller order, we might as well use them and dispense with the bootstrap.

- **Example 3.** Extend (i) to
$$\lambda_n(F) = P_F\left(\frac{\sqrt{n}\left(\hat{\theta}_n - h(F)\right)}{\tau(F)} \leq a\right),$$
for some scale functional $\tau(F)$. We typically have $\lambda_n(F) \to \Phi(a) \overset{def}{=} \lambda$, say, and this holds as well if $\tau(F)$ is replaced by a consistent estimate such as $\tau(\hat{F}_n)$. In such a situation we could use λ as an estimate of $\lambda_n(F)$. This is certainly easier than estimating $\lambda_n(F)$ by $\lambda_n(\hat{F}_n)$ and approximating the latter by $\lambda^*_{B,n}$. Which is more accurate?

- More generally, we typically have $\lambda_n(F) \to \lambda(F)$, and then the theory of the last few lectures tells us that this limit $\lambda(F)$ can be estimated by the plug-in estimate $\lambda\left(\hat{F}_n\right)$, with error $O_P\left(n^{-1/2}\right)$. In this case the error in estimating $\lambda_n(F)$ is
$$\begin{aligned} E_0 &\overset{def}{=} \lambda\left(\hat{F}_n\right) - \lambda_n(F) \\ &= \left[\lambda\left(\hat{F}_n\right) - \lambda(F)\right] - \left[\lambda_n(F) - \lambda(F)\right] \\ &= O_P\left(n^{-1/2}\right) - \left[\lambda_n(F) - \lambda(F)\right] \\ &\overset{pr}{\to} 0. \end{aligned}$$

Even if the bootstrap 'works', it is not clear that it works better than the more easily computed estimate $\lambda\left(\hat{F}_n\right)$ of $\lambda_n(F)$ (or merely λ, as above), i.e. whether or not the bootstrap error E_B is any smaller than E_0. This is the subject of considerable recent work – see *The Bootstrap and Edgeworth Expansion* (Peter Hall, 1992).

- First suppose that

$$\lambda_n(F) \to \lambda,$$

where λ does not depend on F. More precisely, suppose that we can expand $\lambda_n(F)$ as

$$\lambda_n(F) = \lambda + \frac{a(F)}{\sqrt{n}} + o(n^{-1/2}), \tag{18.2}$$

(e.g. Edgeworth expansions) so that

$$\lambda_n(\hat{F}_n) = \lambda + \frac{a(\hat{F}_n)}{\sqrt{n}} + o_P(n^{-1/2}). \tag{18.3}$$

Suppose also that

$$a(\hat{F}_n) = a(F) + o_P(1). \tag{18.4}$$

(True for functionals $a(\cdot)$ continuous at F, as in Lecture 17.) Then in particular, the bootstrap works: the bootstrap error E_B is

$$\begin{aligned} E_B &= \lambda_n(\hat{F}_n) - \lambda_n(F) \\ &= \frac{a(\hat{F}_n) - a(F)}{\sqrt{n}} + o_P(n^{-1/2}) \\ &= o_P(n^{-1/2}). \end{aligned}$$

Furthermore, it works better than the limiting estimate $\lambda\left(\hat{F}_n\right) = \lambda$ of $\lambda_n(F)$, which by (18.2) has error

$$E_0 = \lambda\left(\hat{F}_n\right) - \lambda_n(F) = \lambda - \lambda_n(F) = O(n^{-1/2}).$$

- The assumption that λ be independent of F is crucial here – if it is violated then E_B nmeed not be of a smaller order than the error E_0 of the more easily computed estimate $\lambda(\hat{F}_n)$.

- **Example 4**. In this example, and in the notation above, $\lambda(F)$ depends on F. The bootstrap still 'works', but no better than $\lambda(\hat{F}_n)$.

  Let $X_1, ..., X_n \overset{i.i.d.}{\sim} F$ with mean ξ_F and variance $\sigma^2 = h(F)$. The plug-in estimate is

  $$h(\hat{F}_n) = n^{-1}\sum\left(X_i - \bar{X}\right)^2 = M_2,$$

  say. Suppose $\lambda_n(F) = VAR\left[\sqrt{n}M_2\right]$. Put

  $$\mu_k = \mu_k(F) = E_F\left[(X-\xi)^k\right]$$

  (assumed finite for $k \leq 8$); then (you should verify)

  $$\lambda_n(F) = \left(\mu_4 - \mu_2^2\right) - \frac{2\left(\mu_4 - 2\mu_2^2\right)}{n} + \frac{\left(\mu_4 - 3\mu_2^2\right)}{n^2}$$

  and so $\lambda_n(F) \to \lambda(F)$, where

  $$\lambda(F) = \mu_4 - \mu_2^2 = VAR_F\left[(X - \xi_F)^2\right]$$

  depends on F. Thus we cannot expect the bootstrap error E_B to be smaller than E_0.

- Indeed, since $\mu_k(F)$ has plug-in estimate $M_k = n^{-1}\sum\left(X_i - \bar{X}\right)^k$, we have

  $$\begin{aligned} E_B &= \lambda_n(\hat{F}_n) - \lambda_n(F) \\ &= \left(M_4 - M_2^2\right) - \left(\mu_4 - \mu_2^2\right) + O_P(n^{-1}), \\ E_0 &= \lambda(\hat{F}_n) - \lambda_n(F) \\ &= \left(M_4 - M_2^2\right) - \left(\mu_4 - \mu_2^2\right) + O(n^{-1}), \end{aligned}$$

  and both are $O_P(n^{-1/2})$. (Reason:

  $$\sqrt{n}\binom{M_2 - \mu_2}{M_4 - \mu_4} = \frac{1}{\sqrt{n}}\sum\binom{(X_i - \xi)^2 - \mu_2}{(X_i - \xi)^4 - \mu_4} + o_P(1)$$

  is asymptotically normal, hence by the delta method so is $\sqrt{n}\left[\left(M_4 - M_2^2\right) - \left(\mu_4 - \mu_2^2\right)\right]$.)
  Thus we might as well use

  $$\lambda(\hat{F}_n) = M_4 - M_2^2$$

  to estimate

  $$\lambda_n(F) = VAR\left[\sqrt{n}M_2\right].$$

  - Why the condition on the first *eight* moments?

- **Example 5**. In the same framework as the previous example, suppose instead that

$$\begin{aligned}\lambda_n(F) &= |\text{bias of } M_2| \\ &= \left| E\left[\frac{n-1}{n}S^2\right] - \sigma^2 \right| \\ &= \sigma^2(F)/n.\end{aligned}$$

Then (18.2) and (18.4) are trivially satisfied, with $\lambda = a(F) = 0$, hence the bootstrap works. It has error

$$\begin{aligned}E_B &= \lambda_n(\hat{F}_n) - \lambda_n(F) \\ &= \frac{\sigma^2(\hat{F}_n) - \sigma^2(F)}{n} \\ &= \frac{M_2 - \sigma^2(F)}{n} \\ &= O_P(n^{-3/2}),\end{aligned}$$

since $\sqrt{n}\left(M_2 - \sigma^2\right) = O_P(1)$. In contrast, the error incurred by using $\lambda\left(\hat{F}_n\right) = \lambda = 0$ to estimate $\lambda_n(F)$ is of a larger order:

$$\begin{aligned}E_0 &= \lambda\left(\hat{F}_n\right) - \lambda_n(F) \\ &= -\sigma^2(F)/n \\ &= O(n^{-1}).\end{aligned}$$

Part V

LIKELIHOOD METHODS

19. Maximum likelihood: regularity, consistency

- Let $\mathbf{X}$ be the data, with density or mass function $p_\theta(\mathbf{x})$. Viewed as a function of θ – $L(\theta|\mathbf{x}) = p_\theta(\mathbf{x}_{observed})$ – this is the 'likelihood' function. The parameter value $\hat{\theta}$ which maximizes $L(\theta|\mathbf{x})$ is viewed as that which makes the observed data 'most likely to have occurred'. This value

$$\hat{\theta} = \text{arg max}\, L(\theta|\mathbf{x})$$

  is the Maximum Likelihood Estimator (MLE). We sometimes omit the dependence on $\mathbf{x}$. Put $l\,(\theta) = \log\ L(\theta)$, the *log-likelihood*. Maximizing $l\,(\theta)$ is generally easier than, and is of course equivalent to, maximizing $L(\theta)$.

- A more quantitative justification for the use of the MLE is as follows.
  Lemma: Assume the X_i are i.i.d. with density or p.m.f. $f_\theta(x)$, and that θ_0 is the true value. Define $S_n\,(\theta) = \{\mathbf{X}\,|L(\theta_0|\mathbf{X}) > L(\theta|\mathbf{X})\ \}$. Then

$$P_{\theta_0}\,(S_n\,(\theta)) \to 1 \text{ as } n \to \infty, \text{ if } \theta \neq \theta_0. \tag{19.1}$$

By this, for large samples and with high probability, $L(\theta|\mathbf{X})$ is maximized by the true parameter value, hence the maximizer of $L(\theta|\mathbf{x})$ ought to be a good estimate of this true value.

Proof of (19.1): The inequality

$$L(\theta_0|\mathbf{X}) = \prod_{i=1}^{n} f_{\theta_0}(X_i) > \prod_{i=1}^{n} f_{\theta}(X_i) = L(\theta|\mathbf{X})$$

is equivalent to

$$-\frac{1}{n}\sum_{i=1}^{n} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} > 0. \tag{19.2}$$

By the WLLN this average tends in probability to

$$\begin{aligned} & E_{\boldsymbol{\theta}_0}\left[-\log \frac{f_\theta(X)}{f_{\theta_0}(X)}\right] \\ > \quad & -\log E_{\boldsymbol{\theta}_0}\left[\frac{f_\theta(X)}{f_{\theta_0}(X)}\right] \text{ by Jensen's Inequality} \\ = \quad & -\log \int \frac{f_\theta(x)}{f_{\theta_0}(x)} f_{\theta_0}(x)dx \\ = \quad & -\log \int f_\theta(x)dx = -\log 1 = 0. \end{aligned}$$

(You should complete the proof that the probability of (19.2) → 1.)

- Maximum likelihood is the most widely used estimation method. It is often applicable when there is no other clear approach. It yields reasonable estimates which can then be modified to suit the situation at hand (e.g. assume a normal likelihood, get the MLE, then ‘robustify’ it).

- Under suitable conditions the MLE is consistent, asymptotically normal, minimum variance. But see examples 7.1.2, 7.1.3 for cases in which the MLE may not exist, or may be inconsistent.

- Assume:
  (C1) Identifiability: $\theta_1 \neq \theta_2 \Rightarrow$ the distributions $P_{\theta_1}, P_{\theta_2}$ are distinct. (This is violated if, e.g., $X_1, ..., X_n \sim N(\theta, 1)$ but only $Y = |X|$ is observed, then $\pm\theta$ both lead to the same distributions of Y: $f_\theta(y) = [\phi(y-\theta) + \phi(y+\theta)]\, I(y > 0)$.)
  (C2) The parameter space is an open interval: $\Omega = \left(\underline{\theta}, \bar{\theta}\right)$ (possibly infinite). (Ensures that maxima of the likelihood function correspond to critical points, assuming differentiability.)
  (C3) The observations $X_1, ..., X_n$ are i.i.d. with density or p.m.f. $f_\theta(x)$. (Then $l(\theta) = \sum \log f_\theta(x_i)$.)
  (C4) The support $A = \{x \mid f_\theta(x) > 0\}$ is independent of θ. (This excludes cases like $X_i \sim U(0, \theta)$, which can be handled using other techniques.)
  (C5) For all $x \in A$, $f_\theta(x)$ is differentiable w.r.t. θ, with derivative $f'_\theta(x)$. (Then maxima of l satisfy the *likelihood equation* $l'(\theta) = \sum \frac{f'_\theta(x_i)}{f_\theta(x_i)} = 0$.)

Theorem: Under (C1)-(C5) there exists a sequence $\hat{\theta}_n = \hat{\theta}_n (X_1, ..., X_n)$ of roots of the likelihood equation such that
(i) $P\left(\hat{\theta}_n \text{ is a local maximum of } l_n(\theta)\right) \to 1$ and
(ii) $\hat{\theta}_n \overset{pr}{\to} \theta_0$.

Proof: Let $a > 0$ be small enough that $\underline{\theta} < \theta_0 - a < \theta_0 + a < \bar{\theta}$, but otherwise arbitrary. (Is this possible? Why?) In the notation of the lemma, define $S_n = S_n (\theta_0 - a) \cap S_n (\theta_0 + a)$. Then $P_{\theta_0} (\mathbf{X} \in S_n) \to 1$. For $\mathbf{x} \in S_n$ there exists $\theta_n^* \in (\theta_0 - a, \theta_0 + a)$ at which $l(\theta|\mathbf{x})$ has a local maximum. This establishes the existence of a sequence $\theta_n^* = \theta_n^*(a)$ of roots of $l'(\theta|\mathbf{x}) = 0$, corresponding to local maxima, with

$$P_{\theta_0} (|\theta_n^* - \theta_0| < a) \to 1.$$

We cannot yet conclude that $\theta_n^* \overset{pr}{\to} \theta_0$, because of the dependence on a. However, define $\hat{\theta}_n$ to be that root, corresponding to a local maximum, which is *closest to* θ_0. Then (with probability approaching 1) $\left|\hat{\theta}_n - \theta_0\right| < a$, and $\hat{\theta}_n$ does not depend on the choice of a, so that $\hat{\theta}_n \overset{pr}{\to} \theta_0$. □

- Under stronger conditions one can show the existence and consistency of the MLE, i.e. of a *global* maximum of $l_n(\theta)$. However, one typically computes a root $\hat{\theta}_n$ of the likelihood equation which may be only a local maximum.

  Corollary: With conditions as above, if $l_n'(\theta)$ has a unique zero $\hat{\theta}_n$ for all (sufficiently large) n and all $x_1, ..., x_n$ then:
  (i) $\hat{\theta}_n \overset{pr}{\rightarrow} \theta_0$;
  (ii) $P\left(\hat{\theta}_n \text{ is the MLE}\right) \rightarrow 1$ as $n \rightarrow \infty$.
  Proof: (i): The uniqueness implies that the zeros $\hat{\theta}_n$ of $l_n'(\theta)$ correspond to the $\hat{\theta}_n$ in the statement of the theorem.
  (ii): With probability tending to 1, $\hat{\theta}_n$ is at least a local maximum, not a saddlepoint or a minimum. Suppose that $\hat{\theta}_n$ is only a local maximum, and not the MLE. Then we must have $l_n(\theta') > l_n\left(\hat{\theta}_n\right)$ for some $\theta' \neq \hat{\theta}_n$; then there is a local minimum between them at which $l' = 0$, contradicting the uniqueness of $\hat{\theta}_n$. Thus (with probability tending to 1) $\hat{\theta}_n$ is the global maximizer, i.e. is the MLE.

20. Fisher information; information inequality

- **Fisher Information**. Under suitable conditions it turns out that the MLE $\hat{\theta}_n$ (more generally, any consistent sequence of roots of the likelihood equation) satisfies

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{L} N\left(0, I^{-1}\left(\theta_0\right)\right), \tag{20.1}$$

where

$$I\left(\theta\right) = E_\theta\left[\left(\frac{\partial \log f_\theta(X)}{\partial\theta}\right)^2\right]$$

is the 'Fisher information' in one observation. In the following we assume f is the density; proofs for p.m.f.'s are identical. Suppose

(C6) The function $f_\theta(x)$ is three times continuously differentiable w.r.t. θ, and the lhs of

$$\int f_\theta(x)dx = 1$$

can be differentiated thrice under the integral sign. See Lemma 7.3.1 for mild conditions ensuring this.

Then (here we use only twice; thrice required later in proof of asymptotic normality):

1. Differentiating once gives

$$0 = \int f_\theta'(x)dx = \int \frac{f_\theta'}{f_\theta}(x) f_\theta(x) dx,$$

so that $E_\theta\left[\frac{\partial \log f_\theta(X)}{\partial\theta}\right] = 0$ and

$$I(\theta) = VAR_\theta\left[\frac{\partial \log f_\theta(X)}{\partial\theta}\right].$$

The 'information in a sample' of size n is then

$$\begin{aligned} VAR\left[\frac{\partial \log L(\theta|\mathbf{X})}{\partial\theta}\right] &= VAR\left[\sum \frac{\partial \log f_\theta(X_i)}{\partial\theta}\right] \\ &= nI(\theta). \end{aligned}$$

(Also $E_\theta\left[\frac{\partial \log L(\theta|\mathbf{X})}{\partial\theta}\right] = 0$.)

2. Differentiating again:

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int \frac{\partial \log f_\theta(x)}{\partial\theta} f_\theta(x)dx \\ &= \int \frac{\partial^2 \log f_\theta(x)}{\partial\theta^2} f_\theta(x)dx + \int \frac{\partial \log f_\theta(x)}{\partial\theta} f'_\theta(x)dx \\ &= E_\theta\left[\frac{\partial^2 \log f_\theta(X)}{\partial\theta^2}\right] + E_\theta\left[\left(\frac{\partial \log f_\theta(X)}{\partial\theta}\right)^2\right], \end{aligned}$$

so that

$$I(\theta) = E_\theta\left[-\frac{\partial^2 \log f_\theta(X)}{\partial\theta^2}\right].$$

3. $I(g(\theta)) = I(\theta)\Big/[g'(\theta)]^2$ – proof outlined in text. The MLE of $g(\theta)$ is $g\left(\hat{\theta}_n\right)$. If (20.1) holds then by the delta method

$$\begin{aligned} &\sqrt{n}\left(g\left(\hat{\theta}_n\right) - g(\theta_0)\right) \xrightarrow{L} \\ &N\left(0, \left[g'(\theta_0)\right]^2 I^{-1}(\theta_0) = I^{-1}(g(\theta_0))\right). \end{aligned}$$

- **Information Inequality**. (Cramér-Rao Inequality.) Assume (C1) - (C6). Let $\delta(\mathbf{X})$ be an unbiased estimator of $g(\theta)$, so that

  $$g(\theta) = \int \delta(\mathbf{x}) L(\theta|\mathbf{x})\, d\mathbf{x}.$$

  Then, assuming that the following interchange of operations is permissible (the main requirement is the continuity, in θ, of the integrand and of its absolute value, after differentiating),

  $$\begin{aligned} g'(\theta) &= \int \delta(\mathbf{x}) \frac{L'(\theta|\mathbf{x})}{L(\theta|\mathbf{x})} L(\theta|\mathbf{x})\, d\mathbf{x} \\ &= E_\theta \left[\delta(\mathbf{X}) \frac{\partial \log L(\theta|\mathbf{X})}{\partial \theta} \right] \\ &= COV \left[\delta(\mathbf{X}), \frac{\partial \log L(\theta|\mathbf{X})}{\partial \theta} \right]. \end{aligned}$$

  Then

  $$\begin{aligned} \left[g'(\theta) \right]^2 &\leq VAR_\theta [\delta(\mathbf{X})]\, VAR \left[\frac{\partial \log L(\theta|\mathbf{X})}{\partial \theta} \right] \\ &= VAR_\theta [\delta(\mathbf{X}) - g(\theta)] \cdot nI(\theta) \\ &= VAR_\theta \left[\sqrt{n} (\delta(\mathbf{X}) - g(\theta)) \right] \cdot I(\theta), \end{aligned}$$

and so

$$\begin{aligned} VAR\left[\sqrt{n}\left(\delta(\mathbf{X})-g\left(\theta\right)\right)\right] &\geq \left[g'\left(\theta\right)\right]^2/I\left(\theta\right) \\ &= I^{-1}\left(g\left(\theta\right)\right), \end{aligned}$$

i.e. the variance of any unbiased estimate exceeds the asym. var. of $\sqrt{n}\left(g\left(\hat{\theta}_n\right)-g\left(\theta\right)\right)$, the MLE. (Note however that there are examples in which it does not exceed the *limiting* variance.)

Harald Cramér (September 25, 1893 - October 5, 1985)

- **Example**: $X_1, ..., X_n$ the indicators of n Bernoulli trials: $\theta = P(X_i = 1) \in (0,1)$. Then $f_\theta(x) = \theta^x (1-\theta)^{1-x}$ for $x \in \{0,1\}$ with

$$\begin{aligned}
\log f_\theta(x) &= x \log\theta + (1-x)\log(1-\theta),\\
\frac{\partial \log f_\theta(x)}{\partial\theta} &= \frac{x}{\theta} - \frac{1-x}{1-\theta},\\
\frac{\partial^2 \log f_\theta(x)}{\partial\theta^2} &= \frac{-x}{\theta^2} - \frac{1-x}{(1-\theta)^2},\\
E_\theta\left[\frac{\partial^2 \log f_\theta(X)}{\partial\theta^2}\right] &= \frac{-1}{\theta(1-\theta)};
\end{aligned}$$

hence $I^{-1}(\theta) = \theta(1-\theta) =$ asym. var. of $\sqrt{n}\left(\hat{\theta}_n - \theta\right)$ where (you should verify) $\hat{\theta}_n = \sum X_i/n$. (Note that (C2) – Ω an open interval – excludes the values $\theta = 0, 1$. But in these cases the MLEs exist and are $\hat{\theta} = 0, 1$, with $I^{-1}(\theta) = 0$.)

21. Asymptotics of likelihood estimation

- **Asymptotic normality**. Assume (C1) - (C6) and (C7) There exists $c(\theta_0) > 0$ and a function $M_{\theta_0}(x)$ such that $\left|\frac{\partial^3}{\partial\theta^3} \log f_\theta(x)\right| \le M_{\theta_0}(x)$ for all $x \in A$ and all θ satisfying $|\theta - \theta_0| < c(\theta_0)$, and $E_{\theta_0}\left[M_{\theta_0}(X)\right] < \infty$.
  Read Example 7.3.1 to see how these conditions can be checked in practice. The purpose of (C7) is:

  Lemma: Under (C1)-(C7) , if $\hat{\theta}_n \overset{pr}{\to} \theta_0$ and θ_n^* is between $\hat{\theta}_n$ and θ_0, then $l_n'''(\theta_n^*)/n$ is $O_P(1)$.
  Proof:
  $$\left|\frac{l_n'''(\theta_n^*)}{n}\right| = \left|\frac{1}{n}\sum\left(\frac{\partial^3}{\partial\theta^3}\log f_{\theta_n^*}(X_i)\right)\right| \le \frac{1}{n}\sum M_{\theta_0}(X_i)$$
  as long as $|\theta_n^* - \theta_0| < c(\theta_0)$, hence (since $|\theta_n^* - \theta_0| \le \left|\hat{\theta}_n - \theta_0\right|$) with probability tending to 1. Since
  $$\frac{1}{n}\sum M_{\theta_0}(X_i) \overset{pr}{\to} E_{\theta_0}\left[M_{\theta_0}(X)\right] < \infty,$$
  $l_n'''(\theta_n^*)/n$ is $O_P(1)$. ($P(|A_n| \le B_n) \to 1$ and $B_n \overset{pr}{\to} b < \infty \Rightarrow A_n = O_P(1)$ – sample midterm.)

Theorem: Under (C1)-(C7) and the condition $0 < I\left(\theta_0\right) < \infty$ we have that

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{L} N\left(0, I^{-1}\left(\theta_0\right)\right),$$

where $\left\{\hat{\theta}_n\right\}$ is *any* consistent sequence of roots of the likelihood equation.

Proof: By Taylor's Theorem,

$$\begin{aligned} 0 &= l_n'\left(\hat{\theta}_n\right) \\ &= l_n'\left(\theta_0\right) + \left(\hat{\theta}_n - \theta_0\right) l_n''\left(\theta_0\right) + \frac{1}{2}\left(\hat{\theta}_n - \theta_0\right)^2 l_n'''\left(\theta_n^*\right) \\ &= l_n'\left(\theta_0\right) + \left(\hat{\theta}_n - \theta_0\right) \cdot \left[\begin{array}{c} l_n''\left(\theta_0\right) + \\ \frac{1}{2}\left(\hat{\theta}_n - \theta_0\right) l_n'''\left(\theta_n^*\right) \end{array}\right], \end{aligned}$$

for some θ_n^* between $\hat{\theta}_n$ and θ_0. Thus

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \frac{l_n'\left(\theta_0\right)/\sqrt{n}}{-\frac{l_n''\left(\theta_0\right)}{n} - \frac{\left(\hat{\theta}_n - \theta_0\right) l_n'''\left(\theta_n^*\right)}{2n}}.$$

By the preceding lemma and the consistency of $\hat{\theta}_n$ we have that $\left(\hat{\theta}_n - \theta_0\right) l_n'''\left(\theta_n^*\right)/n \xrightarrow{pr} 0$, so that it suffices (Slutsky!) to establish:

$$\begin{array}{ll}\text{(i)} & l_n'(\theta_0)/\sqrt{n} \xrightarrow{L} N(0, I(\theta_0)), \\ \text{(ii)} & -\dfrac{l_n''(\theta_0)}{n} \xrightarrow{pr} I(\theta_0) > 0.\end{array}$$

Proof of (i): $l_n'(\theta_0)/\sqrt{n} = \frac{1}{\sqrt{n}} \sum \frac{\partial}{\partial\theta} \log f_\theta(X_i)_{|\theta=\theta_0}$, a normalized sum of i.i.d. r.v.s with mean 0 and variance $I(\theta_0)$.

Proof of (ii):

$$\begin{aligned}-\tfrac{l_n''(\theta_0)}{n} = \; & \frac{1}{n}\sum\left(-\frac{\partial^2}{\partial\theta^2}\log f_\theta(X_i)\right)_{|\theta_0} \\ & \xrightarrow{pr} E_{\theta_0}\left[-\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\right] = I(\theta_0).\end{aligned}$$

- There may be multiple roots of the likelihood equation, in which case the theory so far does not tell us which one to choose. This can be dealt with as follows. Assume that we have some estimator $\tilde{\theta}_n$ which is $\sqrt{n}$-consistent, i.e. $\sqrt{n}\left(\tilde{\theta}_n - \theta_0\right)$ is $O_P(1)$. This holds in particular if $\sqrt{n}\left(\tilde{\theta}_n - \theta_0\right)$

has a limit distribution, so that for instance a plug-in estimator could be used. Now perform one step of the Newton-Raphson method to solve $l_n'(\theta) = 0$, starting with $\tilde{\theta}_n$:

$$\delta_n = \tilde{\theta}_n - \frac{l_n'\left(\tilde{\theta}_n\right)}{l_n''\left(\tilde{\theta}_n\right)}. \tag{21.1}$$

We will show that, under these conditions (and the assumptions of the previous theorem),

$$\sqrt{n}\left(\delta_n - \theta_0\right) \xrightarrow{L} N\left(0, I^{-1}\left(\theta_0\right)\right). \tag{21.2}$$

Note that by this, δ_n is itself $\sqrt{n}$-consistent and so can be used as the starting point for one more iteration. A common prescription is the '3-step' estimate – do three steps of Newton-Raphson, starting with $\tilde{\theta}_n$. (But why not 'Three steps if I like what I see, a few more otherwise'?)

Proof of (21.2): By Taylor's Theorem,

$$l_n'\left(\tilde{\theta}_n\right) =$$

$$l_n'(\theta_0) + \left(\tilde{\theta}_n - \theta_0\right) l_n''(\theta_0) + \frac{1}{2}\left(\tilde{\theta}_n - \theta_0\right)^2 l_n'''(\theta_n^*),$$

with θ_n^* between θ_0 and (the consistent estimate) $\tilde{\theta}_n$. Substituting into (21.1) gives

$$\sqrt{n}(\delta_n - \theta_0) = \sqrt{n}\left(\tilde{\theta}_n - \theta_0\right) - \frac{\sqrt{n}\left\{\begin{array}{c} l_n'(\theta_0) + \left(\tilde{\theta}_n - \theta_0\right) l_n''(\theta_0) + \\ \frac{1}{2}\left(\tilde{\theta}_n - \theta_0\right)^2 l_n'''(\theta_n^*) \end{array}\right\}}{l_n''\left(\tilde{\theta}_n\right)}$$

$$= \sqrt{n}\left(\tilde{\theta}_n - \theta_0\right)\left[1 - \frac{l_n''(\theta_0)}{l_n''\left(\tilde{\theta}_n\right)} - \frac{1}{2}\left(\tilde{\theta}_n - \theta_0\right)\frac{l_n'''(\theta_n^*)}{l_n''\left(\tilde{\theta}_n\right)}\right] - \frac{l_n'(\theta_0)/\sqrt{n}}{l_n''\left(\tilde{\theta}_n\right)/n}.$$

Claim:

$$\text{(i)} \qquad \frac{l_n'(\theta_0)/\sqrt{n}}{-l_n''(\theta_0)/n} \xrightarrow{L} N\left(0, I^{-1}(\theta_0)\right),$$

$$\text{(ii)} \qquad \frac{l_n''\left(\tilde{\theta}_n\right)}{l_n''(\theta_0)} \xrightarrow{pr} 1,$$

$$\text{(iii)} \qquad \left[1 - \frac{l_n''(\theta_0)}{l_n''\left(\tilde{\theta}_n\right)} - \frac{1}{2}\left(\tilde{\theta}_n - \theta_0\right)\frac{l_n'''(\theta_n^*)}{l_n''\left(\tilde{\theta}_n\right)}\right] \xrightarrow{pr} 0.$$

By (iii) and the $\sqrt{n}$-consistency of $\tilde{\theta}_n$, then (i), (ii) and Slutsky's Theorem,

$$\begin{aligned}
\sqrt{n}(\delta_n - \theta_0) &= \frac{l_n'(\theta_0)/\sqrt{n}}{-l_n''\left(\tilde{\theta}_n\right)/n} + o_P(1) \\
&= \frac{l_n'(\theta_0)/\sqrt{n}}{-l_n''(\theta_0)/n} \Bigg/ \frac{l_n''\left(\tilde{\theta}_n\right)}{l_n''(\theta_0)} + o_P(1) \\
&\xrightarrow{L} N\left(0, I^{-1}(\theta_0)\right).
\end{aligned}$$

Proof of (i): As above.
Proof of (ii): $l_n''\left(\tilde{\theta}_n\right) = l_n''(\theta_0) + \left(\tilde{\theta}_n - \theta_0\right) l_n'''(\theta_n^{**})$, and θ_n^{**} satisfies the condition of the lemma, so that

$$\frac{l_n''\left(\tilde{\theta}_n\right)}{l_n''(\theta_0)} = 1 + \left(\tilde{\theta}_n - \theta_0\right)\frac{l_n'''(\theta_n^{**})/n}{l_n''(\theta_0)/n},$$

where $\tilde{\theta}_n - \theta_0 = o_P(1)$, $l_n''(\theta_0)/n \overset{pr}{\to} -I(\theta_0) < 0$, and $l_n'''(\theta_n^{**})/n$ is $O_P(1)$.

Proof of (iii): By (ii) and the $\sqrt{n}$-consistency of $\tilde{\theta}_n$ we need to show that

$$\frac{l_n'''(\theta_n^*)}{l_n''\left(\tilde{\theta}_n\right)} = \frac{l_n'''(\theta_n^*)/n}{l_n''(\theta_0)/n} \Big/ \frac{l_n''\left(\tilde{\theta}_n\right)/n}{l_n''(\theta_0)/n}$$

is $O_P(1)$. But by (ii) the denominator on the rhs $\overset{pr}{\to} 1$, and the numerator is $O_P(1)$ by the lemma and the assumption $I(\theta_0) > 0$.

Remark: We have also shown that if $\left\{\hat{\theta}_n\right\}$ is a consistent sequence of roots of the likelihood equation, then

$$\sqrt{n}(\delta_n - \theta_0) = \sqrt{n}\left(\hat{\theta}_n - \theta_0\right) + o_P(1).$$

The reason is that both $= \sqrt{n} l_n'(\theta_0)/[-l_n''(\theta_0)] + o_P(1)$.

- Example. $X_1, ..., X_n \sim$ Logistic, with

$$f_\theta(x) = \frac{e^{-(x-\theta)}}{\left(1 + e^{-(x-\theta)}\right)^2} = \frac{1}{4\cosh^2\left(\frac{x-\theta}{2}\right)}.$$

Define the 'scores'

$$\psi\left(x-\theta\right)=\frac{\partial}{\partial\theta}\log f_\theta(x)=\tanh\left(\frac{x-\theta}{2}\right).$$

Then the likelihood equation is

$$0=l_n'\left(\theta\right)=\sum\psi\left(x_i-\theta\right),$$

and since $l_n'\left(\theta\right)$ is a strictly decreasing function of θ with $l_n'\left(\mp\infty\right)=\pm n$, there is a unique root which is necessarily the MLE. The one-step estimate obtained above is

$$\delta_n=\tilde{\theta}_n-\frac{l_n'\left(\tilde{\theta}_n\right)}{l_n''\left(\tilde{\theta}_n\right)}=\tilde{\theta}_n+\frac{\sum\psi\left(x_i-\tilde{\theta}_n\right)}{\sum\psi'\left(x_i-\tilde{\theta}_n\right)}.$$

One can iterate to convergence to obtain the MLE. By Problem 3.5 (assigned) one could instead replace $\sum\psi'\left(x_i-\tilde{\theta}_n\right)/n$ by $I(\tilde{\theta}_n)=1/3$:

$$\delta_n=\tilde{\theta}_n+\frac{3}{n}\sum\psi\left(x_i-\tilde{\theta}_n\right).$$

In either case $\sqrt{n}\left(\delta_n-\theta_0\right)\xrightarrow{L}N\left(0,I^{-1}\left(\theta_0\right)=3\right)$. A possible choice of $\tilde{\theta}_n$ is the sample median.

22. Efficiency; multiparameter estimation; method of moments

- **Efficiency**. Recall that if $\hat{\theta}_n$ is the MLE (or a one-step approximation) then
$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{L} N\left(0, I^{-1}(\theta)\right)$$
and $I^{-1}(\theta) \leq VAR[\sqrt{n}(\delta_n - \theta)]$ for any *unbiased* δ_n. Consider now estimates δ_n of θ, not necessarily unbiased in finite samples but satisfying
$$\sqrt{n}(\delta_n - \theta) \xrightarrow{L} N(0, v(\theta))$$
with $v(\theta)$ *continuous* and $0 < v(\theta) < \infty$. Under (C1)-(C7) it can be shown that, if $I(\theta)$ is also continuous, then $v(\theta) \geq I^{-1}(\theta)$. An estimator attaining this lower bound is called *efficient*. In particular, the MLE is efficient if $I(\theta)$ is continuous. By the delta method, if δ_n is efficient for θ then $g(\delta_n)$ is efficient for $g(\theta)$ at all points θ where $g'(\theta) \neq 0$, since
$$\sqrt{n}(g(\delta_n) - g(\theta)) \xrightarrow{L} N\left(0, \frac{[g'(\theta)]^2}{I(\theta)} = I^{-1}(g(\theta))\right).$$

- The MLE is not the only efficient estimator, for instance '$MLE + o_P\left(n^{-1/2}\right)$' is also efficient. A less trivial example is if $X_1, ..., X_n \sim N\left(\mu = a\sigma, \sigma^2\right)$ for a *known* a $(= 1/cv)$. To obtain the MLE $\hat{\sigma}$ it is convenient to first obtain $\hat{\gamma}$, where $\gamma = 1/\sigma$. We define $m_2 = \sum X_i^2/n$ and calculate

$$\begin{aligned}
L(\sigma) &\propto \sigma^{-n} \operatorname{exp}\left\{-\frac{1}{2}\sum\frac{(X_i - a\sigma)^2}{\sigma^2}\right\} \\
&= \gamma^n \operatorname{exp}\left\{-\frac{1}{2}\sum(\gamma X_i - a)^2\right\}, \\
l(\gamma) &= const. + n\log\gamma - \frac{1}{2}\gamma^2 n m_2 + \gamma a n\bar{X}, \\
l'(\gamma) &= n\left\{\frac{1}{\gamma} - \gamma m_2 + a\bar{X}\right\},
\end{aligned}$$

and then

$$\begin{aligned}
\hat{\gamma} &= \frac{a\bar{X} + \sqrt{\left(a\bar{X}\right)^2 + 4m_2}}{2m_2}, \\
\hat{\sigma} &= \frac{1}{\hat{\gamma}} = \sqrt{\left(\frac{a}{2}\bar{X}\right)^2 + m_2} - \frac{a}{2}\bar{X}.
\end{aligned}$$

The Fisher information about γ is

$$I(\gamma) = \frac{1}{n} E\left[-l''(\gamma)\right] = \sigma^2\left(2 + a^2\right),$$

and so the information about $\sigma = g(\gamma)$ (with $g(\gamma) = 1/\gamma$ and $g'(\gamma) = -\sigma^2$) is

$$I(\sigma) = I(\gamma) / \left[g'(\gamma)\right]^2 = \frac{2 + a^2}{\sigma^2}.$$

Thus $\sqrt{n}(\hat{\sigma} - \sigma) \xrightarrow{L} N\left(0, I^{-1}(\sigma) = \sigma^2/\left(2 + a^2\right)\right)$. Two other estimates of σ $(= \mu/a)$ are $\delta_1 = \bar{X}/a$ (with $\sqrt{n}(\delta_1 - \sigma) \xrightarrow{L} N\left(0, \sigma^2/a^2\right)$) and $\delta_2 = S$:

$$\begin{aligned} &\sqrt{n}\left(S^2 - \sigma^2\right) \xrightarrow{L} N\left(0, 2\sigma^4\right) \\ \Rightarrow\ &\sqrt{n}(S - \sigma) \xrightarrow{L} N\left(0, \sigma^2/2\right). \end{aligned}$$

Of course each is asymptotically more highly varied than the MLE, but consider $\delta_\alpha = (1 - \alpha)\delta_1 + \alpha\delta_2$.

Since δ_1 and δ_2 are independent, we have

$$\begin{aligned} &\sqrt{n}(\delta_\alpha - \sigma) = (1 - \alpha)\sqrt{n}(\delta_1 - \sigma) + \alpha\sqrt{n}(\delta_2 - \sigma) \\ &\xrightarrow{L} N\left(0, (1 - \alpha)^2 \frac{\sigma^2}{a^2} + \alpha^2 \frac{\sigma^2}{2}\right). \end{aligned}$$

The variance-minimizing choice of α is $\alpha^* = 2/\left(2+a^2\right)$, and

$$\delta_{\alpha^*} = \frac{a\bar{X} + 2S}{2+a^2}$$

has variance $\sigma^2/\left(2+a^2\right) = I^{-1}(\sigma)$. Thus δ_{α^*} is also efficient (and simpler).

- This idea of reducing the variance through a linear combination of estimators is a special case ($\rho = 0$) of the following.

  Lemma: Suppose that $\hat{\theta}_1, \hat{\theta}_2$ are unbiased estimates of θ, with covariance matrix

  $$\mathbf{\Sigma} = \left(\begin{array}{cc} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ * & \sigma_2^2 \end{array} \right).$$

  Then

  $$\hat{\theta}_\alpha = (1-\alpha)\,\hat{\theta}_1 + \alpha\hat{\theta}_2$$

  is unbiased, with variance

  $$\sigma_\alpha^2 = (1-\alpha)^2\,\sigma_1^2 + 2\alpha\,(1-\alpha)\,\rho\sigma_1\sigma_2 + \alpha^2\sigma_2^2.$$

This is a convex function of α which is minimized by

$$\alpha^* = \frac{\left(\frac{\sigma_1}{\sigma_2} - \rho\right)}{\left(\frac{\sigma_1}{\sigma_2} - \rho\right) + \left(\frac{\sigma_2}{\sigma_1} - \rho\right)};$$

so $\alpha^* \in [0,1]$ iff

$$\rho \leq \min\left(\sigma_1/\sigma_2, \sigma_2/\sigma_1\right),$$

and in particular if $\rho \leq 0$.
The minimum variance is, with $\mathbf{1} = (1,1)'$,

$$\sigma^2_{\alpha^*} = \frac{1}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}} = \frac{\left(1-\rho^2\right)\sigma_1^2\sigma_2^2}{\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2} \leq \min\left(\sigma_1^2, \sigma_2^2\right),$$

The reduction in variance is shown by

$$\begin{aligned} \sigma^2_{\alpha^*} &= \sigma_1^2 - \frac{\sigma_1^2\sigma_2^2\left(\frac{\sigma_1}{\sigma_2} - \rho\right)^2}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}\cdot|\boldsymbol{\Sigma}|}, \\ \sigma^2_{\alpha^*} &= \sigma_2^2 - \frac{\sigma_1^2\sigma_2^2\left(\frac{\sigma_2}{\sigma_1} - \rho\right)^2}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}\cdot|\boldsymbol{\Sigma}|}. \end{aligned}$$

(Why is this not a contradiction if $\hat{\theta}_1$ is already a minimum variance estimator?)

- **Multiparameter likelihood estimation**. Assume $X_1, ..., X_n \overset{i.i.d.}{\sim} f(x, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is $p \times 1$. The X_i may be vector-valued as well. We make assumptions similar to (C1)-(C7), e.g. support independent of the parameters, differentiation under integrals is permissible, etc. Define

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum \log f(X_i, \boldsymbol{\theta}), \\ \dot{l}(\boldsymbol{\theta})_{p\times 1} &= \sum \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_i, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}), \\ \mathbf{I}(\boldsymbol{\theta})_{p\times p} &= E\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(X, \boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(X, \boldsymbol{\theta})\right)'\right]. \end{aligned}$$

The Fisher Information matrix $\mathbf{I}(\boldsymbol{\theta})$ will be assumed to be positive definite. Under the regularity conditions we obtain, in much the same manner as before,

$$\begin{aligned} E\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f(X, \boldsymbol{\theta})\right] &= \mathbf{0} = E\left[\dot{l}(\boldsymbol{\theta})\right], \\ COV\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f(X, \boldsymbol{\theta})\right] &= \mathbf{I}(\boldsymbol{\theta}) = \frac{1}{n} COV\left[\dot{l}(\boldsymbol{\theta})\right], \end{aligned}$$

and

$$\mathbf{I}(\boldsymbol{\theta}) = E\left[-\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f(X, \boldsymbol{\theta})\right] = \frac{1}{n} E\left[-\ddot{l}(\boldsymbol{\theta})\right],$$

where the Hessian $\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f\left(X,\boldsymbol{\theta}\right)$ has $(j,k)^{th}$ element $\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f\left(x,\boldsymbol{\theta}\right)$ and

$$\ddot{l}\left(\boldsymbol{\theta}\right)_{p\times p}=\frac{\partial}{\partial\boldsymbol{\theta}}\dot{l}\left(\boldsymbol{\theta}\right)=\sum\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f\left(X_i,\boldsymbol{\theta}\right).$$

As before, it can be shown that with probability approaching 1 there exists a consistent sequence of roots of the likelihood equations $\dot{l}\left(\boldsymbol{\theta}\right)=\mathbf{0}$, corresponding to local maxima of the likelihood function. Any such sequence satisfies

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\xrightarrow{L}N\left(0,\mathbf{I}^{-1}\left(\theta_0\right)\right).$$

The estimate

$$\hat{\boldsymbol{\theta}}_n=\widetilde{\boldsymbol{\theta}}_n-\left[\ddot{l}\left(\widetilde{\boldsymbol{\theta}}_n\right)\right]^{-1}\dot{l}\left(\widetilde{\boldsymbol{\theta}}_n\right),$$

obtained by doing one step of Newton-Raphson, starting with a $\sqrt{n}$-consistent initial estimate $\widetilde{\boldsymbol{\theta}}_n$, has this same limit distribution.

- Often $\mathbf{I}\left(\boldsymbol{\theta}_0\right)$ is replaced by its consistent estimate $-\ddot{l}\left(\hat{\boldsymbol{\theta}}_n\right)/n$ (‘observed information’) in order to make inferences; by Slutsky’s Theorem this does not affect the asymptotic properties.

- **Efficiency**. Under appropriate conditions, if $\boldsymbol{\delta}_n$ is a sequence of estimators satisfying

$$\sqrt{n}\left(\boldsymbol{\delta}_n - \boldsymbol{\theta}_0\right) \xrightarrow{L} N\left(\mathbf{0}, \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)\right)$$

then

$$\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right) \geq \mathbf{I}^{-1}\left(\boldsymbol{\theta}_0\right)$$

in the sense that $\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right) - \mathbf{I}^{-1}\left(\boldsymbol{\theta}_0\right)$ is positive semi-definite. (The conditions include the requirement that $\boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right)$ and $\mathbf{I}\left(\boldsymbol{\theta}\right)$ be continuous in $\boldsymbol{\theta}$.) Thus for any $\mathbf{c}_{p\times 1}$, both $\mathbf{c}'\hat{\boldsymbol{\theta}}_n$ and $\mathbf{c}'\boldsymbol{\delta}_n$ are AN estimators of $\mathbf{c}'\boldsymbol{\theta}_0$, but the asymptotic variance of $\sqrt{n}\mathbf{c}'\boldsymbol{\delta}_n$ exceeds the asymptotic variance of $\sqrt{n}\mathbf{c}'\hat{\boldsymbol{\theta}}_n$:

$$\mathbf{c}'\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)\mathbf{c} \geq \mathbf{c}'\mathbf{I}^{-1}\left(\boldsymbol{\theta}_0\right)\mathbf{c}.$$

An estimator attaining the lower bound $\mathbf{I}^{-1}\left(\boldsymbol{\theta}_0\right)$ is *efficient*.

- Example. A process produces units $X \sim N(0,1)$ when in control (w.p. p), and $X \sim N(\xi,1)$ ($\xi \neq 0$) when not in control (w.p. $1-p$). Then

  $X \sim f(x;\boldsymbol{\theta}) = p\phi(x)+(1-p)\phi(x-\xi)$, $\boldsymbol{\theta} = (p,\xi)'$.

  We calculate

  $\dfrac{\partial f}{\partial p} = \phi(x)-\phi(x-\xi)$, $\dfrac{\partial f}{\partial \xi} = (1-p)(x-\xi)\phi(x-\xi)$,

  so

  $$\dot{l}\left(\boldsymbol{\theta}\right) = \sum_{i=1}^{n}\left\{\left(\begin{array}{c}\phi(x_i)-\phi(x_i-\xi)\\(1-p)(x_i-\xi)\phi(x_i-\xi)\end{array}\right)/f(x_i;\boldsymbol{\theta})\right\}$$

  and we seek a solution to $\dot{l}\left(\boldsymbol{\theta}\right) = \mathbf{0}$. If $\widetilde{\boldsymbol{\theta}}_n$ is $\sqrt{n}$-consistent and $\hat{\boldsymbol{\theta}}_n$ is the one-step improvement, then $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{L} N\left(0, \mathbf{I}^{-1}\left(\theta_0\right)\right)$. To get $\widetilde{\boldsymbol{\theta}}_n$ we can use the 'method of moments': if $\boldsymbol{\theta} = \mathrm{g}\left(\mu_1, ...\mu_r\right)$ is a continuous function of the moments $\mu_k = E_F\left[X^k\right]$, and if $m_k = E_{\hat{F}_n}\left[X^k\right] = \sum x_i^k/n$ is the corresponding sample moment, then $\widetilde{\boldsymbol{\theta}}_n = \mathrm{g}\left(m_1, ...m_r\right)$ is consistent, and is $\sqrt{n}$-consistent if g is differentiable. This follows from the joint asymptotic normality of the sample moments: by the CLT,

$$\sqrt{n}\left(m_1-\mu_1,\cdots,m_r-\mu_r\right)' \xrightarrow{L} N\left(0,\Sigma\right),$$

where $\sigma_{kl} = COV\left[X^k, X^l\right] = \mu_{k+l} - \mu_k\mu_l$. In this example

$$\begin{aligned}
\mu_X &= (1-p)\xi, \\
E\left[X^2\right] &= p+(1-p)(1+\xi^2) = 1+(1-p)\xi^2, \\
\sigma_X^2 &= E\left[X^2\right] - \mu_X^2 = 1+p(1-p)\xi^2.
\end{aligned}$$

Solving for p, ξ gives

$$p = \frac{\sigma_X^2 - 1}{\mu_X^2+\sigma_X^2-1},\ \xi = \frac{\mu_X^2+\sigma_X^2-1}{\mu_X}$$

and so

$$\tilde{p} = \frac{S^2-1}{\bar{X}^2+S^2-1},\ \tilde{\xi} = \frac{\bar{X}^2+S^2-1}{\bar{X}}.$$

Although $\tilde{p}$ may not be in $(0,1)$, it is with probability $\to 1$. To see this let ε be such that $\varepsilon < p < 1-\varepsilon$. Then

$$P\left(0<\tilde{p}<1\right) \geq P\left(\left|\tilde{p}-p\right| \leq \varepsilon\right) \to 1.$$

23. Likelihood ratio, Wald's and Scores tests

- **Single parameter inferences**. Throughout, assume (C1)-(C7). We test $H : \theta = \theta_0$ vs. a two-sided alternative $K : \theta \neq \theta_0$. First consider **Wald's test**. Under H we have that
$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{L} N\left(0, I^{-1}\left(\theta_0\right)\right),$$
whence
$$W_n = \sqrt{n}\left(\hat{\theta}_n - \theta_0\right)\sqrt{I\left(\theta_0\right)} \xrightarrow{L} N\left(0,1\right)$$
and a level α test rejects if $|W_n| > u_{\alpha/2}$ or $W_n^2 > \chi_1^2\left(\alpha\right)$. (Obvious modification for 1-sided alternatives.) Similarly, a $100\left(1-\alpha\right)\%$ CI is
$$\hat{\theta}_n \pm u_{\alpha/2}/\sqrt{n\hat{I}_n}$$
for $\hat{I}_n \xrightarrow{pr} I\left(\theta_0\right)$. If $I\left(\theta\right)$ is continuous at θ_0 one can use $\hat{I}_n = I\left(\hat{\theta}_n\right)$. More commonly we use the 'observed information' $\hat{I}_n = -\frac{1}{n}l_n''\left(\hat{\theta}_n\right)$. (You should show that $-\frac{1}{n}l_n''\left(\hat{\theta}_n\right) \xrightarrow{pr} I\left(\theta_0\right)$.)

- **Likelihood ratio test**. Put

$$\Delta_n = l_n\left(\hat{\theta}_n\right) - l_n(\theta_0) = \log \frac{L_n\left(\hat{\theta}_n\right)}{L_n(\theta_0)}.$$

Under K this is expected to be large.
We will prove
Lemma: Under H, $2\Delta_n = W_n^2 + o_P(1)$.

By this, (recall sample midterm exam question, or Lemma 7.7.1) the tests based on W_n and $2\Delta_n$ are asymptotically equivalent. That based on $2\Delta_n$ rejects for

$$2\Delta_n > \chi_1^2(\alpha).$$

(For one-sided alternatives the 'signed likelihood ratio' $\sqrt{2\Delta_n} \cdot sign(\hat{\theta}_n - \theta_0)$ can be used.) An asymptotic $1-\alpha$ confidence region is

$$CR = \left\{\theta_0 \,|\, \ 2\Delta_n \le \chi_1^2(\alpha)\right\}.$$

Proof of Lemma: Expand

$$2\Delta_n = 2\left[\left(\hat{\theta}_n - \theta_0\right) l_n'\left(\theta_0\right) + \left(\hat{\theta}_n - \theta_0\right)^2 l_n''\left(\theta_n^*\right)/2\right]. \tag{23.1}$$

Since $0 = l_n'\left(\hat{\theta}_n\right) = l_n'\left(\theta_0\right) +$

$$\left(\hat{\theta}_n - \theta_0\right) l_n''\left(\theta_0\right) + \left(\hat{\theta}_n - \theta_0\right)^2 l_n'''\left(\theta_n^{**}\right)/2,$$

we have

$$l_n'\left(\theta_0\right) = -\left(\hat{\theta}_n - \theta_0\right) l_n''\left(\theta_0\right) - \left(\hat{\theta}_n - \theta_0\right)^2 l_n'''\left(\theta_n^{**}\right)/2.$$

This in (23.1) gives

$$\begin{aligned} 2\Delta_n &= \left\{\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)\right\}^2 \times \\ &\quad \left\{-2\frac{l_n''\left(\theta_0\right)}{n} + \frac{l_n''\left(\theta_n^*\right)}{n} - \left(\hat{\theta}_n - \theta_0\right)\frac{l_n'''\left(\theta_n^{**}\right)}{n}\right\}. \end{aligned}$$

Now

$$\frac{l_n''\left(\theta_0\right)}{n} \xrightarrow{pr} -I\left(\theta_0\right);$$

as in the proof of asymptotic normality of the MLE $l_n''\left(\theta_n^*\right)/l_n''\left(\theta_0\right) \xrightarrow{pr} 1$ so that

$$l_n''\left(\theta_n^*\right)/n \xrightarrow{pr} -I\left(\theta_0\right),$$

and $l_n'''\left(\theta_n^{**}\right)/n$ is $O_P(1)$. Thus

$$\begin{aligned} 2\Delta_n &= \left\{\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)\right\}^2 \left\{I\left(\theta_0\right) + o_P(1)\right\} \\ &= W_n^2 + o_P(1). \end{aligned}$$

- **Rao Scores Test**. Recall from the proof of asymptotic normality of the MLE that an expansion of $0 = l_n'\left(\hat{\theta}_n\right)$ around θ_0 resulted in

  $$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \frac{l_n'\left(\theta_0\right)/\sqrt{n}}{I\left(\theta_0\right)} + o_P(1),$$

  so that

  $$R_n \overset{def}{=} \frac{l_n'\left(\theta_0\right)}{\sqrt{nI\left(\theta_0\right)}} = W_n + o_P(1)$$

  and rejecting H for $|R_n| > u_{\alpha/2}$ is asymptotically equivalent to Wald's test.

- The three tests are asymptotically equivalent. The Scores test has the advantage of not requiring the computation of $\hat{\theta}_n$. The Scores and LR tests are invariant under reparameterizations; i.e. they have the same value if one tests $H : \psi = \psi_0$ for a $1 - 1$ (differentiable) function $\psi = \psi(\theta)$. Wald's test has this property (in finite samples) only for *linear* transformations (nonlinear transformations as well, asymptotically). You should verify these statements.

- The behaviour under K can be quite different. See Example 7.7.3 for instance, where for the Scores test of a Cauchy median θ, the power $\to 0$ as $|\theta - \theta_0| \to \infty$ for fixed n. There are analogous examples with respect to Wald's test. The LR test is most often preferable in terms of power.

- Example. $X_1, ..., X_n \sim$ Logistic, with

$$\begin{aligned} f_\theta(x) &= \frac{e^{-(x-\theta)}}{\left(1+e^{-(x-\theta)}\right)^2} = \frac{1}{4\cosh^2\left(\frac{x-\theta}{2}\right)}, \\ l_n'(\theta) &= \sum \psi(x_i - \theta) \text{ with} \\ \psi(x-\theta) &= \tanh\left(\frac{x-\theta}{2}\right) \text{ and } \psi' = 1 - \psi^2. \end{aligned}$$

  Then

$$W_n = \left(\hat{\theta}_n - \theta_0\right)\sqrt{n\hat{I}_n}.$$

  We can replace $\hat{\theta}_n$ by the one-step estimate

$$\delta_n = \tilde{\theta}_n + \frac{\sum \psi\left(x_i - \tilde{\theta}_n\right)}{\sum \psi'\left(x_i - \tilde{\theta}_n\right)},$$

  and as $\hat{I}_n$ we could use $I\left(\hat{\theta}_n\right)$ $(= I(\theta_0)) = 1/3$, or

$$\hat{I}_n = -\frac{1}{n} l_n''\left(\hat{\theta}_n\right) = \frac{1}{n}\sum \psi'\left(x_i - \hat{\theta}_n\right).$$

The LR test is based on

$$\begin{aligned} 2\Delta_n &= 2\left\{\sum \log f_{\hat{\theta}_n}(x_i) - \sum \log f_{\theta_0}(x_i)\right\} \\ &= -4\sum \log \frac{\cosh\left(\frac{x_i-\hat{\theta}_n}{2}\right)}{\cosh\left(\frac{x_i-\theta_0}{2}\right)}. \end{aligned}$$

The Scores test uses

$$R_n = \frac{l_n'(\theta_0)}{\sqrt{nI(\theta_0)}} = \sqrt{\frac{3}{n}}\sum \tanh\left(\frac{x_i-\theta_0}{2}\right).$$

- **Multiparameter inferences**. Test (simple null) H : $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs. K : $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}$ is $p \times 1$ and $X_1, ..., X_n \overset{i.i.d.}{\sim} f(x, \boldsymbol{\theta})$. The multiparameter versions of the tests given above are as follows. (Note: <u>all</u> evaluations from now on will assume that H is true.)

  - Wald's test:

    $$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{L} N\left(0, \mathbf{I}^{-1}(\theta_0)\right)$$
    $$\Rightarrow \quad W_n^2 = n\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)' \mathbf{I}(\theta_0)\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{L} \chi_p^2$$

    and H is rejected if $W_n^2 > \chi_p^2(\alpha)$. The corresponding confidence region is the p-dimensional ellipsoid

    $$\left\{\boldsymbol{\theta}_0 \mid \; n\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)' \hat{\mathbf{I}}_n\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \leq \chi_p^2(\alpha)\right\},$$

    where $\hat{\mathbf{I}}_n = \mathbf{I}\left(\hat{\boldsymbol{\theta}}_n\right)$ or $-\frac{1}{n}\ddot{l}_n\left(\hat{\boldsymbol{\theta}}_n\right)$.

  - Likelihood ratio test:

    $$2\Delta_n = 2\left(l_n\left(\hat{\boldsymbol{\theta}}_n\right) - l_n(\boldsymbol{\theta}_0)\right) \xrightarrow{L} \chi_p^2.$$

- Scores test:

$$\begin{aligned} &\frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\theta}_0) \xrightarrow{L} N_p(\mathbf{0}, \mathbf{I}(\theta_0)) \\ \Rightarrow\quad & R_n^2 = \frac{1}{n}\left[\dot{l}_n(\boldsymbol{\theta}_0)\right]' \mathbf{I}^{-1}(\theta_0)\left[\dot{l}_n(\boldsymbol{\theta}_0)\right] \xrightarrow{L} \chi_p^2. \end{aligned}$$

Typically there are nuisance parameters. Partition $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = \binom{\boldsymbol{\theta}_1}{\boldsymbol{\theta}_2}\begin{matrix} r \\ p-r \end{matrix}$ and test $H : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1,0}$ vs. $K : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_{1,0}$. Partition $\mathbf{I}(\boldsymbol{\theta})$ and $\mathbf{I}^{-1}(\boldsymbol{\theta})$ compatibly as

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{21} & \mathbf{I}_{22} \end{pmatrix}, \ \mathbf{I}^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{I}^{11} & \mathbf{I}^{12} \\ \mathbf{I}^{21} & \mathbf{I}^{22} \end{pmatrix}.$$

- Wald's test:

$$\begin{aligned} &\sqrt{n}\left(\begin{pmatrix} \hat{\boldsymbol{\theta}}_1 \\ \hat{\boldsymbol{\theta}}_2 \end{pmatrix} - \begin{pmatrix} \boldsymbol{\theta}_{1,0} \\ \boldsymbol{\theta}_2 \end{pmatrix}\right) \xrightarrow{L} N\left(0, \mathbf{I}^{-1}(\theta_0)\right) \\ \Rightarrow\quad & \sqrt{n}\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{1,0}\right) \xrightarrow{L} N\left(0, \mathbf{I}^{11}(\theta_0)\right) \\ \Rightarrow\quad & W_n^2 = n\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{1,0}\right)'\left[\mathbf{I}^{11}(\theta_0)\right]^{-1}\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{1,0}\right) \\ & \xrightarrow{L} \chi_r^2. \end{aligned}$$

For testing, $\mathbf{I}(\theta_0)$ is replaced by

$$\hat{\mathbf{I}}_n = \mathbf{I}\left(\boldsymbol{\theta}_{1,0}, \hat{\boldsymbol{\theta}}_2\right) \text{ or } \hat{\mathbf{I}}_n = -\frac{1}{n}\ddot{l}_n\left(\boldsymbol{\theta}_{1,0}, \hat{\boldsymbol{\theta}}_2\right).$$

We could instead use the 'restricted' MLE $\left(\boldsymbol{\theta}_{1,0}, \hat{\hat{\boldsymbol{\theta}}}_2\right)$ computed assuming H to be true:

$$\hat{\hat{\boldsymbol{\theta}}}_2 = \arg\max l_n\left(\boldsymbol{\theta}_{1,0}, \boldsymbol{\theta}_2\right).$$

For confidence regions use $\hat{\mathbf{I}}_n^{11}$, with $\hat{\mathbf{I}}_n = \mathbf{I}\left(\hat{\boldsymbol{\theta}}_n\right)$ or $-\frac{1}{n}\ddot{l}_n\left(\hat{\boldsymbol{\theta}}_n\right)$. Occasionally useful is the identity

$$\left[\mathbf{I}^{11}\right]^{-1} = \mathbf{I}_{11} - \mathbf{I}_{12}\mathbf{I}_{22}^{-1}\mathbf{I}_{21}.$$

- Likelihood ratio test: Let $\hat{\boldsymbol{\theta}}_n = \binom{\hat{\boldsymbol{\theta}}_1}{\hat{\boldsymbol{\theta}}_2}$ be the 'unrestricted' MLE, computed ignoring the hypothesis. Let $\hat{\hat{\boldsymbol{\theta}}}_n = \binom{\boldsymbol{\theta}_{1,0}}{\hat{\hat{\boldsymbol{\theta}}}_2}$ be the restricted MLE. Then

$$\begin{aligned} 2\Delta_n &= 2\left(l_n\left(\hat{\boldsymbol{\theta}}_n\right) - l_n\left(\hat{\hat{\boldsymbol{\theta}}}_n\right)\right) \\ &= 2\left(l_n\left(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2\right) - l_n\left(\boldsymbol{\theta}_{1,0}, \hat{\hat{\boldsymbol{\theta}}}_2\right)\right) \\ &\xrightarrow{L} \chi^2_r. \end{aligned}$$

- Scores test: Let $\dot{l}_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ be the first r elements of $\dot{l}_n(\boldsymbol{\theta})$. Since

$$\frac{1}{\sqrt{n}}\dot{l}_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \xrightarrow{L} N\left(\mathbf{0}, \mathbf{I}_{11}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right)$$

we have that under H,

$$\begin{aligned} R_n^2(\boldsymbol{\theta}_2) &= \frac{1}{n}\dot{l}_1'\left(\boldsymbol{\theta}_{1,0}, \boldsymbol{\theta}_2\right)\mathbf{I}_{11}^{-1}\left(\boldsymbol{\theta}_{1,0}, \boldsymbol{\theta}_2\right)\dot{l}_1\left(\boldsymbol{\theta}_{1,0}, \boldsymbol{\theta}_2\right) \\ &\xrightarrow{L} \chi_r^2. \end{aligned}$$

Finding appropriate estimates $\widetilde{\boldsymbol{\theta}}_2$ of $\boldsymbol{\theta}_2$ for which $R_n^2\left(\widetilde{\boldsymbol{\theta}}_2\right) \xrightarrow{L} \chi_r^2$ can be problematic. (Exercise on Assignment 3.)

- Note: For a 1-1 reparameterization $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta})$, we have $\mathbf{I}(\boldsymbol{\psi}) = \left[\mathbf{J}_{\boldsymbol{\psi}}^{-1}\right]'\mathbf{I}(\boldsymbol{\theta})\mathbf{J}_{\boldsymbol{\psi}}^{-1}$. This can be seen by equating the asymptotic covariance $\mathbf{I}^{-1}(\boldsymbol{\psi})$ of $\sqrt{n}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right)$ to that obtained from writing $\sqrt{n}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right) = \mathbf{J}_{\boldsymbol{\psi}} \cdot \sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) + o_P(1)$, etc.

24. Examples

Example: $X_1, ..., X_m \sim P(\lambda)$, $Y_1, ..., Y_n \sim P(\mu)$. Test $\lambda = \mu$ by defining $\mu = \lambda + \xi$ and testing $\xi = 0$ vs. $\xi \neq 0$; then $\boldsymbol{\theta} = (\xi, \lambda)$. Assume $m, n \to \infty$ with $n/(m+n) \to p \in (0,1)$. See Theorem 7.6.3 – the asymptotic normality of consistent sequences of local maxima of the likelihoods continues to hold, with

$$\begin{aligned}\mathbf{I}(\boldsymbol{\theta}) &= (1-p)\mathbf{I}_{\mathbf{X}}(\boldsymbol{\theta}) + p\mathbf{I}_{\mathbf{Y}}(\boldsymbol{\theta}) \\ &= \lim \frac{1}{m+n} E\left[-\ddot{l}(\boldsymbol{\theta})\right],\end{aligned}$$

where $\mathbf{I}_{\mathbf{X}}(\boldsymbol{\theta})$ is the information matrix based on the distribution of $X_1, ..., X_m$, etc. and $\ddot{l}(\boldsymbol{\theta})$ is computed from all $N = m + n$ observations. The basic idea is that the log-likelihood is a convex combination of that from X and that from Y, and its averages and moments converge to the corresponding convex combinations; e.g.

$$\begin{aligned} & \frac{1}{N} E\left[-\ddot{l}_N(\boldsymbol{\theta})\right] \\ = \; & E\left[\frac{-\ddot{l}_X(\boldsymbol{\theta})}{m}\frac{m}{m+n} + \frac{-\ddot{l}_Y(\boldsymbol{\theta})}{n}\frac{n}{m+n}\right] \\ \to \; & (1-p)\mathbf{I}_{\mathbf{X}}(\boldsymbol{\theta}) + p\mathbf{I}_{\mathbf{Y}}(\boldsymbol{\theta}).\end{aligned}$$

In this example the likelihood and log-likelihood are

$$\begin{aligned} L\left(\boldsymbol{\theta}\right) &= \prod_{i=1}^{m} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} \cdot \prod_{j=1}^{n} e^{-(\lambda+\xi)}\frac{(\lambda+\xi)^{y_j}}{y_j!}, \\ l_N\left(\boldsymbol{\theta}\right) &= -m\lambda + m\bar{X}\,\mathsf{log}\,\lambda - n\left(\lambda+\xi\right) \\ & \quad + n\bar{Y}\,\mathsf{log}\left(\lambda+\xi\right) + const. \end{aligned}$$

with

$$\begin{aligned} \dot{l}\left(\boldsymbol{\theta}\right) &= \begin{pmatrix} -n + \frac{n\bar{Y}}{\lambda+\xi} \\ -(m+n) + \frac{m\bar{X}}{\lambda} + \frac{n\bar{Y}}{\lambda+\xi} \end{pmatrix}, \\ -\ddot{l}\left(\boldsymbol{\theta}\right) &= \begin{pmatrix} \frac{n\bar{Y}}{(\lambda+\xi)^2} & \frac{n\bar{Y}}{(\lambda+\xi)^2} \\ \frac{n\bar{Y}}{(\lambda+\xi)^2} & \frac{m\bar{X}}{\lambda^2} + \frac{n\bar{Y}}{(\lambda+\xi)^2} \end{pmatrix}. \end{aligned}$$

Thus $\hat{\boldsymbol{\theta}} = \left(\hat{\xi}, \hat{\lambda}\right) = \left(\bar{Y} - \bar{X}, \bar{X}\right)$. Under H all $m+n$ observations have Poisson parameter λ and so $\hat{\hat{\boldsymbol{\theta}}} = \left(0, \hat{\hat{\lambda}}\right) = \left(0, \frac{m\bar{X}+n\bar{Y}}{m+n}\right)$. The information matrix is

$$\begin{aligned} \mathbf{I}\left(\boldsymbol{\theta}\right) &= \lim \frac{1}{N} E\left[-\ddot{l}\left(\boldsymbol{\theta}\right)\right] \\ &= \begin{pmatrix} \frac{p}{(\lambda+\xi)} & \frac{p}{(\lambda+\xi)} \\ \frac{p}{(\lambda+\xi)} & \frac{1-p}{\lambda} + \frac{p}{(\lambda+\xi)} \end{pmatrix}, \end{aligned}$$

with

$$\mathbf{I}\left(\boldsymbol{\theta}_0\right) = \frac{1}{\lambda}\left(\begin{array}{cc} p & p \\ p & 1 \end{array}\right)$$

and

$$\mathbf{I}^{11}\left(\boldsymbol{\theta}_0\right) = \frac{\lambda}{p\left(1-p\right)}.$$

- Wald's statistic is, with $N = m + n$,

$$\begin{aligned} W_N^2 &= N\left(\hat{\theta}_1 - \theta_{1,0}\right)'\left[\mathbf{I}^{11}\left(\boldsymbol{\theta}_0\right)\right]^{-1}\left(\hat{\theta}_1 - \theta_{1,0}\right) \\ &= \frac{N\left(\bar{Y} - \bar{X}\right)^2}{\mathbf{I}^{11}\left(\boldsymbol{\theta}_0\right)}. \end{aligned} \tag{24.1}$$

  Evaluation of $\mathbf{I}^{11}$ at $\hat{\hat{\boldsymbol{\theta}}}$ gives

$$\mathbf{I}^{11}\left(\hat{\hat{\boldsymbol{\theta}}}\right) = N\left(\frac{\bar{X}}{n} + \frac{\bar{Y}}{m}\right)$$

  and then

$$W_N^2 = \left(\frac{\bar{Y} - \bar{X}}{\sqrt{\frac{\bar{X}}{n} + \frac{\bar{Y}}{m}}}\right)^2.$$

If instead of $\mathbf{I}\left(\hat{\hat{\boldsymbol{\theta}}}\right)$ we use the observed information

$$-\frac{1}{N}\ddot{l}\left(\hat{\boldsymbol{\theta}}\right)=\frac{1}{N}\left(\begin{array}{cc}\frac{n}{\bar{Y}} & \frac{n}{\bar{Y}}\\ \frac{n}{\bar{Y}} & \frac{m}{\bar{X}}+\frac{n}{\bar{Y}}\end{array}\right),$$

with

$$\left[-\frac{1}{N}\ddot{l}\left(\hat{\boldsymbol{\theta}}\right)\right]^{11}=N\left(\frac{\bar{X}}{m}+\frac{\bar{Y}}{n}\right),$$

we obtain

$$W_N^2=\left(\frac{\bar{Y}-\bar{X}}{\sqrt{\frac{\bar{X}}{m}+\frac{\bar{Y}}{n}}}\right)^2.$$

This is equivalent to evaluating $\mathbf{I}^{11}\left(\boldsymbol{\theta}_0\right)$ at

$$\tilde{\lambda}=\frac{n\bar{X}+m\bar{Y}}{m+n}.$$

Recall that it was shown in Lecture 5 that these two forms of W_N^2 are asymptotically equivalent.

- The likelihood ratio statistic $2\Delta_N$ is

$$\begin{aligned}
&= 2\left[l_N\left(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2\right) - l_N\left(\boldsymbol{\theta}_{1,0}, \hat{\hat{\boldsymbol{\theta}}}_2\right)\right] \\
&= 2\left[l_N\left(\bar{Y} - \bar{X}, \bar{X}\right) - l_N\left(0, \frac{m\bar{X} + n\bar{Y}}{m+n}\right)\right] \\
&= 2\left[\begin{array}{c} -m\lambda + m\bar{X}\log\lambda \\ -n\left(\lambda+\xi\right) + n\bar{Y}\log\left(\lambda+\xi\right) \end{array}\right]_{|\lambda=\bar{X},\lambda+\xi=\bar{Y}} \\
&\quad -2\left[\begin{array}{c} -\left(m+n\right)\lambda \\ +\left(m\bar{X}+n\bar{Y}\right)\log\lambda \end{array}\right]_{|\lambda=\hat{\hat{\lambda}}=\frac{m\bar{X}+n\bar{Y}}{m+n}} \\
&= 2\left[m\bar{X}\log\frac{\bar{X}}{\hat{\hat{\lambda}}} + n\bar{Y}\log\frac{\bar{Y}}{\hat{\hat{\lambda}}}\right].
\end{aligned}$$

  How is this related to Wald? Some algebra shows that $2\Delta_N = 2\bar{X}f(R)$, where $R = \bar{Y}/\bar{X}$ and

  $$f(r) = (m+nr)\left[\log N - \log(m+nr)\right] + nr\log r$$

  satisfies $f(1) = f'(1) = 0$, $f''(1) = mn/N$. Thus under H,

$$\begin{aligned} 2\Delta_N &= 2\bar{X}\left[f''(1)\frac{(R-1)^2}{2} + o_p\left((R-1)^2\right)\right] \\ &= \left(\frac{\bar{Y}-\bar{X}}{\sqrt{\bar{X}\left(\frac{1}{m}+\frac{1}{n}\right)}}\right)^2 + o_p(1)\,. \end{aligned}$$

We can replace $\bar{X}$ in the denominator by any other consistent estimate of λ without altering the asymptotic properties; the two estimates

$$\begin{aligned} \hat{\hat{\lambda}} &= \frac{m\bar{X}+n\bar{Y}}{m+n}, \\ \tilde{\lambda} &= \frac{n\bar{X}+m\bar{Y}}{m+n} \end{aligned}$$

give the two versions of Wald's test discussed above.

- Scores test – Exercise on Assignment 3.

- If instead of testing $\xi = \lambda - \mu = 0$ one tests $\kappa = \mu/\lambda = 1$, both $2\Delta_N$ and R_N^2 are unchanged. But representing μ as $\mu = \kappa\lambda$, writing out the likelihood for $\theta = (\kappa, \lambda)'$ and proceeding as before gives $\hat{\kappa} = \bar{Y}/\bar{X}$, $\hat{\lambda} = \bar{X}$ and under H,
$$\sqrt{N}\left(\hat{\kappa} - 1\right) \xrightarrow{L} N\left(0, \frac{1}{\lambda p\left(1-p\right)}\right)$$
so that Wald's statistic is
$$\tilde{W}_N^2 = N\hat{\lambda}_1 p(1-p)\left(\frac{\bar{Y}}{\bar{X}} - 1\right)^2,$$
for a consistent estimate $\hat{\lambda}_1$. For the original null hypothesis (recall (24.1))
$$W_N^2 = \frac{Np(1-p)\left(\bar{Y} - \bar{X}\right)^2}{\hat{\lambda}_2},$$
for a consistent estimate $\hat{\lambda}_2$. The ratio between the two is
$$\frac{\tilde{W}_N^2}{W_N^2} = \frac{\hat{\lambda}_1\hat{\lambda}_2}{\bar{X}^2} \xrightarrow{pr} 1,$$
so that the two statistics are asymptotically equal. Whether or not Wald's test remains the same

under the two parameterizations for *finite* N depends on how the nuisance parameter λ is estimated. In particular, using the restricted MLE $\hat{\hat{\lambda}}$ in both instances results in different forms.

- **Regression**: In (linear or nonlinear) regression with Normal errors the LR test of a linear hypothesis becomes the usual 'drop in sum of squares' F-test:

$$F = \frac{SSE_H - SSE}{r} \Big/ \frac{SSE}{n-p} \ ,$$

where the 'full' model contains p regression parameters, the hypothesized model contains $p-r$ and SSE, SSE_H are the minimum sums of squares of residuals in the two models. (In a linear model $F \sim F^r_{n-p}$.)

To see this consider a general regression model

$$\mathbf{Y}_{n\times 1} = \boldsymbol{\eta}\left(\boldsymbol{\beta}\right) + \boldsymbol{\varepsilon} \sim N\left(\boldsymbol{\eta}\left(\boldsymbol{\beta}\right), \sigma^2 \mathbf{I}_n\right).$$

(In linear regression $\boldsymbol{\eta}\left(\boldsymbol{\beta}\right) = \mathbf{X}\boldsymbol{\beta}$.)

The parameter vector is $\boldsymbol{\theta} = \binom{\boldsymbol{\beta}}{\sigma^2}\begin{matrix}\leftarrow p \\ \leftarrow 1\end{matrix}$ and the likelihood function is

$$L(\boldsymbol{\theta}|\mathbf{y}) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{||\mathbf{y}-\boldsymbol{\eta}(\boldsymbol{\beta})||^2}{2\sigma^2}\right\}$$

with log-likelihood (apart from additive constants)

$$l(\boldsymbol{\theta}) = -\frac{n}{2}\log\sigma^2 - \frac{||\mathbf{y}-\boldsymbol{\eta}(\boldsymbol{\beta})||^2}{2\sigma^2}.$$

This is maximized over $\boldsymbol{\beta}$ by the LSE estimator $\hat{\boldsymbol{\beta}}$ minimizing

$$SS(\boldsymbol{\beta}) = ||\mathbf{y}-\boldsymbol{\eta}(\boldsymbol{\beta})||^2$$

and then over σ^2 by

$$\hat{\sigma}^2 = \frac{SS\left(\hat{\boldsymbol{\beta}}\right)}{n} = \frac{SSE}{n}.$$

The maximized log-likelihood is

$$l\left(\hat{\boldsymbol{\theta}}\right) = -\frac{n}{2}\left(\log\hat{\sigma}^2 + 1\right).$$

A linear hypothesis on the β's can always be written (after a linear reparameterization) in the form

$$\boldsymbol{\beta} = \begin{pmatrix}\boldsymbol{\beta}_{(1)} \\ \boldsymbol{\beta}_{(2)}\end{pmatrix} \begin{matrix}\leftarrow p-r \\ \leftarrow r\end{matrix} = \begin{pmatrix}\boldsymbol{\beta}_{(1)} \\ \mathbf{0}\end{pmatrix}.$$

The restricted MLE $\hat{\hat{\boldsymbol{\theta}}}$ is obtained by first finding the LSE in the restricted model:

$$\hat{\hat{\boldsymbol{\beta}}}_{(1)} = \arg\min ||\mathbf{y} - \boldsymbol{\eta}\left(\boldsymbol{\beta}_{(1)}, \mathbf{0}\right)||^2.$$

The MLE of σ^2 in the restricted model is

$$\hat{\hat{\sigma}}^2 = \frac{SS\left(\hat{\hat{\boldsymbol{\beta}}}\right)}{n} = \frac{SSE_H}{n}$$

and then

$$l\left(\hat{\hat{\boldsymbol{\theta}}}\right) = -\frac{n}{2}\left(\log \hat{\hat{\sigma}}^2 + 1\right).$$

The likelihood ratio statistic is

$$\begin{aligned} 2\Delta_n &= 2\left(l\left(\hat{\boldsymbol{\theta}}\right) - l\left(\hat{\hat{\boldsymbol{\theta}}}\right)\right) = n\log\frac{SSE_H}{SSE} \\ &= n\log\left(1 + \frac{r}{n-p}F\right), \end{aligned}$$

which is an increasing function of F. In linear models F is exactly distributed as F^r_{n-p}, and as χ^2_r/r as $n \to \infty$.

- Wald's test and the Scores test of a linear hypothesis coincide with the LR test in *linear* regression models; they differ for nonlinear regression.

25. Higher order asymptotics

- Statistics such as MLEs, likelihood ratio statistics, etc. can very often be represented as (nonlinear) functions of averages, yielding distributional approximations via techniques such as Edgeworth expansions.

- **Example 1**. Consider the simple case of the MLE $\hat{\alpha}$ from the exponential density $\alpha e^{-\alpha x}$ $(x > 0)$. We have

  $$\hat{\alpha} = \frac{1}{\bar{x}} = \frac{\alpha}{1 + (\alpha\bar{x} - 1)} = \frac{\alpha}{1 + Z_n/\sqrt{n}},$$

  where

  $$Z_n = \sqrt{n}\left(\alpha\bar{X} - 1\right) = \sqrt{n}\bar{Y}$$

  for $Y_i = \alpha X_i - 1$.

  The first three moments of Y_i are

  $$E\left[Y\right] = 0, \text{var}\left[Y\right] = 1, E\left[Y^3\right] = 2.$$

In particular, Z_n is $AN(0,1)$, and this mean and variance are exact. We are interested in the distribution of $T_n = \sqrt{n}\left(\frac{\hat{\alpha}-\alpha}{\alpha}\right)$, which by the delta method applied to $\sqrt{n}\left(\bar{X} - \alpha^{-1}\right)$ is $AN(0,1)$. To get a more refined approximation we start by expanding this as

$$\begin{aligned} T_n &= \sqrt{n}\left(\frac{1}{1+Z_n/\sqrt{n}} - 1\right) \\ &= -Z_n + \frac{Z_n^2}{\sqrt{n}} + O_p\left(n^{-1}\right). \end{aligned}$$

In Lectures 3 and 4 we studied Edgeworth expansions of averages, obtaining approximating densities in terms of the cumulants. The method relied on an expansion of the cumulant generating function (c.g.f.), and can be applied to functions of averages as well, as long as one can compute the cumulants:

$$\log E\left[e^{tT_n}\right] = \beta_1 t + \frac{\beta_2}{2!}t^2 + \frac{\beta_3}{3!}t^3 + \ldots$$

For T_n as above, the first cumulant is

$$\beta_1 = E\left[T_n\right] = \frac{E\left[Z_n^2\right]}{\sqrt{n}} + O\left(n^{-1}\right) = \frac{1}{\sqrt{n}} + O\left(n^{-1}\right).$$

As in Lecture 10,

$$\begin{aligned}\text{cov}\left[Z_n, Z_n^2\right] &= E\left[Z_n^3\right] = E\left[Y^3\right]/\sqrt{n} = 2/\sqrt{n},\\ \text{var}\left[Z_n^2\right] &= E\left[Z_n^4\right] - 1 = 2 + O\left(n^{-1}\right),\end{aligned}$$

so that

$$\begin{aligned}\beta_2 &= \text{var}\left[T_n\right] = \text{var}\left[Z_n\right] - 2\frac{\text{cov}\left[Z_n, Z_n^2\right]}{\sqrt{n}} + O\left(n^{-1}\right)\\ &= 1 + O\left(n^{-1}\right),\end{aligned}$$

and similarly

$$\begin{aligned}\beta_3 &= E\left[\left(T_n - E\left[T_n\right]\right)^3\right]\\ &= E\left[\left(-Z_n + \frac{Z_n^2 - E\left[Z_n^2\right]}{\sqrt{n}}\right)^3\right] + O\left(n^{-1}\right)\\ &= -E\left[Z_n^3\right] + 3E\left[Z_n^2\left\{\frac{Z_n^2 - E\left[Z_n^2\right]}{\sqrt{n}}\right\}\right] + O\left(n^{-1}\right)\\ &= \frac{-2}{\sqrt{n}} + \frac{3}{\sqrt{n}}\text{var}\left[Z_n^2\right] + O\left(n^{-1}\right)\\ &= \frac{4}{\sqrt{n}} + O\left(n^{-1}\right).\end{aligned}$$

All higher cumulants are $O\left(n^{-1}\right)$. Thus

$$E\left[e^{tT_n}\right] = \exp\left\{\frac{t}{\sqrt{n}} + \frac{t^2}{2!} + \frac{4}{3!\sqrt{n}}t^3 + O\left(n^{-1}\right)\right\},$$

hence $E\left[e^{t\left(T_n - \frac{1}{\sqrt{n}}\right)}\right] = e^{t^2/2}\exp\left\{\frac{2t^3}{3\sqrt{n}} + O\left(n^{-1}\right)\right\}$

$$= e^{t^2/2}\left\{1 + \frac{2t^3}{3\sqrt{n}}\right\} + O\left(n^{-1}\right).$$

The density of $U_n = T_n - \frac{1}{\sqrt{n}}$ is then

$$f_n(u) = \phi(u)\left[1 + \frac{2}{3\sqrt{n}}H_3(u)\right] + O\left(n^{-1}\right).$$

From this an expansion of the density of T_n is

$$\begin{aligned} g_n(t) &= f_n(t - \frac{1}{\sqrt{n}}) \\ &= \phi(t - \frac{1}{\sqrt{n}})\left[1 + \frac{2}{3\sqrt{n}}H_3(t - \frac{1}{\sqrt{n}})\right] + O\left(n^{-1}\right) \\ &= \phi(t)\left(1 + \frac{t}{\sqrt{n}}\right)\left[1 + \frac{2}{3\sqrt{n}}H_3(t)\right] + O\left(n^{-1}\right) \\ &= \phi(t)\left[1 + \frac{t + \frac{2}{3}H_3(t)}{\sqrt{n}}\right] + O\left(n^{-1}\right). \end{aligned}$$

- The normal approximation, arising from using only the first term, is most accurate near the mean $(t = 0)$: $H_3(0) = 0$ and so $g_n(0) = \phi(0) + O\left(n^{-1}\right)$. Elsewhere the error is $O\left(n^{-1/2}\right)$. In other words, just using the leading term – the basic normal approximation – is as accurate as the two-term Edgeworth expansion, at the mean. This is the idea behind a modification of this method known as 'exponential tilting'. Here, to get an expansion at a point s, the density of interest is modified ('tilted') so as to have a mean of s. Then the basic normal approximation, very accurate at s, is computed. Finally the density is 'untilted'. The development here is from Barndorff-Nielsen & Cox (1989). It is presented for the approximation of the distribution of a sample average, but can be extended to functions of averages as in Example 1.

- Let $S_n = \sum_1^n X_i$ for i.i.d. r.v.s X_i with density $f(x)$, m.g.f. $M(t)$ and c.g.f. $K(t) = \log M(t)$. Define a 'tilted' density
$$f(x;\lambda) = \frac{e^{\lambda x} f(x)}{M(\lambda)} = e^{\lambda x - K(\lambda)} f(x).$$
Note that under this density the m.g.f. of S_n (whose density is written $f_{S_n}(s;\lambda)$) is
$$\begin{aligned} E\left[e^{tS_n};\lambda\right] &= \int e^{ts} f_{S_n}(s;\lambda)\,ds, \\ \text{and also} &= \left\{\int e^{tx} f(x;\lambda)\,dx\right\}^n \\ &= \left\{\frac{M(t+\lambda)}{M(\lambda)}\right\}^n \\ &= \frac{\int e^{(t+\lambda)s} f_{S_n}(s;0)\,ds}{\int e^{\lambda s} f_{S_n}(s;0)\,ds} \\ &= \int e^{ts} \left\{\frac{e^{\lambda s} f_{S_n}(s;0)}{\int e^{\lambda s} f_{S_n}(s;0)\,ds}\right\} ds. \end{aligned}$$

By uniqueness of m.g.f.s,

$$\begin{aligned} f_{S_n}(s;\lambda) &= \frac{e^{\lambda s} f_{S_n}(s;0)}{\int e^{\lambda s} f_{S_n}(s;0)\,ds} \\ &= \frac{e^{\lambda s} f_{S_n}(s;0)}{M^n(\lambda)} \\ &= e^{\{\lambda s - nK(\lambda)\}} f_{S_n}(s;0). \end{aligned}$$

Thus the density of interest is

$$f_{S_n}(s;0) = e^{\{-\lambda s + nK(\lambda)\}} f_{S_n}(s;\lambda). \tag{25.1}$$

To get an accurate normal approximation at s we choose λ such that the tilted sum has a mean of s:

$$s = E[S_n;\lambda] = \frac{d}{dt}\left\{\frac{M(t+\lambda)}{M(\lambda)}\right\}^n_{|t=0} = nK'(\lambda).$$

Define $\hat{\lambda}$ by

$$K'\left(\hat{\lambda}\right) = s/n. \tag{25.2}$$

The normal approximation to the density of the tilted S_n is

$$f_{S_n}(s;\lambda) \approx \frac{1}{\sigma(S_n;\lambda)}\phi\left(\frac{s - E[S_n;\lambda]}{\sigma(S_n;\lambda)}\right).$$

A calculation as above gives $\sigma^2(S_n;\lambda) = nK''(\lambda)$, thus

$$f_{S_n}\left(s;\hat{\lambda}\right) = \frac{\phi(0)}{\sqrt{nK''\left(\hat{\lambda}\right)}} = \left(2\pi nK''\left(\hat{\lambda}\right)\right)^{-1/2}.$$

Since the normal approximation is being evaluated at the mean, the error is $O(n^{-1})$:

$$f_{S_n}\left(s;\hat{\lambda}\right) = \left(2\pi nK''\left(\hat{\lambda}\right)\right)^{-1/2}\left\{1+O(n^{-1})\right\}.$$

From (25.1), the density of the 'untilted' sum is

$$\begin{aligned} f_{S_n}(s;0) &= e^{\left\{-\hat{\lambda}s+nK\left(\hat{\lambda}\right)\right\}} f_{S_n}\left(s;\hat{\lambda}\right) \\ &= \frac{e^{\left\{-\hat{\lambda}s+nK\left(\hat{\lambda}\right)\right\}}}{\sqrt{2\pi nK''\left(\hat{\lambda}\right)}}\left\{1+O(n^{-1})\right\}. \end{aligned}$$

This is called the 'saddlepoint approximation' to the density of S_n. Note that it is guaranteed to be non-negative; it is often normalized to have an integral of 1 for finite n (but note $\hat{\lambda} = \hat{\lambda}(s)$).

- **Example 2**. Let $S_n = \sum_1^n X_i$ with X_i exponentially distributed with mean α^{-1}, as above. The exact ('Erlang') density is

$$f_{S_n}^{\text{exact}}(s) = \frac{(\alpha s)^{n-1}\,\alpha e^{-\alpha s}}{(n-1)!} I(s>0).$$

The normal approximation is

$$f_{S_n}^{\text{Normal}}(s) = \frac{\alpha}{\sqrt{n}}\phi\left(\frac{\alpha s - n}{\sqrt{n}}\right).$$

The two-term Edgeworth expansion (Lecture 4) results in

$$\begin{aligned} & f_{S_n}^{\text{Edgeworth}}(s) \\ = \;& f_{S_n}^{\text{Normal}}(s) \cdot \left(1 + \frac{H_3\left(\frac{\alpha s - n}{\sqrt{n}}\right)}{3\sqrt{n}} + O\left(n^{-1}\right)\right). \end{aligned}$$

Note that this is negative if s is sufficiently far from the mean n/α.

To get the saddlepoint approximation note that

$$K(t) = -\log\left(1 - \frac{t}{\alpha}\right),$$

and so (25.2) yields

$$\hat{\lambda} = \alpha - \frac{n}{s},$$

with $K\left(\hat{\lambda}\right) = \log\left(\frac{s\alpha}{n}\right)$ and

$$K''\left(\hat{\lambda}\right) = \left(\alpha - \hat{\lambda}\right)^{-2} = (s/n)^2 .$$

The saddlepoint approximation is thus

$$\begin{aligned} f_{S_n}^{\text{Saddlepoint}}(s) &= \frac{e^{\left\{-\hat{\lambda}s + nK\left(\hat{\lambda}\right)\right\}}}{\sqrt{2\pi nK''\left(\hat{\lambda}\right)}} \\ &= \frac{e^{-(\alpha s - n)}\left(\frac{\alpha s}{n}\right)^n}{\frac{s}{n}\sqrt{2\pi n}} \\ &= f_{S_n}^{\text{exact}}(s) \cdot \frac{(n-1)!e^n}{n^{n-\frac{1}{2}}\sqrt{2\pi}}. \end{aligned}$$

The saddlepoint approximation gives the exact answer, apart from Stirling's approximation of $(n-1)!$:

$$\frac{(n-1)!e^n}{n^{n-\frac{1}{2}}\sqrt{2\pi}} = 1 + O\left(n^{-1}\right).$$

Even this error would vanish if $f_{S_n}^{\text{Saddlepoint}}(\cdot)$ were normalized to have an integral of 1.

- The error of $f_{S_n}^{\text{Saddlepoint}}(s)$ is $O\left(n^{-1}\right)$ <u>for each s</u>. In contrast, the normal and Edgeworth approximations achieve this same degree of accuracy only at the mean, i.e. $s = s_0 = n/\alpha$ (with $H_3\left(\frac{\alpha s_0 - n}{\sqrt{n}}\right) = 0$):

  $$f_{S_n}^{\text{Edgeworth}}(s_0) = f_{S_n}^{\text{Normal}}(s_0) \cdot \left(1 + O\left(n^{-1}\right)\right)$$

  and

  $$f_{S_n}^{\text{Normal}}(s_0) = f_{S_n}^{\text{exact}}(s_0) \cdot \left(1 + O\left(n^{-1}\right)\right).$$

- The saddlepoint approximation method has been extended to quite general statistics – M-estimators, for example – and often yields fantastically accurate approximations, even for n as small as 2. For this reason the technique is sometimes called ‘small sample asymptotics’. A detailed development is given in the IMS monograph of this name, by C. Field and E. Ronchetti.