*Chapters 1-6:*
Population vs. sample → Parameter vs. statistic
Statistics: Descriptive vs. inferential

*Types of variables*

Quantitative vs. Qualitative
/ |
Discrete Continuous

*Tables, charts & graphs*
- frequency tables
- qualitative: bar graph/pie chart
- stem-and-leaf plot/dot plot
- time plot
- histogram (modality)
  - traits: # of modes, tail weight, overall shape (symmetry, skewness)
  - identify skewness by TAIL
- boxplot (skewness)
  - outliers, overall shape (symmetry, skewness)
  - identify skewness inside box or entire graph

*Measures of center/spread/position*
- center: mean, median, mode
  → Outlier effect? Skewness effect?
- spread: range, variance, standard deviation, IQR
  → Why use squared and ($n - 1$)? Ever negative? Empirical Rule?
- position: min, max, percentiles (quartiles)
  → recall that we INCLUDE the median when determining quartiles
  → 5-number summary, boxplot, types of outliers

*Chapters 7-10:*
*Displaying bivariate data*
- scatterplot: visual aid to see form/strength/direction of relationship and/or outliers (large residual, high leverage, influential)
- correlation: numerical aid to see strength/direction of relationship (range?)
  → Warning: assumes linearity, sensitive to outliers

*Simple linear regression analysis*
- regression line: $\hat{y} = b_0 + b_1 x$
- least-squares estimation gives $b_1 = r\left(\frac{s_y}{s_x}\right)$ and $b_0 = \bar{y} - b_1\bar{x}$
- estimation: interpolation vs. extrapolation (BAD!)
- R-squared: $r^2$ = proportion of variation in $y$ explained by $x$
- causation: association does NOT imply causation
- residual plots: observed vs. theoretical appearance
- transformation of a variable can help improve linearity

*Chapter 11-13:*

- observational/retrospective/prospective study, experiment/controlled clinical trial
  - → population and causal inferences (what needs to be present for each?)
- types of bias (response, undercoverage, nonresponse)
- types of sampling: with/without replacement, SRS/stratified/cluster/ voluntary/convenience/systematic
- controlling factors: randomization, blocking, direct control, replication
- more experiment design definitions

*Chapters 14-15:*

- types of events: marginal, conditional, union, intersection, complement,
  - What common words identify them?
- relating events: dependent vs. disjoint vs. independent
  - Do these relations affect the rules below? If so, how?
  - Do they allow certain rules to be easily extended?
- probability laws:
  - conditional probability: $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
  - complement rule: $P(A^C) = 1 - P(A)$
  - multiplication rule: $P(A \cap B) = P(A \text{ and } B) = P(A) \times P(B \mid A) = P(B) \times P(A \mid B)$
  - addition rule: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
  - total probability rule: $P(A) = P\left(A \cap B\right) + P\left(A \cap B^C\right)$
  - recall examples where we combined a few of these together

*Chapter 16-17:*

*Distributions*

- discrete (exact probability or intervals) vs. continuous (only intervals)
  - Discrete: If $P(X = a) > 0$, then $P(X \le a) \neq P(X < a)$
  - Continuous: If $P(X = a) = P(X = b) = 0$, then $P(a \le X \le b) = P(a < X < b)$
- discrete distributions:
  - determine probability distribution (values of $X$ and corresponding probabilities)
  - mean: $\mu = \sum x_i P(X = x_i)$
  - variance: $\sigma^2 = \sum (x_i - \mu)^2 P(X = x_i)$
- continuous distributions:
  - uniform distribution: finding an area of a rectangle (with a twist!)
  - normal distribution: symmetric, 2 parameters: $\mu$ and $\sigma$, other properties

*Standard Normal Distribution (and its applications)*
- $\mu = 0$ and $\sigma = 1$
- Table Z only gives areas to left of value $z$, conversion to these values required
  → use diagrams, complements, symmetry, etc.
- standardizing: $P(X \le x)$ → $P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = P(Z \le z)$
- identifying values for a given probability: $x = \mu + z\sigma$

*Combinations and Functions of Random Variables*
For any constants $a$ and $b$,

| *Means*: | *Variances*: |
|---|---|
| 1. $E(a) = a$ | 1. $V(a) = 0$ |
| 2. $E(aX) = aE(X)$ | 2. $V(aX) = a^2V(X)$ |
| 3. $E(aX + b) = aE(X) + b$ | 3. $V(aX + b) = a^2V(X)$ |
| 4. $E(aX \pm bY) = aE(X) \pm bE(Y)$ | 4. $V(aX \pm bY) = a^2V(X) + b^2V(Y) \pm$ ~~$2ab\text{cov}(X, Y)$~~ |

$Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n + b$, $E(Y) = a_1E(X_1) + a_2E(X_2) + \ldots + a_nE(X_n) + b$
If $X_1, X_2, \ldots, X_n$ are independent, $V(Y) = a_1^2V(X_1) + a_2^2V(X_2) + \ldots + a_n^2V(X_n)$

<u>*Chapter 18:*</u>
*Sampling Distributions*
- sample proportion:

  *Rule 1*: $\mu_{\hat{p}} = p$. *Rule 2*: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}}$.

  *Rule 3*: If $np$ and $n(1 - p)$ are both $\ge 10$, then $\hat{p}$ has an approx. normal dist'n.

  All 3 rules → If rule 3 holds, $Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)$

- sample mean:

  *Rule 1*: $\mu_{\bar{y}} = \mu$ *Rule 2*: $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$

  *Rule 3*: When the population distribution is normal, the sampling distribution of $\bar{y}$ is also normal for any sample size $n$.

  *Rule 4* (CLT): When $n > 30$, the sampling distribution of $\bar{y}$ is well approximated by a normal curve, even when the population distribution is not itself normal.

  All 4 rules → If $n$ is large OR the population is normal, $Z = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1)$

*Chapter 19:*
- how to interpret CI?
- generic CI: point estimate ± (critical value) × (standard error)

→ confidence level increases, *ME* increases

→ *n* increases, *ME* decreases
- sample proportion:

Assumptions: random sample, $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$.

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- choosing *n*:

$$n \approx p^*(1-p^*)\left(\frac{CV}{ME}\right)^2 \qquad \text{where } CV = z_{\alpha/2}$$

No previous info? Use $p^* = 0.5$ for safety.

*Chapters 20/21:*

Null vs. alternative hypotheses; 2-tailed vs. 1-tailed (lower or upper)

Rules for hypothesis construction:
1. No statistics, only parameters.
2. Parameter symbols and claimed values appear the same in $H_0$ and $H_A$.
3. The signs must be different between $H_0$ and $H_A$.
4. Equality signs ONLY appear in the null.

| | | Actual situation | |
|---|---|---|---|
| | | $H_0$ is true | $H_0$ is false |
| Decision | Do not reject $H_0$ | Correct decision | Type II or $\beta$ error |
| | Reject $H_0$ | Type I or $\alpha$ error | Correct decision |

Test statistic, *P*-value → judgment approach (JA) OR significance level approach (SLA)

Significance Level Approach:

*P*-value ≤ $\alpha$ → reject $H_0$;

*P*-value > $\alpha$ → do not reject $H_0$

Judgment Approach:

0 < *P*-value < 0.01 indicates convincing to strong evidence against $H_0$

0.01 < *P*-value < 0.05 indicates strong to moderate evidence

0.05 < *P*-value < 0.1 indicates moderate to suggestive, but inconclusive evidence

0.1 < *P*-value < 1 indicates weak evidence

Steps in a Hypothesis-Testing Analysis:

1. Assumptions; 2. Hypotheses (select $\alpha$); 3. Test statistic; 4. *P*-value; 5. Conclusion

*Test for Population Proportion*

Assumptions: categorical variable, random sample, $np_0 \geq 10$ and $n(1 - p_0) \geq 10$.

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

<u>*Chapter 22:*</u>

*Two Independent Population Proportions*

Assumptions:

independent random samples, $n_1$ & $n_2$ large ($n_i p_i \geq 10$ & $n_i(1 - p_i) \geq 10$ for $i = 1,2$)

CI: $\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

$H_0$: $p_1 - p_2 = 0$ → Because of $H_0$, we use $\hat{p}_{pooled} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{y_1 + y_2}{n_1 + n_2}$

Then, under same assumptions, $z_0 = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\hat{p}_{pooled}(1 - \hat{p}_{pooled})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

<u>*Chapter 23:*</u>

- introduce $t$-distribution (same as $z$ except for parameter ($df$) and NOT knowing $\sigma$)
- sample mean:

Assumptions: random sample, $n \geq 30$ OR population is normal, $\sigma$ unknown

$$\bar{y} \pm t_{\alpha/2,\, n-1} \times \left(\frac{s}{\sqrt{n}}\right)$$

- choosing $n$:

$$n \approx \left(\frac{CV}{ME}\right)^2 \hat{\sigma}^2 \qquad \text{where } CV = z_{\alpha/2}$$

No $\hat{\sigma}$? Use $\hat{\sigma} \approx$ range/6 for approximately normal data.

*Test for Population Mean*

Assumptions: num. var., random sample, $n \geq 30$ OR population is normal, $\sigma$ unknown

$$t_0 = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}}$$

*Chapters 24/25:*

*Independent samples*

Assumptions:

independent random samples, $n_1$ & $n_2 \geq 30$ OR both pop'ns are normal, unknown and unequal standard deviations ($\sigma_1 \neq \sigma_2$) → how do you check?

$H_0$: $\mu_1 - \mu_2 = 0$

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad t_0 = \frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{SE(\bar{y}_1 - \bar{y}_2)} \qquad (df \geq \min\{n_1 - 1, n_2 - 1\})$$

CI (same assumptions): $\bar{y}_1 - \bar{y}_2 \pm t_{\alpha/2,\, df} \times SE(\bar{y}_1 - \bar{y}_2)$

Assumptions:

independent random samples, $n_1$ & $n_2 \geq 30$ OR both pop'ns are normal, unknown and equal standard deviations ($\sigma_1 = \sigma_2$) → how do you check?

$H_0$: $\mu_1 - \mu_2 = 0$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \qquad SE(\bar{y}_1 - \bar{y}_2) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad t_0 = \frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{SE(\bar{y}_1 - \bar{y}_2)}$$

CI (same assumptions): $\bar{y}_1 - \bar{y}_2 \pm t_{\alpha/2,\, df} \times SE(\bar{y}_1 - \bar{y}_2)$ ($df = n_1 + n_2 - 2$)

*Paired samples*

Assumptions: paired samples, random sample of $d$'s, $n \geq 30$ OR pop'n dist'n is normal, $\sigma_d$ unknown

$H_0$: $\mu_d = 0$ (define '$d$' first)

$$t_0 = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} \qquad (df = n - 1)$$

CI (same assumptions): $\bar{d} \pm t_{\alpha/2,\, df} \times \left(\frac{s_d}{\sqrt{n}}\right)$

*Chapter 26:*

$$\chi^2 = \sum_{cells} \frac{(Obs - Exp)^2}{Exp}$$ follows $\chi^2$-dist'n

| Goodness-of-fit test | Test of homogeneity | Test of independence |
|---|---|---|
| - single random sample<br>- one categorical variable | - indep. random samples<br>- one variable per sample | - single random sample<br>- two categorical variables |
| - every expected count $\geq 5$ | - every expected count $\geq 5$ | - every expected count $\geq 5$ |
| - $df = \#Categories - 1$ | - $df = (R - 1)(C - 1)$ | - $df = (R - 1)(C - 1)$ |
| - expected counts: $np_i$ | $\frac{\text{(row marginal total)(column marginal total)}}{\text{grand total}}$ | |

$R$ = # of rows; $C$ = # of columns

*Chapter 28:*
ANOVA Assumptions:
independent and random samples, similar $\sigma$, normal distributions
- $y_{ij}$ = observation for $i^{th}$ subject in $j^{th}$ group
- $\bar{y}_j$ vs. $\bar{\bar{y}}$ to predict $y_{ij}$
- $H_0$: $\mu_1 = \ldots = \mu_k$
- $H_A$: the $\mu_i$ are not all equal (OR, at least 2 $\mu_i$ are different)

$SS_T$ = (variability between samples)
$SS_E$ = (variability within samples)

$$F_0 = \frac{MS_T}{MS_E} = \frac{SS_T/(k-1)}{SS_E/(N-k)} \sim F_{N-k}^{k-1}$$

Reject $H_0$ when $F_0$ is large (greater than 10 works for most values of $\alpha$) or use $P$-value.

*ANOVA Table*:

| **Source** | **df** | **SS** | **MS** | ***F*** | ***P*-value** |
|---|---|---|---|---|---|
| Between | $k-1$ | $SS_T$ | $MS_T$ | $MS_T/MS_E$ | ? |
| Within | $N-k$ | $SS_E$ | $MS_E$ | | |
| **Total** | **$N-1$** | **$SS_Y$** | | | |

Note that $SS_Y = SS_T + SS_E$; df(total) = df(Between) + df(Within);
For each of the top two rows, MS = SS/df.

Also note that the estimate for common variance $= \hat{\sigma}^2 = MS_E$.

**How to determine data structure?**

1. Are you dealing with proportions or means?

2. How many samples are there?

3a. Are the samples independent or paired? 3b. How many variables/levels are there?

4a. Are the variances equal/unequal?