Topic 2 – One or Two Population Means

T-tools: A General Approach

Inferential procedure in words:
1. Indicate parameter in question.
2. Indicate estimator used to estimate said parameter.
3. Find sampling distribution of estimate (or standardized estimate).
4. i) Confidence intervals: calculate a margin of error for estimate based on some specified level of confidence.
   ii) Hypothesis testing: calculate likelihood of observed estimate under some null hypothesis, given the distribution of the estimator.

...In formulae:
For a general parameter θ, the estimator is $\hat{\theta}$. Preferably, $\hat{\theta}$ follows a NORMAL (or approximately normal) distribution: $\hat{\theta} \dot{\sim} N(\mu_{\hat{\theta}}, \sigma_{\hat{\theta}})$. If $\hat{\theta}$ is an unbiased estimator for θ, then $\mu_{\hat{\theta}} = \theta$, and $\hat{\theta} \dot{\sim} N(\theta, \sigma_{\hat{\theta}})$.

Thus, the statistic Z has an (approx.) standard normal distribution such that

$$Z = \frac{Estimate - Parameter}{S.D.(Estimate)} = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim N(0,1)$$

Usually, $\sigma_{\hat{\theta}}$ is unknown. The estimate of $\sigma_{\hat{\theta}}$ (or, standard error of $\hat{\theta}$) is denoted by $S.E.(\hat{\theta})$. The extra variability requires a different distribution, so

$$T = \frac{Estimate - Parameter}{S.E.(Estimate)} = \frac{\hat{\theta} - \theta}{S.E.(\hat{\theta})} \sim t_{df}$$

"Step 4.i)": A $(1 - \alpha)100\%$ Confidence Interval (CI) for θ

Estimate ± Margin of Error
Estimate ± (Critical value) × $S.E.(Estimate)$
$\hat{\theta} \pm [t_{df,\, \alpha/2}] \times [S.E.(\hat{\theta})]$

where the critical value comes from the t-distribution at the $(1 - \alpha)$ confidence level.

"Step 4.ii)": Testing H_0: $\theta = \theta_0$

I. Assume H_0 is true and calculate $t_0 = \frac{\hat{\theta} - \theta_0}{S.E.(\hat{\theta})}$.

II. Determine how likely it was to observe $\hat{\theta}$ if, in fact, $\theta = \theta_0$. In other words, find the *p-value* of the test.

Assume H_0 is true, the *p-value* is the probability of obtaining T as extreme or more extreme than t_0. If the p-value is small, the data is inconsistent with H_0.

1. If H_a: $\theta > \theta_0$, then p-value = $P(t_{df} > t_0)$.
2. If H_a: $\theta < \theta_0$, then p-value = $P(t_{df} < t_0)$.
3. If H_a: $\theta \neq \theta_0$, then p-value = $P(t_{df} > t_0) + P(t_{df} < -t_0) = 2P(t_{df} > | t_0 |)$.

One-Sample t-tools for the Population Mean:
Consider the sample $Y_1, Y_2, \ldots, Y_n$ from some population with mean μ_Y and standard deviation σ_Y.

1. Parameter: μ_Y

2. Estimator: The sample mean $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

3. Recall C.L.T.: If the defined sample above is also random, then, for large n, the sample mean $\overline{Y}$ is approximately normal such that $\overline{Y} \dot{\sim} N(\mu_Y, \sigma_Y / \sqrt{n})$. Since the estimator is approximately normal, then t-tools may be applied. Thus,

$$T_0 = \frac{\overline{Y} - \mu_Y}{S_Y / \sqrt{n}} \sim t_{df} = t_{n-1}$$

Also, a $(1 - \alpha)100\%$ CI for μ_Y

Estimate ± (Critical value) × *S.E.*(*Estimate*)
$\bar{y} \pm [t_{n-1,\, \alpha/2}] \times [\, s_Y / \sqrt{n}\,]$

And, for testing H_0: $\mu_Y = \mu_0$,

test statistic $= t_0 = \frac{\bar{y} - \mu_0}{s_Y / \sqrt{n}}$

For H_a: $\mu_Y > \mu_0$, p-value = $P(t_{n-1} > t_0)$.
For H_a: $\mu_Y < \mu_0$, p-value = $P(t_{n-1} < t_0) = P(t_{n-1} > | t_0 |)$.
For H_a: $\mu_Y \neq \mu_0$, p-value = $2P(t_{n-1} > | t_0 |)$.

The above procedures can and will be applied throughout this course. As a result, once the overall procedures for the general t-tools are covered, development for inferential procedures of appropriate parameters requires only the identification of the estimate, the standard error of the estimate, and the sampling distribution of T.

Return to the *p*-value

The *p-value* is already defined, but how does it direct one's conclusion?

Well, the significance level (denoted by α) is a number such that we reject H_0 if the *p*-value is less than or equal to that number. Recall α is also the Type I error.

Using the "significance level approach", reject H_0 if

$$p\text{-value} \leq \alpha$$

and do not reject H_0 if

$$p\text{-value} > \alpha$$

Common choices for α are 0.01, 0.05 and 0.1, depending on the nature of the test.

Notes:
1. NEVER "accept" H_0 or H_a…for different reasons.
2. Notice the difference between the 2 following statements. The first is correct, while the second is not.

a) Since the *p*-value < α, I reject H_0 at the significance level α.

b) Since the *p*-value < α, I reject H_0.

The 1st statement takes into account that rejecting H_0 has a chance of error.

PROBLEMS:
a) If you're comparing to α and do not mention the *p*-value, there is no way of knowing if the *p*-value is 0.045 or 0.000001.
b) If we use a "cut-off" like $\alpha = 0.05$, does it make sense to conclude differently between *p*-values of 0.049 and 0.051?
Solution: ALWAYS report your *p*-value! This way, a reader may draw their own conclusions. Moreover, use the "judgment approach" for rejection. Here, there's a tendency of avoiding "cut-off" points and going toward some "acceptable" guidelines:
- *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
- *p*-value > 0.1 indicates weak evidence

Example:
A scientist claims the average survival time of a particular bacterial strain is 21.104 seconds. A bored student takes a random sample of 7 strains and found their mean survival time to be 21.16 s with a standard deviation of 0.20 s. It is known that the survival time of all the strains has an approximate normal distribution. Test the above claim with $\alpha = 0.01$.

H_0: $\mu = 21.104$ $\quad$ H_a: $\mu \neq 21.104$

The sample is random, the population is normally distributed, and *n* is small. Consequently, we may use a *t*-distribution; specifically, t_6. Also, the test statistic is

$$t_0 = \frac{\bar{Y} - \mu_0}{s_Y / \sqrt{n}} = \frac{21.16 - 21.104}{0.20 / \sqrt{7}} = 0.741$$

Using the row of $df = 6$ in the t-table, the value of 0.741 is between the values of 0.906 and 0.718, which correspond to areas of 0.80 and 0.75 in the left tail of the t-dist'n. The area in the right tail of the t-dist'n is, therefore (0.20, 0.25). Since this test is two-sided, the range for our p-value is (0.40, 0.50). In fact, software produces an exact p-value of 0.487. Using the "judgment approach", we have weak evidence against H_0. Using the "significance level approach" and $\alpha = 0.01$, we do not reject H_0. Either way, we conclude that the scientist's claim may indeed be valid.

(Note that one-tailed tests with this example were discussed in class.)

Example (continued): Construct a 99% confidence interval for the average length of all the strains for these particular bacteria. Interpret the interval.

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

$$21.16 \pm t_{0.005,6}\left(\frac{0.2}{\sqrt{7}}\right)$$

$$21.16 \pm (3.707)(0.076)$$

$$21.16 \pm 0.280$$

$$(20.880,\ 21.440)$$

It is estimated with 99% confidence that the population mean survival time lies between 20.880 and 21.440 seconds.

Two-sample Procedures

Usually, the goal is to see if the two samples differ from each other in some way.

Independent samples:

With one sample, the parameter was μ_Y. With two independent samples, it is $\mu_1 - \mu_2$. Thus,

$H_0: \mu_1 - \mu_2 = 0$ → Estimate = $\bar{y}_1 - \bar{y}_2$

$$S.E.(\bar{y}_1 - \bar{y}_2) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \qquad \text{where } s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S.E.(\bar{Y}_1 - \bar{Y}_2)} \sim t_{n_1+n_2-2}$$

Also, all the necessary parts for a CI are present (estimate, std. error, distribution for critical value).

Paired samples:

Due to the data structure, two samples have been reduced to one sample. Thus,

$H_0: \mu_d = 0$ → Estimate = $\bar{d} = \dfrac{\sum d_i}{n}$

$$S.E.(\bar{d}) = \frac{s_d}{\sqrt{n}} \qquad T = \frac{\bar{D} - \mu_d}{S.E.(\bar{D})} \sim t_{n-1}$$

Assumptions (for 2 independent sample *t*-tools)

1a. What are they?

1. Normality: Each sample comes from a *normal* population.

2. Equal variability: The two populations have the same but unknown standard deviation (or variance). There is a 2-sample *t*-test that doesn't assume equal variability, but we won't consider it in this course.

3. Independence: The two samples are independent. All observations are independent within each sample and across samples.

1b. How to check for them?

1. Normality:

   i) *Normal Probability Plots*: For normal populations, the N.P.P. should show an increasing linear relationship. In SPSS, it's called the Q-Q plot or P-P plot. Don't worry about how the plot is constructed.

   ii) *Boxplots*: See section 3.1 in the text, though these plots are not the best to assess normality. Still, look for either symmetry or strong skewness and/or outliers.

   iii) *Histograms*: See page 60 – 62 in the text. This graph represents an estimate of the actual distribution of the population. Look for a *normal* ("bell-shaped") curve. Histograms are not useful for small sample sizes.

2. Equal variability:

   i) *Boxplots*: Useful to see if samples have equal spread by looking at side-by-side boxplots. Compare the spreads (a.k.a. IQR), or ranges, of the data. If they are roughly the same, the assumption of equal variability is satisfied.

   ii) *Compare Standard Deviation*: An informal rule says that if the larger of the two is less than twice the other, then the assumption is reasonable.

   iii) *Formal Tests*: These exist to assess evidence of unequal variances. For the scope of this course, however, the first two options will suffice.

3. Independence: No graphical/numerical method to satisfy this assumption. Only a close look at the study design and how the data were obtained is appropriate. Unless there is a clear violation of independence, one assumes it (cautiously) holds. Still, there are two common types of dependence to consider:

   i) *Cluster Dependence*: Data collected in sub-groups.

   ii) *Serial Dependence*: Measurements close together (in time) tend to be more similar than observations taken a lengthy time apart.

2. Robustness of 2 independent sample t-tools

Def'n: A statistical procedure is <u>robust</u> to departures from a particular assumption if the procedure remains valid even when the assumption is violated.

1. *Departures from Normality*: Recall the assumption that each sample is from a normal population such that group 1 has n_1 observations, while group 2 has n_2. Then, we have $Y_{11}, Y_{12}, \ldots, Y_{1n_1}, Y_{21}, Y_{22}, \ldots, Y_{2n_2}$.

   Recall also the C.L.T. If both n_1 and n_2 are large, then the sample means will both have approx. normal distributions. Thus, normality is not a very serious assumption. Moreover, the *t*-tools are said to be robust to departures from normality; that is, they are valid even when the data is non-normal.
2. *Differing Standard Deviations*: Most crucial assumption. If the standard deviations are considerably unequal, s_p does not estimate its intended parameter. Thus, T does not follow a *t*-distribution. If n_1 and n_2 are roughly equal, the *t*-tools remain fairly valid even if the standard deviations differ. Still, it is best to deal with this problem (will be done shortly).
3. *Departures from Independence*: If independence is not present, the information is biased. As a result, the *t*-tools may provide misleading results.
4. *Outliers*: The sample mean is sensitive (or, non-resistant) to outliers. Since the *t*-tools are based on sample means, they are non-resistant, too. Outliers, therefore, affect the results of *t*-tools.
   Note: Medians are resistant (perhaps they might be of use?)

3. What to do if assumptions are violated?

If violations occur, here are the steps during analysis to reduce their effects:

1. *Check Assumptions*: Before any analysis, *ALWAYS* perform a preliminary analysis, which includes graphical analysis to check assumptions of the model to be used.
   a) Assumptions OK? Continue with *t*-tools on original data.
   b) Assumptions violated? Go to Step 2.

2. *Transformation*: If serious violations of assumptions exist on the original scale, consider transforming the data to a new scale where assumptions are satisfied.
   a) If a (reasonable) transformation is easily found, apply *t*-tools on the transformed data and proceed to Step 3.
   b) If no transformation can be found, proceed to Step 4.

   <u>Common transformations</u>:
   - Logarithmic Transformation (natural log): Beneficial when both groups are right-skewed and group with larger center has larger spread. (Main transformation considered in STAT 252.)
   - Square Root: Useful for count responses.
   - Reciprocal: Useful for waiting times.
   - Logit: Useful for proportions (might be encountered).

3. *Back Transformation*: Once analysis is complete on the transformed scale, back transform (if possible) the results and interpret them on the original scale. Usually, this step is the most difficult. The natural log transformation, though, offers a fairly straightforward back transformation; hence, it will receive priority.

4. *Alternatives*: If no (reasonable) transformation can be found, non-parametric methods are available.

_T_-tools Assumptions via Graphs
(see graphical examples in different file)

(Natural) Log Transformations
Original scale:
Consider two populations Y_1 and Y_2 with means μ_1 and μ_2. Thus, the samples consist of $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$, and $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$, respectively. The additive model looks at the difference between the two population means, $\mu_1 - \mu_2$. This parameter of interest is also known as the additive (treatment) effect, $\delta = \mu_1 - \mu_2$. For example, if $\delta = 0$, there is no additive (treatment) effect, or, no difference between the means.

Effect of log transformation:
New variables:
$\ln(Y_1)$ – natural log of population 1
$\ln(Y_2)$ – natural log of population 2
Thus, the datasets become $\ln(Y_{11}), \ldots, \ln(Y_{1n_1})$, and $\ln(Y_{21}), \ldots, \ln(Y_{2n_2})$

Parameter: $\delta = \mu_{\ln(Y_1)} - \mu_{\ln(Y_2)}$. The additive (treatment) effect now refers to the difference in the means of the logged responses.

Estimate: $\hat{\delta} = \overline{\ln Y_1} - \overline{\ln Y_2} = \frac{1}{n_1}\sum_{i=1}^{n_1}\log(Y_{1i}) - \frac{1}{n_2}\sum_{i=1}^{n_2}\log(Y_{2i})$

This estimate is *back-transformed* by taking the *anti-log*. The back-transformed estimate is $e^{\hat{\delta}}$, where 'e' is the mathematical constant roughly equal to 2.718.

Problem: $e^{\hat{\delta}} = e^{\overline{\ln Y_1} - \overline{\ln Y_2}}$ doesn't estimate $\mu_{Y_1} - \mu_{Y_2}$ (due to mathematical properties).
Simply put, taking the antilog of the additive effect on the log-scale does NOT estimate the additive effect on the original scale. Also, means are affected by outliers, so should we go back to them or something else?

Solution: Since $Med(\ln(Y)) = \ln(Med(Y))$ and $Med(Y) = \mu_Y$ when Y is normal,
$$\mu_{\ln(Y)} = Med(\ln(Y)) = \ln(Med(Y))$$

Thus, $\hat{\delta} = \overline{\ln Y_1} - \overline{\ln Y_2}$ estimates $\ln(Med(Y_1)) - \ln(Med(Y_2))$ and

$$e^{\ln(Med(Y_1)) - \ln(Med(Y_2))} = \frac{e^{\ln(Med(Y_1))}}{e^{\ln(Med(Y_2))}} = \frac{Med(Y_1)}{Med(Y_2)}$$

Conclusion: If δ is the additive (treatment) effect for the means on the log-scale, then e^{δ} is the *multiplicative (treatment) effect* for the medians on the original scale. Furthermore, the concluding statement would be:

"It is estimated that $Med(Y_1)$ is $e^{\hat{\delta}}$ times that of $Med(Y_2)$."

Inference after a log transformation:

Confidence Intervals: If a CI for $\mu_{\ln Y_1} - \mu_{\ln Y_2}$ is (L, U), then a CI for $\frac{Med(Y_1)}{Med(Y_2)}$ is (e^L, e^U).

Hypothesis Testing:

H_0: $\delta = \mu_{\ln Y_1} - \mu_{\ln Y_2} = 0$ may be interpreted as H_0: $e^{\delta} = \frac{Med(Y_1)}{Med(Y_2)} = 1$