Last Time:
- Numerical and graphical summaries for quantitative data
  - Discuss shape (symmetric vs. skewed), center (e.g., mean, median), spread or variability (e.g., standard deviation \( s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \), IQR = interquartile range = width of middle 50% of the distribution)
- Statistical inference (assuming a simple shift between the two distributions)
  - Tests of significance vs. confidence interval about \( \mu_1 - \mu_2 \) (difference in population means or difference in long-run means or difference in treatment means)
  - Pooled vs. unpooled standard deviation
    - Welch’s \( t \) test: \( t_0 = [\overline{x}_1 - \overline{x}_2 - \delta_0] / \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \) with crazy degrees of freedom
    - Student’s \( t \) test: \( t_0 = \frac{\overline{x}_1 - \overline{x}_2}{s_p \sqrt{1/n_1 + 1/n_2}} \) with \( df = n_1 + n_2 - 2 \) where \( s_p \) is a “pooled” estimate of the common standard deviation
  - \( p \)-value and margin-of-error are very much a function of sample size

Example 1: Lactate dehydrogenase (LD) is an enzyme that may show elevated activity following damage to the heart muscle or other tissues. A large study of serum LD levels (“units per liter”) in healthy young people found the following results (Williams et al., 1978).

<table>
<thead>
<tr>
<th></th>
<th>( n )</th>
<th>Mean (( \bar{x} ))</th>
<th>SD (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>270</td>
<td>60 U/l</td>
<td>11 U/l</td>
</tr>
<tr>
<td>Females</td>
<td>264</td>
<td>57 U/l</td>
<td>10 U/l</td>
</tr>
</tbody>
</table>

(a) Is the difference in mean LD levels significantly different between males and females? (Obtain the \( t \) statistic and \( p \)-value, and state a one-sentence conclusion.)

(b) Is the difference in mean bodyweight significantly different between males and females? (Obtain the \( t \) statistic and \( p \)-value, and state a one-sentence conclusion.)

(c) Is it reasonable to conclude that the difference of 3 U/l is large or important? Is it reasonable to conclude that the different of 32 lbs is small and unimportant?
Here is where the subject matter expert comes in. If the expert tells you the day-to-day fluctuation in a person’s LD level is around 6.5 U/l, this helps tell you that the difference is negligible from the medical standpoint. In the LD example, the confidence interval is (1.2, 4.8) U/l and so the physician may conclude that while the difference is statistically significant, we probably don’t need to differentiate between males and females in establishing clinical thresholds for diagnosis or illness.

When comparing two means, it is increasingly recommended to report the effect size of the study. One measure of effect size is $\frac{|x_1 - x_2|}{s}$. One reasonable choice for $s$ is the larger of the two sample standard deviations, or you can use $\sqrt{\frac{s_1^2 + s_2^2}{2}}$. While pretty ad hoc, Cohen has suggested that effect sizes of 0.2, 0.5, and 0.8 represent small, median, and large effect sizes respectively.

(d) Calculate, compare, and interpret the two effect sizes for these two studies.

Another common choice of “effect size” is $\frac{|x_1 - x_2|}{x_2}$, so reporting the size of the difference as a percentage of one of the means. In this case, the males are about 22% heavier than the females.

**Example 2:** A group of scientists wants to investigate whether or not an Internet-based intervention program would help women lose weight after giving birth. The program will provide weekly exercise and dietary guidance and establish an online forum for nutrition and exercise discussion with other recent moms. A second group will be given traditional written dietary and exercise guidelines by their doctors. The response variable will be the amount of weight lost at 12 months postpartum (kg). Previous studies have shown that at 12 months postpartum, the mean weight lost is about 3.6 kg with a standard deviation of 4.0 kg. The research team wanted to show at least a 50% improvement in weight loss for the internet group, that is that the internet group lose at least 1.8 kg more weight than the controls, at the 5% level of significance, with 80% power.

(a) So what is the effect size of interest?

(b) What sample size is necessary to achieve the desired power?

See the `power.t.test` command in the Day 10 R Script file. Note that you can enter either the desired difference in means and a (common) standard deviation estimate, or you can enter the effect size with $sd = 1$.

(c) Any other sample size considerations with this particular study?
**Example 3:** Researchers Dumas and Dunbar wanted to see whether stereotypes related to creativity can influence the creativity of a subject (2016). The 64 subjects were randomly placed into two groups of 32 each. One group was told to imagine that they were rigid librarians and another group was to imagine they were eccentric poets. As a measure of their creativity, they were then shown 10 items and asked to generate as many different uses for these items as possible. The “rigid librarians” averaged 60.34 total uses for the objects (SD = 24.38 uses) and the “eccentric poets” averaged 92.16 total uses (SD = 36.99 uses).

(a) Would it be reasonable to apply a pooled two-sample t-test to these data?

---

**Validity conditions for Student’s t test**

- **Study design:** The observations within each sample must be independent (e.g., should be reasonable to regard the data as random samples from the populations, in which case the populations must be large relative to the sample sizes, or random assignment). The two samples are independent of each other (e.g., not matched pairs).
- **Population distributions:** We need the sampling distributions of the sample means to be approximately normal. We can assume this if either the population distributions are normal or the sample sizes are large (like both bigger than 20).
- **Homoscedasticity:** The population variances are the same.

---

**Checking the conditions**

- No biases in the study design, no “nesting” of the data, not paired samples
- Context may imply normally distributed populations. Otherwise look at graphs including probability plots. Can use a Shapiro-Wilk test for normality but t-test is rather “robust” against departures from normality.
- There are “tests” for $H_0: \sigma_1^2 = \sigma_2^2$ (e.g., “Bartlett’s Test”) but the t-test is rather “robust” against departures from equal variances as well, at least with equal sample sizes. So many people advocate using the Welch procedure all the time and not bothering with this additional condition. The loss in power is small if the variances really are equal but the power of the tests of equal variance are even worse. (“It’s like sending out a rowboat to check the waters for an ocean liner.”)
  - One simplified informal check for this condition is that the ratio of the sample standard deviations (largest over smallest) is less than 2.

(b) What is your estimate for the average score on this task? What is your estimate for the average “effect” of telling someone to imagine they are a rigid librarian? For telling someone they are an eccentric poet?
Definitions: Instead of focusing in the difference in treatment means $\mu_1 - \mu_2$, so can focus on “treatment effects” where $\mu_i = \mu + \alpha_i$ and $\alpha_1 + \alpha_2 = 0$. But there will still be some “random variation” in the individual responses, so our model equation is $E(Y_{ij}) = \mu + \alpha_i + \epsilon_{ij}$. Then our validity conditions actually become about the $\epsilon_i \sim N(0, \sigma^2)$.

(c) Instead of writing $H_0: \mu_1 - \mu_2 = 0$, how can we write the null hypothesis in terms of the $\alpha_i$?

(d) How can we estimate $\alpha_i$ from the sample data?

(e) How can we estimate $\epsilon_i$?

Definition: The residual is the difference between the observed $y_{ij}$ and the predicted $\hat{y}_{ij}$

$$e_{ij} = y_{ij} - \hat{y}_{ij}$$

We will think of the residuals as telling us about the "unexplained variation" in the data.

So rather than checking the normality of each sample individually, you can just check the normality of the residuals. We can also look at a plot of the residuals across the two groups. The graphs should look very similar, especially in their variability. If one set of residuals is much less variable than the other, we do not consider the equal variance condition met. If you decide either condition is violated, you should take remedial action like a transformation (see Ch. 4).

“...The discussion that follows in this chapter will continue to use the terminology of samples rather than of treatment groups, even though the methods also apply to data from randomized experiments with I treatment groups. An additive treatment effect model asserts that an experimental unit that would produce a response of $Y_1$ on treatment 1 would produce a response of $Y_1 + \delta_1$ on treatment 2, $Y_1 + \delta_2$ on treatment 3 and so on. As before, exact randomization tests are available for inferences about the delta’s, but approximations based on the tools developed for random samples from populations are usually adequate. The practical upshot is that data from randomized experiments will be analyzed in exactly the same way as samples from populations, but concluding statements will be worded in terms of treatment effects rather than differences in population means.”
Some R Reminders

Loading in quantitative data
1. Copy and paste tab delimited data
   Select the data (and headers) and copy to your clipboard
   PC: mydata = read.table("clipboard", header=TRUE)
   MAC: mydata = read.table(pipe("pbpaste"), header=TRUE)
2. In RStudio, copy in a web URL
   Select Import Dataset > From Web URL, paste the URL
   Works best for tab delimited .txt files
3. In RStudio, import a txt file
   Select Import Dataset > From Text File, select a .txt file on your computer
   You can also get a dialog box using
   mydata = read.table(file.choose(), header=T)
I recommend then “attaching” your data set (attach(mydata)), otherwise, you can refer to
individual variables using the format mydata$variable.
To check that your data loaded in correctly, use View(mydata) or head(mydata)

Two-sample t-tests
1. Summary data
   • iscamtwosamplet(x1 = , sd1 = , n1 = , x2= , sd2 = , n2 = , alt = "greater")
2. Raw data
   • t.test(y~x, alt="two.sided", var.equal=FALSE) #stacked data
     Note: Look at the means to see the order of subtract or use levels(x) to see how R has
     ordered the categories, will then subtract first group listed minus second group listed
   • t.test(g1, g2, alt="greater", conf.level = .95) #unstacked data
     Note: R subtracts first group minus second group

To unstack data
> regular = ounces[group == "regular"]
> secret = ounces[group=="secret"]

Normal probability plots
> qqnorm(ounces)

To “panel” the graph window
> par(mfrow=c(2,2)) # graph window is now 2x2