Chapters 4-6: Inference with two samples

Read sections 4.2.5, 5.2, 5.3, 6.2

COMPARING TWO POPULATION MEANS

When presented with two samples that you wish to compare, there are two possibilities:
I. independent samples and **II. paired samples**.

I. INDEPENDENT SAMPLES:

- Select a SRS of size n_1 from population 1 which has mean μ_1 and standard deviation σ_1. Independently from the first sample, select a SRS of size n_2 from population 2 which has mean μ_2 and standard deviation σ_2.
- If sampling a finite population without replacement, then $.05N_1 \geq n_1$ and $.05N_2 \geq n_2$.
- The difference between the two population means is the parameter of interest, $\mu_1 - \mu_2$.
- The statistic $\overline{X}_1 - \overline{X}_2$ is a point estimator of the parameter $\mu_1 - \mu_2$.

Note about Study Design:

- Completely Randomized Design Experiment
  - Two treatment groups from a CRD are independent.
  - In a CRD, it is possible to claim that the treatment caused the difference in means.
  - If the individuals in the CRD were chosen from a SRS, then conclusions about $\mu_1 - \mu_2$ can be extended to the populations from which the individuals were drawn. However, if the individuals were not a SRS, then conclusions to larger populations are dubious.
- Observational Study
  - Two samples to be compared in an observational study are considered to be independent if individuals were randomly chosen from each respective population.
  - Do not claim that an explanatory variable caused the difference in means.

Facts about the sampling distribution of $\overline{X}_1 - \overline{X}_2$:

- $\mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2$, so $\overline{X}_1 - \overline{X}_2$ is an unbiased point estimator of $\mu_1 - \mu_2$.
- $\sigma^2_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ and $\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
- If the sampling distributions of $\overline{X}_1$ and $\overline{X}_2$ are approximately normal, then the sampling distribution of $\overline{X}_1 - \overline{X}_2$ is approximately normal, i.e., $\overline{X}_1 - \overline{X}_2 \dot{\sim} N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$.

Un-pooled Approach to estimating and testing $\mu_1 - \mu_2$ (5.3.1 and 5.3.2):

- A **Confidence Interval for** $\mu_1 - \mu_2$ is $\boxed{\overline{X}_1 - \overline{X}_2 \ \pm \ t_{1-\frac{\alpha}{2},\text{df}}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$
  - $df = \frac{(V_1+V_2)^2}{\frac{V_1^2}{n_1-1}+\frac{V_2^2}{n_2-1}}$ and $V_1 = \frac{S_1^2}{n_1}$ and $V_2 = \frac{S_2^2}{n_2}$.
  - A conservative degree of freedom estimate is $df = \min(n_1 - 1, n_2 - 1)$.
  - Check the assumptions that the SRS's are independent and that $\overline{X}_1$ and $\overline{X}_2$ are normal!
- A **Hypothesis Test for** $\mu_1 - \mu_2$ is:
  1. Hypotheses:

     $\text{H}_0 : \mu_1 - \mu_2 = \Delta_0$ where Δ_0 is a specific value (usually 0).

     Choose one: $\text{H}_a : \mu_1 - \mu_2 > \Delta_0$ / $\text{H}_a : \mu_1 - \mu_2 < \Delta_0$ / $\text{H}_a : \mu_1 - \mu_2 \neq \Delta_0$

     NOTE: Perform steps 2-4 assuming that H_0 is true!
  2. Assumptions:
     (a) SRS's from each population.
     (b) The two SRS's are independent of each other.
     (c) If sampling finite populations without replacement, then $.05N_1 \geq n_1$ and $.05N_2 \geq n_2$.
     (d) The sampling distributions of $\overline{X}_1$ and $\overline{X}_2$ must be at least approximately normal (so the data is normal OR $n_1, n_2 \geq 15$ for symmetric non-normal data OR $n_1, n_2 \geq 30$ for skewed data). If H_0 is true, then $\left(\overline{X}_1 - \overline{X}_2\right) \ \dot{\sim} \ N\left(\Delta_0, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$.
  3. Test Statistic:
     - $T = \frac{\overline{X}_1 - \overline{X}_2 \ - \ \Delta_0}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$
     - When H_0 is true, $T \sim t(df)$ where $df = \frac{(V_1+V_2)^2}{\frac{V_1^2}{n_1-1}+\frac{V_2^2}{n_2-1}}$.
     - A conservative degree of freedom estimate is $df = \min(n_1 - 1, n_2 - 1)$.
  4. The *p*-value is the probability of obtaining a sample statistic that is as extreme or more extreme than what was actually observed assuming that H_0 is true.

Alternative Hypothesis	*p*-value
$\text{H}_a : \mu_1 - \mu_2 > \Delta_0$	$P(T > t)$
$\text{H}_a : \mu_1 - \mu_2 < \Delta_0$	$P(T < t)$
$\text{H}_a : \mu_1 - \mu_2 \neq \Delta_0$	$2P(T > \lvert t \rvert)$

  5. Decision: Given a predetermined *significance level* α:
     - **Reject** H_0 if *p*-value $\leq \alpha$.
     - **Fail to Reject** H_0 if *p*-value $> \alpha$.
  6. Conclusion: a statement of how much evidence there is for the alternative hypothesis.
     - If you reject H_0, then you conclude "The evidence suggests H_a".

– If you fail to reject H_0, then you conclude "The evidence fails to suggest H_a".

In each case, specify H_a in terms of the problem.

EXAMPLE: To compare the mean cholesterol content (in milligrams) of grilled chicken sandwiches from Arby's and McDonald's, you randomly select several sandwiches from each restaurant and measure their cholesterol contents. From Arby's, $n_1 = 15$ sandwiches are randomly selected, $\bar{x}_1 = 60$ mg and $s_1 = 3.59$ mg. From McDonald's, $n_2 = 12$ sandwiches are randomly selected, $\bar{x}_2 = 70$ mg and $s_2 = 2.41$ mg.

Construct an 90% confidence interval for the difference in mean cholesterol content of grilled chicken sandwiches.

Is there sufficient evidence to conclude the McDonald's grilled chicken sandwiches have a higher cholesterol content, on average, compared to Arby's?

1. Hypotheses:

2. Check assumptions:

3. Test statistic:

4. p-value:

5. Decision:

6. Conclusion:

Pooled Approach to testing and estimating $\mu_1 - \mu_2$ (5.3.6):

- Assume that $\sigma_1^2 = \sigma_2^2$, called the **homogeneity of variance** assumption.
- The pooled approach is more powerful than the un-pooled approach (i.e. Power= $1 - \beta$ is larger) if σ_1^2 and σ_2^2 are truly equal since the degrees of freedom for the pooled approach is larger than the degrees of freedom for the un-pooled approach.
- If $\sigma_p^2 = \sigma_1^2 = \sigma_2^2$, then $\sigma_{\overline{X}_1 - \overline{X}_2} = \sigma_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ where σ_p^2 is the **pooled population variance**.
- **Confidence Interval for** $\mu_1 - \mu_2$ is $\boxed{\overline{X}_1 - \overline{X}_2 \ \pm \ t_{1-\frac{\alpha}{2},\text{df}} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$.
  - The **pooled sample variance** is $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$, an unbiased estimator of σ_p^2.
  - $df = n_1 + n_2 - 2$.
  - Check the assumptions that the SRS's are independent and that $\overline{X}_1$ and $\overline{X}_2$ are normal!

- **Hypothesis Testing for** $\mu_1 - \mu_2$:

1. Hypotheses:

   $\text{H}_0 : \mu_1 - \mu_2 = \Delta_0$ where Δ_0 is a specific value (usually 0).

   Choose one:
   $\text{H}_a : \mu_1 - \mu_2 > \Delta_0$
   $\text{H}_a : \mu_1 - \mu_2 < \Delta_0$
   $\text{H}_a : \mu_1 - \mu_2 \neq \Delta_0$

   NOTE: Perform steps 2-4 assuming that H_0 is true!

2. Assumptions:
   (a) SRS's from each population.
   (b) The two SRS's are independent of each other.
   (c) If sampling finite populations without replacement, then $.05N_1 \geq n_1$ and $.05N_2 \geq n_2$.
   (d) The sampling distributions of $\overline{X}_1$ and $\overline{X}_2$ must be at least approximately normal (so the data is normal OR $n_1, n_2 \geq 15$ for symmetric non-normal data OR $n_1, n_2 \geq 30$ for skewed data). If H_0 is true, then $\left(\overline{X}_1 - \overline{X}_2\right) \ \dot{\sim} \ N\left(\Delta_0, \sigma_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$.
   (e) Homogeneity of variance: $\sigma_1^2 = \sigma_2^2$. If one sample standard deviation is more than twice the other, then this assumption is suspect.

3. Test Statistic:
   - $T = \frac{\overline{X}_1 - \overline{X}_2 \ - \ \Delta_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
   - If H_0 is true, $T \sim t(df)$ where $\text{df} = n_1 + n_2 - 2$.

4. , 5. and 6. Find the p-value, make a Decision, and give a Conclusion as in the un-pooled case.

EXAMPLE: Redo the grilled chicken problem using a pooled variance this time.

II. PAIRED SAMPLES (5.2.1 and 5.2.2):

- Samples can be paired when:
  - From a SRS of size n, two measurements (X_i, Y_i) are recorded for each individual, such as when there are before-treatment and after-treatment measurements.
  - From two SRS's, each of size n, two measurements (X_i, Y_i) are recorded for two similar individuals, such as when there are blocks of two individuals, and each individual gets a different treatment.
- The paired measurements are subtracted, $d_i = X_i - Y_i$. The differences $d_1, d_2, ..., d_n$ form a new set sample of data, which is analyzed using the one sample procedures from Chapter 10.
- The sample mean difference, $\bar{d}$, is an unbiased point estimator of the parameter, μ_{d}, the mean of the population of differences. Your book uses the notation $\bar{x}_d$ instead of $\bar{d}$.
- The **sampling distribution of** d has mean $\mu_{\bar{d}} = \mu_d$ and sampling variability $\sigma_{\bar{d}}^2 = \frac{\sigma_d}{\sqrt{n}}$.
- A **Confidence Interval for** μ_d is $\boxed{\bar{d} \pm t_{1-\frac{\alpha}{2},\mathrm{n}-1}\frac{S_{\mathrm{d}}}{\sqrt{n}}}$. Check the Assumption that $\bar{d}$ is normal!
- A **Hypothesis Test for** μ_d:

1. Hypotheses:

   $\mathrm{H}_0 : \mu_d = \delta_0$ where δ_0 is a specific value (usually 0).

   Choose one:
   $\mathrm{H}_a : \mu_d > \delta_0$
   $\mathrm{H}_a : \mu_d < \delta_0$
   $\mathrm{H}_a : \mu_d \neq \delta_0$

   NOTE: Perform steps 2-4 assuming that H_0 is true!

2. Assumptions: Same as for hypothesis testing using one sample:
   (a) The data must be a SRS.
   (b) If sampling a finite population without replacement, then $.05N \geq n$.
   (c) The sampling distribution of $\bar{d}$ must be at least approximately normal (so the differences are normal OR $n \geq 15$ for symmetric non-normal differences OR $n \geq 30$ for skewed differences). If H_0 is true, then $\bar{d} \dot{\sim} N\left(\delta_0, \frac{\sigma_d}{\sqrt{n}}\right)$.

3. The Test Statistic is $T = \frac{\bar{d} - \delta_0}{\frac{S_d}{\sqrt{n}}}$ where $T \sim t(n-1)$ when H_0 is true.

4. p-value

Alternative Hypothesis	p-value
$\mathrm{H}_a : \mu_d > \delta_0$	$P(T > t)$
$\mathrm{H}_a : \mu_d < \delta_0$	$P(T < t)$
$\mathrm{H}_a : \mu_d \neq \delta_0$	$2P(T > \lvert t \rvert)$

5. and 6. Make a Decision and give a Conclusion.

EXAMPLE: An herbal medicine is tested on 14 randomly selected patients with sleeping disorders. The table shows the hours of sleep patients got during one night without using the herbal medicine and during another night after the herbal medicine was administered.

Patient	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Without medicine	1.0	1.4	3.4	3.7	5.1	5.1	5.2	2.3	5.5	5.8	4.2	4.8	2.9	4.5
With medicine	2.9	3.3	3.5	4.4	5.0	5.0	5.2	5.3	6.0	6.5	4.4	4.7	3.1	4.7
Difference	-1.9	-1.9	-0.1	-0.7	0.1	0.1	0	0	-0.5	-0.7	-0.2	0.1	-0.2	-0.2

The mean difference in sleep is $\bar{d}$ = -0.436 and the sample standard deviation of the differences is s_d = 0.677

Construct an 95% confidence interval for μ_d.

Is there sufficient evidence to suggest the herbal medicine provides a longer night's sleep, on average?

1. Hypotheses:

2. Check assumptions:

3. Test statistic:

4. p-value:

5. Decision:

6. Conclusion:

COMPARING POPULATION PROPORTIONS FROM INDEPENDENT SAMPLES

- Collect binary data from a SRS of size n_1 from population 1, which has mean p_1. Collect binary data from a SRS of size n_2 from population 2, which has mean p_2.
- If sampling a finite population without replacement, then $.05N_1 \geq n_1$ and $.05N_2 \geq n_2$.
- When comparing two population proportions, the difference between the two population proportions is the parameter of interest, $p_1 - p_2$.
- The statistic $\hat{p}_1 - \hat{p}_2$ is a point estimator of the parameter $p_1 - p_2$.

Note about Study Design: See notes for comparing two population means.

Facts about the sampling distribution of $\hat{p}_1 - \hat{p}_2$ (6.2.1):

- $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$ so $\hat{p}_1 - \hat{p}_2$ is an unbiased point estimator of $p_1 - p_2$.
- $\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$ and $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
- If $n_1 p_1 \geq 10$, $n_1(1-p_1) \geq 10$, $n_2 p_2 \geq 10$, and $n_2(1-p_2) \geq 10$ then

$$(\hat{p}_1 - \hat{p}_2) \dot{\sim} N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right).$$

Estimating and Testing $p_1 - p_2$ (6.2.2 and 6.2.3):

- A **Confidence Interval for** $p_1 - p_2$ is $\boxed{\hat{p}_1 - \hat{p}_2 \ \pm \ z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$ Check the assumptions that the SRS's are independent and that $\hat{p}_1$ and $\hat{p}_2$ are normal!
- A **Hypothesis Test for** $p_1 - p_2$:
  1. Hypotheses:
     $\mathrm{H}_0 : p_1 - p_2 = \Delta_0$ where Δ_0 is a specific value (usually 0).

     Choose one: $\mathrm{H}_a : p_1 - p_2 > \Delta_0$ / $\mathrm{H}_a : p_1 - p_2 < \Delta_0$ / $\mathrm{H}_a : p_1 - p_2 \neq \Delta_0$

     NOTE: Perform steps 2-4 assuming that H_0 is true!
  2. Assumptions:
     Under H_0, if $\Delta_0 = 0$, then $p = p_1 = p_2$, and so an estimate of p is

$$\hat{p} = \frac{\text{total successes}}{\text{total sample size}} = \frac{\text{count}_1 + \text{count}_2}{n_1 + n_2}.$$

     (a) SRS's from each population.
     (b) The two SRS's are independent of each other.
     (c) If sampling a finite population without replacement, then $.05N_1 \geq n_1$ and $.05N_2 \geq n_2$.
     (d) The sampling distributions for $\hat{p}_1$ and $\hat{p}_2$ must be approximately normal.

- If $\Delta_0 \neq 0$, then check the sample size like this: $n_1\hat{p}_1 \geq 10$, $n_1(1-\hat{p}_1) \geq 10$, $n_2\hat{p}_2 \geq 10$, and $n_2(1-\hat{p}_2) \geq 10$.
- If $\Delta_0 = 0$, use $\hat{p}$ to check the sample size: $n_1\hat{p} \geq 10$, $n_1(1-\hat{p}) \geq 10$, $n_2\hat{p} \geq 10$, and $n_2(1-\hat{p}) \geq 10$.

For large sample sizes, if H_0 is true,

$$(\hat{p}_1 - \hat{p}_2) \dot{\sim} N\left(\Delta_0, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right).$$

For small sample sizes, consider *Fisher's exact test* (6.5.1 and 6.5.3).

3. Test Statistic:
   - If $\Delta_0 \neq 0$, then $Z = \frac{\hat{p}_1 - \hat{p}_2 \ - \ \Delta_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$.
   - If $\Delta_0 = 0$, then $p = p_1 = p_2$ and now $Z = \frac{\hat{p}_1 - \hat{p}_2 \ - \ \Delta_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$.
   - $Z \sim N(0,\ 1)$ if H_0 is true.
4. *p*-value:

Alternative Hypothesis	*p*-value
$H_a : p_1 - p_2 > \Delta_0$	$P(Z > z)$
$H_a : p_1 - p_2 < \Delta_0$	$P(Z < z)$
$H_a : p_1 - p_2 \neq \Delta_0$	$2P(Z > \lvert z \rvert)$

5. and 6. Make a Decision and give a Conclusion.

EXAMPLE: Among 2200 randomly selected male car occupants over the age of 8, 72% wear seat belts. Among 2380 randomly selected female car occupants over the age of 8, 84% wear seat belts. Test the claim that both genders have the same rate of seat belt use. Use $\alpha = 0.05$. Does there appear to be a gender gap?

1. Hypotheses:

2. Check assumptions:

3. Test statistic:

4. p-value:

5. Decision:

6. Conclusion:

Construct a 90% CI to estimate the true difference in the proportions of seat belt users in the male and female populations.

R code

```
# Comparing Two Population Means using UNPOOLED Variance:
# Arby's and McDonald's chicken sandwiches Example

> t.test(Arby,McD,conf.level=0.90,var.equal=FALSE,alternative="two.sided")

        Two Sample t-test

data:  Arby and McD
t = -8.62834, df = 24.37, p-value = 3.556461e-09
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -11.981841  -8.018159
sample estimates:
mean of x mean of y
 60       70


# Comparing Two Population Means using POOLED Variance:
# Arby's and McDonald's chicken sandwiches Example

> t.test(Arby,McD,conf.level=0.90,var.equal=TRUE,alternative="two.sided")

        Two Sample t-test

data:  Arby and McD
t = -8.64354, df = 25, p-value = 2.786212e-09
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -11.976205  -8.023795
sample estimates:
mean of x mean of y
 60        70


# Comparing Two Population Means using Matched Pairs t-test:
```

```
# Herbal Medicine Example

# Here's the data given as a table:
> nomed=c(1,1.4,3.4,3.7,5.1,5.1,5.2,2.3,5.5,5.8,4.2,4.8,2.9,4.5)
> med=c(2.9,3.3,3.5,4.4,5,5,5.2,2.3,6,6.5,4.4,4.7,3.1,4.7)
> rbind(nomed,med,nomed-med)
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
nomed  1.0  1.4  3.4  3.7  5.1  5.1  5.2  2.3  5.5   5.8   4.2   4.8   2.9   4.5
med    2.9  3.3  3.5  4.4  5.0  5.0  5.2  2.3  6.0   6.5   4.4   4.7   3.1   4.7
      -1.9 -1.9 -0.1 -0.7  0.1  0.1  0.0  0.0 -0.5  -0.7  -0.2   0.1  -0.2  -0.2

> t.test(nomed,med,conf.level=0.95,paired=TRUE,alternative="less")

        Paired t-test

data:  nomed and med
t = -2.4094, df = 13, p-value = 0.01576
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.1154539
sample estimates:
mean of the differences
             -0.4357143


# Comparing Two Population Proportions:
# Seat Belt Example
> prop.test(c(.72*2200,.84*2380),c(2200,2380),conf.level=.95,alternative="two.sided",
            correct=F)

        2-sample test for equality of proportions without continuity correction

data:  c(0.72 * 2200, 0.84 * 2380) out of c(2200, 2380)
X-squared = 96.6826, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1438526 -0.0961474
sample estimates:
prop 1 prop 2
  0.72   0.84
```

Exercises

Two independent means on p263: 5.25, 5.26 (answers: Yes because n is large; test statistic is -5.78, p-value is tiny ($< 10^{-6}$), so the evidence suggests that the age difference is not due to chance), 5.31 - 5.35 odd, 5.38 (answers: T (if first sample is symmetric and the second is skewed) TF)

Paired means on p259: 5.15, 5.16 (answers: TTTF), 5.17, 5.18 (answers: YYN), 5.19 - 5.23 odd, 5.24 (answer: F), 5.27, 5.29

Two independent proportions on p317: 6.23 - 6.33 odd, 6.37