Stat 324 – Introduction to Statistics for Engineers

LECTURE 12: SAMPLING DISTRIBUTION, HYPOTHESIS TESTING, AND CONFIDENCE INTERVALS FOR TWO INDEPENDENT POPULATIONS

OTT AND LONGNECKER 6.2, 6.3

Comparing Two Independent Populations

Up to this point, we have had **a single random sample drawn from one population**, from which we either

(1) made estimate about a population mean or proportion using a confidence interval

or

(2) performed a hypothesis test to determine if our population parameter of interest "matched" some pre-known population median, mean or proportion.

Often, we are interested in comparing the central tendencies or proportion sucesses across **two independent populations.**

We will look at options for testing based on different assumptions we want to make about the populations.

(1) Do they have equal _____?

(2)Do we have known variance for each/either _____?

(3) Is each population roughly normally distributed?

(4) Is each sample size large enough that the CLT will kick in and make the sampling distribution of means approximately Normal?

Example 1: Comparing populations of lizards

The horned lizard *Phrynosoma mcalli* is named for the fringe of spikes around the back of the head. It was thought that the spikes may provide the lizard protection from its primary predator, the loggerhead shrike, Lanius ludovicanus, though there was not much existing quantitative evidence to support this.

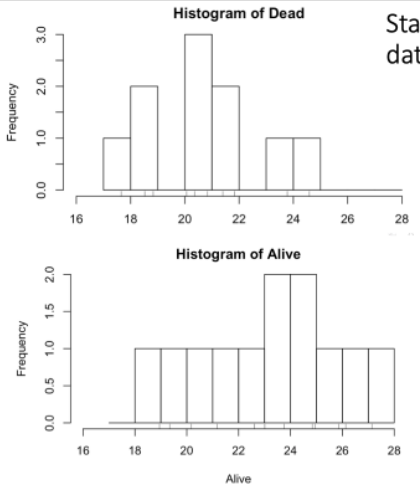
Researchers questioned "Is there any difference in the size of the spikes between those lizards killed by shrikes and those who are alive?"

In order to compare the two populations, random samples were taken from (D) dead lizards known to be killed by shrikes, and the population of (L) live lizards from the same geographic location. The longest spike was measured on each sampled lizard, in mm.

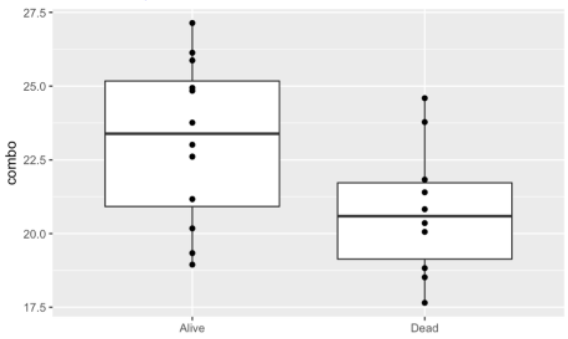
Dead Group: 17.65, 20.83, 24.59, 18.52, 21.40, 23.78, 20.36, 18.83, 21.83, 20.06

Live Group: 23.76, 21.17, 26.13, 20.18, 23.01, 24.84, 19.34, 24.94, 27.14, 25.87, 18.95, 22.61

Notice, even if we find a statistical difference in means, we cannot conclude based on this data that spike length directly caused a difference in survival chances…



Start with graphical and numeric summaries for quantatative data…histogram, boxplot, dotplot, stem and leaf



From the graphs and summary measures we see:

both roughly symmetric

	Dead	Alive
Sample Size (n)	10	12
Mean	20.79	23.16
Median	20.59	23.39
Standard Deviation	2.22	2.76

slightly higher, is it high enough to say significant difference?

Comparing means in two independent populations

L = alive
D = dead

Is there a difference in the two population means μ_L and μ_D ?

Some necessary notation:

	RV	Mean	Variance	Sample Size	Sample Mean	Sample Variance
Population L	X_L	μ_L	σ_L^2	n_L	$\overline{X_L}$	s_L^2
Population D	X_D	μ_D	σ_D^2	n_D	$\overline{X_D}$	s_D^2

We would like to test whether $\mu_L = \mu_D$

1. Make hypotheses: $H_0: \mu_L = \mu_D$ or $H_0: \mu_L - \mu_D = 0$

We can equivalently write the hypothesis in terms of the difference in means (which will make our work easier later…)

$H_A: \mu_L \neq \mu_D$ or $H_0: \mu_L - \mu_D \neq 0$

2. Choose a significance level α to control type 1 Error Rate.

Comparing means in two independent populations

3. Choose a Test Statistic that reflect the relationship in the null hypothesis.

With 1 sample, when we were interested in hypothesis tests about the mean and we did not know population sd, we used a T test statistic: $t_{n-1} = \frac{\text{Sample mean} - H_0: \mu}{\text{Estimate SE mean}} = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$

For inference about diff $= \mu_L - \mu_D$, we can use the statistic $\widehat{diff} = \overline{X_L} - \overline{X_d}$, but we need to know its distribution.

$E(\widehat{diff}) = E(\overline{X_L} - \overline{X_D}) =$ $E(\bar{x}_L) - E(\bar{x}_D) = \mu_L - \mu_D$. so

$\widehat{diff} = \overline{X_L} - \overline{X_D}$ is an unbiased estimator of the true difference in means.

For independent populations, X_L and X_D, and also independence within each population:

$Var(\widehat{diff}) = Var(\overline{X_L} - \overline{X_D}) =$ $Var(\bar{x}_L) + Var(\bar{x}_D) = \frac{\sigma_L^2}{n_L} + \frac{\sigma_D^2}{n_D}$

$SE(\widehat{diff}) = SE(\overline{X_L} - \overline{X_D}) =$ $\sqrt{\frac{\sigma_L^2}{n_L} + \frac{\sigma_D^2}{n_D}}$

Comparing means in two independent populations

3. Choose a Test Statistic that reflect the relationship in the null hypothesis.

For inference about diff $= \mu_L - \mu_D$, we can use the statistic $\widehat{diff} = \overline{X_L} - \overline{X_D}$, after we know its distribution.

If X_L and $X_D \sim N$, $\overline{X_L} \sim$ N , $\overline{X_D} \sim$ N and therefore $\overline{X_L} - \overline{X_D} \sim$ N

(each is linear combo from N) (Linear Combo of Normal)

So, $\overline{X_L} - \overline{X_D} \sim N(\mu_L - \mu_D, \frac{\sigma_L^2}{n_L} + \frac{\sigma_D^2}{n_D})$

If we knew σ_L^2 and σ_D^2, then under H_o: $(\mu_L - \mu_D)_0 = diff_0$

$\frac{(\overline{X_L} - \overline{X_D}) - (\mu_L - \mu_D)_0}{\sqrt{\frac{\sigma_L^2}{n_L} + \frac{\sigma_D^2}{n_D}}} \sim N(0,1)$ but typically, we won't know σ_L^2 or σ_D^2, so we need to estimate them from s_L^2 or s_D^2

Comparing means in two independent populations (Assuming $\sigma_L^2 = \sigma_D^2$)

Since we don't know σ_1^2 or σ_2^2 and they appear to be *similar* based on our sample data, we assume they are equal and calculate a "pooled" estimate by doing a weighted average of the sample standard deviations:

$s_p^2 = \frac{(n_L - 1)s_L^2 + (n_D - 1)s_D^2}{n_L + n_D - 2} =$ $s_p^2 = \frac{\sum_{i=1}^{n_L} (x_{L,i} - \bar{x}_L)^2 + \sum_{j=1}^{n_D} (x_{D,j} - \bar{x}_D)^2}{n_L + n_D - 2}$

- Numerator is a total sum of squared deviations from each mean
- Denominator is a degrees of freedom (we had to estimate the population means with our sample means)

Similar guideline: when the sample sizes are close, if $0.5 < \frac{s_L}{s_D} < 2$, we can be comfortable assuming $\sigma_L^2 = \sigma_D^2$ equal. The more sample sizes differ, the more carefully we need to compare variance.

In our case, $\frac{s_L}{s_D} = \frac{2.76}{2.22} = 1.24$ so we'll assume $\sigma_L^2 = \sigma_D^2 = s_p^2$

*There are formal tests for "equal variance" but often we will just compare sample SD and graphs of data.

Comparing means in two independent populations (Assuming $\sigma_L^2 = \sigma_D^2$)

$SE = \sqrt{\frac{\sigma_L^2}{n_L} + \frac{\sigma_D^2}{n_D}}$

For inference about $\mu_1 - \mu_2$, we can use the statistic $\overline{X_1}$- $\overline{X_2}$ with

$$E(\overline{X_L} - \overline{X_D}) = \mu_L - \mu_D \text{ and } SE(\overline{X_L} - \overline{X_D}) = \sqrt{\frac{s_p^2}{n_L} + \frac{s_p^2}{n_D}} = s_p\sqrt{\frac{1}{n_L} + \frac{1}{n_D}}$$

We would like to assume that our distribution of $\overline{X_L}$- $\overline{X_D} \approx N$ so we can say

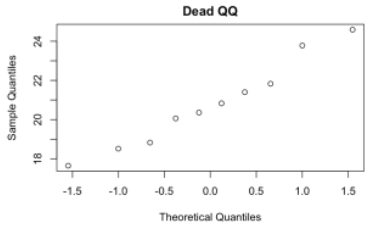
$$\frac{(\bar{x}_A - \overline{x_D}) - (\mu_L - \mu_D)_0}{\sqrt{\frac{s_p^2}{n_L} + \frac{s_p^2}{n_D}}} = \frac{(\bar{x}_L - \overline{x_D}) - (\mu_L - \mu_D)_0}{s_p\sqrt{\frac{1}{n_L} + \frac{1}{n_D}}}$$

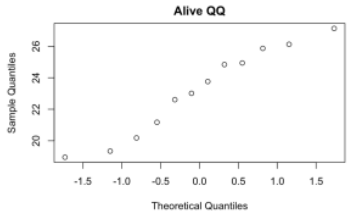
is a T statistic with $n_L + n_D - 2$ df

Check normality assumption of each population with seperate QQ plots (also refer back to histograms).

From these, we would say:

normality of pops reasonable based on linearity of QQ plot





Comparing means in two independent populations (Assuming $\sigma_1^2 = \sigma_2^2$)

For inference about $\mu_1 - \mu_2$, we can use the statistic $\overline{X_1}$- $\overline{X_2}$ and after making the assumptions:

(1) Population 1 and 2 are independent and observations within each sample are independent
(2) Both Population 1 and Population 2 follow normal distributions (or both sample sizes are large enough that the CLT could apply to each and therefore the combination)
(3) Population variances are equal (from graphs of samples), [rule of thumb: plausible that population variances are equal if the larger sample variance is less than twice the smaller]

We can do a t test where $H_o: \mu_1 - \mu_2 = \delta_0$, $t_{obs} = \frac{(\bar{x}_1 - \overline{x_2}) - \delta_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

and pvalue based on H_A:

$$H_A: \mu_1 - \mu_2 > \delta_0 \implies pvalue = P(T_{df=n_1+n_2-2} > t_{obs})$$
$$H_A: \mu_1 - \mu_2 < \delta_0 \implies pvalue = P(T_{df=n_1+n_2-2} < t_{obs})$$
$$H_A: \mu_1 - \mu_2 \neq \delta_0 \implies pvalue = 2 * P(T_{df=n_1+n_2-2} > |t_{obs}|)$$

Or a $100(1-\alpha)\%$ CI interval for $\mu_1 - \mu_2$: $(\bar{X}_1 - \overline{X_2}) \pm t_{(n_1+n_2-2,\frac{\alpha}{2})} s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Comparing means in two independent Lizard populations (Assuming $\sigma_L^2 = \sigma_D^2$)

Dead Group: 17.65, 20.83, 24.59, 18.52, 21.40, 23.78, 20.36, 18.83, 21.83, 20.06
Live Group: 23.76, 21.17, 26.13, 20.18, 23.01, 24.84, 19.34, 24.94, 27.14, 25.87, 18.95, 22.61

Test whether the lizard populations have the same mean spike lengths at the 5% level and find a 95% confidence interval for the difference in mean lengths .

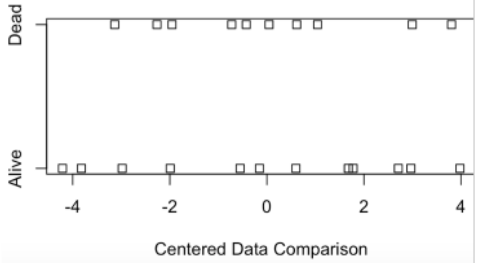
	Dead	Alive
Sample Size (n)	10	12
Mean	20.79	23.16
Median	20.59	23.39
Standard Deviation	2.22	2.76

Hypotheses: $H_0: \mu_L - \mu_D = 0$ $H_A: \mu_L - \mu_D \neq 0$

Check Assumptions:

(1) Independence within and between populations

(2) Check normality of both population: Check QQ Plots



(3) Check Equal Variance: (Centered graphs also help show spread)

$\frac{s_L}{s_D} = \frac{2.22}{2.76} = 0.80$ and sample sizes are close

Comparing means in two independent Lizard populations (Assuming $\sigma_L^2 = \sigma_D^2$)

$H_0: \mu_L - \mu_D = 0$ vs $H_A: \mu_L - \mu_D \neq 0$

Sample Stats	Dead	Alive
Size (n)	10	12
Mean	20.79	23.16
Median	20.59	23.39
SD	2.22	2.76

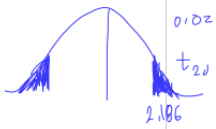
Calculate Test Statistic and P value

$t_{obs} = \frac{\text{Obs diff} - \text{Expec diff}}{\text{SE diff}} = \frac{(\bar{X}_L - \bar{X}_D) - 0}{s_p\sqrt{\frac{1}{n_L} + \frac{1}{n_D}}} = \frac{23.16 - 20.79}{2.531\sqrt{\frac{1}{12} + \frac{1}{10}}} = 2.186$

$s_p^2 = \frac{(10-1)2.22^2 + (12-1)2.76^2}{(10-1)+(12-1)} = \frac{126.1492}{20} = 6.407$ $\quad s_p = 2.531$

Df= $10 + 12 - 2 = 20$

Pvalue: $2 * P(t_{20} > 2.186) = 2 * (0.02, 0.025) = 0.04 < \text{pvalue} < 0.05$



$0.025 < \text{pvalue} < 0.02$ one sided double to get 2-sided

because pvalue < 0.05 Rstudio: 0.0409

We have sufficient evidence at the 5% level to suggest there is a difference in the mean spike length of Alive and Dead Horned Lizards. Reject the null

Comparing means in two independent Lizard populations (Assuming $\sigma_L^2 = \sigma_D^2$)

$H_0: \mu_L - \mu_D = 0$ vs $H_A: \mu_L - \mu_D \neq 0$

95%

Find a confidence interval for the difference in mean lengths .

Sample Stats	Dead	Alive
Size (n)	10	12
Mean	20.79	23.16
Median	20.59	23.39
SD	2.22	2.76

$Obs_{diff} = \bar{X}_L - \bar{X}_D = 23.16 - 20.79 = 2.37$

$SE_{diff} = s_p\sqrt{\frac{1}{n_L}+\frac{1}{n_D}} = 2.531\sqrt{\frac{1}{12}+\frac{1}{10}} = 1.084$

df= $10+12-2 = 20$ $t_{20,0.025} = 2.086$

CI: Obs diff $\pm$ $t_{20,0.025}$ SE diff $= 2.37 \pm 2.086(1.084) = (0.109, 4.631)$

Interpretation: we are 95% confident that the interval from 0.109 to 4.631 covers the true difference in mean spike length

Notice this interval does not contain 0, so again we have sufficient evidence at the 5% level to suggest there is a difference in the mean spike length of Alive and Dead Horned Lizards

Example 2: Comparing concrete breaking force.

Concrete is often reinforced with steel "rebar" ("reinforcing bar"). Steel is strong, but tends to corrode over time. An experiment tested two corrosion-resistant materials, one fiberglass and the other carbon. The experimenter wants to know: Is there a difference in (population) mean strength between the two types of beams?

Eight concrete beams with fiberglass reinforcement, and 11 with carbon reinforcement, were poured. Each was subjected to a load test, with the breaking force measured in kN (kiloNewtons):

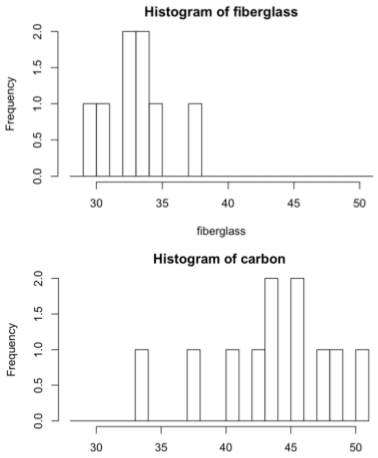
Fiberglass: 37.3, 29.6, 33.4, 33.6, 30.7, 32.7, 34.6, 32.3
Carbon: 48.8, 38.0, 42.2, 45.1, 33.8, 47.2, 50.6, 44.0, 43.9, 40.4, 45.8

We would like to test: $H_o: \mu_F - \mu_C = 0$ vs $H_A: \mu_F - \mu_C \neq 0$ at $\alpha = 0.05$

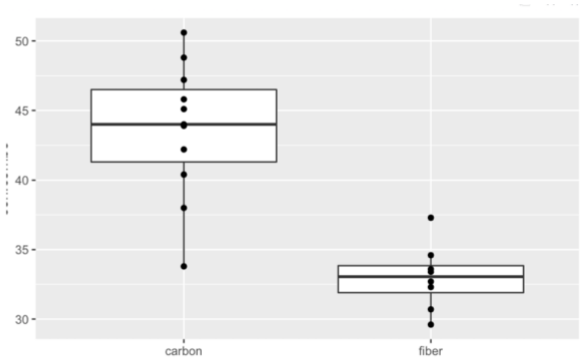
(or make a 95% confidence interval)

Where should we start?

graph



Start with graphical and numeric summaries



From the graphs and summary measures we see:

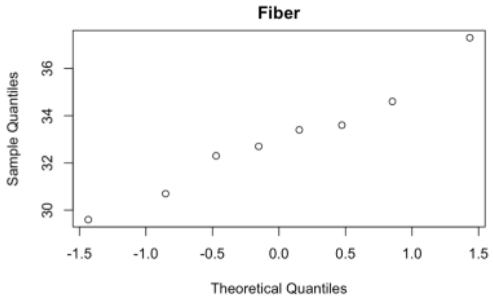
	Fiber	Carbon
Sample Size (n)	8	11
Mean	33.02	43.62
Median	33.05	44.00
Standard Deviation	2.36	4.86

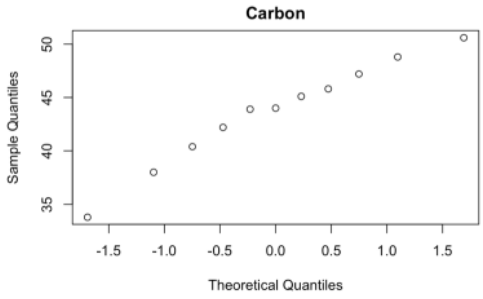
Comparing means of two independent populations

For inference about $\mu_1 - \mu_2$, we can use the statistic $\overline{X_1}$- $\overline{X_2}$ and after making the assumptions:

(1) Population 1 and 2 are independent and observations within each sample are independent

(2) Both Population 1 and Population 2 follow normal distributions (or both sample sizes are large enough that the CLT could apply to each and therefore the combination)





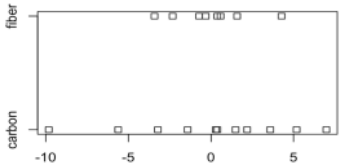
(3) Population variances are equal (from graphs of samples), [rule of thumb: plausible that population variances are equal if the larger sample S D is less than twice the smaller]

Comparing means of two independent populations assuming $\sigma_1^2 \neq \sigma_2^2$

(3) Population variances are equal?

a. Graph centered samples

b. $\frac{s_F}{s_C} = \frac{2.36}{4.86} = 0.48$



If it is not appropriate/less desirable to have the assumption that the variances are equal, we just leave the variance estimates from the two groups' sample standard deviations which gives us:

$$SE(\widehat{\overline{X_F} - \overline{X_C}}) = \sqrt{\frac{s_F^2}{n_F} + \frac{s_C^2}{n_C}}$$ and related statistic:

$$\frac{\bar{X}_F - \bar{X}_C - \delta_0}{\sqrt{\frac{s_F^2}{n_F} + \frac{s_C^2}{n_C}}} \approx t_v$$

Where degrees of freedom $v = \frac{\left(\frac{s_F^2}{n_F} + \frac{s_C^2}{n_C}\right)^2}{\frac{\left(\frac{s_F^2}{n_F}\right)^2}{n_F - 1} + \frac{\left(\frac{s_C^2}{n_C}\right)^2}{n_C - 1}}$, rounded down.

V is called the Satterthwaite approximation to the df and the resulting test is called the **Welch t-test** (for unequal variances).

Comparing means of two independent populations assuming $\sigma_1^2 \neq \sigma_2^2$

Sample	Fiber	Carbon
Size (n)	8	11
Mean	33.03	43.62
Median	33.05	44.00
SD	2.36	4.86

Fiberglass: 37.3, 29.6, 33.4, 33.6, 30.7, 32.7, 34.6, 32.3
Carbon: 48.8, 38.0, 42.2, 45.1, 33.8, 47.2, 50.6, 44.0, 43.9, 40.4, 45.8

We test: $H_o: \mu_F - \mu_C = 0$ vs $H_A: \mu_F - \mu_C \neq 0$ at $\alpha = 0.05$

Test whether the concrete populations have the same mean strength

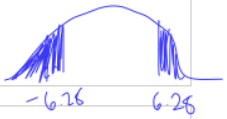
Check Assumptions: Independence, Normality, Equal Variance

pop ✓ sample ✓ QQ ✓ X

Calculate a test statistic under the null for the concrete strengths:

$$t_v = \frac{\bar{X}_F - \bar{X}_C - 0}{\sqrt{\frac{s_C^2}{n_C} + \frac{s_F^2}{n_F}}} = \frac{33.03 - 43.62 - 0}{\sqrt{\frac{4.86^2}{11} + \frac{2.36^2}{8}}} = -6.28 \qquad v = \frac{\left(\frac{2.36^2}{8} + \frac{4.86^2}{11}\right)^2}{\left(\frac{2.36^2}{8}\right)^2 + \left(\frac{4.86^2}{11}\right)^2} = 15.7 \text{ round down } 15$$

$pvalue =$ $2 * P(t_{15} < -6.28) < 0.001$. There is strong evidence to reject the null and suggest the mean strength of the Fiber and Carbon reinforced concrete are not equal.



Comparing means of two independent populations assuming $\sigma_1^2 \neq \sigma_2^2$

Sample	Fiber	Carbon
Size (n)	8	11
Mean	33.03	43.62
Median	33.05	44.00
SD	2.36	4.86

Fiberglass: 37.3, 29.6, 33.4, 33.6, 30.7, 32.7, 34.6, 32.3
Carbon: 48.8, 38.0, 42.2, 45.1, 33.8, 47.2, 50.6, 44.0, 43.9, 40.4, 45.8

Construct a 95% CI equivalent to :$H_o: \mu_F - \mu_C = 0$ vs $H_A: \mu_F - \mu_C \neq 0$ at $\alpha = 0.05$

Check Assumptions: Independence ✓, Normality ✓, Equal Variance ✗

$\widehat{SE_{(\bar{X}_F - \bar{X}_C)}} = \sqrt{\frac{2.36^2}{8} + \frac{4.86^2}{11}} = 1.664$ and $v = 15.7$, $t_{v=15.7,.025} = 2.123$

Df rounded down to 15, $t_{v=15,.025} =$ 2.131

$\bar{X}_F - \bar{X}_C$

$33.03 - 43.62 = -10.59$

Obs diff $\pm$ t * SE diff

$-10.59 \pm 2.131(1.664)$

$(-14.17, -7.01)$

Since the interval does not cover zero, we would reject the null hypothesis that the difference is zero, just as we concluded using the two sided test.

T Tests in R (?t.test)

1 population mean:

to test: $H_0: \mu_F = 30, H_A: \mu_F \neq 30$

```
t.test(fiber, mu=30)

data:  fiber
t = 3.6251, df = 7, p-value = 0.008453
alternative hypothesis: true mean is not equal to 30
95 percent confidence interval:
 31.05179 34.99821
```

2 populations' means with equal variance:

to test: $H_0: \mu_F - \mu_c = -10,$
$H_A: \mu_F - \mu_c < -10$
$(\sigma_F^2 = \sigma_C^2)$

```
t.test(fiber, carbon, var.equal=TRUE, mu=-10, alternative = "less")

        Two Sample t-test

data:  fiber and carbon
t = -0.31744, df = 17, p-value = 0.3774
alternative hypothesis: true difference in means is less than -10
95 percent confidence interval:
      -Inf -7.342502
```

2 populations' means with unequal variance:

to test: $H_0: \mu_F - \mu_c = -10,$
$H_A: \mu_F - \mu_c < -10$
$(\sigma_F^2 \neq \sigma_C^2)$

```
t.test(fiber, carbon, var.equl=FALSE, mu=-10, alternative = "less")
```

```
        Welch Two Sample t-test

data:  fiber and carbon
t = -0.35191, df = 15.251, p-value = 0.3649
alternative hypothesis: true difference in means is less than -10
95 percent confidence interval:
      -Inf -7.641444
```

Choosing between equal and unequal variance tests:

- If the variances are truly equal, but are allowed to differ (by using Welch's test), the test loses some power but is still a good test.
- If the variances are truly different, but they are assumed equal (so original 2 sample t test is used), the test can make wildly incorrect conclusions.
- There is usually more to lose in the second case. Therefore, if there is any doubt about the equality of the variances, it's generally safer to allow them to differ by using Welchs test and using the approximate DF give by the Satterthwaite approximation .

Example 3: Cricket Data

When sage crickets mate, the male allows the female to eat part of his hind wings. Does female hunger influence desire to mate? Do starved females have the same mean time to mating than normally fed females?

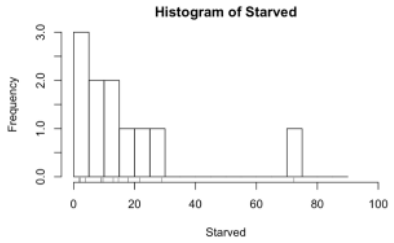
An experiment randomly assigned 24 females to two groups. One group of 11 was starved for two days, while the other group of 13 was fed normally. Each female was presented with a male and the time to mating (in hours) was recorded. The primary research question was: "Do starved females attempt mating more quickly than normally fed females?"

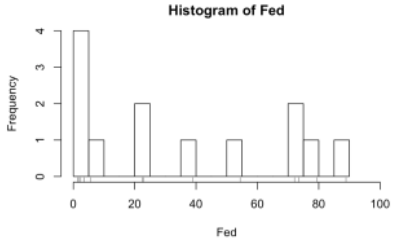
We test: $H_0: \mu_S - \mu_F = 0$ vs $H_A: \mu_S - \mu_F < 0$ at $\alpha = 0.05$

Here are the data from a study:

Starved: 1.9, 2.1, 3.8, 9.0, 9.6, 13.0, 14.7, 17.9, 21.7, 29.0, 72.3

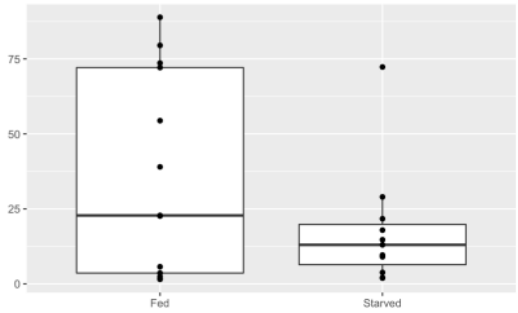
Fed: 1.5, 1.7, 2.4, 3.6, 5.7, 22.6, 22.8, 39.0, 54.4, 72.1, 73.6, 79.5, 88.9





	Starved	Fed
Sample Size (n)	11	13
Mean	17.73	35.98
Median	13.00	22.80
Standard Deviation	19.96	33.63

Start with graphical and numeric summaries



From the graphs and summary measures we see:

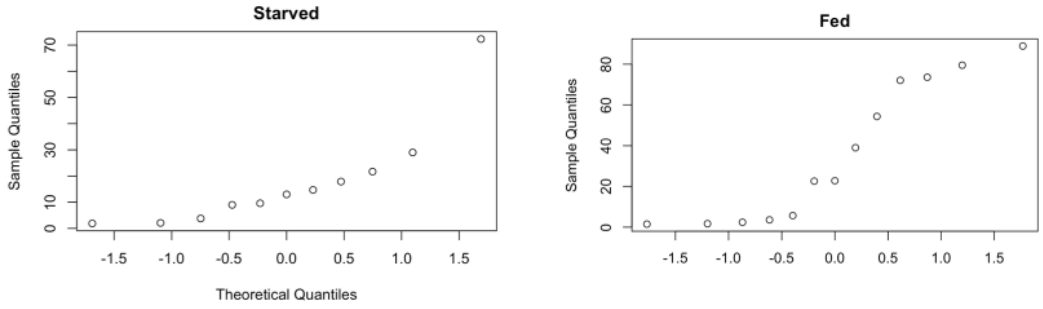
mean of starved is lower

looks skewed

Comparing means from two independent populations

For inference about $\mu_1 - \mu_2$, we can use the statistic $\overline{X_1}$- $\overline{X_2}$ and after making the assumptions:

(1) Population 1 and 2 are independent and and observations within each sample are independent
(2) Both Population 1 and Population 2 follow normal distributions (or both sample sizes are large enough that the CLT could apply to each and therefore the combination)



Since normality assumption of population is not supported by our sample data, we have a few options: (1) Bootstrap , (2) Wilcoxon Rank Sum (or Mann-Whitney Test), (3) Randomization

Bootstrap analysis for $\mu_1 - \mu_2$=0.

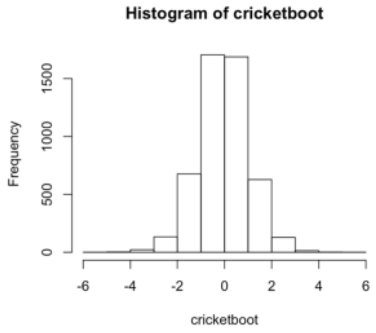
1. Draw a (original) simple random samples $x_{1,1}, x_{1,2}, \dots, x_{1,n_1}$ and $x_{2,1}, x_{2,2}, \dots, x_{2,n_2}$ of size n_1 and n_2 from the two populations. Compute $\overline{x_1}, s_1^2, \overline{x_2}, s_2^2$. Find $t_{obs} = \frac{(\overline{x_1}-\overline{x_2})-0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$.
2. Draw simple random samples with replacement: $x^*_{1,1}, x^*_{1,2}, \dots, x^*_{1,n_1}$ from sample 1 and $x^*_{2,1}, x^*_{2,2}, \dots, x^*_{2,n_2}$ from sample 2.
3. Compute the means and variances of the resampled data for each group separately. Call these $\overline{x_1^*}, s_1^{2*}, \overline{x_2^*}, s_2^{2*}$
4. Compute the statistic: $\hat{t} = \frac{(\overline{x_1^*}-\overline{x_2^*})-(\overline{x_1}-\overline{x_2})}{\sqrt{\frac{s_1^{2*}}{n_1}+\frac{s_2^{2*}}{n_2}}}$
5. Repeat steps 2-4 a large number, B times to get a collection of $\hat{t}$ that approximate the sampling distribution of $T = \frac{(\overline{X_1}-\overline{X_2})}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$ under the null hypothesis
6. Find the p value, an area under the approximate sampling distribution given by p-value=$\frac{M}{B}$

Where m depends on H_A:

$H_A: \mu_1 - \mu_2 > 0 \implies m_u$ is the number of values of $\hat{t}$ for which $\hat{t} > t_{obs}$
$H_A: \mu_1 - \mu_2 < 0 \implies m_l$ is the number of values of $\hat{t}$ for which $\hat{t} < t_{obs}$
$H_A: \mu_1 - \mu_2 \neq 0 \implies$ m= $2 * \min(m_u, m_l)$ (Example code in lecture Code)

```r
# Here's one way to do the bootstrap for a difference of two means in R:
# dat1 and dat2 are data from the two groups. nboot is the number of resamples.
#Notice obsdiff is computed as mean(dat1)-mean(dat2) - order matters!
boottwo = function(dat1, dat2, nboot) {
  bootstat = numeric(nboot)
  obsdiff = mean(dat1) - mean(dat2)
  n1 = length(dat1)
  n2 = length(dat2)
  for(i in 1:nboot) {
    samp1 = sample(dat1, size = n1, replace = T)
    samp2 = sample(dat2, size = n2, replace = T)
    bootmean1 = mean(samp1)
    bootmean2 = mean(samp2)
    bootvar1 = var(samp1)
    bootvar2 = var(samp2)
    bootstat[i] = ((bootmean1 - bootmean2) - obsdiff)/sqrt((bootvar1/n1) + (bootvar2/n2))
}
  return(bootstat)
}
```

```r
B = 5000
set.seed(1)
cricketboot = boottwo(Starved, Fed, B)  #Notice Starved put in first
hist(cricketboot)

(t.obs = (mean(Starved) - mean(Fed)) /
         sqrt(var(Starved) / length(Starved)  + var(Fed) / length(Fed))) #-1.64476

(m.equal=sum(cricketboot == t.obs))   #0
(m.low  = sum(cricketboot < t.obs))  #317
(m.high = sum(cricketboot > t.obs))  #4683
summary(cricketboot>t.obs)
(p.val = m.low/ B)  #0.0634
```



	Starved	Fed
Sample Size (n)	11	13
Mean	17.73	35.98
Median	13.00	22.80
Standard Deviation	19.96	33.63

Bootstrap Findings:

$H_o: \mu_S - \mu_F = 0$ vs $H_A: \mu_S - \mu_F < 0$ at $\alpha = 0.05$

For the starved/fed cricket data, we find

$t_{obs} = \frac{17.73 - 35.98}{\sqrt{\frac{19.96^2}{11} + \frac{33.36^2}{13}}} = -\frac{18.25}{11.1003} = 1.6441$

Total Bootstrap T	5000
$t^* < t_{obs}$	317
$t^* > t_{obs}$	4683

Pvalue= $\frac{317}{5000} = 0.0634$ (for 1 sided alternative given above)

insufficient evidence at 5% level

Compare with Welch's T: p= 0.0589

t.test(Starved, Fed, var.equal=FALSE, alternative="less")

Bootstrap is slightly more conservative than Welch's T test. Histogram of generated t gives indication that p values will be similar

Wilcoxon Rank Sum or Mann-Whitney Test for 2 independent populations

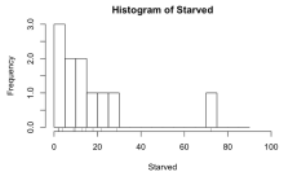
A second test for two independent populations that may not be normal is the Wilcoxon Rank Sum Test or Mann-Whitney Test.

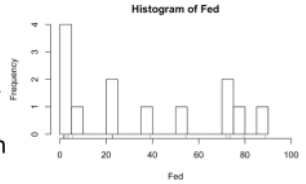
Instead of the hypothesis being a statement about means, this test is looking at the distribution for the two populations:

H_o: The distributions of the two groups are identical

H_A: The distributions of the two groups have the same shape but are shifted relative to each other.

Consider again the cricket data:
Starved: 1.9, 2.1, 3.8, 9.0, 9.6, 13.0, 14.7, 17.9, 21.7, 29.0, 72.3 ($n_s = 11$)
Fed: 1.5, 1.7, 2.4, 3.6, 5.7, 22.6, 22.8, 39.0, 54.4, 72.1, 73.6, 79.5, 88.9 ($n_F = 13$)





We must assume: (1) independence between and (2) within groups, and
(3) the distributions of the two populations have about the same shape

(Just comparing the sample histograms is sufficient. If the graphs are symmeric, then Wilcoxon can be used as a test for equality of means.)

Wilcoxon Rank Sum or Mann-Whitney Test for 2 independent populations

The test statistic is related to the ranks of the observations in the samples (as opposed to raw value), so we

1. rank the data without regard for sample, while retaining sample lables. (For tied observations, use the average rank. Eg. If the samples had Two 1.2s, they would be #1 and #2, so each would get rank 1.5)

We then find:

2. R=sum of sample A ranks ,

3. $R_{min} = \frac{n_A(n_A+1)}{2}$ = minimum possible sum of ranks in sample 1

4. $U =$ R − Rmin (U_{obs} : value of U for particular data set)

We then compare our observed test statistic U to the distribution of rank sums that would be generated if the ranks had been randomly assigned to the groups

The larger the sum of the ranks, the more evidence there is that the group is larger than the other one (or vice versa)

rank	time	sample	starved ranks
1	1.5	fed	
2	1.7	fed	
3	1.9	starved	3
4	2.1	starved	4
5	2.4	fed	
6	3.6	fed	
7	3.8	starved	7
8	5.7	fed	
9	9.0	starved	9
10	9.6	starved	10
11	13.0	starved	11
12	14.7	starved	12
13	17.9	starved	13
14	21.7	starved	14
15	22.6	fed	
16	22.8	fed	
17	29.0	starved	17
18	39.0	fed	
19	54.4	fed	
20	72.1	fed	
21	72.3	starved	21
22	73.6	fed	
23	79.5	fed	
24	88.9	fed	

n = 11

R = 3+4+7+9+10+11+12+13+14+17+21 = 121
R_{min} (11*12)/2 = 66
U_{obs} R − Rmin = 121 − 66 = 55

Wilcoxon Rank Sum or Mann-Whitney Test for 2 independent populations

P-value (again depends on H_a):

H_a: population A is shifted left of B : $\Rightarrow pvalue = P(U \leq U_{obs})$

H_a: population A is shifted right of B : $\Rightarrow pvalue = P(U \geq U_{obs})$

H_a: population A is shifted from B : $\Rightarrow pvalue = 2 * \min[P(U \leq U_{obs}), P(U \geq U_{obs}), ½]$

check if starved is less, shifted to left of fed

For the cricket data,

There are $\binom{11+13}{11} = 2496144$ ways to choose 11 ranks from {1, 2, …24}, so we'll not compute this p value by hand.

R gives us p value: 0.1804 with U_{obs}=W=55 with wilcox.test(starved, fed, altnernative="less")

There is insufficient evidence to reject the null at $\alpha = 0.05$.
We have insufficient evidence to suggest the mating time of starved crickets is lower than the mating time of fed crickets.

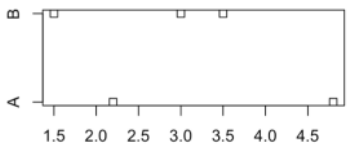
```
        Wilcoxon rank sum test

data:  starved and fed
W = 55, p-value = 0.1804
alternative hypothesis: true location shift is less than 0
```

*Notice, this conclusion isn't talking about the mean – it is talking about the whole distribution!

Wilcoxon Rank Sum or Mann-Whitney Test for 2 independent populations

Here's a simpler example for which it is not hard to calculate the p value by hand.
Let H_o: The distributions of the two groups are identical H_a: population A is shifted from B

Suppose sample A from population A is 4.8, 2.2 and sample B from population B is 3.0, 1.5, 3.5



Value	1.5	2.2	3.0	3.5	4.8
Rank	1	2	3	4	5
Group	B	A	B	B	A

Sample A's ranks are: 2, 5

So R= $2+5=7$ $R_{min} = \frac{2*(2+1)}{3} = 3$ and U_{obs}= $7-3=4$

To get a p-value we need a distribution to compare our U_{obs} to, so we will generate it.
Under H_0: equal distributions, ranks (1-5) are randomly assigned to the two samples.

There are the $\binom{5}{2} = 10$ possible sample A ranks.
(A has 2 spots)

Wilcoxon Rank Sum or Mann-Whitney Test for 2 independent populations

Enumerating ways that 2 ranks can be randomly assigned to Group A (since it had 2 observations and there were 5 total ranks/observations) allows us to know the distribution of the test statistic U.

H_0 Sample A ranks:	1,2	1,3	1,4	1,5	2,3	2,4	2,5	3,4	3,5	4,5
R	3	4	5	6	5	6	7	7	8	9
R_{min}	3	3	3	3	3	3	3	3	3	3
U	0	1	2	3	2	3	4	4	5	6

U	P(u=U)
0	1/10
1	1/10
2	2/10
3	2/10
4	2/10
5	1/10
6	1/10

The pvalue for H_a: population A is shifted from B $2 * \min[P(U \leq U_{obs}), P(U \geq U_{obs}), ½]$

$P(U \leq 4) = P(u=0) + P(u=1) + P(u=2) + P(u=3) + P(u=4) = 8/10$
$1/10 + 1/10 + 2/10 + 2/10 + 2/10$

$P(U \geq 4) = P(u=4) + P(u=5) + P(u=6) = 4/10$

Pvalue= $2 * 0.40 = 0.80$; Which is no evidence against the null. There is insufficient evidence at the 5% level to suggest population A is shifted from population B.

Permutation Test for comparing 2 Independent Populations' means

Another test for the situation when we are interested in comparing the means of two populations but the normality of the populations are in question is the Permutation Test. This test is very useful when the sample sizes are small.

The idea is, if the means for the two groups are truly equal, then the group labels are arbitrary and we can just randomly reassign them. We randomly divide them into two groups, and see how often the differences under random labeling is more extreme than the difference we observed.

1. Draw a sample of size n_1 from the first population and compute $\bar{x}_1$. Draw a sample of size n_2 from the second population and compute $\bar{x}_2$. Let $t_{obs} = \bar{x}_1 - \bar{x}_2$

2. Combine all the data from both samples into a single group of size $n_1 + n_2$. Take a random sample of size n_1, without replacement, from this group, and compute the mean of those data, call it $\bar{x}_1^*$. There will be n_2 observations left. Compute the mean of those data, call it $\bar{x}_2^*$. Let $\hat{t} = \bar{x}^*{}_1 - \bar{x}^*{}_2$.

3. Repeat step 2 B times, where B is a large number, and compute $\hat{t}$ from each one. The accumulation of the B $\hat{t}$ s is an approximation to the sample distribution of $T = \bar{X}_1 - \bar{X}_2$ under the null of equal means.

4. Find the p value, an area under the approximate sampling distribution given by p-value=$\frac{M}{B}$

Where m depends on H_A:

$H_A: \mu_1 - \mu_2 > 0 \implies m_u$ is the number of values of $\hat{t}$ for which $\hat{t} > t_{obs}$

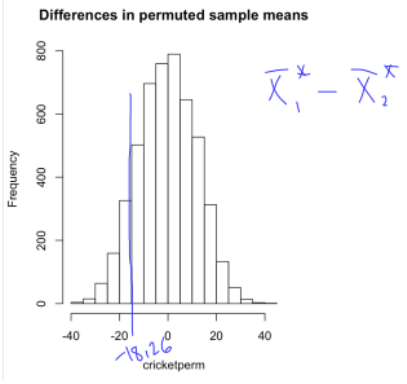
$H_A: \mu_1 - \mu_2 < 0 \implies m_l$ is the number of values of $\hat{t}$ for which $\hat{t} < t_{obs}$

$H_A: \mu_1 - \mu_2 \neq 0 \implies$ m= $2 * \min(m_u, m_l)$ (Example code in lecture Code)

```r
#function to perform permutation test#
#dat1 is data from first group, dat2 is from second#
#nperm is the number of times to do the random splitting#

permtwo <- function(dat1, dat2, nperm) {
  permstat <- NULL
  for(i in 1:nperm) {
    n1 <- length(dat1)  #find length of first sample
    n2 <- length(dat2)  #find length of second sample
    alldat <- c(dat1, dat2)  #combine into 1 large sample
    samp <- sample(alldat, replace=FALSE)  #shuffle the combined vector of values
    pmean1 <- mean(samp[1:n1])   #find mean of first n1 observations as x1*-bar
    pmean2 <- mean(samp[(n1+1):(n1+n2)]) #find mean of last n2 observations as x2*-bar
    permstat[i] <- pmean1 - pmean2
}
  return(permstat)
}
```

```r
#do the permutataion for cricket data#
set.seed(1)
cricketperm <- permtwo(starved, fed, 5000)
hist(cricketperm, main="Differences in permuted sample means")

#observed stat#
mstarved <- mean(starved)
mfed <- mean(fed)
(tobs <- mstarved - mfed)

#find p-value for two-sided#
(mlo <- sum(cricketperm <= tobs))
(mup <- sum(cricketperm >=tobs))
(pval <- mlo/5000)
```

Permutation Test for comparing 2 Independent Populations' means



$\bar{X}_1^* - \bar{X}_2^*$

H_o: $\mu_s = \mu_F$ or $\mu_s - \mu_F = 0$ vs H_A: $\mu_s < \mu_F$ or $\mu_s - \mu_F < 0$

```
> (tobs <- mstarved - mfed)
[1] -18.25734
```

Total Permutated t^*	5000 replications
$t^* < t_{obs} = -18.26$	341
$t^* > t_{obs} = -18.26$	4658
$t^* = t_{obs} = -18.26$	1

If $H_a: \mu_s - \mu_F \neq 0$

$pvalue$: $2 * \min(0.0684, 0.9318)$

$= 0.1368$

$Ha: \mu_s - \mu_F < 0$

$pvalue$: $\frac{341 + 1}{5000} = 0.0684$

If $H_a: \mu_s - \mu_F > 0$

$pvalue$: $\frac{4658}{5000} = 0.9318$

Comparing Bootstrap, Wilcoxon Rank Sum, Permutation for 2 ind. Pops, and Welch's 2 sample t test

We examined the hypotheses: H_o: $\mu_s = \mu_F$ or $\mu_s - \mu_F = 0$ vs H_A: $\mu_s < \mu_F$ or H_A : $\mu_s - \mu_F < 0$ for the cricket data

Test	Assumptions	P value
Bootstrap	(1) Independent between and within groups	0.0677
Welch's T test	(1) Independent between and within groups (2) Both samples come from normal population	0.0589
Wilcoxon Rank Sum	(1) Independent between and within groups (2) pop distributions have about the same shape (3) Symmetric if want to compare means	0.1804
Permutation test	(1) Independent between and within groups	0.0684

We see that the Welch's T test, Bootstrap, and Permutation tests all give close to the same amount of evidence. The T test is robust to deviations from normality so long as the sample size is larger than ~ 5 in each group.

The Wilcoxon is more conservative (larger p value).

Notice, the Wilcoxon Rank Sum will be the least effected by an outlying value since it is the only one based on ranking the data.

Comparing two Independent Population Proportions

e.g. High levels of cholesterol in the blood are associated with higher risk of heart attacks. Will using a drug to lower blood cholesterol reduce heart attacks? The Helsinki Heart Study recruited middle-aged men with high cholesterol but no history of other serious medical problems to investigate this question.

The volunteer subjects were assigned at random to one of two treatments:

> 2051 men took the drug gemfibrozil to reduce their cholesterol levels, and a control group of 2030 men took a placebo.

During the next five years, 56 men in the gemfibrozil group and 84 men in the placebo group had heart attacks. (Assume no patients were lost to follow up and stuck to their treatment regimen…)

Is there statistically significant evidence that gemfibrozil reduces heart attacks?

Our hypotheses: π_G: proportion of gemfibrozil patients who have a heart attack

π_C: " " control / placebo of patient heart attacks

$H_A: \pi_G < \pi_C \Rightarrow \pi_G - \pi_C < 0$ $\quad H_0: \pi_G = \pi_C \Rightarrow \pi_G - \pi_C = 0$

Comparing two Independent Population Proportions

We would like to make inferences about gemfibrozil efficacy in the general population. We would like to better understand the true difference between :

π_g: proportion of gemfibrozil patients that have heart attack and
π_c: proportion of control (non-medicated) patients that have heart attack

Hypothesis tests and confidence intervals for $\pi_g - \pi_c$, the difference in population proportions are desired.

A natural point estimator is $\widehat{p_g} - \widehat{p_C}$

If $\pi_g * n_g > 5, \pi_c * n_c > 5, (1-\pi_g) * n_g > 5$, and $(1-\pi_c) * n_c > 5$,

we can use the CLT to say: $\widehat{p_g} - \widehat{p_C} \sim N(\pi_G - \pi_C, \frac{\pi_G(1-\pi_G)}{n_G} + \frac{\pi_C(1-\pi_C)}{n_C})$

*Since we don't know π_g and π_C, we will approximate them with $\widehat{\pi_g}$ and $\widehat{\pi_C}$

Comparing two Independent Population Proportions

π_{HA}: true in pop

$\hat{\pi}_{HA}$: estimating from sample

We can use the CLT to say:

$$\widehat{p_g} - \widehat{p_C} \sim N(\pi_g - \pi_c, \frac{\pi_g(1-\pi_g)}{n_g} + \frac{\pi_c(1-\pi_c)}{n_c})$$

Since we don't know π_g or π_c, but under H_o, they are equal, $\pi_g = \pi_c = \pi_{HA}$ and we can simplify our distribution. We don't know the common proportion: π_{HA}: proportion with heart attack so we estimate it by pooling

$$\widehat{\pi_{HA}} = \frac{Number\ of\ Successes\ in\ both\ groups}{Total\ size\ of\ both\ groups} = \frac{\hat{p}_G * n_G + \hat{p}_C * n_C}{n_G + n_C} = \frac{x_G + x_C}{n_G + n_C}$$

Under the null: H_o: $\pi_g - \pi_c = 0$

$$\widehat{p_g} - \widehat{p_C} \approx N\left(0, \frac{\widehat{\pi_{HA}}(1-\widehat{\pi_{HA}})}{n_g} + \frac{\widehat{\pi_{HA}}(1-\widehat{\pi_{HA}})}{n_c}\right) = N\left(0, \hat{\pi}_{HA}(1-\hat{\pi}_{HA})\left[\frac{1}{n_G} + \frac{1}{n_C}\right]\right)$$

So p values can be computed using a Z test statistic and the standard normal.

$$Z = \frac{(\hat{p}_G - \hat{p}_C) - 0}{\sqrt{\hat{\pi}_{HA}(1-\hat{\pi}_{HA})\left(\frac{1}{n_G} + \frac{1}{n_C}\right)}} \sim N(0,1)$$

plug in sample data to get Z_{obs}

Comparing two Independent Population Proportions

For the Cholesterol and Heart Attack Data:

Sample Data:

Hypotheses: H_o: $\pi_g - \pi_c = 0, H_A$: $\pi_g - \pi_c < 0$

$\widehat{p_g} = \frac{56}{2051} = 0.02730375, \quad \widehat{p_C} = \frac{84}{2030} = 0.04137931,$

and $\widehat{p_{HA}} = \frac{56+84}{2051+2030} = 0.0343$

Assumptions: independence within samples

randomly selected people dont influence eachother

$\hat{\pi}_G * n_G = 56 \quad (1-\hat{\pi}_G) * n_G = 1995 \quad \hat{\pi}_C * n_C = 84 \quad (1-\hat{\pi}_C) * n_C = 1946$

So, under H_o: $\widehat{p_g} - \widehat{p_C} \sim N(0, 0.0343(1-0.0343)\left(\frac{1}{2051} + \frac{1}{2030}\right))$ (CLT)

Test statistic under H_o: and p value:

$$Z = \frac{(\widehat{p_g} - \widehat{p_C}) - 0}{\sqrt{\widehat{\pi_{HA}}(1-\widehat{\pi_{HA}})\left(\frac{1}{n_g} + \frac{1}{n_c}\right)}}$$

so $Z_{obs} = \frac{(0.0273 - 0.0414) - 0}{\sqrt{0.0343(1-0.0343)\left(\frac{1}{2051} + \frac{1}{2030}\right)}} = -2.47$

P value and Conclusion:

Pvalue = $P(Z \leq -2.47) = 0.0068$

Since Pvalue = 0.0068 < 0.05 we have strong evidence against the null, reject the null. evidence suggests rate of genfibrozal HA is less than rate of control patients

Comparing two Independent Population Proportions

We can also make a CI. It does not come with a H_o, however so we use a more general approximation for the variance of the difference:

$$\widehat{p_g} - \widehat{p_C} \sim N(\pi_g - \pi_c, \frac{\pi_g(1-\pi_g)}{n_g} + \frac{\pi_c(1-\pi_c)}{n_c}) \qquad Var(\widehat{p_g} - \widehat{p_C}) = \frac{\hat{P}_G(1-\hat{P}_G)}{n_G} + \frac{\hat{P}_c(1-\hat{P}_c)}{n_c}$$

So a $(1-\alpha) * 100\%$ CI for the difference $\pi_g - \pi_c$ is $\hat{P}_G - \hat{P}_C \pm Z_{\alpha/2}\sqrt{\frac{\hat{P}_G(1-\hat{P}_G)}{n_G} + \frac{\hat{P}_c(1-\hat{P}_c)}{n_c}}$

Obs dif $\pm Z_{\alpha/2} * SE_{dif}$

$$\widehat{p_g} = \frac{56}{2051} = 0.02730375, \quad \widehat{p_C} = \frac{84}{2030} = 0.04137931,$$

$$and \ \widehat{p_{HA}} = \frac{56+84}{2051+2030} = 0.03430532$$

For the Cholesterol and Heart Attack Data a (2-sided) 95% Confidence interval is:

$$(0.0273 - 0.414) \pm Z_{0.025}\sqrt{\frac{0.0273(1-0.0273)}{2051} + \frac{0.0414(1-0.0414)}{2030}}$$

$$-0.014 \pm 1.96 * (0.00657) = (-0.025, -0.003)$$

Comparing two Independent Population Proportions Summary

Suppose $X \sim Bin(n_x, \pi_x)$ and Y $\sim Bin(n_y, \pi_y)$ are independent, with

$n_x * \pi_x; \quad n_x * (1-\pi_x); \quad n_y * \pi_y; \quad n_y * (1-\pi_y)$ all >5

To test: $H_o: \pi_X - \pi_y = 0$: or CI: $(\widehat{p_X} - \widehat{p_Y}) \pm z_{\alpha/2}\sqrt{\frac{\widehat{p_X}(1-\widehat{p_X})}{n_X} + \frac{\widehat{p_Y}(1-\widehat{p_Y})}{n_Y}}$

1. State null and alternative hypotheses H_o and H_A
2. Check assumptions
3. Calculate $\widehat{p_x} = \frac{X}{n_x}$ and $\widehat{p_Y} = \frac{Y}{n_y}$ and pooled $\widehat{p_{pooled}} = \frac{Y+X}{n_y+n_x}$
4. Find the test statistic: $Z = \frac{(\widehat{p_x} - \widehat{p_y}) - 0}{\sqrt{\widehat{\pi_{pool}}(1-\widehat{\pi_{pool}})\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$
5. Find the p-value, which is an area under the N(0,1) curve depending on H_a.

$H_a: \pi_x - \pi_y > 0, \Longrightarrow$ p-value: $P(Z > z_{obs})$, the area right of z

$H_a: \pi_x - \pi_y < 0, \Longrightarrow$ p-value: $P(Z < z_{obs})$, the area left of z

$H_a: \pi_x - \pi_y \neq 0, \Longrightarrow$ p-value: $2 * P(Z > |z_{obs}|)$, the sum of the two tails.

6. Draw conclusion in context.

For Next Time

- Start/Continue working through posted homework 6. Post questions on Piazza.

- Continue working on Quiz 3