Chapter 6

ST 511 - Inferences Comparing Two Population Central Values

Readings: Chapter 6 (for 6.3-6.5 read if interested)

Our problems so far have dealt with inference for the mean of only **1 population** of interest. In real life this will not usually be the case. We will start with looking at inference regarding the means of **2 populations** and then in later chapters look at what to do with an arbitrary number of populations.

Motivating Example:
Jocko's garage seems to be giving out really high estimates for insurance claims. To investigate insurance fraud, insurance adjusters take 10 damaged cars and take each one to both Jocko's and a repair shop they trust, Jami's repair shop. Then then get the estimates from the repair shop (in the end, 2 for each car). Data are provided below:

Obs	Jocko	Jami
1	450	255
2	699	720
3	670	499
4	800	760
5	401	225
6	1000	700
7	535	300
8	680	350
9	1100	1000
10	850	770

Here we have two populations: all estimates from Jocko's and all estimates from Jami's repair shop.

Therefore, we have 2 random variables:
Y_i= estimate for the i^{th} randomly selected car at Jocko's
X_i= estimate for the i^{th} randomly selected car at Jami's

We now have two sample sizes:
n_1 (or n_Y) = number sampled at Jocko's
n_2 (or n_X)= number sampled at Jami's.
(Here they are equal, but generally for a two sample problem, they need not be.)

We now have two sample mean **random variables**:
$\bar{Y}$= mean estimate for a randomly selected sample of 10 cars at Jocko's
$\bar{X}$= mean estimate for a randomly selected sample of 10 cars at Jami's

We also have 2 sets of summary statistics (1 for each sample):

The UNIVARIATE Procedure
Variable: Jocko

Moments			
N	10	Sum Weights	10
Mean	718.5	Sum Observations	7185
Std Deviation	225.955871	Variance	51056.0556
Skewness	0.27601197	Kurtosis	-0.6296669
Uncorrected SS	5621927	Corrected SS	459504.5
Coeff Variation	31.4482771	Std Error Mean	71.4535202

Basic Statistical Measures			
Location		Variability	
Mean	718.5000	Std Deviation	225.95587
Median	689.5000	Variance	51056
Mode	.	Range	699.00000
		Interquartile Range	315.00000

The UNIVARIATE Procedure
Variable: Jami

Moments			
N	10	Sum Weights	10
Mean	557.9	Sum Observations	5579
Std Deviation	267.400428	Variance	71502.9889
Skewness	0.14751428	Kurtosis	-1.3512691
Uncorrected SS	3756051	Corrected SS	643526.9
Coeff Variation	47.9298132	Std Error Mean	84.55944

Basic Statistical Measures			
Location		Variability	
Mean	557.9000	Std Deviation	267.40043
Median	599.5000	Variance	71503
Mode	.	Range	775.00000
		Interquartile Range	460.00000

Two parameters of interest:
μ_Y (or μ_1) = (true) mean of all estimates at Jocko's
μ_X (or μ_2) = (true) mean of all estimates at Jami's repair shop.

Goal: Investigate $\mu_D = \mu_{diff} = \mu_1 - \mu_2 = \mu_Y - \mu_X$

What are possible methods of inference for $\mu_{diff} = \mu_1 - \mu_2$?

Distribution	Two Samples are Independent	Two Samples are 'Paired'
$\bar{Y} - \bar{X} \sim Normal$	6.2 - Two-sample t-test	6.4 Paired-t-test
$\bar{Y} - \bar{X} \sim Not\ Normal$	6.3 - Wilcoxon Rank Sum Test	6.5 - Wilcoxon Signed Rank Test

6.4 - Inference for Paired Data (Matched Pairs t or Paired t)

What is paired data?
Each 'unit' receives two treatments. The units could be:

1. A single subject (each subject gets both treatments)
2. Two subjects that have been **matched** together (one receives treatment A and the other receives treatment B)

Ex: Auto example - We have paired data because

How to make inference here? Hypothesis test = paired t-test:
Parameter:

Null hypothesis:

Alternative Hypothesis:

Test Statistic:

RR/p-value:

Conclusions same as for all HT. Note that this test is **equivalent to the one-sample t-test on the differences between the paired data.**

Similarly we can create a confidence interval using the test statistic above:

Note: We do not need to know each variable's sample mean and standard deviation, **only the mean and standard deviation of the differences!**.

Both the HT and the CI can be done very easily in SAS:

```
data autodata;
input Jocko Jami;
datalines;
450 255
699 720
670 499
800 760
401 225
1000 700
535 300
680 350
1100 1000
850 770
;

proc ttest data=autodata;
     paired Jocko*Jami;
run;

/* About this code:
The PAIRED VAR1*VAR2 statement requests the paired t-test.
SAS calculates the differences as VAR1-VAR2.
*/
```

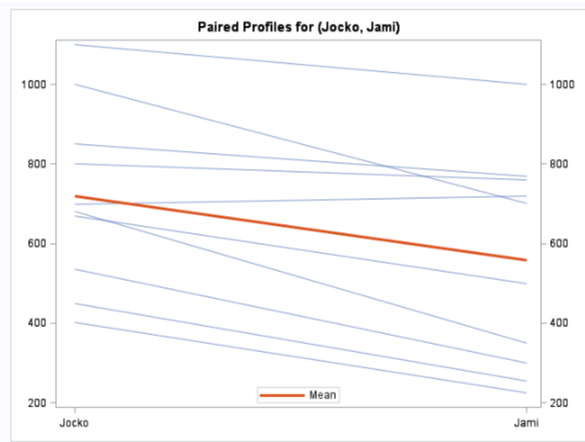
The TTEST Procedure

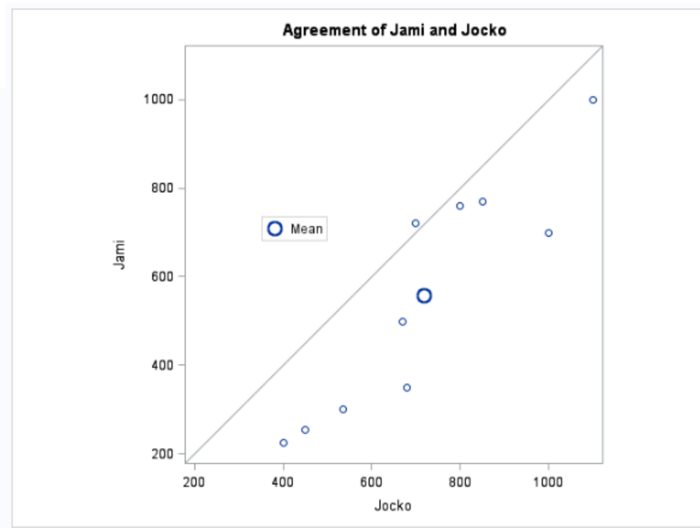
Difference: Jocko - Jami

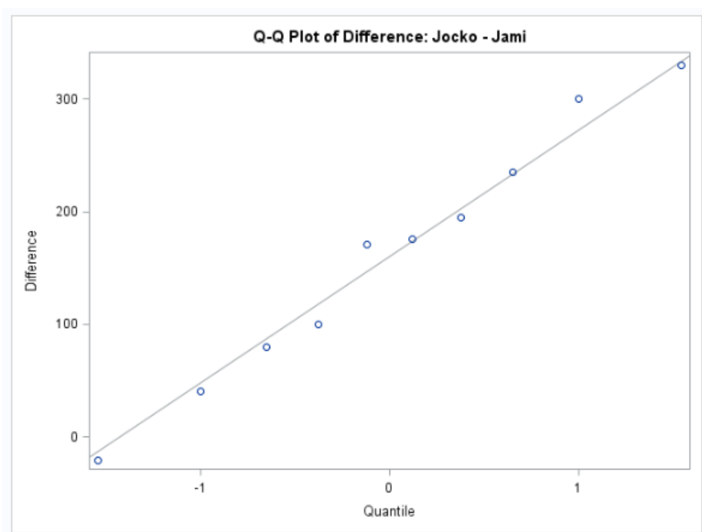
N	Mean	Std Dev	Std Err	Minimum	Maximum
10	160.6	112.0	35.4157	-21.0000	330.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
160.6	80.4841	240.7	112.0	77.0336	204.5

DF	t Value	Pr > |t|
9	4.53	0.0014


Paired Profiles for (Jocko, Jami)
1000
800
600
400
200
Mean
Jocko
Jami


Agreement of Jami and Jocko
Mean
Jami
Jocko
200
400
600
800
1000


Q-Q Plot of Difference: Jocko - Jami
Difference
Quantile
300
200
100
0
-1
0
1

One of the two scenarios below has paired data where looking at paired differences makes sense and on scenario has a case where that does not make any sense (even though paired differences for each are given). Identify which of the two scenarios below has paired data - for the example with paired data find a 95% confidence interval (state assumptions needed on the data, how you would inspect the assumption, and interpret the interval): Some values - $P(T_9 > 1.83) = 0.05$ $P(T_9 > 2.26) = 0.025$ $P(T_{22} > 1.72) = 0.05$ $P(T_{22} > 2.07) = 0.025$

1. A nutrition expert is examining a weight loss program to evaluate its effectiveness (i.e., if participants lose weight on the program). Ten subjects are randomly selected for the investigation. Each subjects initial weight is recorded, they follow the program for 6 weeks, and they are again weighed. Is the program effective?
   The data are given below:

Subject	Initial Weight	Final Weight
1	180	165
2	142	138
3	126	128
4	138	136
5	175	170
6	205	197
7	116	115
8	142	128
9	157	144
10	136	130

The UNIVARIATE Procedure
Variable: FminusI

Moments			
N	10	Sum Weights	10
Mean	-6.6	Sum Observations	-66
Std Deviation	5.8156876	Variance	33.8222222
Skewness	-0.2343677	Kurtosis	-1.1697528
Uncorrected SS	740	Corrected SS	304.4
Coeff Variation	-88.116479	Std Error Mean	1.8390819

2. A manufacturer of cat food wants to assure that the packages being produced at the Tennessee plant have the same average weight as the packages being produced at the Wisconsin plant. Samples of 23 packages each were collected from Tennessee plant and Wisconsin plant respectively. The package weights (in ounces) are given below:

Sample	Tennessee	Wisconsin
1	4.67	4.74
2	4.65	4.65
3	4.68	4.60
4	4.59	4.62
⋮	⋮	⋮
23	4.66	4.62

The UNIVARIATE Procedure
Variable: Tenn_Wisc

Moments			
N	23	Sum Weights	23
Mean	-0.0008696	Sum Observations	-0.02
Std Deviation	0.08564796	Variance	0.00733557
Skewness	-0.327649	Kurtosis	-1.2283775
Uncorrected SS	0.1614	Corrected SS	0.16138261
Coeff Variation	-9849.5154	Std Error Mean	0.01785883

Output from SAS to conduct the paired t-test on the weight example.

```
*Conduct paired t-test;
proc ttest data=weight;
paired Final*Initial;
run;
```

The TTEST Procedure

Difference: final - initial

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	-6.6000	5.8157	1.8391	-15.0000	2.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-6.6000	-10.7603	-2.4397	5.8157	4.0002	10.6172

DF	t Value	Pr > |t|
9	-3.59	0.0059

6.2 - Inference for Two Independent Samples (Two-Sample t)

For the second example on the previous page, we did not have paired data, but rather two samples, one from the Tennessee population and one from the Wisconsin population.

The UNIVARIATE Procedure
Variable: Tenn

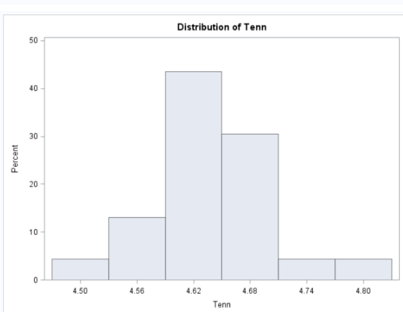
Moments			
N	23	Sum Weights	23
Mean	4.63304348	Sum Observations	106.56
Std Deviation	0.06483789	Variance	0.00420395
Skewness	0.61292717	Kurtosis	1.39541642
Uncorrected SS	493.7896	Corrected SS	0.09248696
Coeff Variation	1.39946657	Std Error Mean	0.01351964

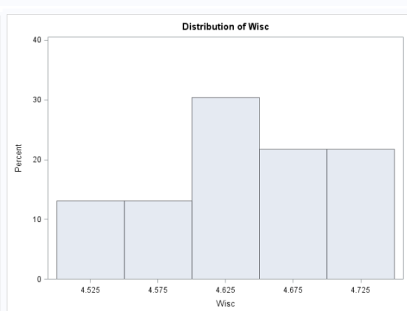
Basic Statistical Measures			
Location		Variability	
Mean	4.633043	Std Deviation	0.06484
Median	4.630000	Variance	0.00420
Mode	4.590000	Range	0.30000
		Interquartile Range	0.08000

The UNIVARIATE Procedure
Variable: Wisc

Moments			
N	23	Sum Weights	23
Mean	4.63391304	Sum Observations	106.58
Std Deviation	0.06058346	Variance	0.00367036
Skewness	-0.142007	Kurtosis	-0.7243252
Uncorrected SS	493.9632	Corrected SS	0.08074783
Coeff Variation	1.30739313	Std Error Mean	0.01263252

Basic Statistical Measures			
Location		Variability	
Mean	4.633913	Std Deviation	0.06058
Median	4.630000	Variance	0.00367
Mode	4.620000	Range	0.22000
		Interquartile Range	0.09000





Define:
Y_i = the weight for the i^{th} randomly selected package from the Tennessee plant
X_i = the weight for the i^{th} randomly selected package from the Wisconsin plant

μ_1 = the mean weights for Tennessee plants
μ_2 = the mean weights for Wisconsin plants

Question of interest (Claim):

What could we do to make inference here?

An 'unbiased' estimate of μ_d is

What is the variance of this quantity?

Let's define the ____________________ between two random variables. $Cov(X, Y)$ is a measure the how

the random variables ____

Mathematically:

$Cov(X, Y) = E(XY) - E(X)E(Y)$ - Similar to $Var(X) = E(X^2) - (E(X))^2 = E(XX) - E(X)E(X)$

Generally, for the random variable $aX + bY$ we have

$$Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$$

Since covariance is a measure of how the RV's vary together. If X is independent of Y that means

This implies that if X is independent of Y then $Cov(X, Y) = 0$.

Now back to our quantity $\bar{D}$, what is the variance of this quantity?

Knowing the mean and variance of this quantity is useful, but to use it for inference we must know the

Theorem: If $Y_i \sim^{iid} N(\mu_1, \sigma^2)$ and $X_i \sim^{iid} N(\mu_2, \sigma^2)$ (both parent populations are normal, same variance) where all Y are **independent** of all X (independent samples) then

We can estimate the common variance by

Thus, the **test statistic** we can use for our HT and CI are ex: Back to the catfood example. Let us assume that $Y_i \sim^{iid} N(\mu_1, \sigma^2)$ and $X_i \sim^{iid} N(\mu_2, \sigma^2)$ where $Y's$ and $X's$ are independent (that is, our parent populations are independent normals with equal variance assumed). Let's conduct a hypothesis test at the 0.01 level to determine if the mean weights differ. Would a 99% CI for μ_{diff} contain 0? Why/why not?

Analysis of cat food data using SAS

```
proc ttest data=catfood2;
*Specify that location is categorical;
class location;
*variable that we want to test on;
var weight;
run;
```

The TTEST Procedure

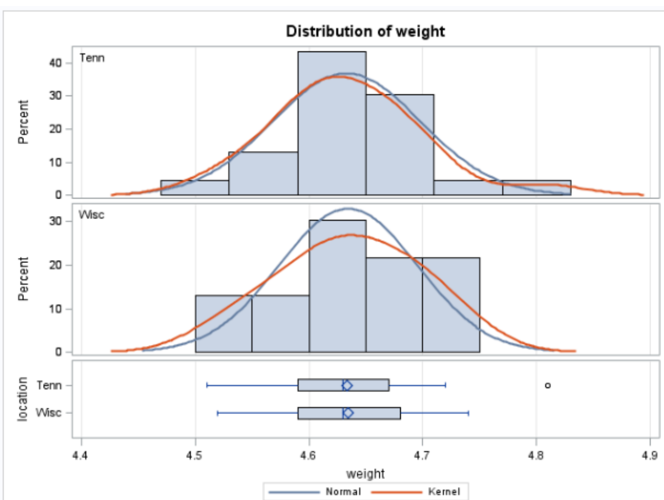
Variable: weight

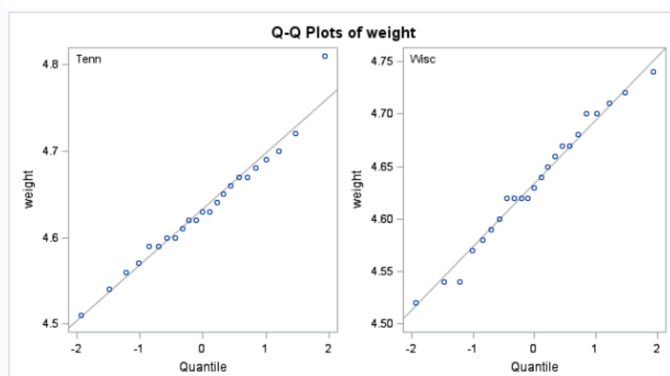
location	N	Mean	Std Dev	Std Err	Minimum	Maximum
Tenn	23	4.6330	0.0648	0.0135	4.5100	4.8100
Wisc	23	4.6343	0.0604	0.0126	4.5200	4.7400
Diff (1-2)		-0.00130	0.0627	0.0185		

location	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Tenn		4.6330	4.6050	4.6611	0.0648	0.0501	0.0918
Wisc		4.6343	4.6082	4.6605	0.0604	0.0467	0.0855
Diff (1-2)	Pooled	-0.00130	-0.0386	0.0359	0.0627	0.0519	0.0792
Diff (1-2)	Satterthwaite	-0.00130	-0.0386	0.0359			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	44	-0.07	0.9441
Satterthwaite	Unequal	43.785	-0.07	0.9441

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	22	22	1.15	0.7447





The equal variance assumption seemed reasonable above. What can we do when it is **not** reasonable?

Theorem: If $Y_i \sim^{iid} N(\mu_1, \sigma_1^2)$ and $X_i \sim^{iid} N(\mu_2, \sigma_2^2)$ (both parent populations are normal, different variance) where all Y are **independent** of all X (independent samples) then

Therefore, $\bar{D} = \bar{Y} - \bar{X}$ still is a good statistic to base our inference on.

Suppose we estimate our standard error using the sample variances:

We can create the test statistic

Issue: What are the degrees of freedom for our test statistic??

Satterthwaite's approximation to degrees of freedom
To approximate the df associated with a t statistic based on a standard error of the form

$$SE = \sqrt{c_1 S_1^2 + c_2 S_2^2 + \cdots + c_k S_k^2}$$

(a linear combination of sample variances), use the **Satterthwaite approximation:**

$$\widehat{df} = \frac{(c_1 S_1^2 + c_2 S_2^2 + \cdots + c_k S_k^2)^2}{(c_1 S_1^2)^2/df_1 + (c_2 S_2^2)^2/df_2 + \cdots + (c_k S_k^2)^2/df_k}$$

Always round down!

Example: Consider an experiment involving the comparison of the mean heart rate following 30 minutes of aerobic exercise among females aged 20 to 24 years (Y variable, group 1) as compared to females aged 30-34 years (X variable, group 2). For this experiment, heart rates are recorded on each participant following 30 minutes of intense aerobic exercise. The sample data and some statistics (not all will be needed) are given below:
$n_1 = 15$, $\bar{y}$=150.22, $s_1^2 = 160$
$n_2 = 10$, $\bar{x} = 141.10$, $s_2^2 = 100$

$$\widehat{SE}\left(\bar{Y} - \bar{X}\right) = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{\frac{(15-1)160 + (10-1)100}{15+10-2}(1/15 + 1/10)} = 4.768$$

$$\widehat{SE}\left(\bar{Y} - \bar{X}\right) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{160}{15} + \frac{100}{10}} = 4.55$$

$$\widehat{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2/(n_1-1) + \left(\frac{s_2^2}{n_2}\right)^2/(n_2-1)} = \frac{\left(\frac{160}{15} + \frac{100}{10}\right)^2}{\left(\frac{160}{15}\right)^2/(15-1) + \left(\frac{100}{10}\right)^2/(10-1)} = 22.20$$

$$P(T_{23} > 2.50) = 0.01 \quad P(T_{23} > 2.81) = 0.005 \quad P(T_{22} > 2.51) = 0.01 \quad P(T_{22} > 2.82) = 0.005$$

Conduct a hypothesis test at the $\alpha = 0.01$ level assuming the variances of the two population are not equal. Be sure to show all steps (use RR, state the assumptions that must be made and how you would check that assumption). Also, create a 99% confidence interval for the difference in means.

Analysis of heart rate data using SAS

```
proc ttest data=heartrate;
*denote group as a categorical variable;
class group;
var rate;
run;
```

The TTEST Procedure

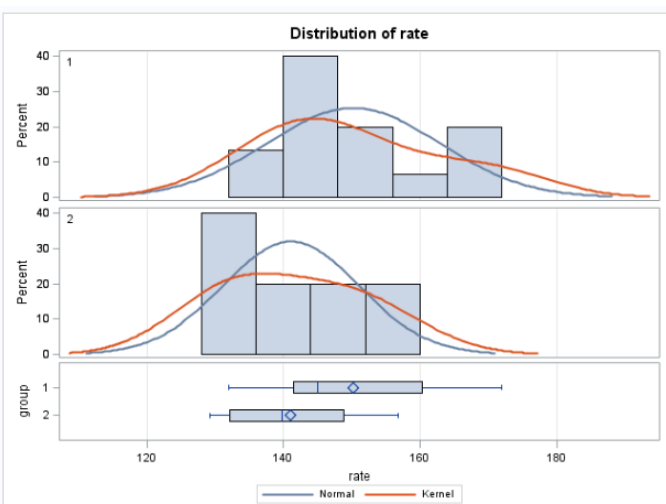
Variable: rate

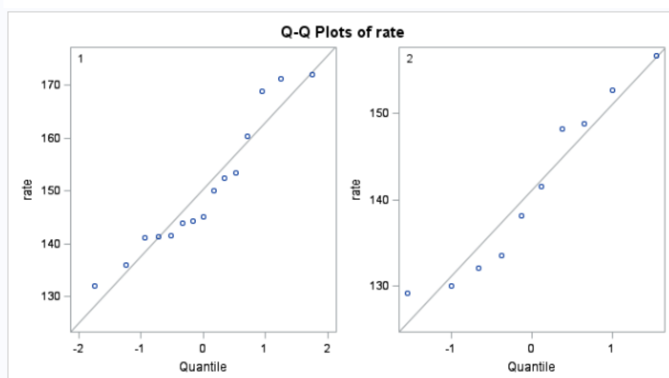
group	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	15	150.2	12.6500	3.2662	132.1	171.9
2	10	141.1	10.0004	3.1624	129.2	156.7
Diff (1-2)		9.1190	11.6849	4.7704		

group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		150.2	143.2	157.2	12.6500	9.2614	19.9503
2		141.1	133.9	148.3	10.0004	6.8786	18.2568
Diff (1-2)	Pooled	9.1190	-0.7492	18.9872	11.6849	9.0817	16.3912
Diff (1-2)	Satterthwaite	9.1190	-0.3045	18.5425			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	23	1.91	0.0685
Satterthwaite	Unequal	22.202	2.01	0.0572

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	14	9	1.60	0.4833





Recap of possible inferences for the difference of means based on the normal distribution:

Paired Data: Assume **differences** are a RS and normally distributed
100(1-α)% CI for μ_d is

$$\overline{D} \pm t_{\alpha/2,n-1} S_D/\sqrt{n} = \bar{Y} - \bar{X} \pm t_{\alpha/2,n-1} S_{\bar{Y}-\bar{X}}/\sqrt{n}$$

HT: for $H_0 : \mu_d = \Delta_0 \;\; vs \;\; H_a : \mu_d > \Delta_0 \;\; or \;\; \mu_d < \Delta_0 \;\; or \;\; \mu_d \neq \Delta_0$

$$\text{Test Statistic: } T = \frac{\bar{Y} - \bar{X} - \Delta_0}{S_d/\sqrt{n}}$$

$$RR : \{t_{obs} : t_{obs} > t_{\alpha,n-1}\} \quad or \quad \{t_{obs} : t_{obs} < -t_{\alpha,n-1}\} \quad or \quad \left\{t_{obs} : |t_{obs}| > t_{\alpha/2,n-1}\right\}$$

$$P-value : P(T_{n-1} > t_{obs}) \quad or \quad P(T_{n-1} < t_{obs}) \quad or \quad 2 * P(T_{n-1} > |t_{obs}|)$$

Independent Samples: Assume populations are independent RS's with each population having a normal distribution
Equal Variance (Pooled Variance):
100(1-α)% CI for μ_d is

$$\bar{Y} - \bar{X} \pm t_{\alpha/2,n_1+n_2-2} \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

HT: for $H_0 : \mu_d = \Delta_0 \;\; vs \;\; H_a : \mu_d > \Delta_0 \;\; or \;\; \mu_d < \Delta_0 \;\; or \;\; \mu_d \neq \Delta_0$

$$\text{Test Statistic: } T = \frac{\bar{Y} - \bar{X} - \Delta_0}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$RR : \{t_{obs} : t_{obs} > t_{\alpha,n_1+n_2-2}\} \quad or \quad \{t_{obs} : t_{obs} < -t_{\alpha,n_1+n_2-2}\} \quad or \quad \left\{t_{obs} : |t_{obs}| > t_{\alpha/2,n_1+n_2-2}\right\}$$

$$P-value : P(T_{n_1+n_2-2} > t_{obs}) \quad or \quad P(T_{n_1+n_2-2} < t_{obs}) \quad or \quad 2 * P(T_{n_1+n_2-2} > |t_{obs}|)$$

Unequal Variance:
100(1-α)% CI for μ_d is

$$\bar{Y} - \bar{X} \pm t_{\alpha/2,\widehat{df}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

HT: for $H_0 : \mu_d = \Delta_0 \;\; vs \;\; H_a : \mu_d > \Delta_0 \;\; or \;\; \mu_d < \Delta_0 \;\; or \;\; \mu_d \neq \Delta_0$

$$\text{Test Statistic: } T = \frac{\bar{Y} - \bar{X} - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$RR : \left\{t_{obs} : t_{obs} > t_{\alpha,\widehat{df}}\right\} \quad or \quad \left\{t_{obs} : t_{obs} < -t_{\alpha,\widehat{df}}\right\} \quad or \quad \left\{t_{obs} : |t_{obs}| > t_{\alpha/2,\widehat{df}}\right\}$$

$$P-value : P(T_{\widehat{df}} > t_{obs}) \quad or \quad P(T_{\widehat{df}} < t_{obs}) \quad or \quad 2 * P(T_{\widehat{df}} > |t_{obs}|)$$

$$\widehat{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2/(n_1-1) + \left(\frac{s_2^2}{n_2}\right)^2/(n_2-1)}$$