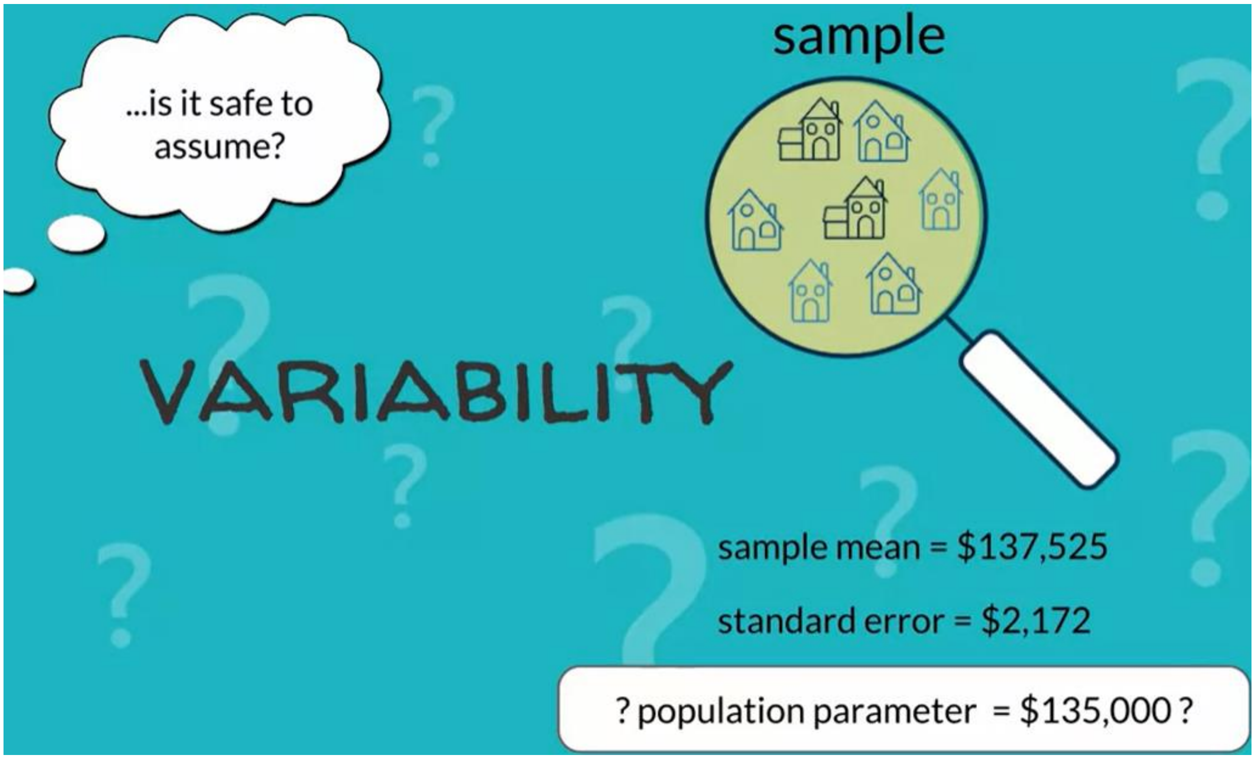
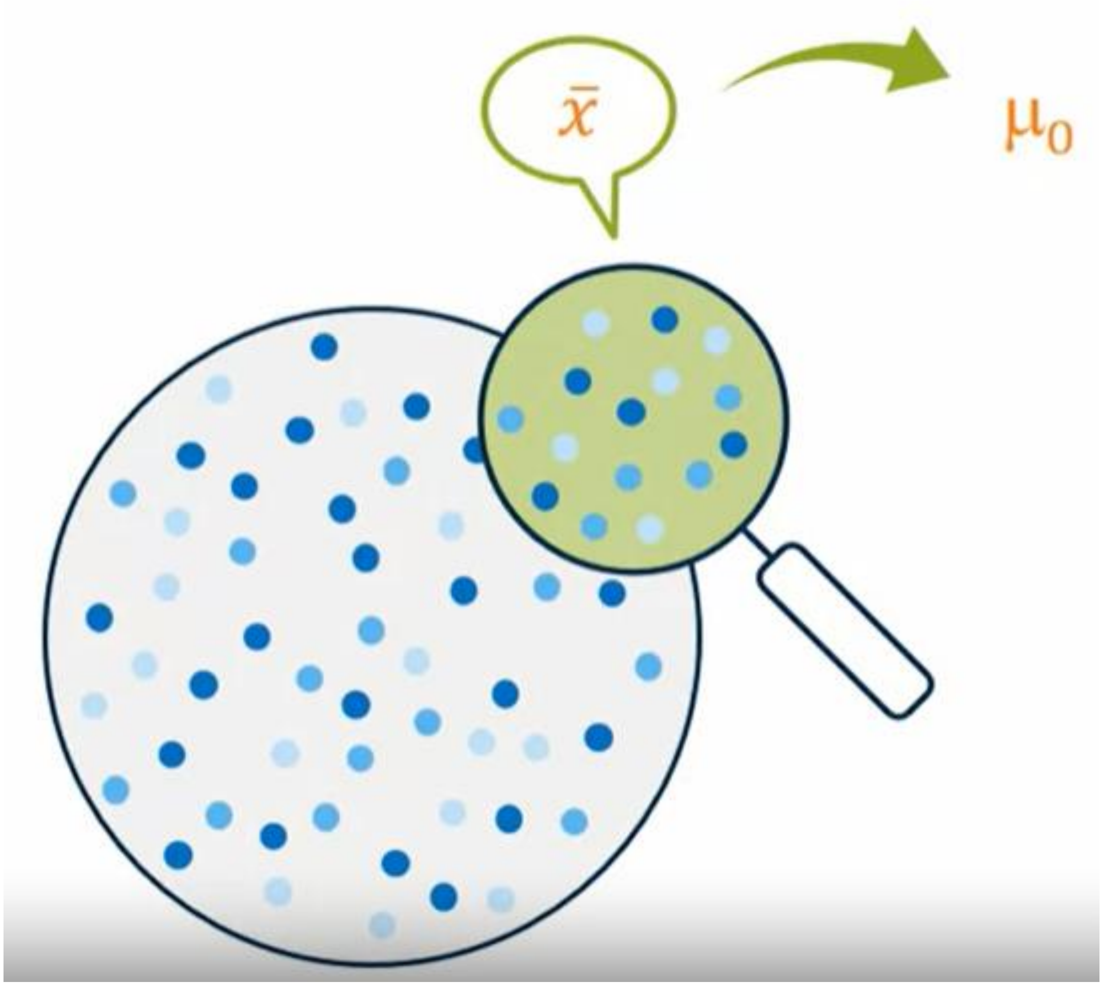
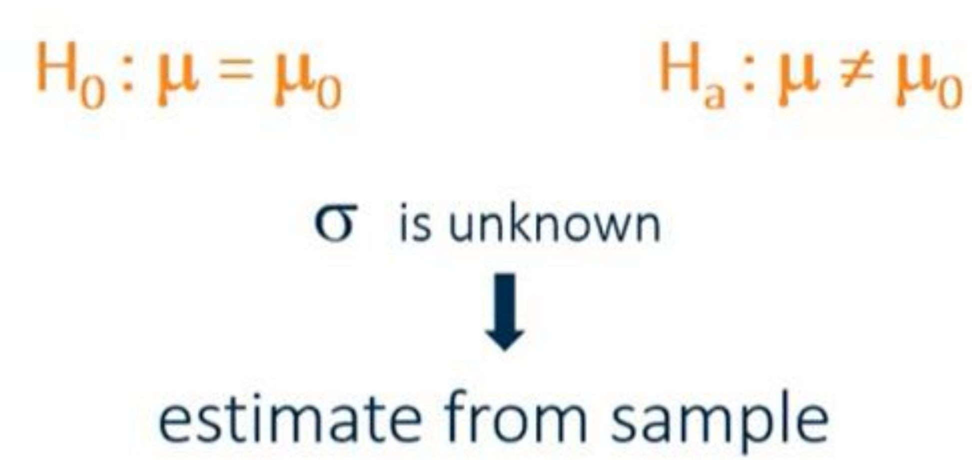
SBA: Statistical Business Analyst with SAS

SBA1: Introduction to Statistical Analysis: Hypothesis Testing

One-Sample t Test using PROC TTEST




$\bar{x}$
μ_0

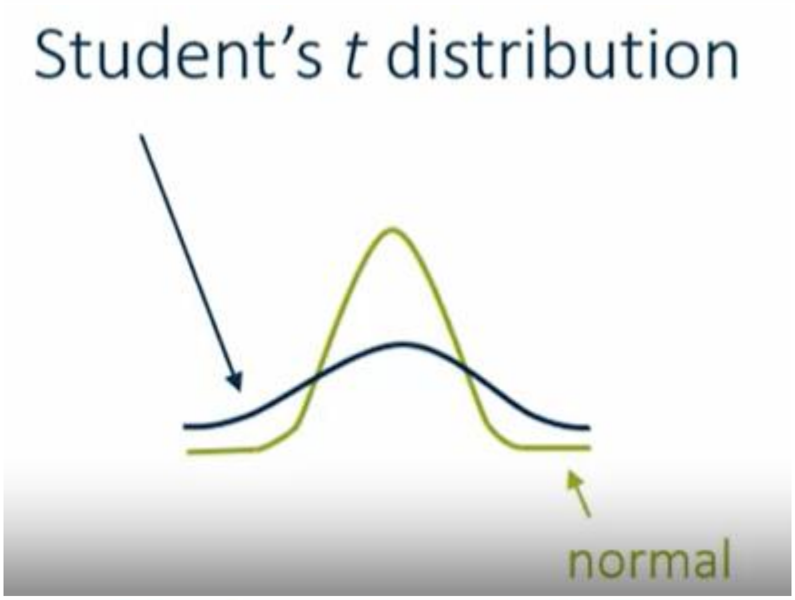

$H_0 : \mu = \mu_0$
$H_a : \mu \neq \mu_0$
σ is unknown
estimate from sample

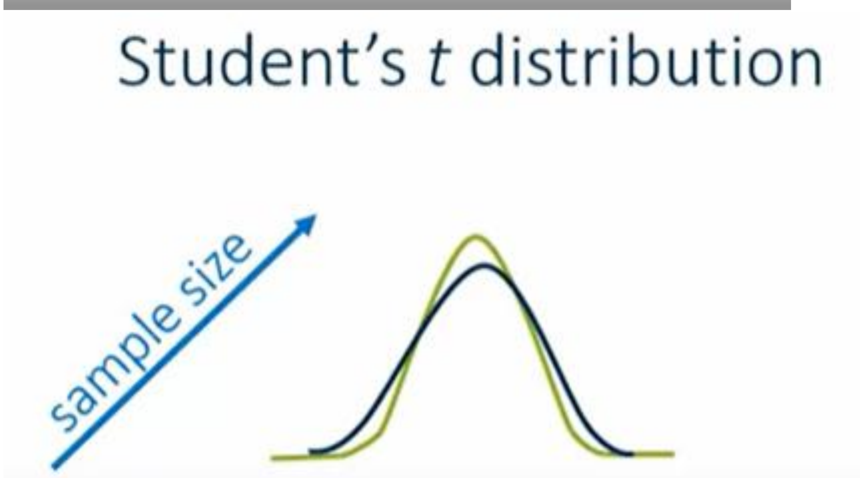
$H_0 : \mu = \mu_0$ $\qquad$ $H_a : \mu \neq \mu_0$

σ is unknown

$\Downarrow$

Student's *t* distribution





Student's t distribution: $t = \dfrac{(\bar{x} - \mu_0)}{s_{\bar{x}}}$

$$t = \frac{(137{,}525 - 135{,}000)}{2{,}172.1} = 1.16$$

$t =$

$t =$

positive

when sample mean > hypothesized mean

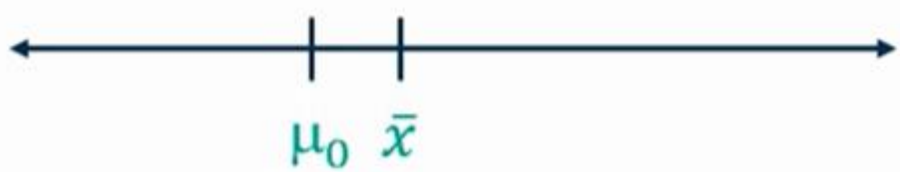
$$t =$$

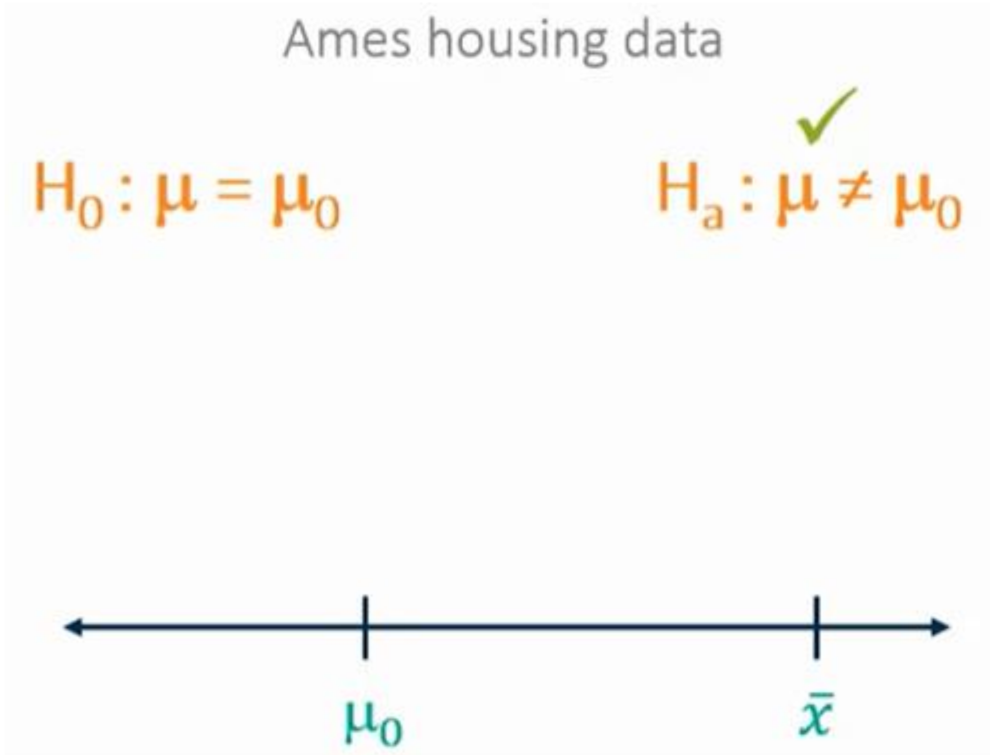
negative

when sample mean < hypothesized mean

Ames housing data

$H_0 : \mu = \mu_0$ ✓ $\quad\quad H_a : \mu \neq \mu_0$




Ames housing data
$H_0 : \mu = \mu_0$
$H_a : \mu \neq \mu_0$
μ_0
$\bar{x}$

$H_0 : \mu = \mu_0$ $\qquad$ $H_a : \mu \neq \mu_0$

probability

$t = 1.16$

$H_0 : \mu = \mu_0$ $\qquad$ $H_a : \mu \neq \mu_0$

probability

$\Downarrow$

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

✓

$H_0 : \mu = \mu_0$ $\qquad$ $H_a : \mu \neq \mu_0$

probability

⬇

p-value $> \alpha$

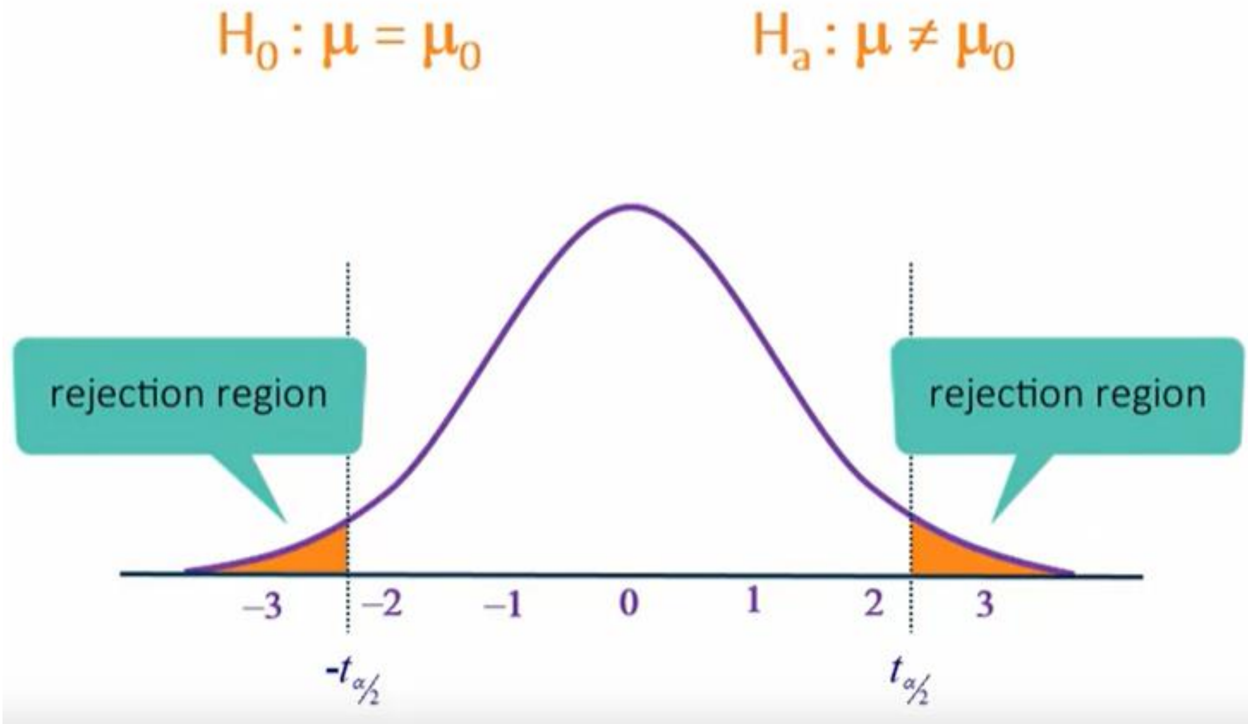
$H_0 : \mu = \mu_0$ $\qquad$ $H_a : \mu \neq \mu_0$

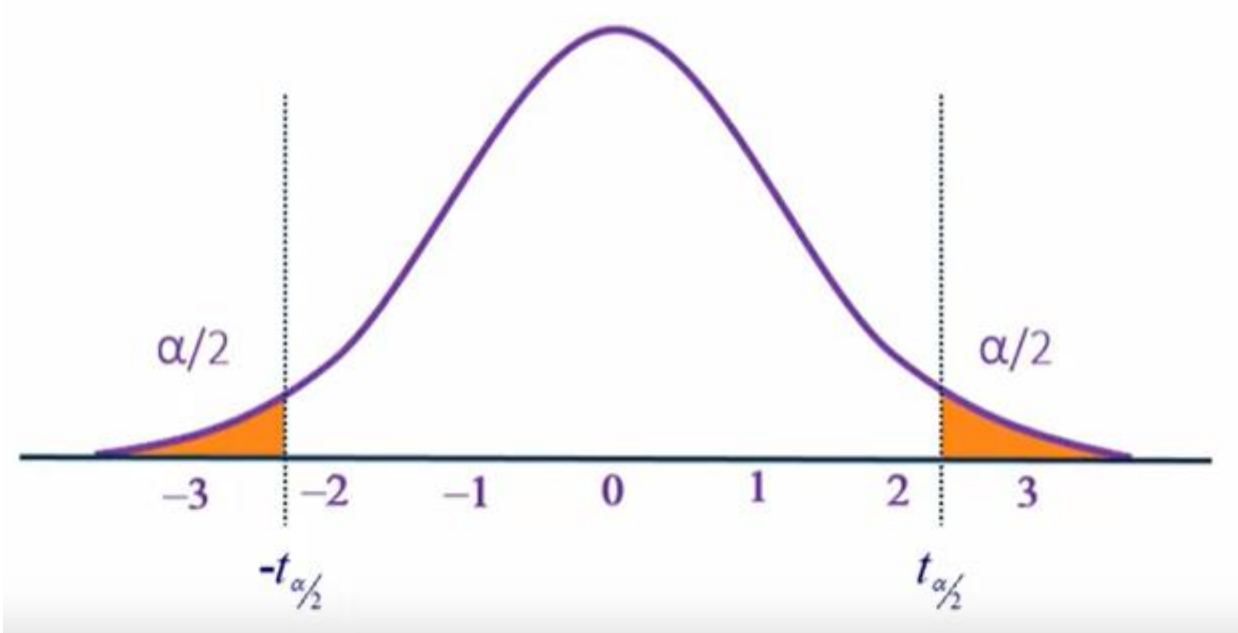
probability

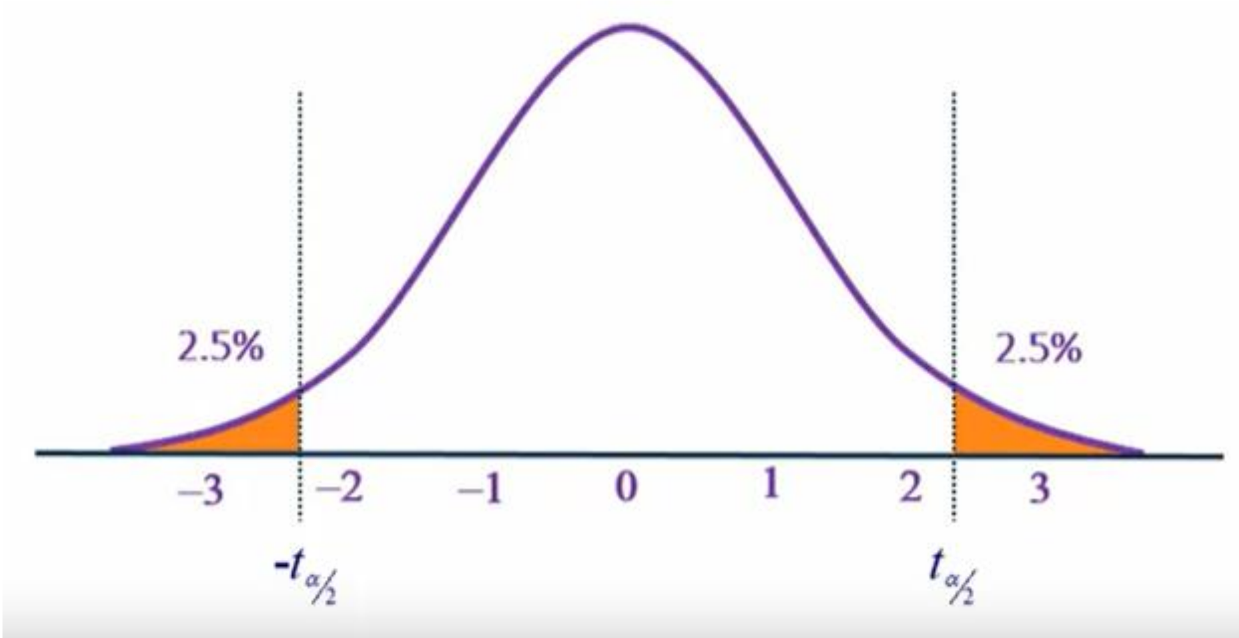
⬇

p-value < 0.05

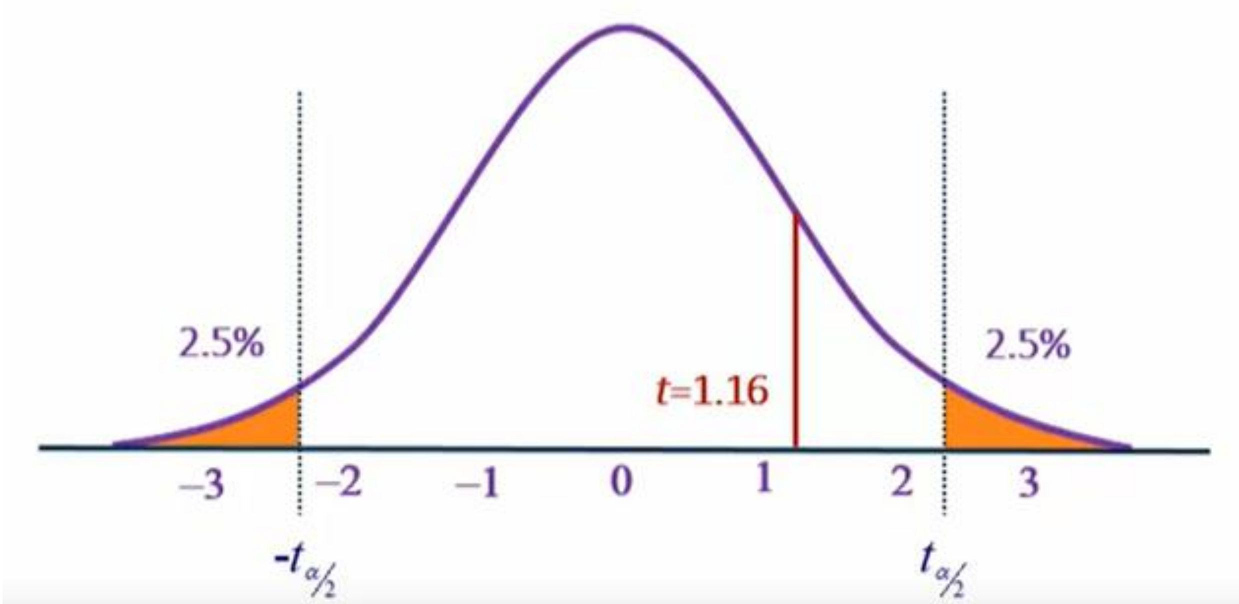
< 5% chance

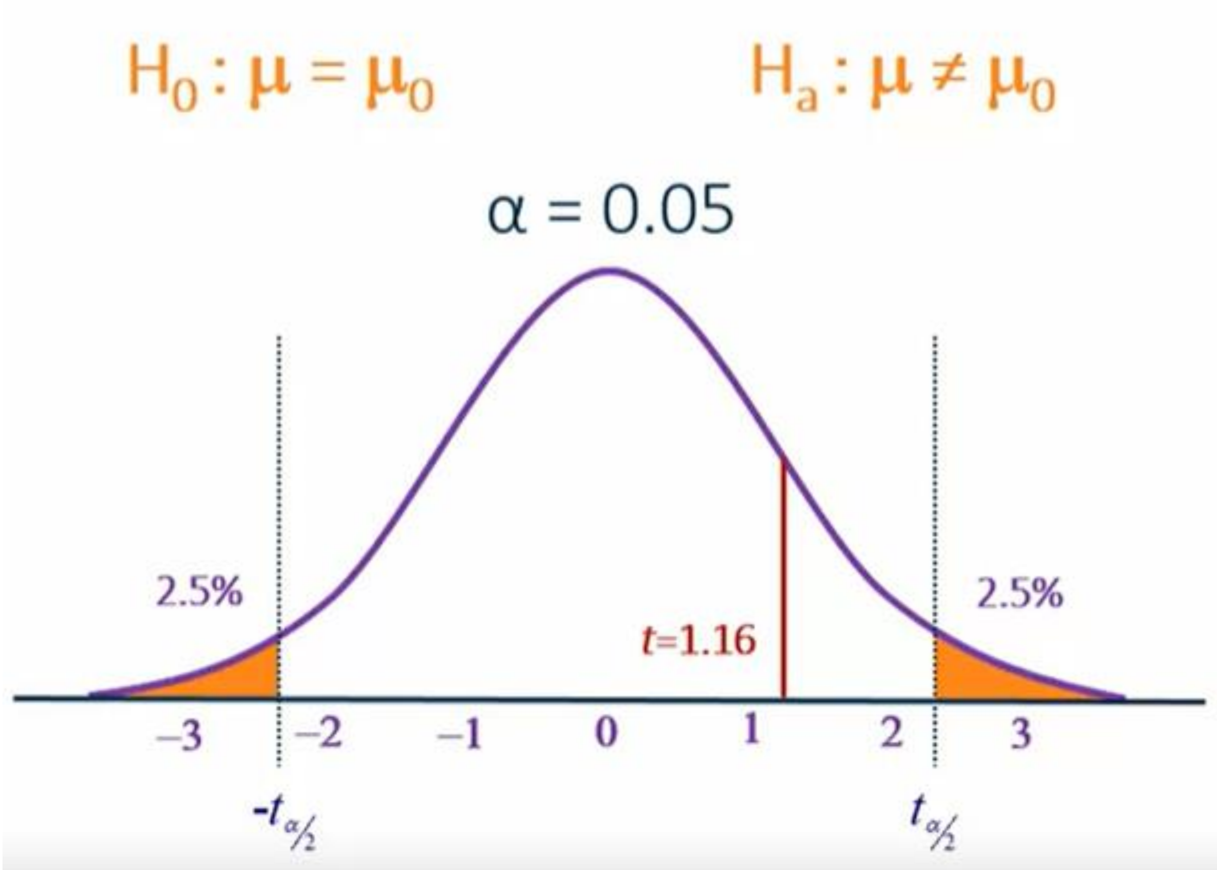

$H_0 : \mu = \mu_0$
$H_a : \mu \neq \mu_0$
rejection region
rejection region
−3
−2
−1
0
1
2
3
$-t_{\alpha/2}$
$t_{\alpha/2}$


α/2
α/2
−3
−2
−1
0
1
2
3
$-t_{\alpha/2}$
$t_{\alpha/2}$

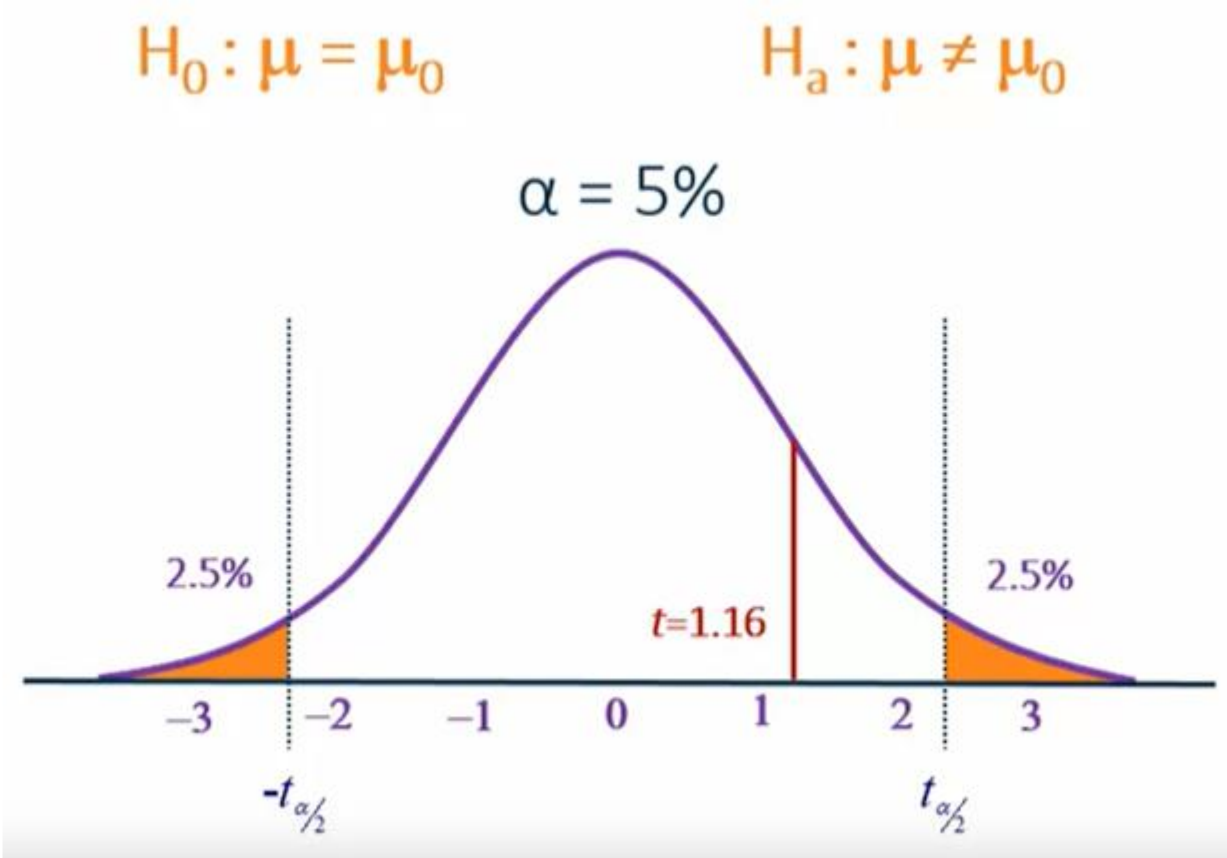


$H_0 : \mu = \mu_0$ $\quad$ $H_a : \mu \neq \mu_0$



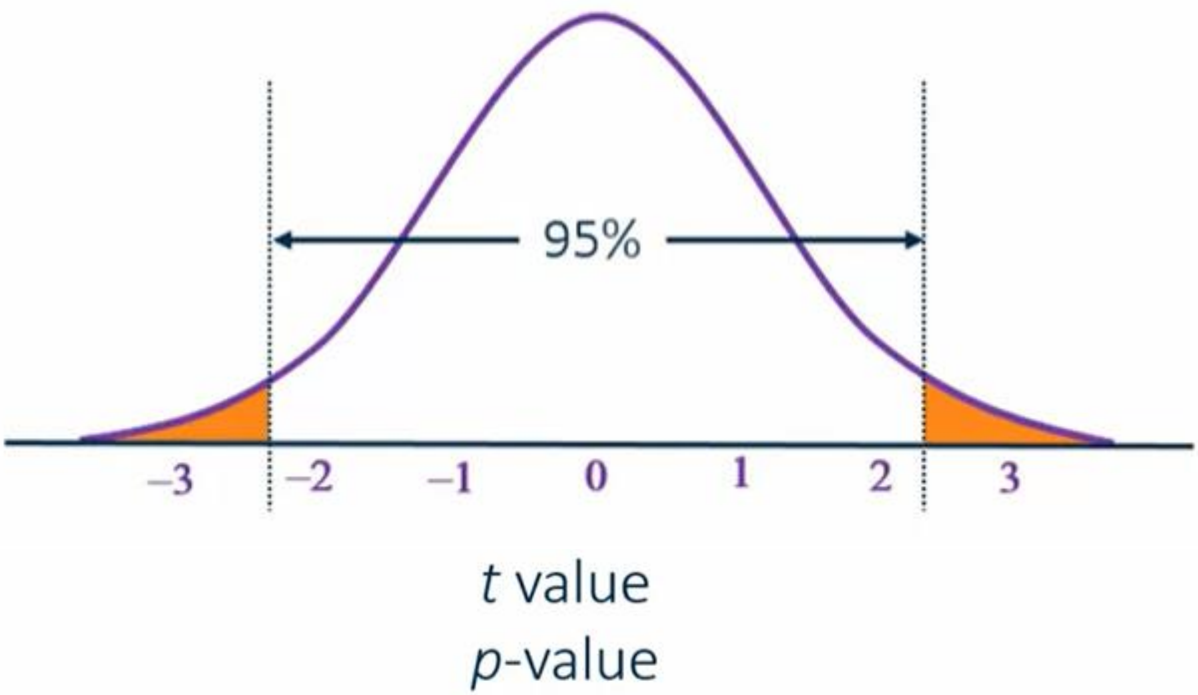


H0 : μ = μ0
Ha : μ ≠ μ0
α = 0.05
2.5%
2.5%
t=1.16
−3
−2
−1
0
1
2
3
-tα/2
tα/2



$H_0 : \mu = \mu_0$
$H_a : \mu \neq \mu_0$
$\alpha = 5\%$
2.5%
2.5%
t=1.16
–3
–2
–1
0
1
2
3
$-t_{\alpha/2}$
$t_{\alpha/2}$

$$H_0 : \mu = \mu_0 \qquad H_a : \mu \neq \mu_0$$



Variable: SalePrice (Sale price in dollars)

N	Mean	Std Dev	Std Err	Minimum	Maximum
300	137525	37622.6	2172.1	35000.0	290000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
137525	133250	141799	37622.6	34833.7	40900.7

DF	t Value	Pr > |t|
299	1.16	0.2460

TTEST Procedure

- performs *t* Tests
- computes confidence limits
- uses ODS graphics to produce:
  - histograms
  - Quantile-Quantile plots
  - box plots
  - confidence limit plots

```
/*st101d02.sas*/

ods graphics;
```

```
proc ttest data=STAT1.ameshousing3
        plots(shownull)=interval
        H0=135000;
   var SalePrice;
   title "One-Sample t-test testing whether mean SalePrice=$135,000";
run;

title;
```

One-Sample t-test testing whether mean SalePrice=$135,000

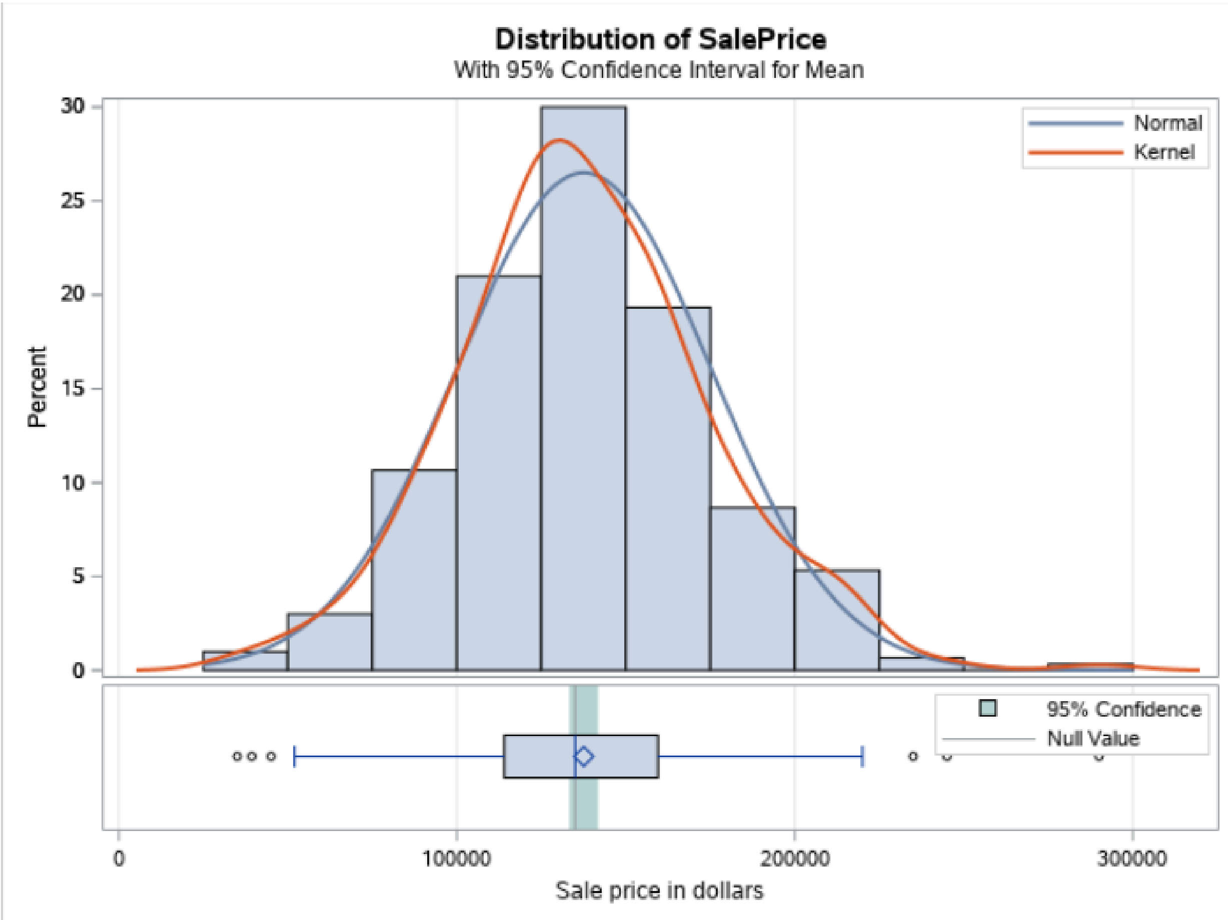
The TTEST Procedure

Variable: SalePrice (Sale price in dollars)

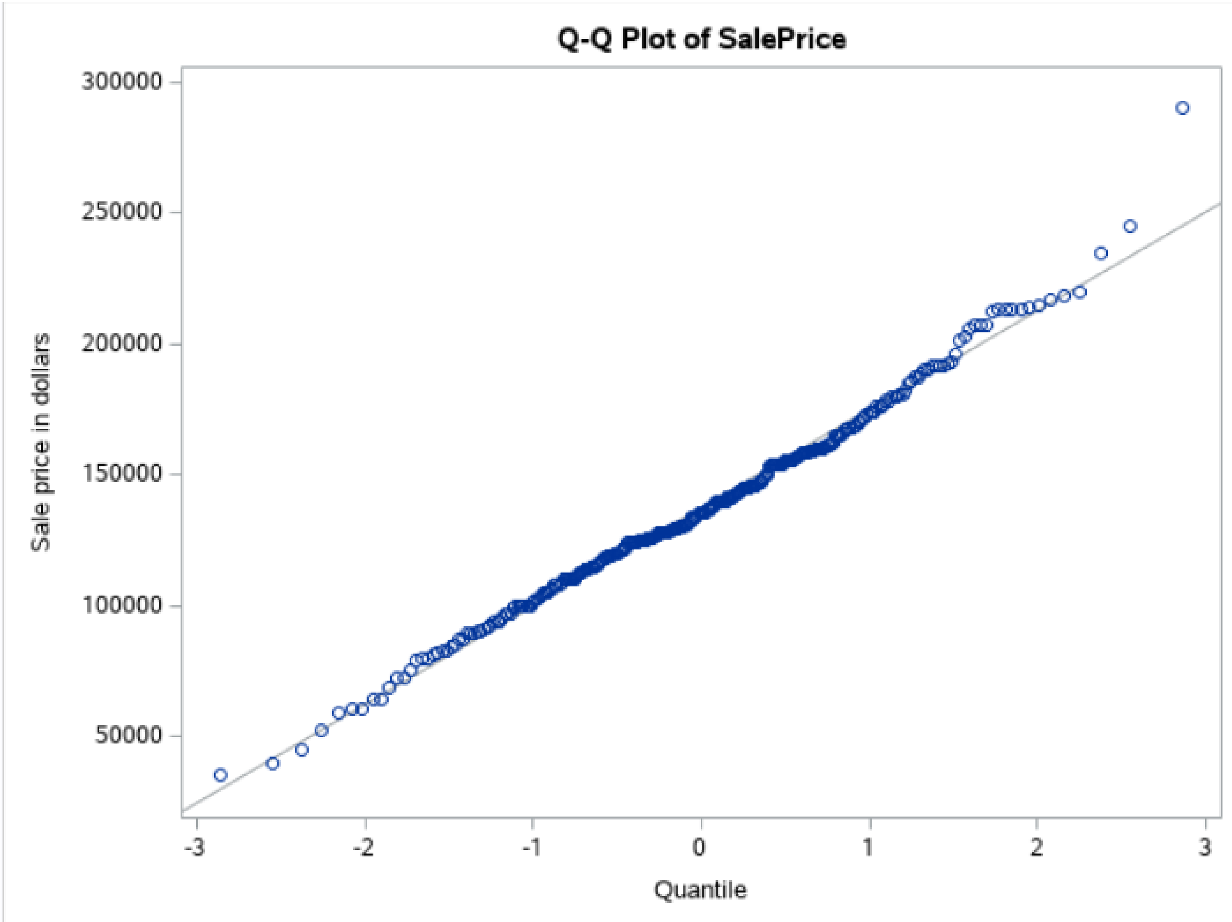
N	Mean	Std Dev	Std Err	Minimum	Maximum
300	137525	37622.6	2172.1	35000.0	290000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
137525	133250	141799	37622.6	34833.7	40900.7

DF	t Value	Pr > |t|
299	1.16	0.2460


Distribution of SalePrice
With 95% Confidence Interval for Mean
Percent
Normal
Kernel
95% Confidence
Null Value
0
100000
200000
300000
Sale price in dollars


Mean of SalePrice
With 95% Confidence Interval
Mean
Null Value
134000
136000
138000
140000
142000
Sale price in dollars



Gender	ID	FatPct
M	1876	13.3
M	1555	22
F	1231	29
F	1977	21.7
M	1194	31

```
proc ttest data=bodyfat;
    class _________ ;
    var ____________;
run;
```

```
proc ttest data=bodyfat;
   class Gender;
   var FatPct;
run;
```

Practice - Using PROC TTEST to Perform a One-Sample t Test

TOTAL POINTS 3

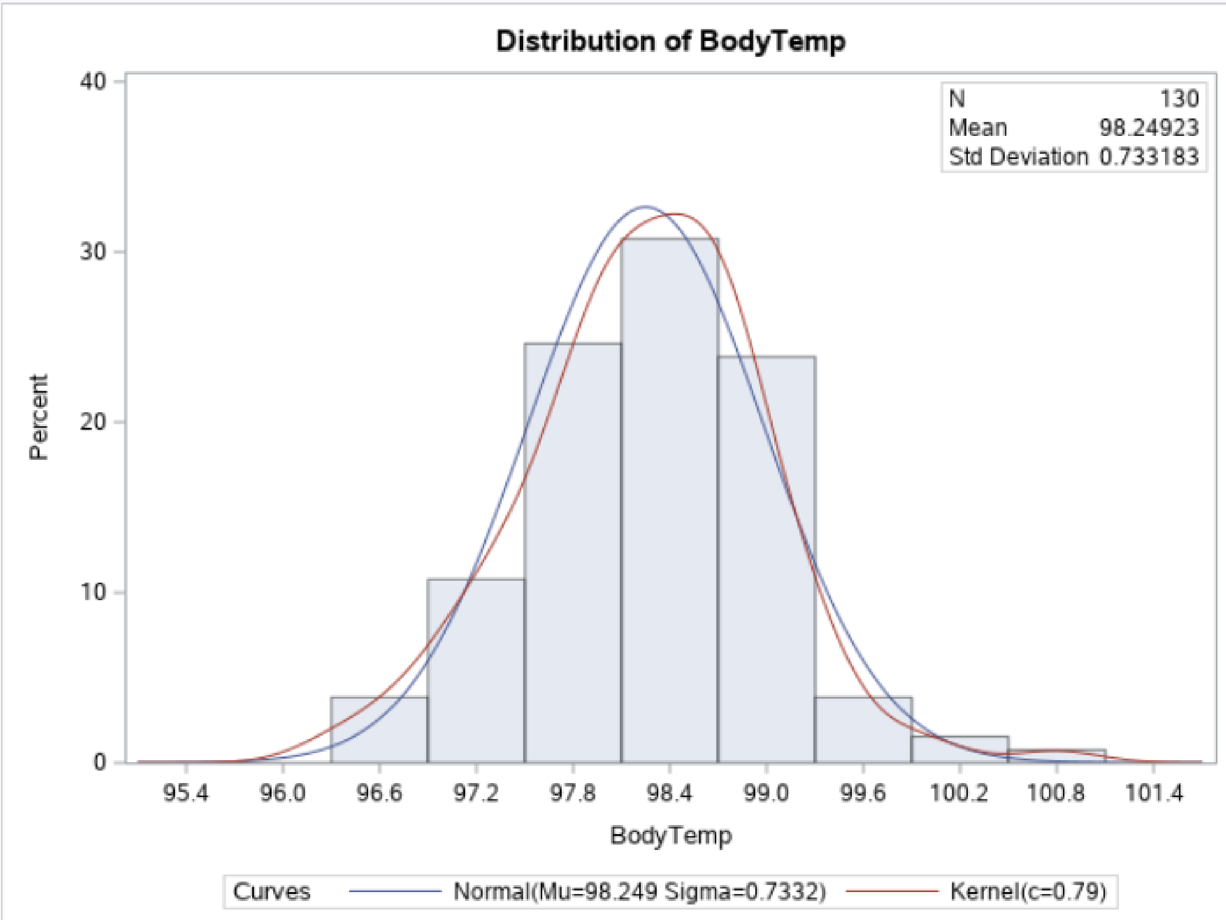
1.
Question 1
The data in **stat1.normtemp** come from an article in the *Journal of Statistics Education* by Dr. Allen L. Shoemaker from the Psychology Department at Calvin College. The data are based on an article in a 1992 edition of *JAMA* (*Journal of the American Medical Association*). The notion that the true mean body temperature is 98.6 is questioned. There are 65 males and 65 females. There is also some doubt about whether mean body temperatures for women are the same as for men.

1. Look at the distribution of the continuous variables in the **stat1.normtemp** data set.Use PROC UNIVARIATE to produce histograms and insets with means, standard deviations, and sample size.

```
/*st101s01.sas*/  /*Part A*/

%let interval=BodyTemp HeartRate;


ods graphics;

ods select histogram;

proc univariate data=STAT1.NormTemp noprint;

   var &interval;

   histogram &interval / normal kernel;

   inset n mean std / position=ne;

   title "Interval Variable Distribution Analysis";

run;

title;
```

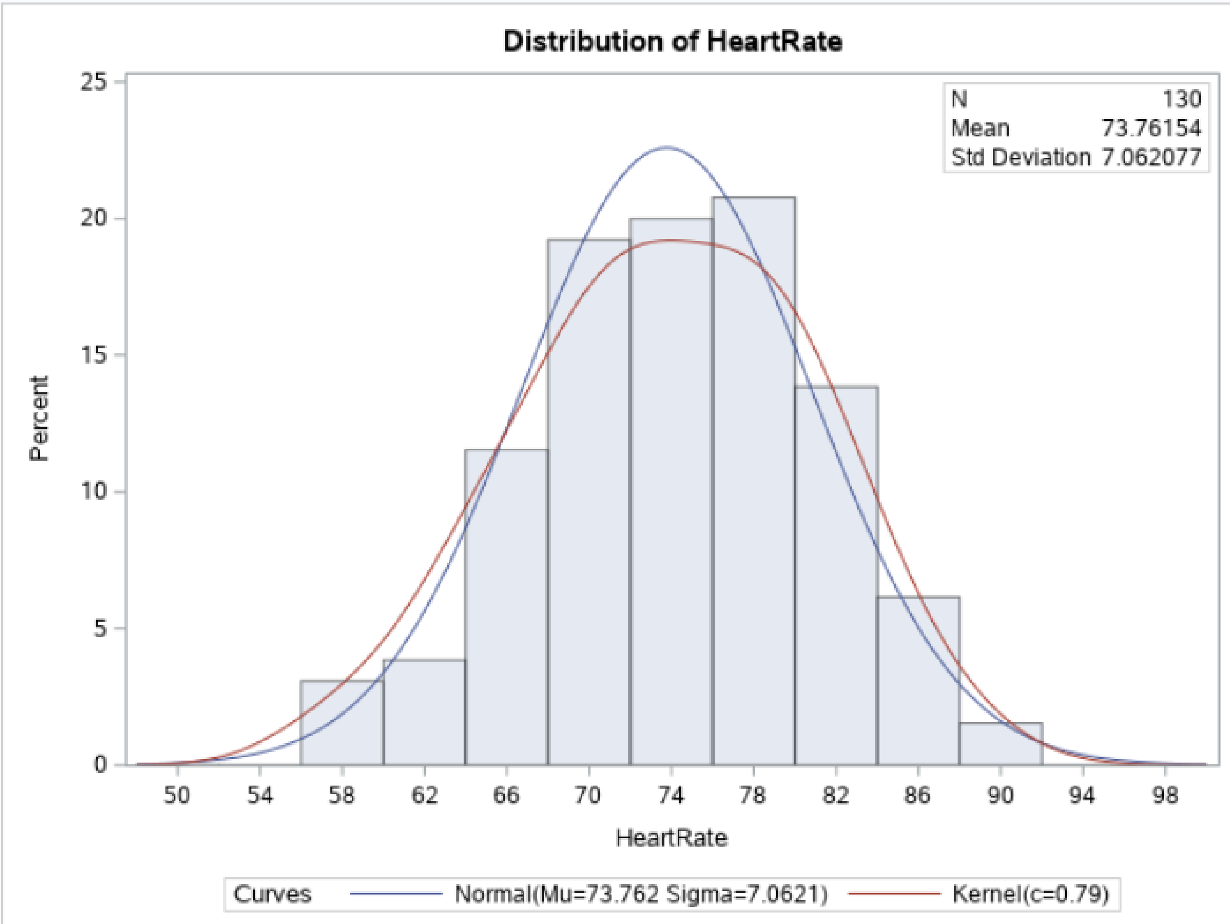
Interval Variable Distribution Analysis

The UNIVARIATE Procedure



Interval Variable Distribution Analysis

The UNIVARIATE Procedure




```
/*st101s01.sas*/  /*Part B*/
proc ttest data=STAT1.NormTemp h0=98.6
         plots(only shownull)=interval;
   var BodyTemp;
   title 'Testing Whether the Mean Body Temperature = 98.6';
run;
title;
```

Testing Whether the Mean Body Temperature = 98.6

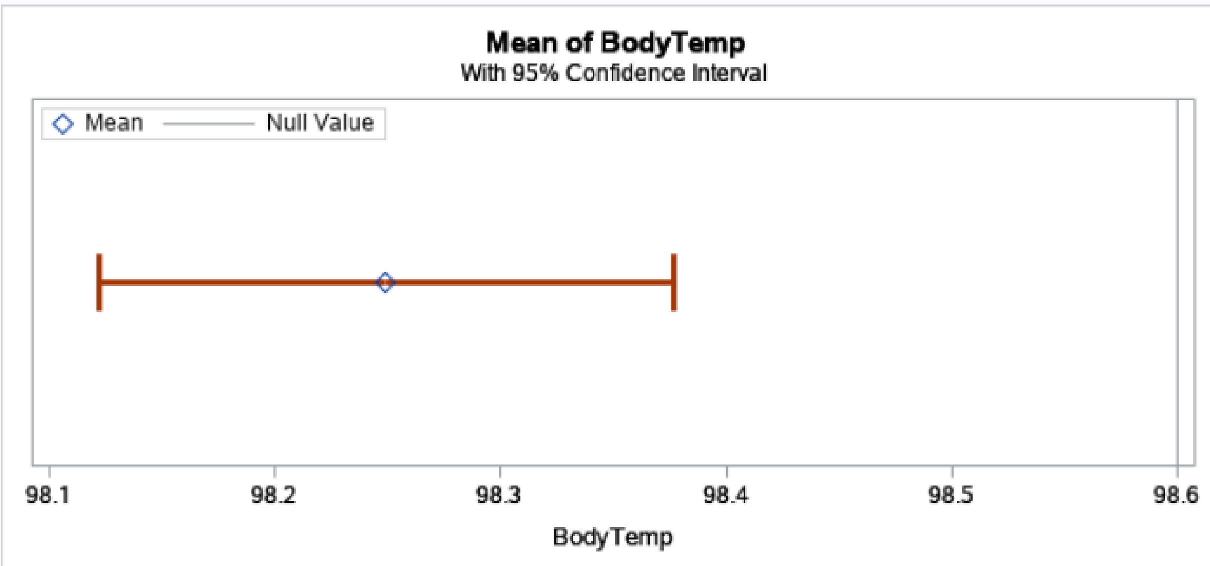
The TTEST Procedure

Variable: BodyTemp

N	Mean	Std Dev	Std Err	Minimum	Maximum
130	98.2492	0.7332	0.0643	96.3000	100.8

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
98.2492	98.1220	98.3765	0.7332	0.6536	0.8350

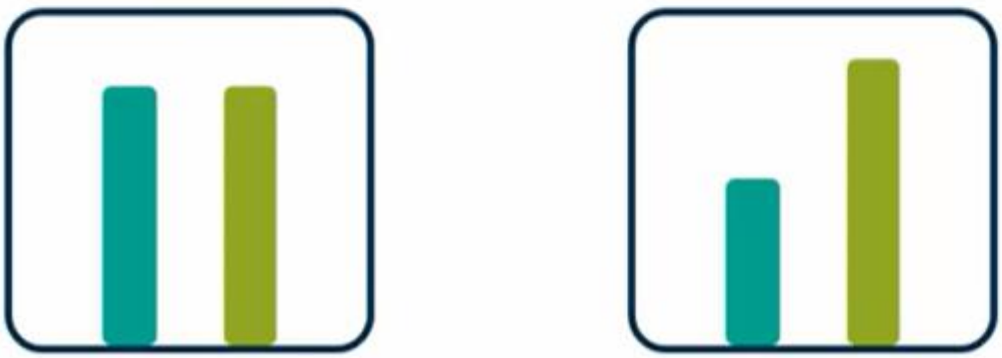
DF	t Value	Pr > |t|
129	-5.45	<.0001



Two-Sample t Tests using PROC TTEST

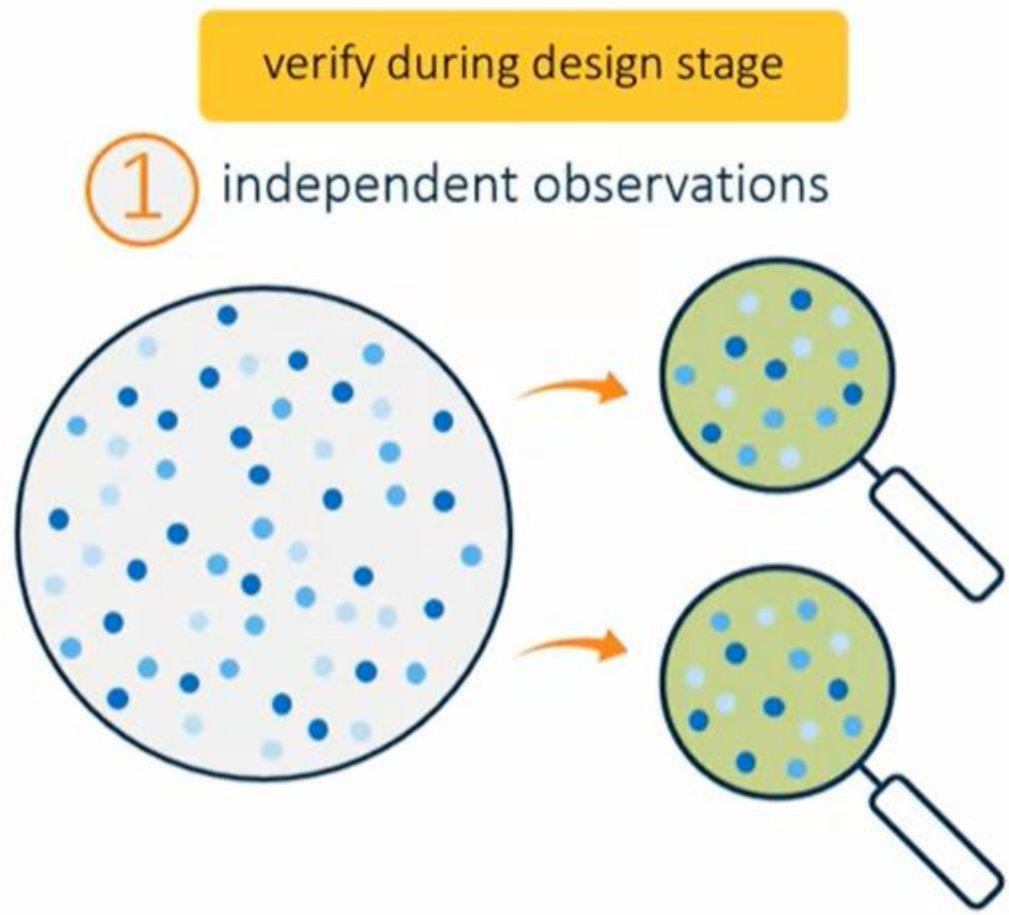
$$H_0 : \mu_1 = \mu_2 \qquad H_a : \mu_1 \neq \mu_2$$

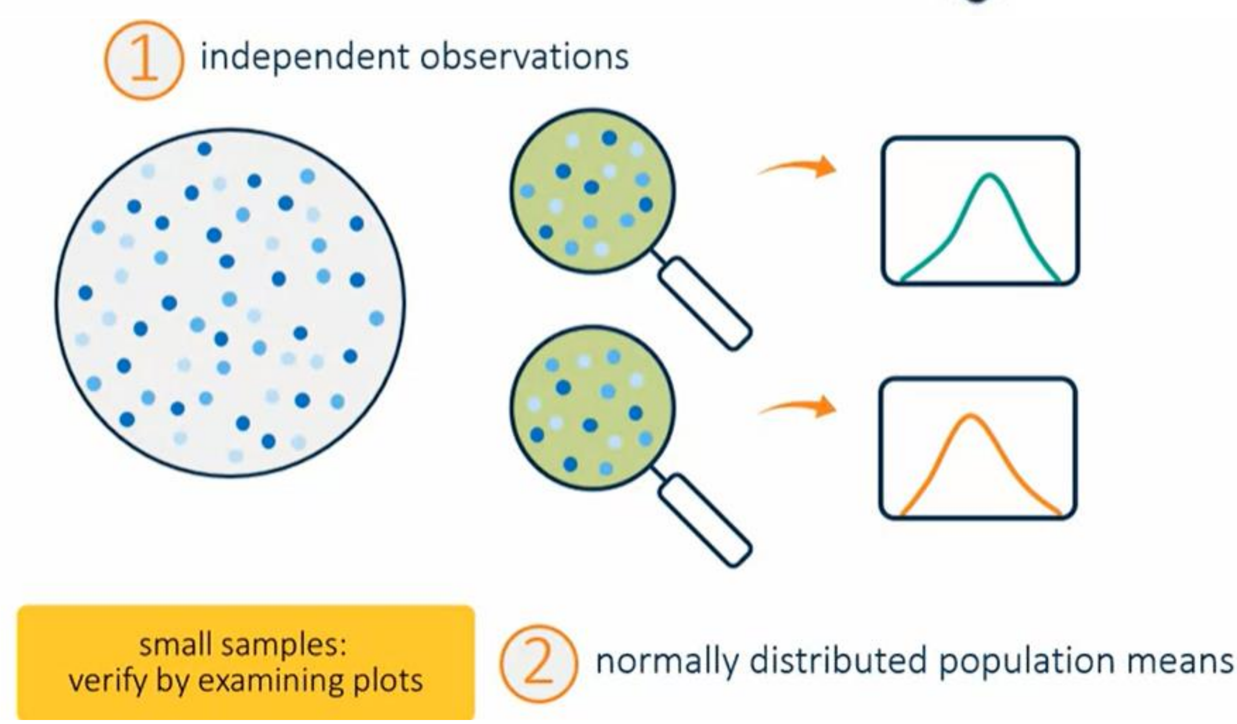
$$H_0 : \mu_1 - \mu_2 = 0 \qquad H_a : \mu_1 - \mu_2 \neq 0$$

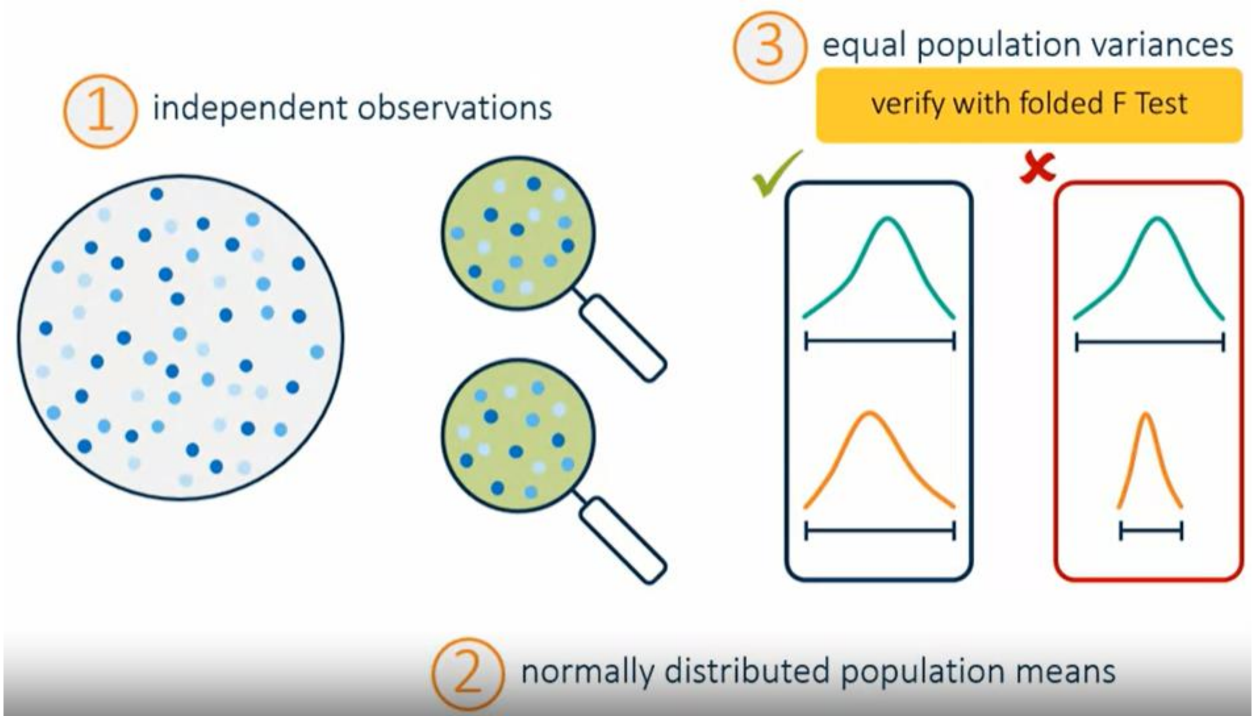


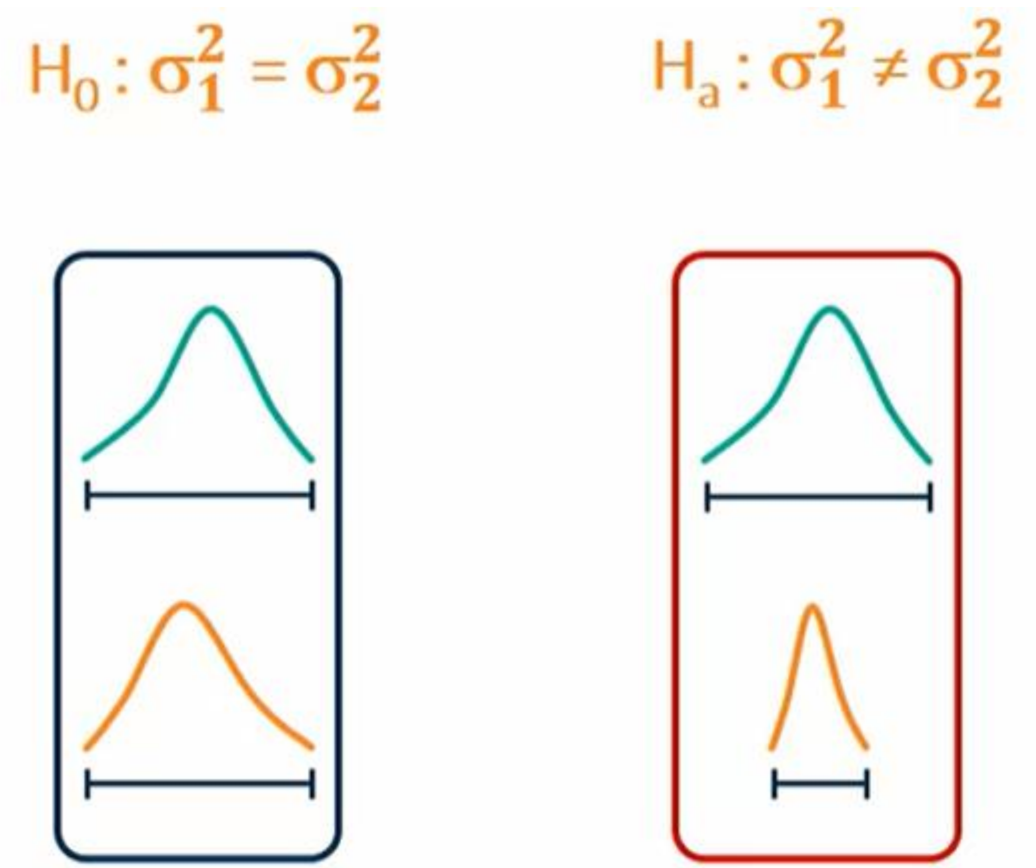
assumptions

1 2 3


verify during design stage
1 independent observations


1 independent observations
small samples:
verify by examining plots
2 normally distributed population means


1 independent observations
3 equal population variances
verify with folded F Test
2 normally distributed population means


$H_0 : \sigma_1^2 = \sigma_2^2$
$H_a : \sigma_1^2 \neq \sigma_2^2$

$H_0: \sigma_1^2 = \sigma_2^2$ $\quad$ $H_a: \sigma_1^2 \neq \sigma_2^2$

F statistic: $\boxed{F = \dfrac{\max(s_1^2, s_2^2)}{min(s_1^2, s_2^2)}}$

$$F \geq 1$$

$H_0: \sigma_1^2 = \sigma_2^2$ $\quad$ $H_a: \sigma_1^2 \neq \sigma_2^2$

F statistic: $\boxed{F = \dfrac{\max(s_1^2, s_2^2)}{min(s_1^2, s_2^2)}}$

$$F \approx 1$$

$H_0: \sigma_1^2 = \sigma_2^2$ $H_a: \sigma_1^2 \neq \sigma_2^2$

F statistic: $\boxed{F = \frac{\max(s_1^2, s_2^2)}{min(s_1^2, s_2^2)}}$

F = large

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	6.0	7.40	0.0003
Satterthwaite	Unequal	5.8	7.40	0.0004

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	3	3	1.51	0.7446

$\alpha = 0.05$

$H_0: \sigma_1^2 = \sigma_2^2$ $H_a: \sigma_1^2 \neq \sigma_2^2$

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	6.0	7.40	0.0003
Satterthwaite	Unequal	5.8	7.40	0.0004

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	3	3	1.51	0.7446

$> \alpha$

$H_0: \sigma_1^2 = \sigma_2^2$ $\quad$ $H_a: \sigma_1^2 \neq \sigma_2^2$

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	6.0	7.40	0.0003
Satterthwaite	Unequal	5.8	7.40	0.0004

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	3	3	1.51	0.7446

$H_0: \sigma_1^2 = \sigma_2^2$ $\quad$ $H_a: \sigma_1^2 \neq \sigma_2^2$

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	6.0	7.40	0.0003
Satterthwaite	Unequal	5.8	7.40	0.0004

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	3	3	1.51	0.7446

✘ $H_0: \sigma_1^2 = \sigma_2^2$ ✔ $H_a: \sigma_1^2 \neq \sigma_2^2$

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	6.0	7.40	$\boxed{0.0003}$
Satterthwaite	Unequal	5.8	7.40	0.0004

< 0.05

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	3	3	1.51	0.7446

✘ $H_0: \sigma_1^2 = \sigma_2^2$ ✔ $H_a: \sigma_1^2 \neq \sigma_2^2$

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	13.0	-1.78	0.0979
Satterthwaite	Unequal	11.1	-2.45	0.0320

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	9	4	15.28	$\boxed{0.0185}$

< 0.05

✗ $H_0 : \sigma_1^2 = \sigma_2^2$ ✓ $H_a : \sigma_1^2 \neq \sigma_2^2$

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	13.0	-1.78	0.0979
Satterthwaite	Unequal	11.1	-2.45	0.0320

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	9	4	15.28	0.0185

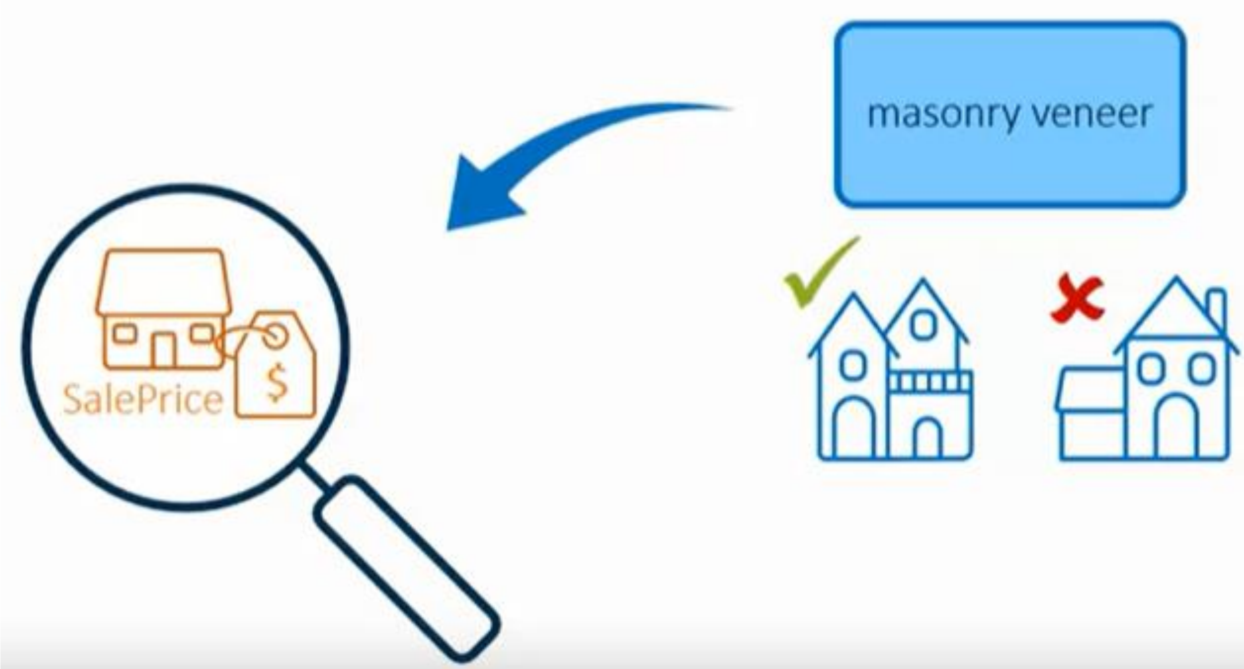
✗ $H_0 : \sigma_1^2 = \sigma_2^2$ ✓ $H_a : \sigma_1^2 \neq \sigma_2^2$

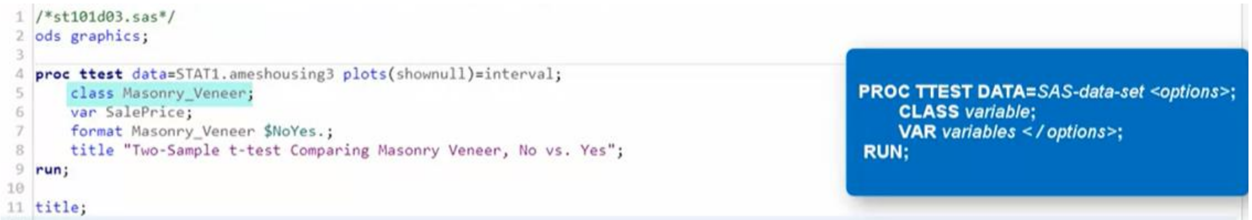
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	13.0	-1.78	0.0979
Satterthwaite	Unequal	11.1	-2.45	0.0320

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	9	4	15.28	0.0185

✓ $H_0: \sigma_1^2 = \sigma_2^2$ ✗ $H_a: \sigma_1^2 \neq \sigma_2^2$

PROC TTEST






```
/*st101d03.sas*/

ods graphics;


proc ttest data=STAT1.ameshousing3 plots(shownull)=interval;

   class Masonry_Veneer;

   var SalePrice;

   format Masonry_Veneer $NoYes.;

   title "Two-Sample t-test Comparing Masonry Veneer, No vs. Yes";

run;


title;
```

Two-Sample t-test Comparing Masonry Veneer, No vs. Yes

The TTEST Procedure

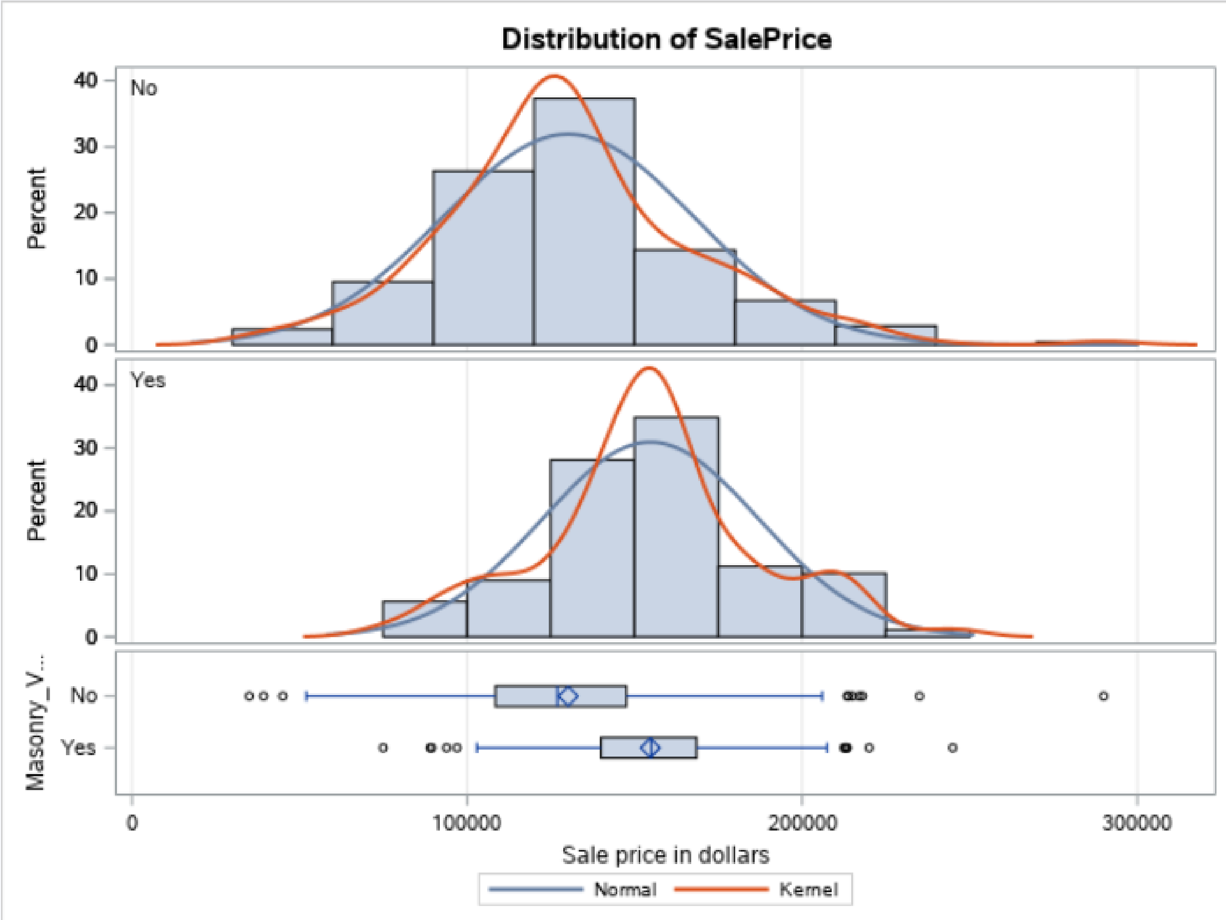
Variable: SalePrice (Sale price in dollars)

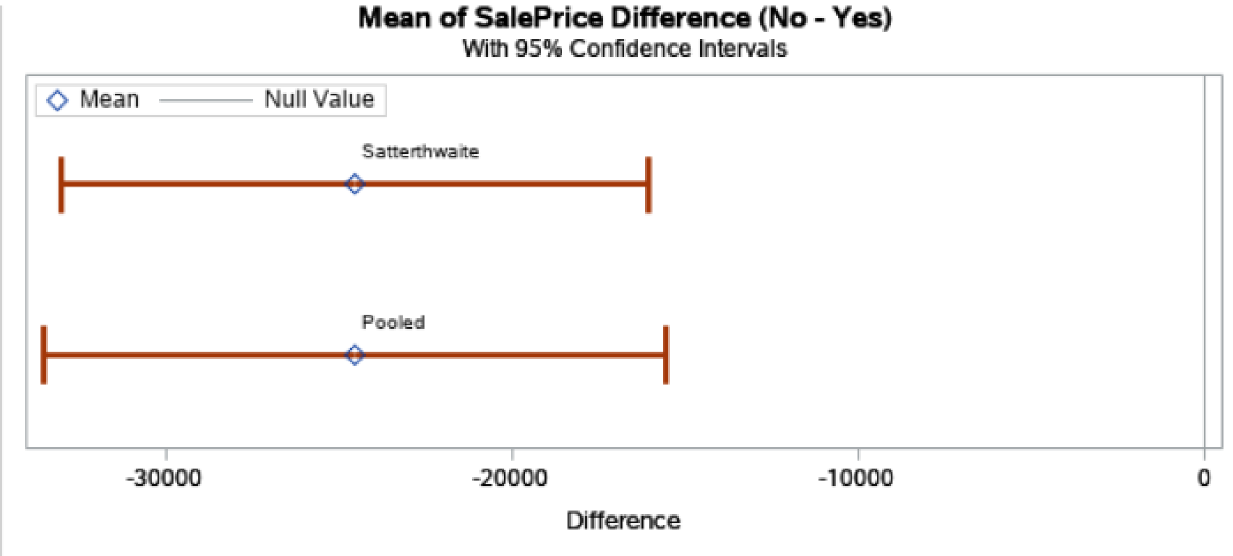
Masonry_Veneer	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
No		209	130172	37531.7	2596.1	35000.0	290000
Yes		89	154705	32239.8	3417.4	75000.0	245000
Diff (1-2)	Pooled		-24533.0	36039.6	4561.6		
Diff (1-2)	Satterthwaite		-24533.0		4291.7		

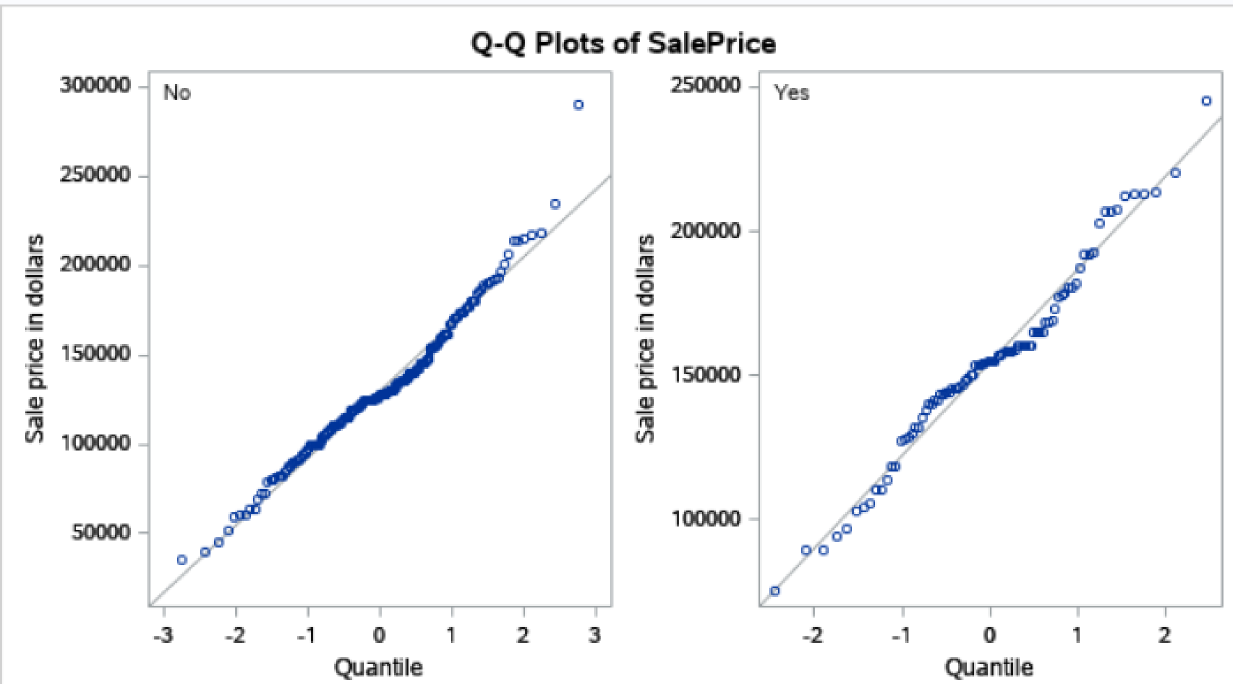
Masonry_Veneer	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
No		130172	125054	135290	37531.7	34245.4	41521.0
Yes		154705	147914	161496	32239.8	28099.7	37821.9
Diff (1-2)	Pooled	-24533.0	-33510.3	-15555.6	36039.6	33355.6	39197.1
Diff (1-2)	Satterthwaite	-24533.0	-32997.9	-16068.0			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	296	-5.38	<.0001
Satterthwaite	Unequal	191.85	-5.72	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	208	88	1.36	0.1039


Distribution of SalePrice
No
Yes
Percent
Percent
40
30
20
10
0
40
30
20
10
0
Masonry_V...
No
Yes
0
100000
200000
300000
Sale price in dollars
Normal
Kernel





Practice - Using PROC TTEST to Compare Groups

TOTAL POINTS 3

1.

Question 1

Elli Sagerman, a Masters of Education candidate in German Education at the University of North Carolina at Chapel Hill in 2000, collected data for a study. She looked at the effectiveness of a new type of foreign language teaching technique on grammar skills. She selected 30 students to receive tutoring. Fifteen received the new type of training during the tutorials and 15 received standard tutoring. Two students moved from the district before completing the study. Scores on a standardized German grammar test were recorded immediately before the 12-week tutorials and again 12 weeks later at the end of the trial. Sagerman wanted to see the effect of the new technique on grammar skills.

1. If necessary, start SAS Studio and redefine the **STAT1** library.
2. Using PROC TTEST, analyze the **stat1.german** data set. Assess whether the treatment group improved more than the control group.
3. Do the two groups seem to be approximately normally distributed?

Question 2

Do the two groups have approximately equal variances?

Question 3

Does the new teaching technique seem to result in significantly different scores compared with the standard technique?

```
/*st101s02.sas*/

ods graphics;

proc ttest data=STAT1.German plots(shownull)=interval;

  class Group;

  var Change;

  title "German Grammar Training, Comparing Treatment to Control";

run;


title;
```

German Grammar Training, Comparing Treatment to Control

The TTEST Procedure

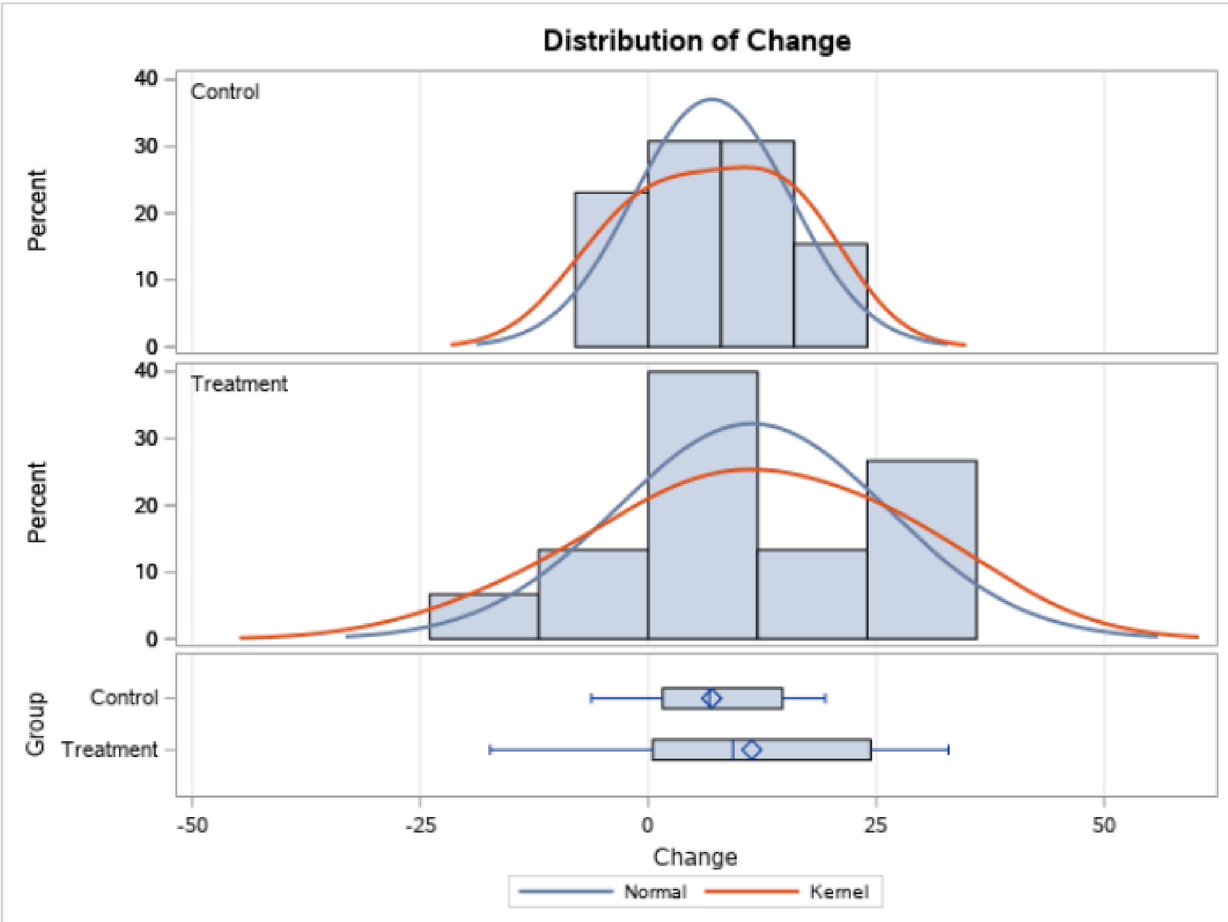
Variable: Change

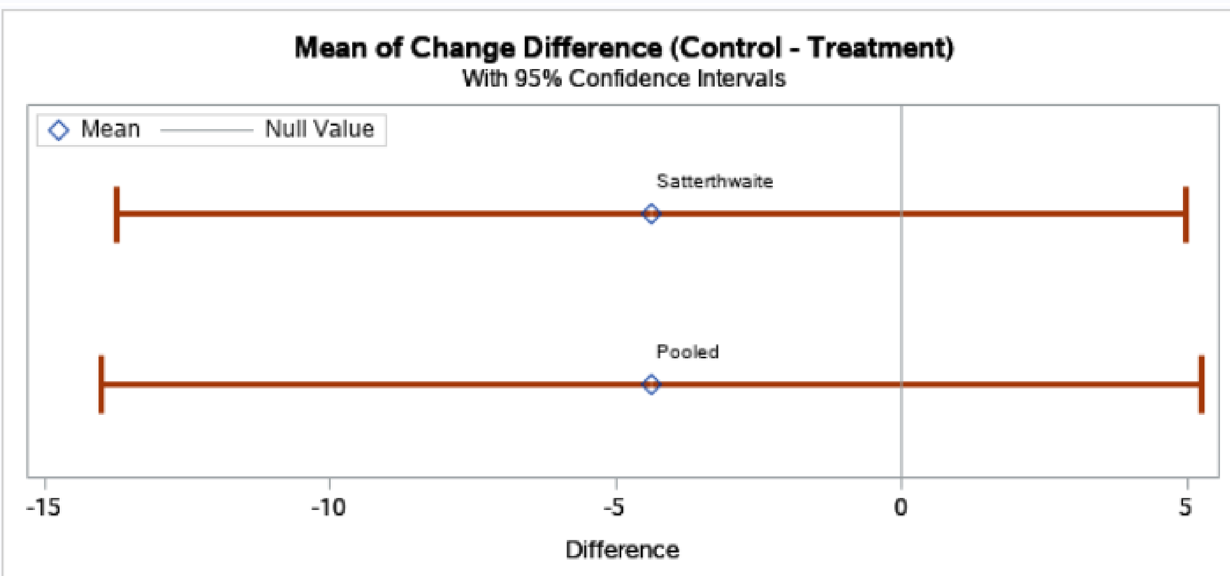
Group	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Control		13	6.9677	8.6166	2.3898	-6.2400	19.4100
Treatment		15	11.3587	14.8535	3.8352	-17.3300	32.9200
Diff (1-2)	Pooled		-4.3910	12.3720	4.6882		
Diff (1-2)	Satterthwaite		-4.3910		4.5188		

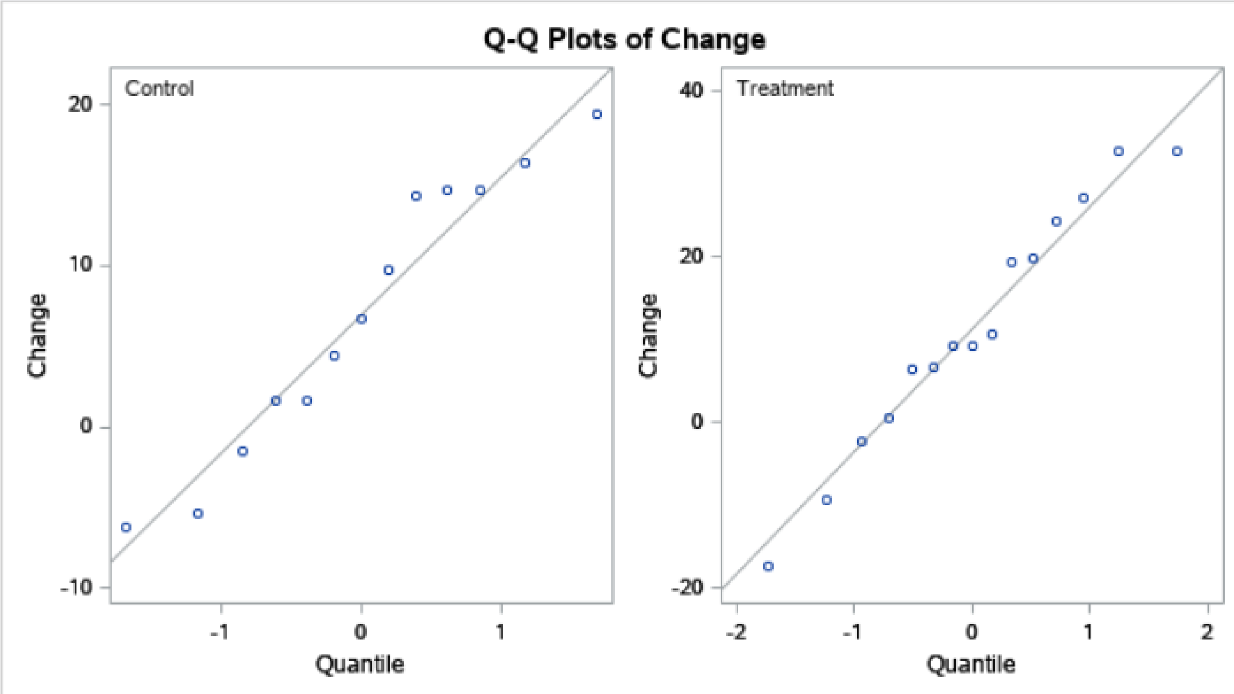
Group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Control		6.9677	1.7607	12.1747	8.6166	6.1789	14.2238
Treatment		11.3587	3.1331	19.5843	14.8535	10.8747	23.4255
Diff (1-2)	Pooled	-4.3910	-14.0276	5.2457	12.3720	9.7432	16.9550
Diff (1-2)	Satterthwaite	-4.3910	-13.7401	4.9581			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	26	-0.94	0.3576
Satterthwaite	Unequal	22.947	-0.97	0.3413

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	14	12	2.97	0.0660


Distribution of Change
Control
Treatment
Percent
Percent
Group
Control
Treatment
-50
-25
0
25
50
Change
Normal
Kernel


Mean of Change Difference (Control - Treatment)
With 95% Confidence Intervals
Mean
Null Value
Satterthwaite
Pooled
-15
-10
-5
0
5
Difference


Q-Q Plots of Change
Control
Treatment
Change
Quantile
20
10
0
-10
-1
0
1
40
20
0
-20
-2
-1
0
1
2