Introductory Inferential Methods

Statistics 426: SAS Programming

Module 13

2021

Iris data

```
filename flower url 'https://webpages.uidaho.edu/~renaes/Data/iris.csv';
data flower;
infile flower dsd missover firstobs=2;
input s_length s_width p_length p_width species$;
run;
```

Old Faithful data

```
filename faith url 'https://webpages.uidaho.edu/~renaes/Data/faithful.csv';
data faithful;
infile faith dsd missover firstobs=2;
input day eruptions waiting;
run;
```

Tree diameter data

```
filename leave url 'https://webpages.uidaho.edu/~renaes/Data/tree_diameter.csv';
data tree;
infile leave dsd missover firstobs=2;
input diameter direction$;
run;
```

Crab data

```
filename eugene url 'https://webpages.uidaho.edu/~renaes/Data/crabs.csv';
data krabs;
infile eugene dsd missover firstobs=2;
input weight species$;
run;
```

Handwashing data

```
filename yuck url 'https://webpages.uidaho.edu/~renaes/Data/handwashing.csv';
data gross;
infile yuck dsd missover firstobs=2;
input bacteria method$;
run;
```

Introduction

Statistical Inference: using information obtained from a proper sample to make an educated judgment about a population

Three types of inference

(1) point estimation
(2) interval estimation (aka confidence intervals (CIs))
(3) statistical tests (aka hypothesis tests)

Terms I

Point Estimation: a point estimate of a parameter (let's just say a generic parameter is called θ and its estimate (statistic) is called $\hat{\theta}$). $\hat{\theta}$ is a single number that can be regarded as a sensible value for θ. It is obtained by selecting a suitable statistic and computing its value from the given sample data. The selected statistic is called the **point estimate** of θ. Examples are: $\overline{X}$ estimates μ, $\hat{\pi}$ estimates π, s estimates σ, and so on.

A point estimate is just a single number and by itself provides no information about the precision and reliability of estimation; it gives no feedback on how close our estimate was to the parameter.

Terms II

Interval estimation: an alternative to reporting a sensible value for the parameter being estimated is to calculate an entire interval of plausible values, called **interval estimation**, specifically we call them **confidence intervals (CIs)**.

Select the level of confidence, it is usually 95% but others are also used often (90%, 98%, 99%). A CI with level 95% implies that 95% of samples would give an interval that contains θ, or that only 5% of samples would not contain θ.

Assumptions

Assumptions: conditions that we need to be true in order for the data to properly fit the model we are using for estimations (1) Independence: observations are independent from one another (2) Randomization: proper randomization was used (takes care of independence issue if there is one) (3) Means need an *approximate* normal distribution (if $n \geq 30$, then it is approximately normal), and proportions need to meet the S/F condition: $np \geq 5$ and $nq \geq 5$ (if $n \geq 60$, S/F condition is met)

If assumptions are violated, the results from the analyses are not as valid nor reliable

General Form

All CIs (even more complex ones) have the same form:

$$point\ estimate \pm bound$$

Where the bound on the error of estimation is $z^\star(se)$ or $t^\star(se)$ and $se_{mean} = \frac{\sigma}{\sqrt{n}}$, $se_\pi = \sqrt{\hat{\pi}(1-\pi)/n}$, or $se_{mean} = \frac{s}{\sqrt{n}}$ (the one you use depends on the situation; explanations to come)

CI forms

CI on μ when σ is known

$$\bar{y} \pm z^\star \left(\frac{\sigma}{\sqrt{n}}\right)$$

CI on μ when σ is unknown

$$\bar{y} \pm t^{\star} \left(\frac{s}{\sqrt{n}} \right)$$

CI on π

$$\hat{\pi} \pm z^{\star} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Hypothesis tests

We have learned about estimating parameters by point estimation and interval estimation (specifically confidence intervals). More often than not, the objective of an investigation is not to estimate a parameter but to decide which of two (or more) contradictory claims about the parameter is correct.

This part of statistics is called *hypothesis testing*

Terms

Statistical hypotheses is a claim or assertion about

(1) The value of a single parameter
(2) The values of several parameters
(3) The form of an entire probability distribution

Hypotheses

(1) Null hypothesis, denoted by H_0, is the claim that is initially assumed to be true (the "prior belief" or "historical" claim)
(2) Alternative hypothesis, denoted by H_a, is the assertion that is contradictory to H_0; it is a researcher's claim, what they are trying to prove (thus the reason behind the study)

Hypothesis Testing Checklist

All tests include the following four steps:

(1) State hypotheses, check assumptions
(2) Calculate the test statistic
(3) Find the rejection region
(4) Results and conclusion of the test

Hypotheses

When stating the hypotheses, the notation used is always population parameter notation; inferences upon populations need population notation (the Greek letters)

μ for the mean and π for the proportion

Hypotheses for μ

Hypotheses for inferences concerning means (regardless of whether or not σ is known

$$H_0 : \mu = \mu_0 \text{ vs. } H_a : \mu \neq \mu_0$$

$$H_0 : \mu \geq \mu_0 \text{ vs. } H_a : \mu < \mu_0$$

$$H_0 : \mu \leq \mu_0 \text{ vs. } H_a : \mu > \mu_0$$

Most often the null hypothesis will have = while the alternative will be one of either $\neq$, $>$, or $<$. μ_0 is a specified value (a number that is given in the problem)

Hypotheses for π

Hypotheses for inferences concerning proportions:

$$H_0 : \pi = \pi_0 \text{ vs. } H_a : \pi \neq \pi_0$$
$$H_0 : \pi \geq \pi_0 \text{ vs. } H_a : \pi < \pi_0$$
$$H_0 : \pi \leq \pi_0 \text{ vs. } H_a : \pi > \pi_0$$

Most often the null hypothesis will have = while the alternative will be one of either $\neq$, $>$, or $<$. π_0 is a specified value (a number that is given in the problem)

Assumptions

(1) Independence: observations are independent from one another
(2) Randomization: proper randomization was used
    - Takes care of independence issue if there is one
(3) Normality
    (a) Means need an *approximate* normal distribution ($n \geq 30$ should take care of it)
    (b) Proportions need $n \geq 60$ (via CLT)

If assumptions are violated, the results from the analyses are not valid nor reliable

Test Statistic

1-sample test of the mean μ when σ is known: Use Z

$$z = \frac{\overline{y} - \mu_0}{se_{mean}} \text{ ; } se_{mean} = \frac{\sigma}{\sqrt{n}}$$

1-sample test of the proportion p (most often a χ^2 test is done in practice): Use Z

$$z = \frac{\hat{\pi} - \pi_0}{se_\pi} \text{ ; } se_\pi = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

1-sample test of the mean μ when σ is unknown: Use t

$$t = \frac{\overline{y} - \mu_0}{se_{mean}} \text{ ; } se_{mean} = \frac{s}{\sqrt{n}}$$

Rejection Region

Is based on significance level α. $\alpha = 1 - CL$ where CL is the confidence level

Always assume $\alpha = 0.05$ unless specified otherwise)

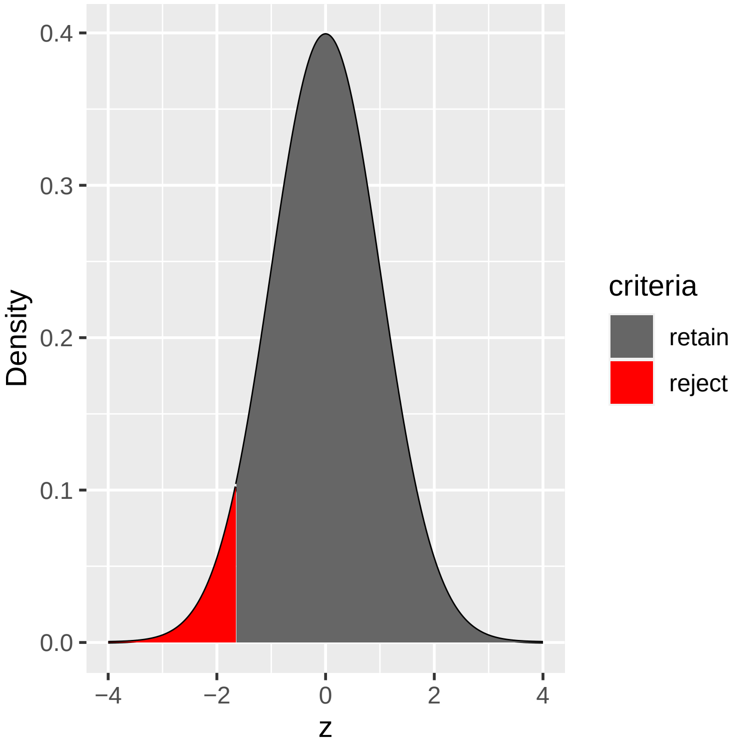
Two methods for rejection:

(1) Critical value approach (not used in this course, only for review)
(2) *pvalue* approach (will be used in this course, including a review)

The alternative hypothesis (H_a) determines rejection based on where you are at on the curve

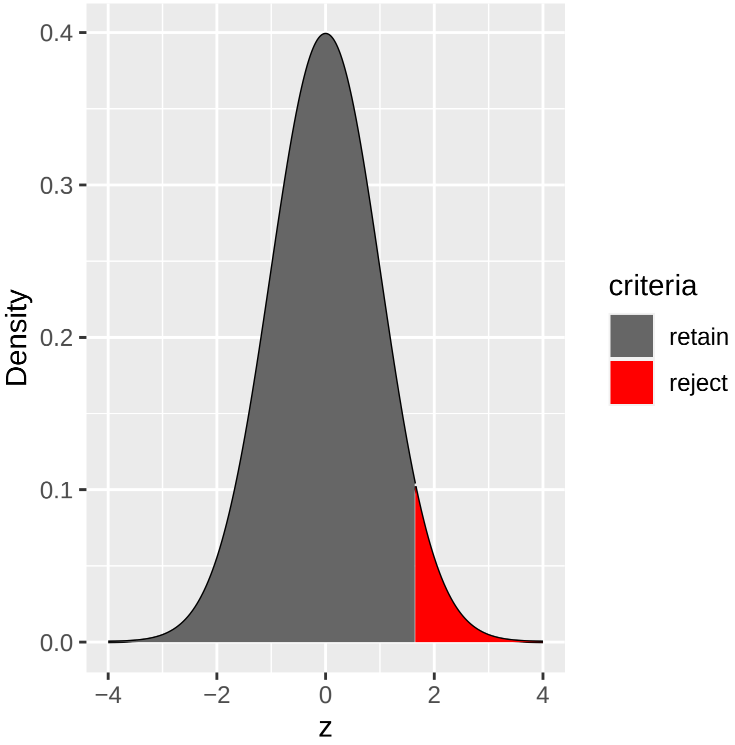
Critical Value Approach $H_a :<$

Reject H_0 iff (if and only if) $z_{calc} \leq z_\alpha$ (z_{calc} will most likely be a negative value and z_α *must* be negative)



Critical Value Approach $H_a :>$

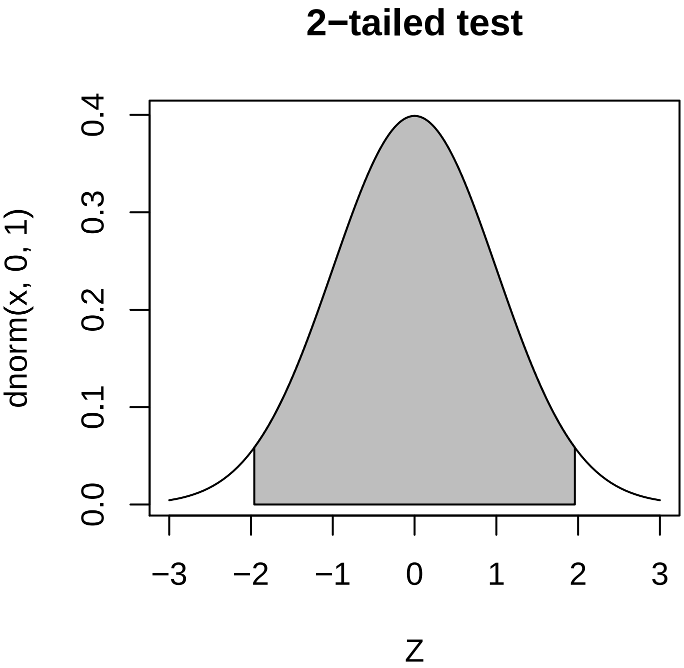
Reject H_0 iff $z_{calc} \geq z_\alpha$ (z_{calc} will most likely be a positive value and z_α *must* be positive)



Critical Value Approach $H_a :\neq$

Reject H_0 iff $|z_{calc}| \geq |z_{\alpha/2}|$ (z_{calc} and z_α can both be either positive or negative, but we will deal with absolute values)

(white area is the rejection region, and yes there are two area here that are *both* the rejection region)



Results and Conclusion

- Results: we either
  - Reject H_0 (rejecting the null hypothesis in favor of the alternative)
  - Fail to reject H_0 (we are not rejecting the null hypothesis so that means that the null hypothesis gives a reasonable explanation of the question at hand) Conclusion: explain what the results did in relation to the actual data

pvalue logistics I

The *pvalue* of a test is the probability that, *given* the null hypothesis (H_0) is true, the results from another random sample will be as or more extreme as the results we observed from our sample.

The *pvalue* of the test is dependent on the type of test you are doing, as in one-tail upper, one-tail lower, or two-tail. The sign of the alternative hypothesis is the determining factor in calculation of the *pvalue*.

pvalue logistics II

The pvalue approach; the null hypothesis can be rejected iff (if and only if) $pvalue \leq \alpha$ (with $\alpha = 0.05$ most often). This does not change, regardless of the sign of the alternative hypothesis. However, the calculation of the *pvalue* is dependent on the sign of the alternative hypothesis. The *pvalue* will be the $P($ the results of the test $|H_0$ is correct), in other words, it is the probability that the results would occur by random chance if the null hypothesis is actually correct.

Assume that $\alpha = 0.05$ unless specified; any rejection of H_0 means that the results (of experiment, survey, etc.) are significant.

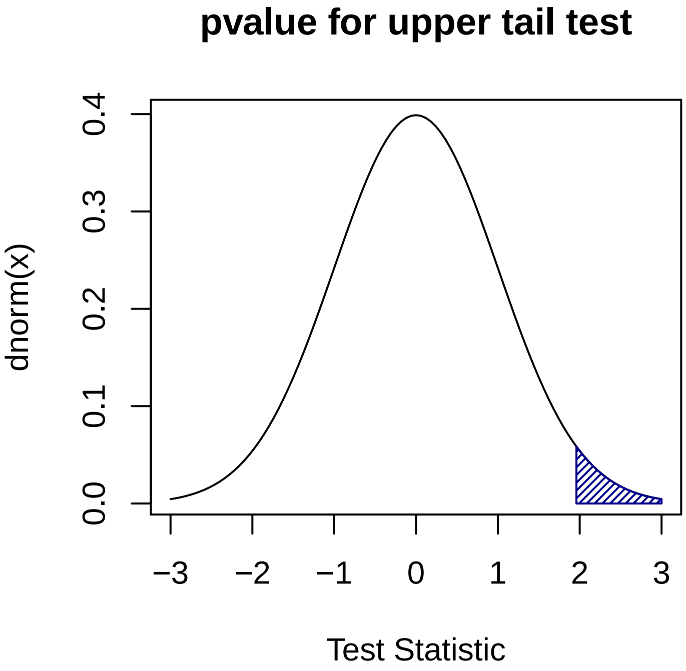
$$pvalue \leq \alpha \Rightarrow Reject\ H_0$$

$H_a :\ >$ **upper tail test**

Note that while all examples are with z, it is interchangeable with t (df is needed)

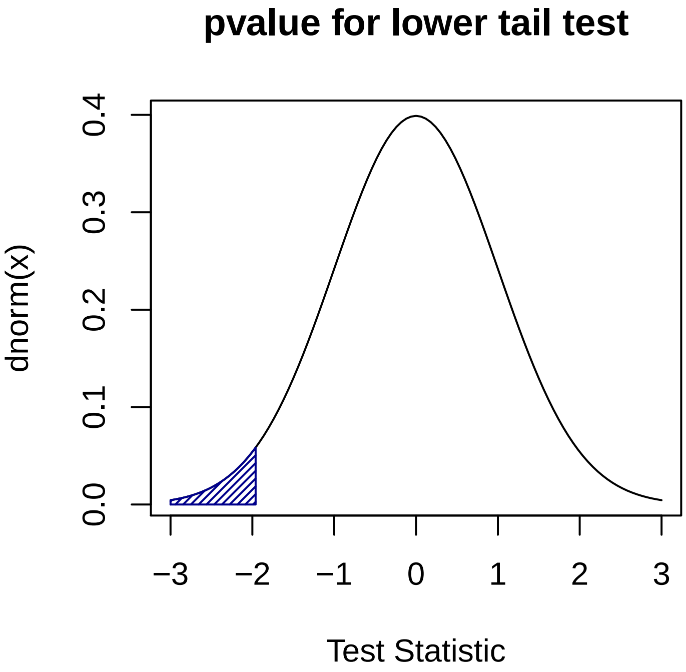
In this case, *pvalue* represents the rejection region in the right tail of the distribution.

$$pvalue = P(Z \geq z_{calc}) = 1 - P(Z \leq z_{calc})$$



$H_a :\ <$ **lower tail test**

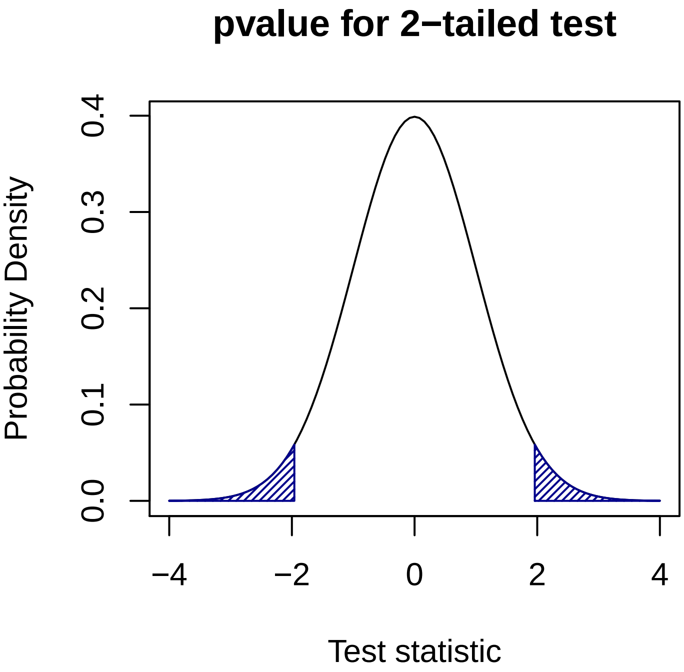
$$pvalue = P(Z \leq z_{calc})$$



$H_a: \;\; \neq$ **two tail test**

$$pvalue = 2[P(Z \leq z_{calc})] \; or \; 2[1 - P(Z \leq z_{calc})]$$

$$= 2[1 - P(Z \leq |z_{calc}|)]$$



pvalue rejection Examples

(1) $pvalue = 0.4$ with $\alpha = 0.05$. Since $pvalue = 0.4 \nleq \alpha(0.05)$, H_0 is not rejected (fail to reject H_0). There is a 40% chance that we would see these results due to random chance (dumb luck) if the null hypothesis is correct; results are not significant.

(2) $pvalue = 0.04$ with $\alpha = 0.05$. Since $pvalue = 0.04 \leq \alpha(0.05)$, H_0 is rejected. There is a 4% chance that we would see these results due to random chance (dumb luck) if the null hypothesis is correct; results are significant.

(3) $pvalue = 0.04$ with $\alpha = 0.01$. Since $pvalue = 0.04 \nleq \alpha(0.01)$, H_0 is not rejected. There is a 4% chance that we would see these results due to random chance (dumb luck) if the null hypothesis is correct; results are not significant.

CIs and tests in SAS

The point of this course is to learn SAS so we will use it for our tests and CIs. The previous slides are for review purposes only and the following will be how we get these done with SAS.

In reality, z tests are not used often, and almost never with CIs and tests for one or more parameters; all examples will only use t-tests.

Tests and CIs for one-sample

General form of PROC TTEST:

```
PROC TTEST data=SASdataset <options>;
CLASS variable;
VAR variable(s);
PAIRED variables;
...;
RUN;
```

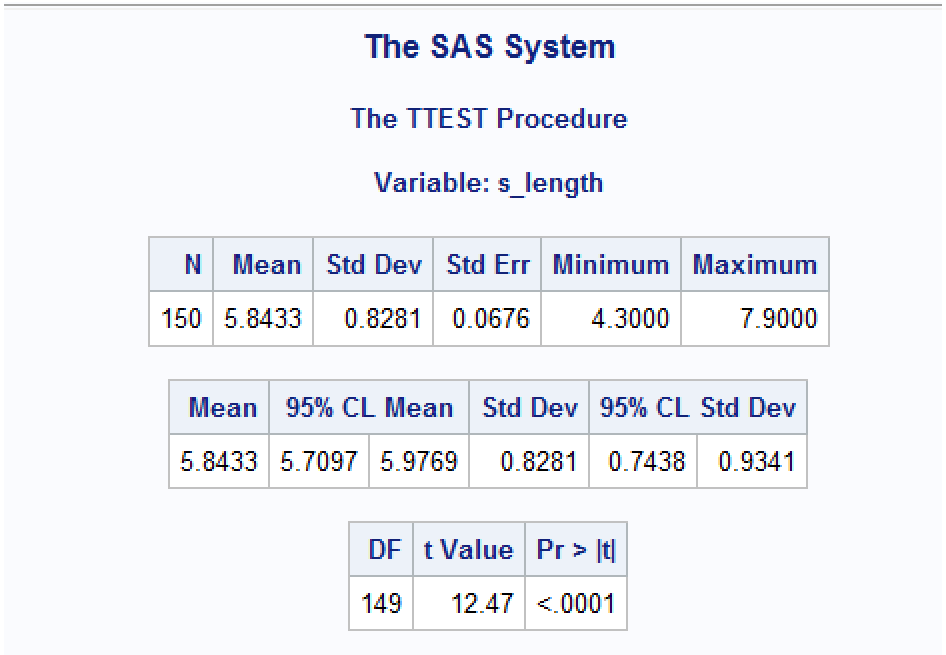
Options for PROC TTEST statement: `DATA=`: SAS-dataset
`ALPHA=`: specifies the significance level
`H0=`: specifies the null value (μ_0, π_0, μ_D, or Δ_0,. . .)
`SIDES=`: specifies one or two-tailed test; 2, U, L
`CI=`: requests confidence intervals for the standard deviation or the coefficient of variation
`PLOTS`: produces statistical graphs (histograms and QQ plot)

A "normal" CI is equivalent to a 2-tailed test. The CIs produced with either a lower- or upper-tail test, respectively, will have intervals that look like $(-\infty, upper)$ or $(lower, \infty)$

$H_a: \quad \neq$

```
* test H0: mu=5 Ha: mu not= 5;
proc ttest data=flower h0=5;
var s_length;
run;
```

2T ttest1



The SAS System

The TTEST Procedure

Variable: s_length

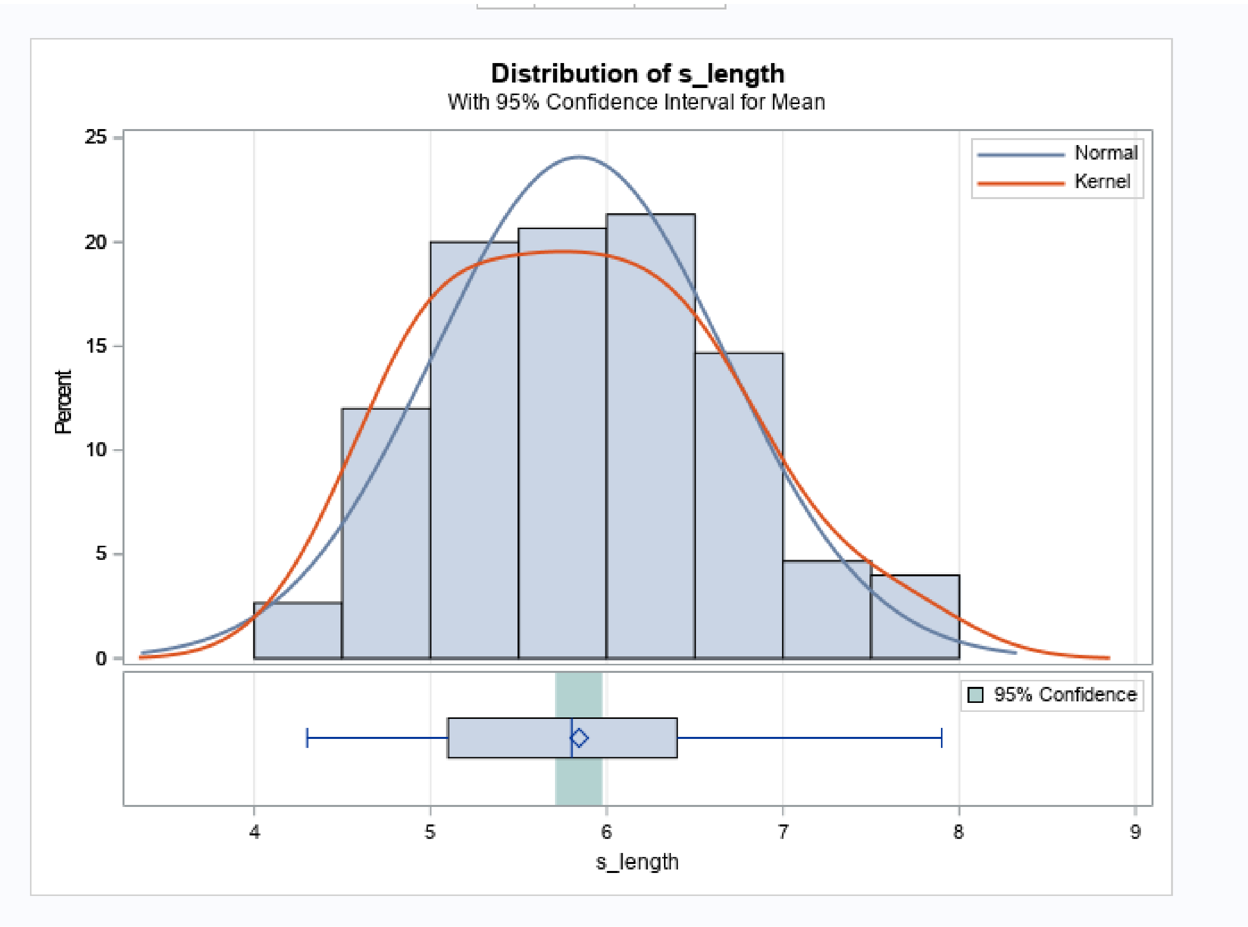
N	Mean	Std Dev	Std Err	Minimum	Maximum
150	5.8433	0.8281	0.0676	4.3000	7.9000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
5.8433	5.7097	5.9769	0.8281	0.7438	0.9341

DF	t Value	Pr > |t|
149	12.47	<.0001

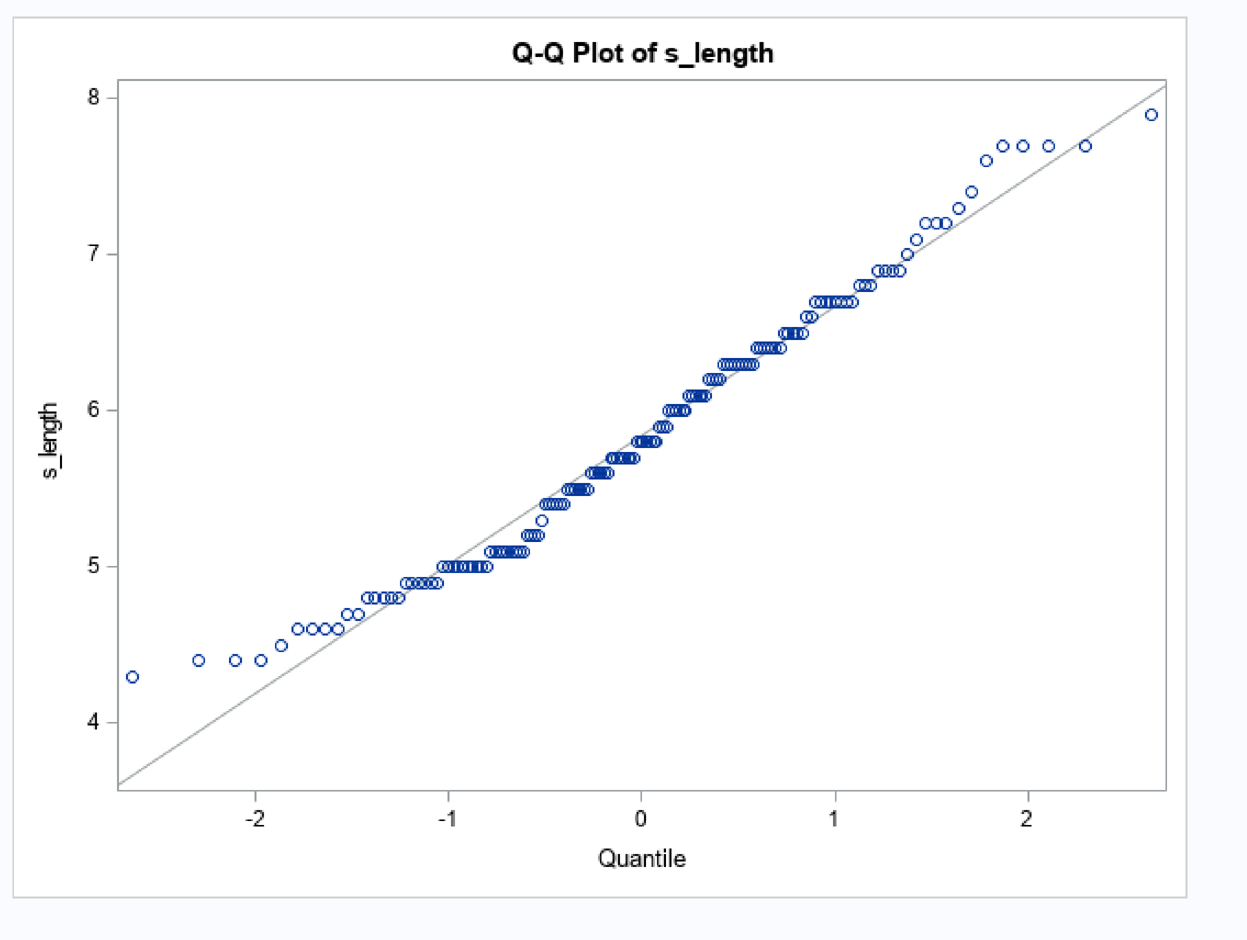
ttest1.png

2T ttest2



ttest2.png

2T ttest3

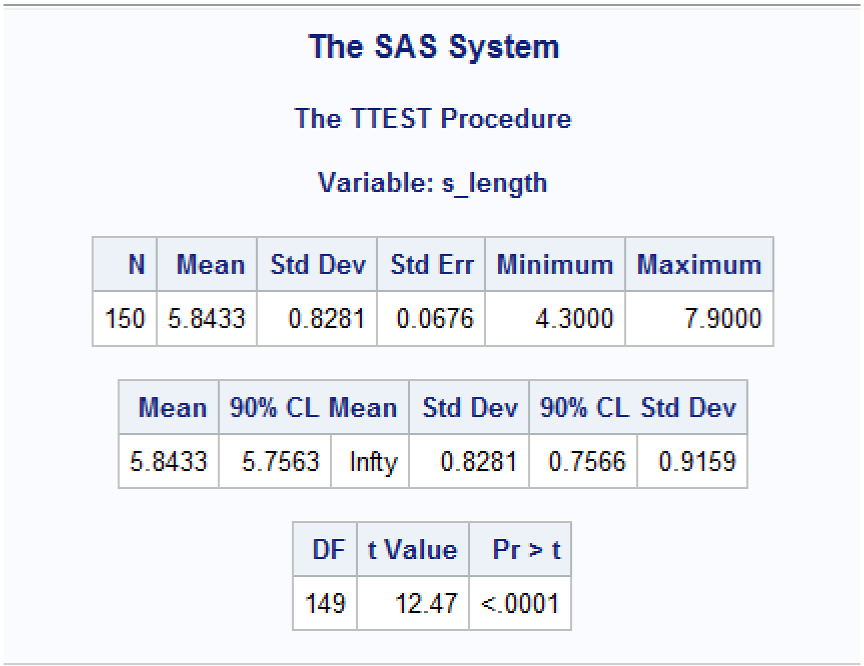


ttest3.png

$H_a : \quad >$

```
* test H0: mu=5 Ha: mu > 5 with alpha=10%;
proc ttest h0=5 data=flower alpha=.1 sides=U;
var s_length;
run;
```

UT ttest1



The SAS System

The TTEST Procedure

Variable: s_length

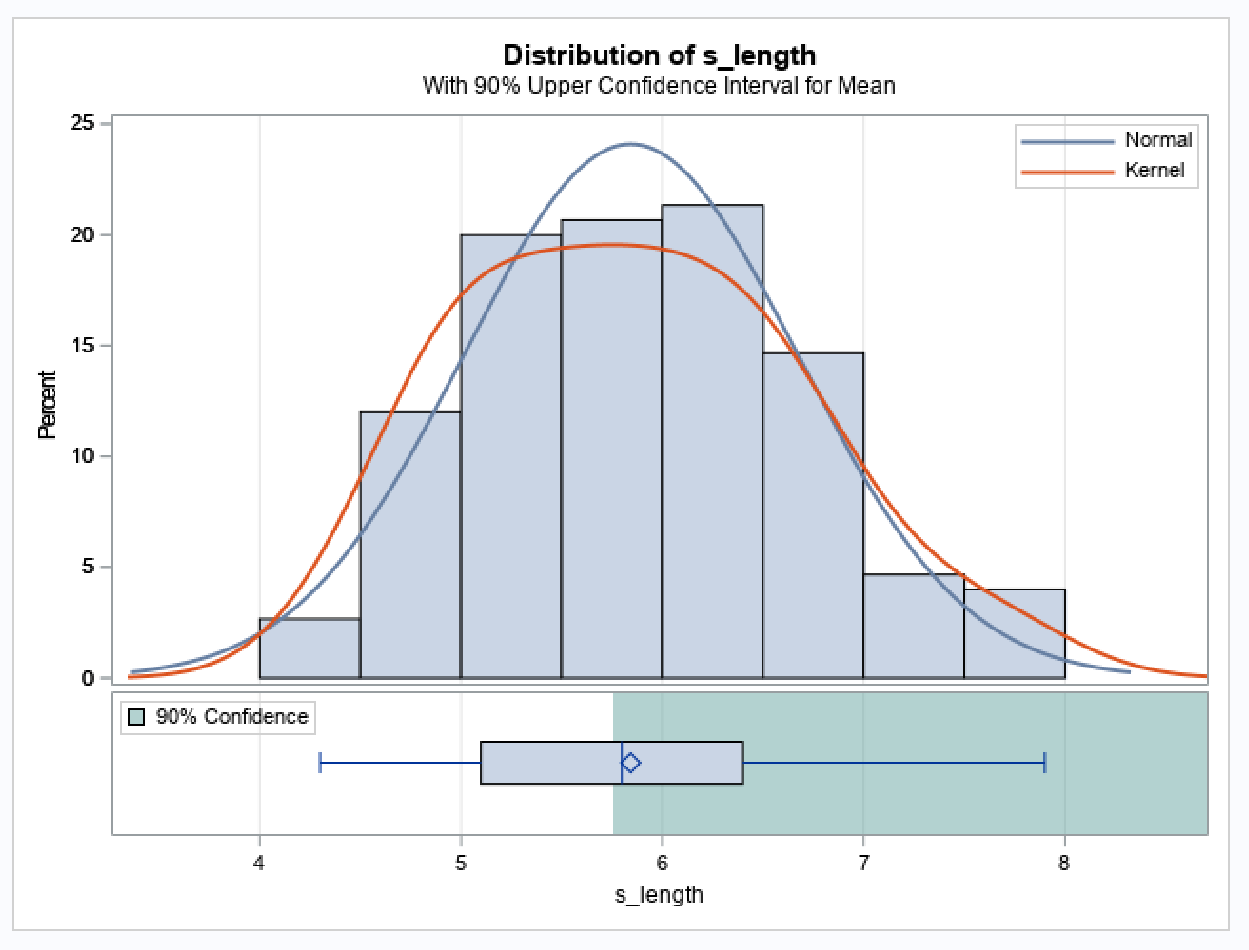
N	Mean	Std Dev	Std Err	Minimum	Maximum
150	5.8433	0.8281	0.0676	4.3000	7.9000

Mean	90% CL Mean		Std Dev	90% CL Std Dev	
5.8433	5.7563	Infty	0.8281	0.7566	0.9159

DF	t Value	Pr > t
149	12.47	<.0001

ttest1.png

UT ttest2

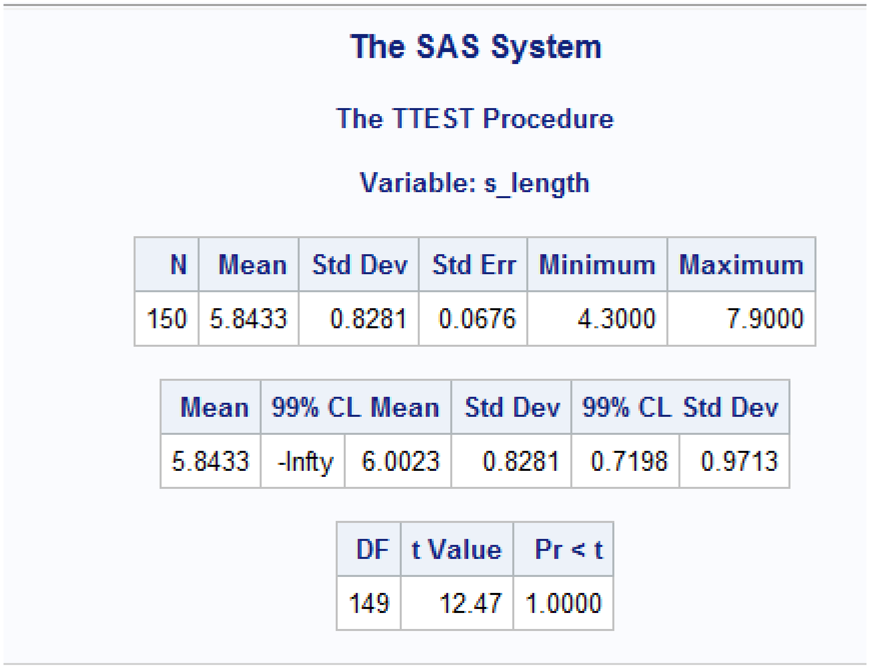


ttest2.png

$H_a: \quad <$

```
* test H0: mu=5 Ha: mu < 5 with alpha=1%;
proc ttest h0=5 data=flower alpha=.01 sides=L;
var s_length;
run;
```

LT ttest1



The SAS System

The TTEST Procedure

Variable: s_length

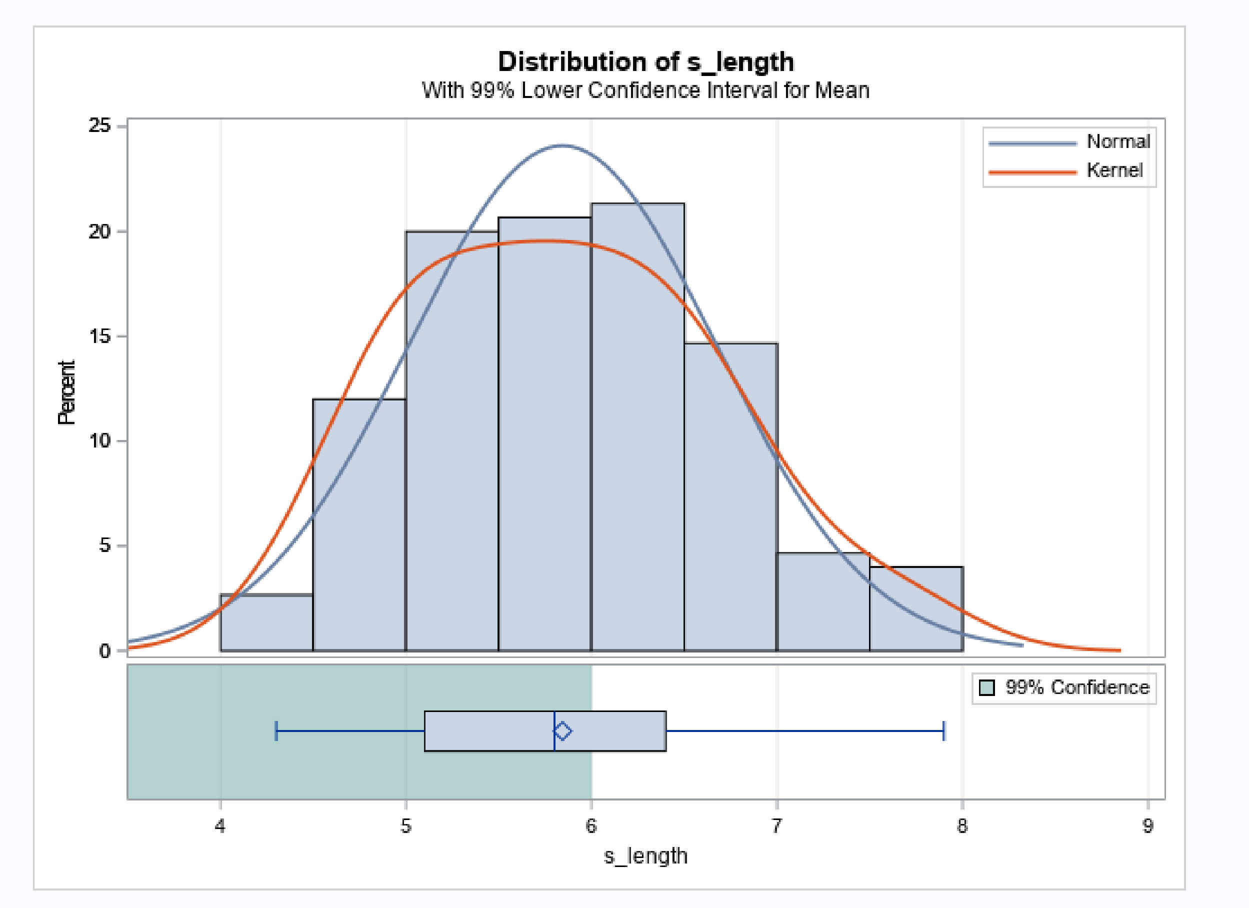
N	Mean	Std Dev	Std Err	Minimum	Maximum
150	5.8433	0.8281	0.0676	4.3000	7.9000

Mean	99% CL Mean		Std Dev	99% CL Std Dev	
5.8433	-Infty	6.0023	0.8281	0.7198	0.9713

DF	t Value	Pr < t
149	12.47	1.0000

ttest1.png

LT ttest2



ttest2.png

Comparing two groups

Comparisons:

(1) Two independent means
  (a) When $\sigma_1^2 \approx \sigma_2^2$: Pooled
  (b) When $\sigma_1^2 \neq \sigma_2^2$: Unpooled (also called a Satterthwaite test in SAS)
(2) Dependent means (mean difference)
(3) Two independent proportions (a χ^2 test is done in practice)

Independent means

This compares the means of two distinct (separate) groups of units or subjects. The wording used is **the difference of two (independent) means**

There are two cases for this (when variances are equal or unequal).

pooled: for when variances are equal ($\sigma_1^2 \approx \sigma_2^2$)
unpooled: for when variances are unequal ($\sigma_1^2 \neq \sigma_2^2$)

The concept of pooled vs. unpooled refers to the standard error and degrees of freedom for the differences of two independent means (the *se*)

Relationship of variances

$\sigma_1^2 \approx \sigma_2^2$ or $\sigma_1^2 \neq \sigma_2^2$?

In order to determine if the variances are equal or not, a variance test needs to be performed first. The "answer" to the test will indicate which method, pooled or unpooled, is most appropriate for the data.

SAS so kindly executes a variance test automatically when performing either of the 2-sample methods (pooled or unpooled). Pooled method output will be labeled as "pooled" and the unpooled method output will bne labeled as "Satterthwaite" (called other names in other programs)

Variance test hypotheses

The hypotheses for this test are (always)

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

Technically, it is NOT a 2-tailed test that is executed in most programs; it is an upper tail test ($H_a : \sigma_1^2 > \sigma_2^2$). A variation on that is to use a ratio of the variances. Divide both sides of the hypotheses equations by σ_2^2. If the two variances are equal (approximately), then the ratio should be close to one, if they are unequal, their ratio will be greater than 1.

$$H_0 : \frac{\sigma_{MAX}^2}{\sigma_{MIN}^2} = 1 \ vs. \ H_a : \frac{\sigma_{MAX}^2}{\sigma_{MIN}^2} > 1$$

Test statistic and pvalue for variance test

The test statistic is an F statistic, commonly used for analysis of variance ($ANOVA$). The F statistic for the variance test is

$$F = \frac{s_1^2}{s_2^2}$$

Then a *pvalue* is calculated as $P(F > F_{calc})$. The reason is it a right tail test is that you can never calculate a negative F statistic because variances can never be negative. Additionally, the F distribution is not a symmetric distribution, but a right skewed distribution.

Variance test pvalue and conclusions

Once you have the *pvalue*

$$Reject\ H_0\ iff\ pvalue \leq \alpha$$

The significance level α is always assumed to be $\alpha = 0.05$ unless specified otherwise.

If H_0 *is rejected*, the variances are not equal and the unpooled (Satterthwaite) method is most appropriate for the data

If H_0 *is not rejected*, the variances are equal (approximately, they do not have to be exactly equal) and the pooled method is most appropriate for the data

Pooled method *se* and *df*

Degrees of freedom for independent means for pooled method, when variances are equal is

$$df = n_1 + n_2 - 2$$

And the standard error for the pooled method is

$$se = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

s_p^2 is called the pooled variance, calculated by

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

Unpooled method *se* and *df*

Degrees of freedom for independent means (unpooled, when variances are unequal) is calculated rather than using $n - 1$ or something similar, and SAS will calculate it for you:

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

And the standard error for the unpooled method is

$$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Formula: CI

CI for the difference of two (independent) means:

$$\overline{y}_1 - \overline{y}_2 \pm t^\star(se)\ where\ t^\star = t_{2T,df}$$

Hypotheses

For the difference of two (independent) means:

$$H_0 : \mu_1 - \mu_2 = \Delta_0 \quad H_a : \mu_1 - \mu_2 \begin{pmatrix} \neq \\ > \\ < \end{pmatrix} \Delta_0$$

Δ_0 is a specified (numerical) value of the hypothesized difference of two independent means.

Assumptions

(1) Independence (if random met, this is met)
(2) Randomization
(3) Each group of observations have an approximate normal distribution

Formula: Test Statistic

$$t = \frac{(\overline{y}_1 - \overline{y}_2) - \Delta_0}{se}$$

se and *df* are dependent on the outcome of the variance test

Dependent means

This compares the mean of the difference between two measurements of the same unit or subject. The wording used is **the mean difference**. This analysis is for comparing measurements on the same subject/unit; once before a treatment and once again after the treatment, to detect if there is a difference due to the treatment.

Examples are weight loss programs, Coke vs. Pepsi, compare GDP of countries at 2 different dates (time is treatment)

Dependent means logistics

The two variables of data need to be subtracted from each other ($before - after$ or $after - before$) to calculate all of the differences between measurements.

d_i: individual differences between measurements

$\overline{y}_d = \frac{\sum d_i}{n}$ sample mean difference (mean of the differences)
$s_d = \sqrt{\frac{\sum (d_i - \overline{y}_d)^2}{n-1}}$: sample standard deviation of the differences

Formula: CI

CI for the mean difference:

$$\overline{y}_d \pm t^\star(se) \ \ where \ se = \frac{s_d}{\sqrt{n}} \ \ and \ \ t^\star = t_{2T,df}, \ df = n - 1$$

Hypotheses

For the mean difference

$$H_0 : \mu_d = \Delta_0 \quad H_a : \mu_d \begin{pmatrix} \neq \\ > \\ < \end{pmatrix} \Delta_0$$

Assumptions

(1) Dependence (two measurements per unit/subject)
(2) Randomization
(3) Differences have approximate normal distribution

Formula: Test Statistic

$$t = \frac{\overline{y}_d - \Delta_0}{se} \; and \; se = \frac{s_d}{\sqrt{n}}$$

Two Proportions

This compares the proportions of two distinct (separate) groups of units or subjects. The wording used is **the difference of two (2) proportions**

The se for the test is different from the se for the CI

Formula: CI

CI for the difference of two (independent) proportions:

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z^\star(se) \text{ where } se = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} \text{ and } z^\star = z_{\alpha/2}$$

Hypotheses

For the difference of two (independent) proportions:

$$H_0 : \pi_1 - \pi_2 = \Delta_0 \quad H_a : \pi_1 - \pi_2 \begin{pmatrix} \neq \\ > \\ < \end{pmatrix} \Delta_0$$

If $\Delta_0 = 0$, then the following will work for the hypotheses:

$$H_0 : \pi_1 = \pi_2 \quad H_a : \pi_1 \begin{pmatrix} \neq \\ > \\ < \end{pmatrix} \pi_2$$

Assumptions

(1) Independent groups (if random met, this is met)
(2) Randomization
(3) success/failure condition to have normality
    (a) **either** $n_1 \geq 60$ **AND** $n_2 \geq 60$ or
    (b) $n_1\pi_1 \geq 5$, $n_1(1-\pi_1) \geq 5$, $n_2\pi_2 \geq 5$, **AND** $n_2(1-\pi_2) \geq 5$[1]

Formula: Test Statistic

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{se} \text{ where } se = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Where $\hat{\pi}$ without subscripts is the pooled proportion used when assuming the difference of proportions is equal to 0.

[1] Annoyingly, depending on the textbook, the values could be 5, 8, 10, 12, or 15... but 5 is good :-)

$$\hat{\pi} = \frac{X_1 + X_2}{n_1 + n_2}$$

X_1, X_2 are the successes from each group. If you are given percents, then successes are calculated by:

$$X_1 = n_1\hat{\pi}_1 \quad X_2 = n_2\hat{\pi}_2$$

Independent means: pooled example

A study of 584 longleaf pine trees in the Wade Tract in Thomas County, Georgia had several purposes. To see if there is a difference in their sizes (in diameters) in two separate areas of the Wade Tract (northern and southern areas), a random sample of 30 trees from the northern area and 30 trees from the southern area was taken. (a) Estimate the true difference in mean tree sizes between the northern and southern parts of the Wade Tract with 95% confidence (b) Is there a significant difference in the mean diameter of trees in the north versus the trees in the south? Conduct hypothesis test(s)

Independent means: pooled example

(1)

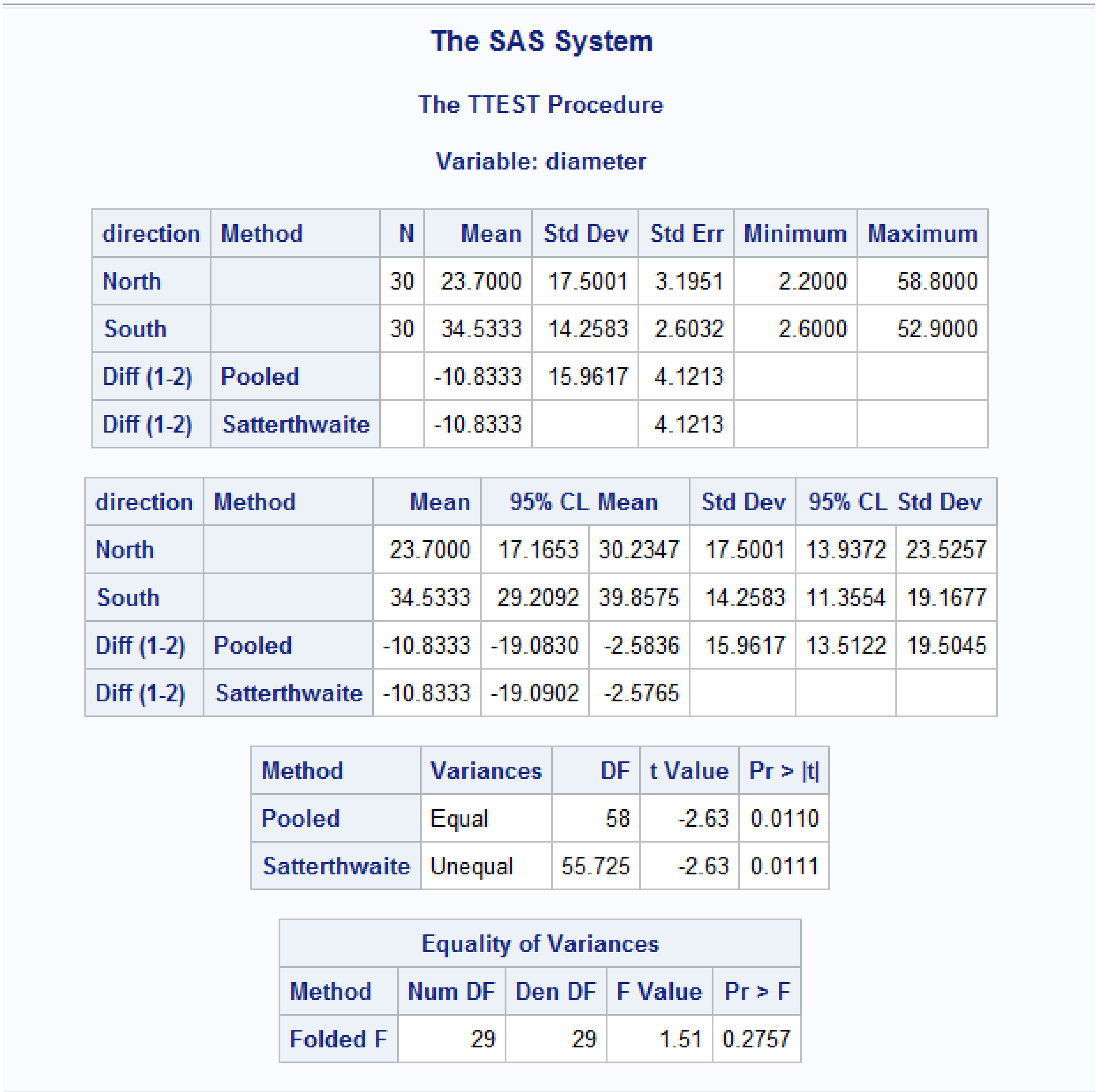
$$H_0 : \sigma_1^2 = \sigma_2^2 \quad H_a : \sigma_1^2 > \sigma_2^2$$

(2)

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

```
proc ttest data=tree;
class direction;
var diameter;
run;
```

pooled1



The SAS System

The TTEST Procedure

Variable: diameter

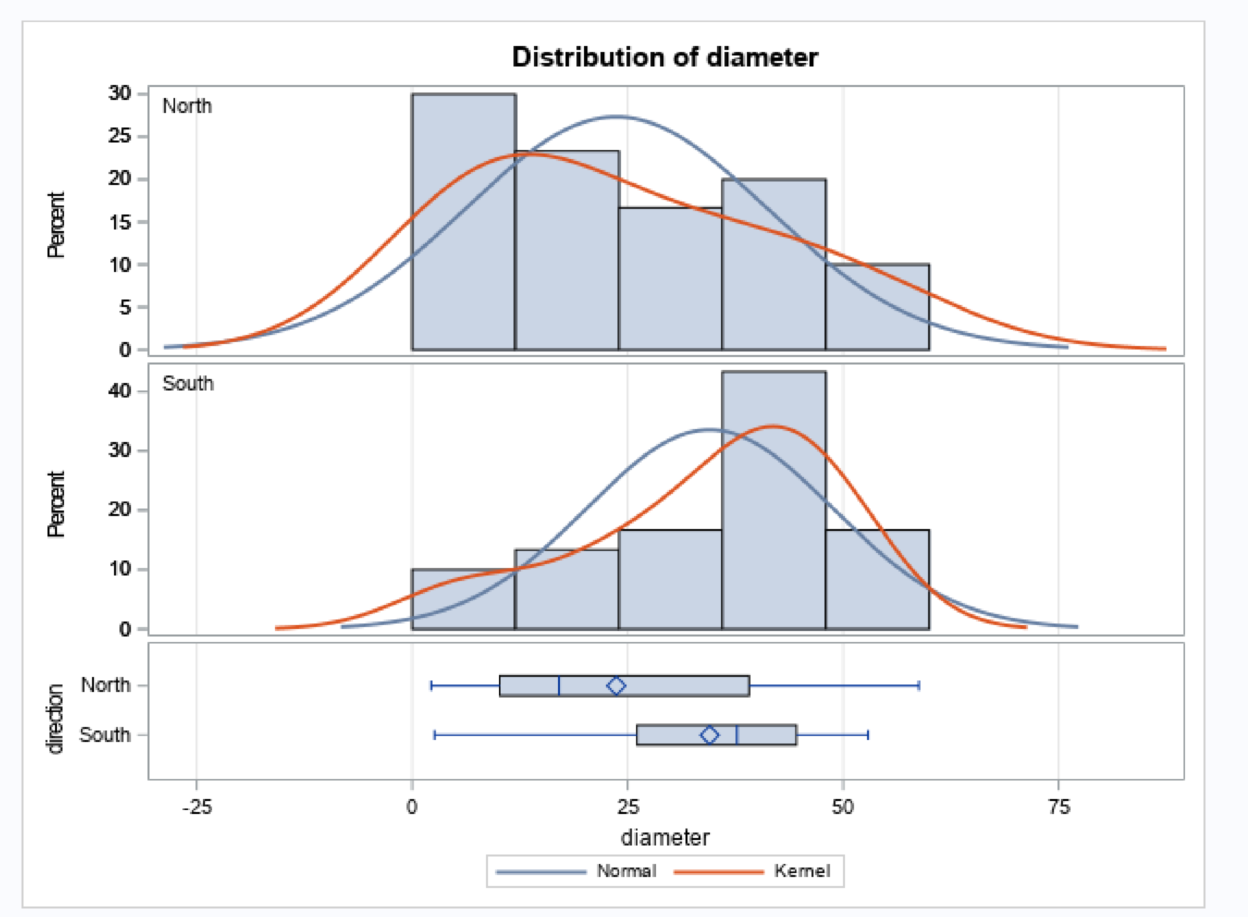
direction	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
North		30	23.7000	17.5001	3.1951	2.2000	58.8000
South		30	34.5333	14.2583	2.6032	2.6000	52.9000
Diff (1-2)	Pooled		-10.8333	15.9617	4.1213		
Diff (1-2)	Satterthwaite		-10.8333		4.1213		

direction	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
North		23.7000	17.1653	30.2347	17.5001	13.9372	23.5257
South		34.5333	29.2092	39.8575	14.2583	11.3554	19.1677
Diff (1-2)	Pooled	-10.8333	-19.0830	-2.5836	15.9617	13.5122	19.5045
Diff (1-2)	Satterthwaite	-10.8333	-19.0902	-2.5765			

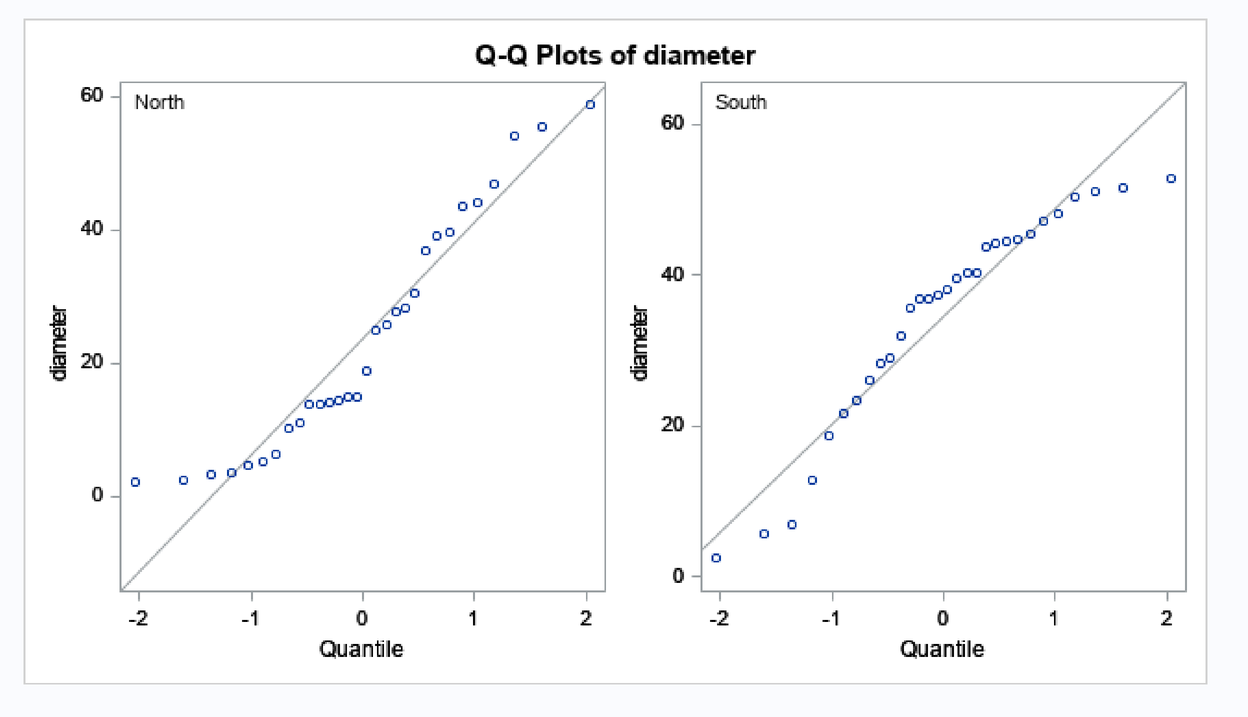
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	58	-2.63	0.0110
Satterthwaite	Unequal	55.725	-2.63	0.0111

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	29	1.51	0.2757

pooled2


Distribution of diameter
North
South
Percent
Percent
0
5
10
15
20
25
30
0
10
20
30
40
direction
North
South
-25
0
25
50
75
diameter
Normal
Kernel

pooled3



Independent means: Satterthwaite (unpooled) example

For this problem, we want to compare the average weights of blue crabs in the two river basins of the Tar-Pamlico and Neuse Rivers. Based on the health of the two rivers, it is thought the crabs in the Neuse will be larger, on average, and will test for this effect. Random samples of 100 blue crabs in each basin were taken. Are blue crabs from the Neuse river significantly larger than the blue crabs from the Tar-Pamlico river? Conduct hypothesis test(s)

Independent means: Satterthwaite (unpooled) example

(1)

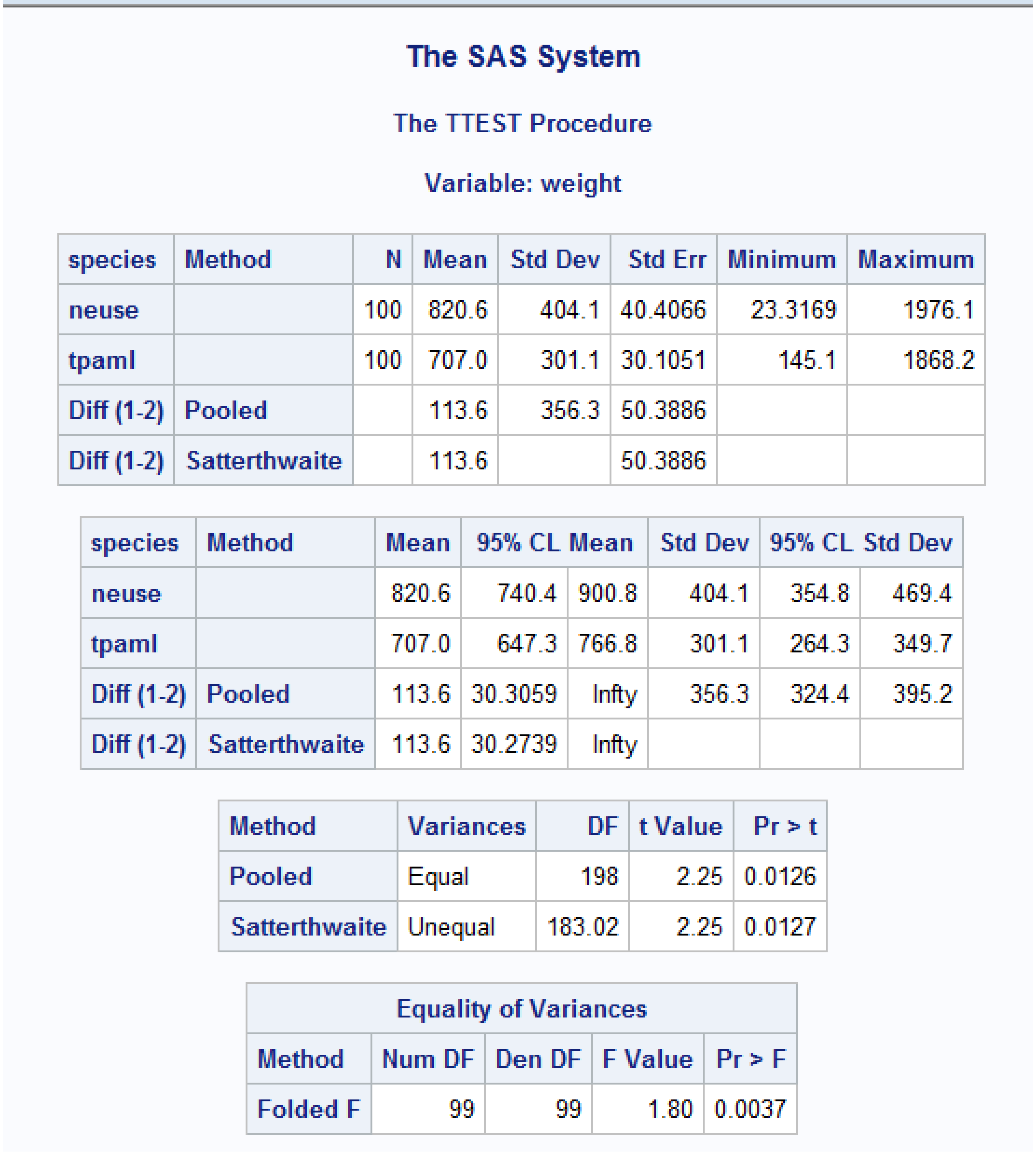
$$H_0 : \sigma_1^2 = \sigma_2^2 \quad H_a : \sigma_1^2 > \sigma_2^2$$

(2)

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 > 0$$

```
proc ttest data=krabs sides=U;
class species;
var weight;
run;
```

unpooled1



The SAS System

The TTEST Procedure

Variable: weight

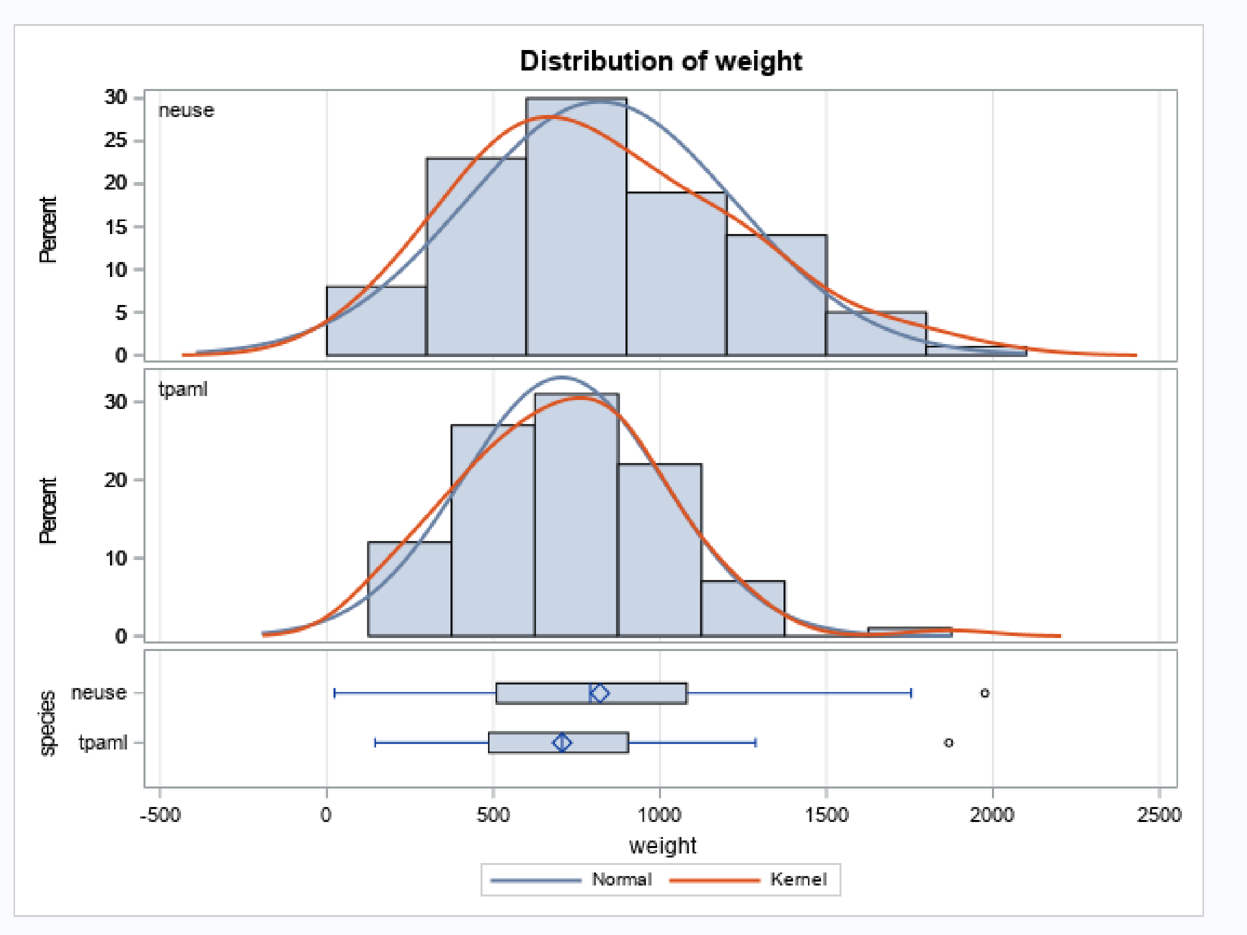
species	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
neuse		100	820.6	404.1	40.4066	23.3169	1976.1
tpaml		100	707.0	301.1	30.1051	145.1	1868.2
Diff (1-2)	Pooled		113.6	356.3	50.3886		
Diff (1-2)	Satterthwaite		113.6		50.3886		

species	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
neuse		820.6	740.4	900.8	404.1	354.8	469.4
tpaml		707.0	647.3	766.8	301.1	264.3	349.7
Diff (1-2)	Pooled	113.6	30.3059	Infty	356.3	324.4	395.2
Diff (1-2)	Satterthwaite	113.6	30.2739	Infty			

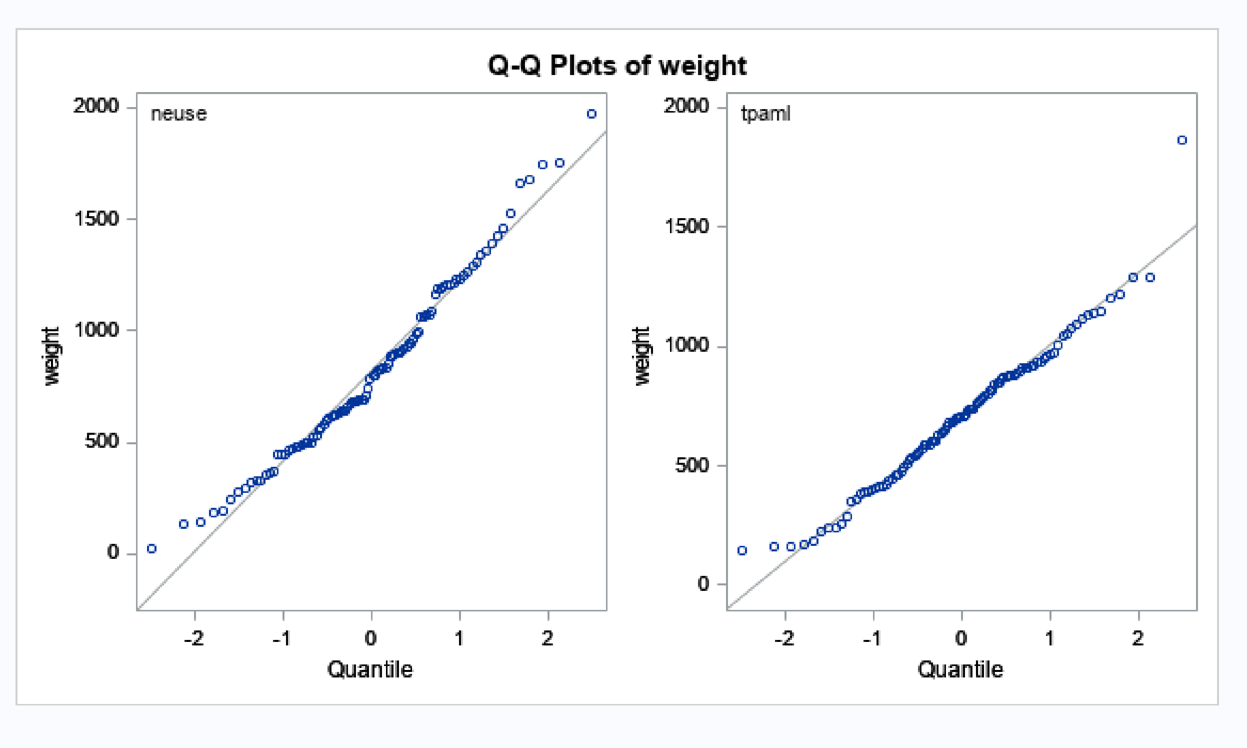
Method	Variances	DF	t Value	Pr > t
Pooled	Equal	198	2.25	0.0126
Satterthwaite	Unequal	183.02	2.25	0.0127

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	99	99	1.80	0.0037

unpooled2


Distribution of weight
neuse
tpaml
Percent
Percent
30
25
20
15
10
5
0
30
20
10
0
species
neuse
tpaml
-500
0
500
1000
1500
2000
2500
weight
Normal
Kernel

unpooled3



Paired (dependent) example

The data are the results to test the effect of a physical fitness course on one's physical ability. A random sample was taken and the number of sit-ups that a person could do in 1 minute, both before and after the fitness course was recorded. Did a significant amount of improvement took place?

$$H_0 : \mu_d = 0 \quad H_a : \mu_d > 0$$

Paired (dependent) example

```
data fitness;
input before after @@;
cards;
29 30 22 26 25 25 29 35 26 33
24 36 31 32 46 54 34 50 28 43
;
run;
proc print data=fitness;
run;

proc ttest data=fitness;
paired after*before;
run;
```

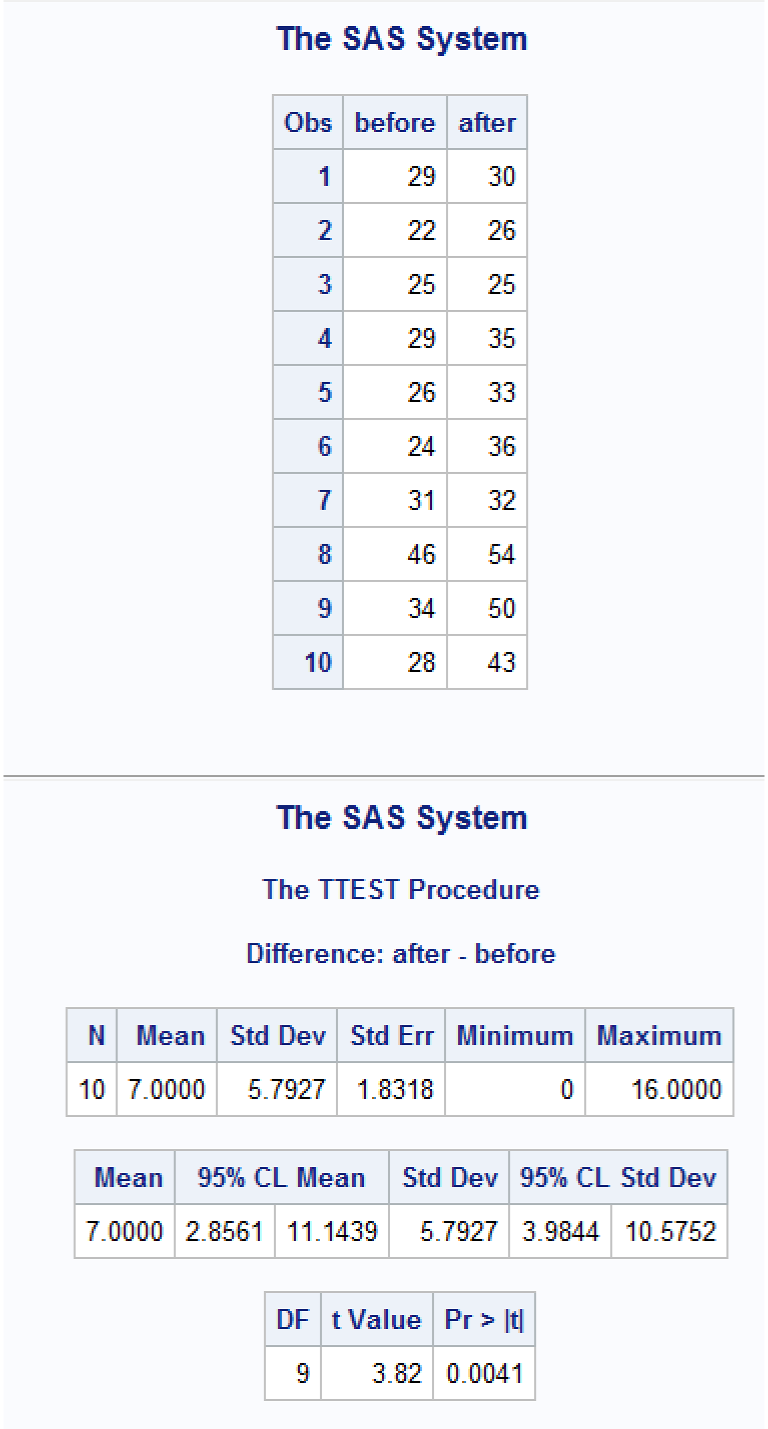
Alternative paired

```
data fitness2;
set fitness;
```

```
diff=after-before;
run;
proc print data=fitness2;
run;

proc ttest data=fitness2;
var diff;
run;
```

paired1



The SAS System

Obs	before	after
1	29	30
2	22	26
3	25	25
4	29	35
5	26	33
6	24	36
7	31	32
8	46	54
9	34	50
10	28	43

The SAS System

The TTEST Procedure

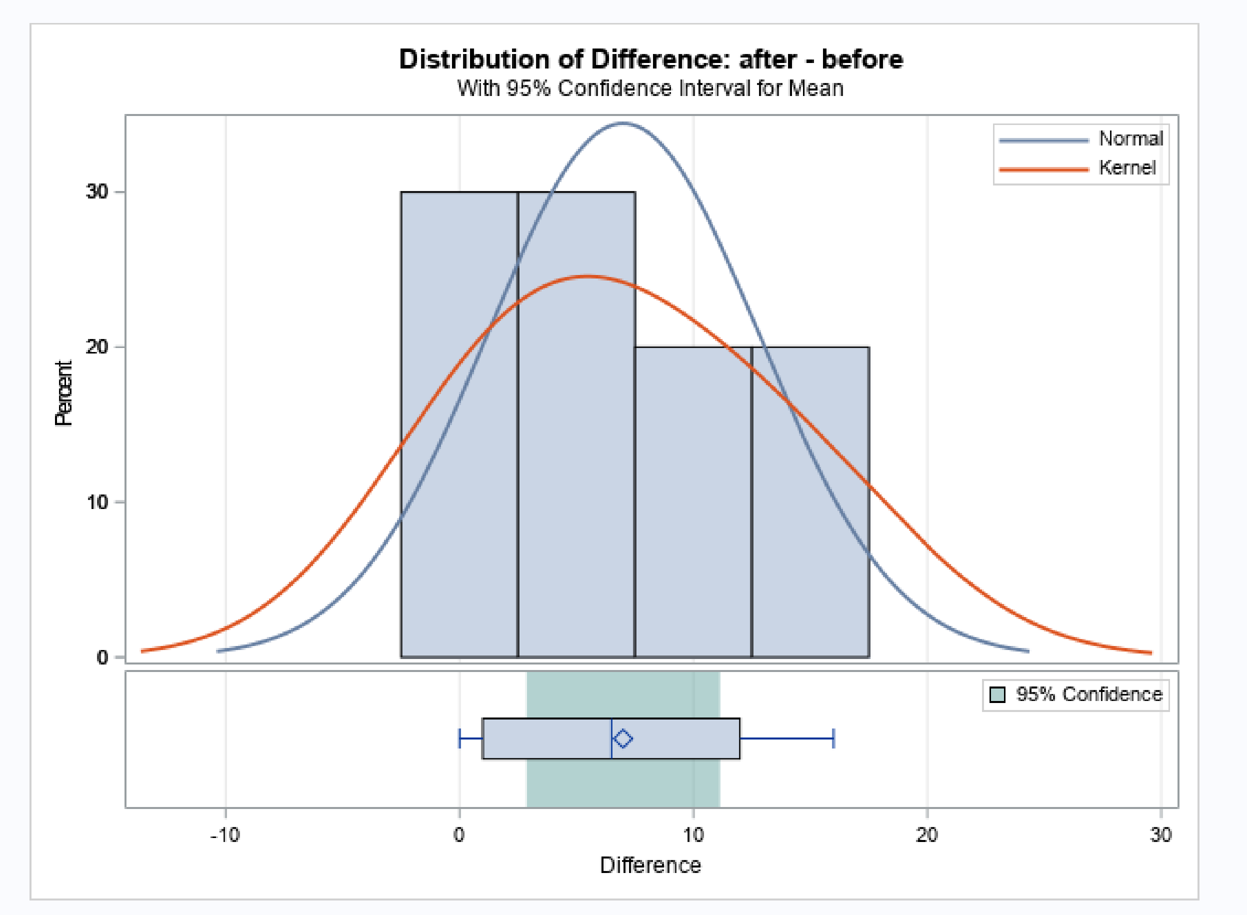
Difference: after - before

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	7.0000	5.7927	1.8318	0	16.0000

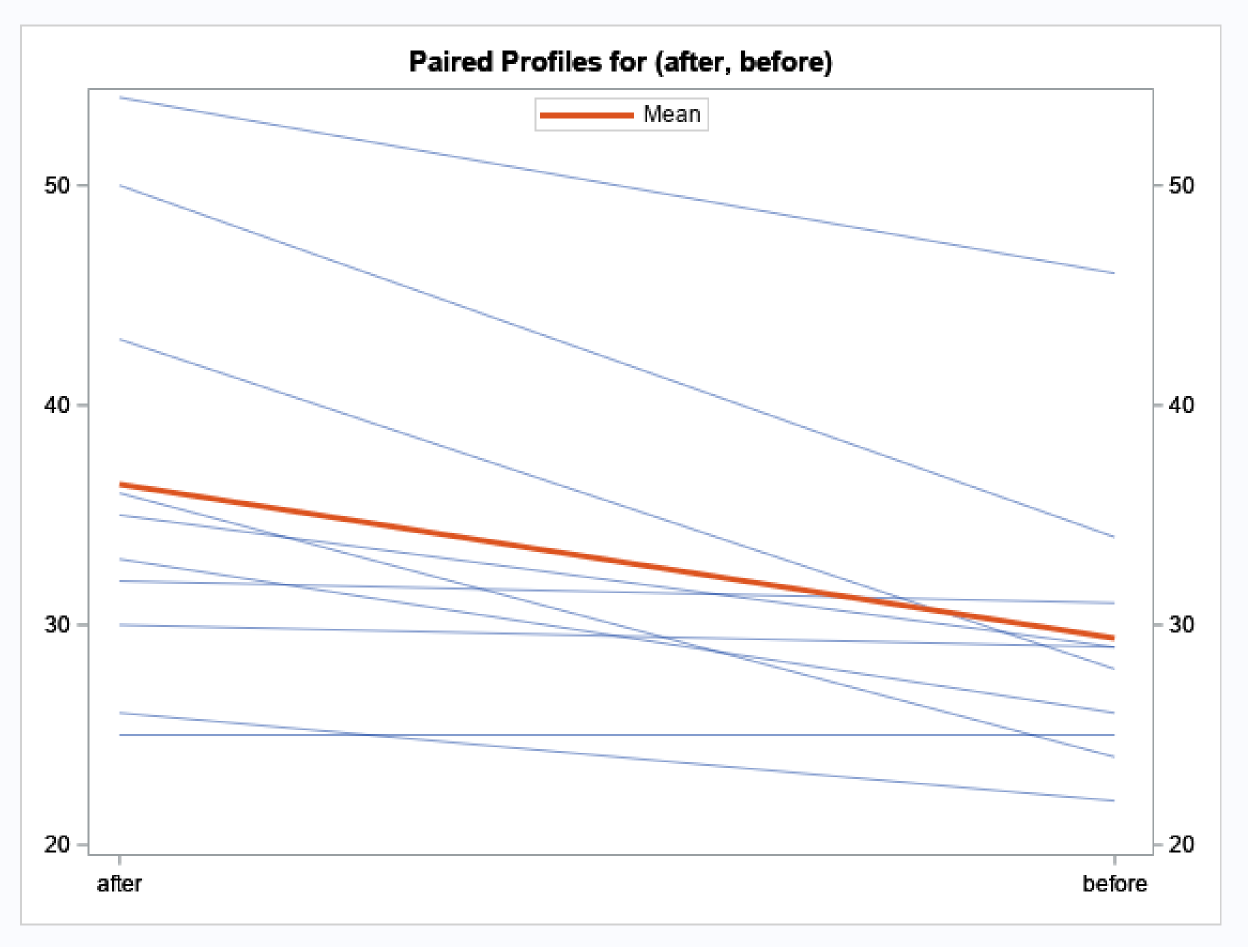
Mean	95% CL Mean		Std Dev	95% CL Std Dev	
7.0000	2.8561	11.1439	5.7927	3.9844	10.5752

DF	t Value	Pr > |t|
9	3.82	0.0041

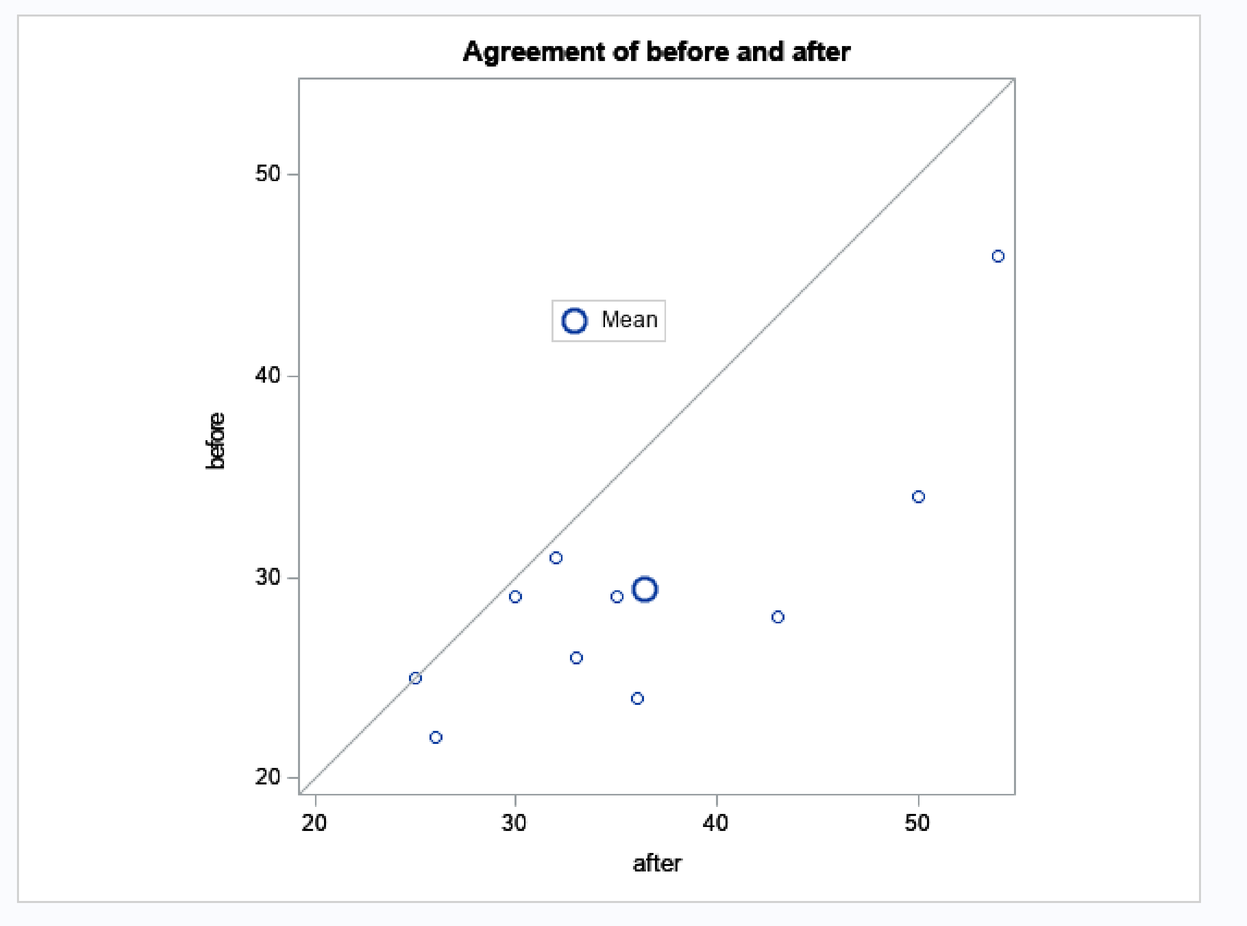
paired2


Distribution of Difference: after - before
With 95% Confidence Interval for Mean
Normal
Kernel
Percent
30
20
10
0
95% Confidence
-10
0
10
20
30
Difference

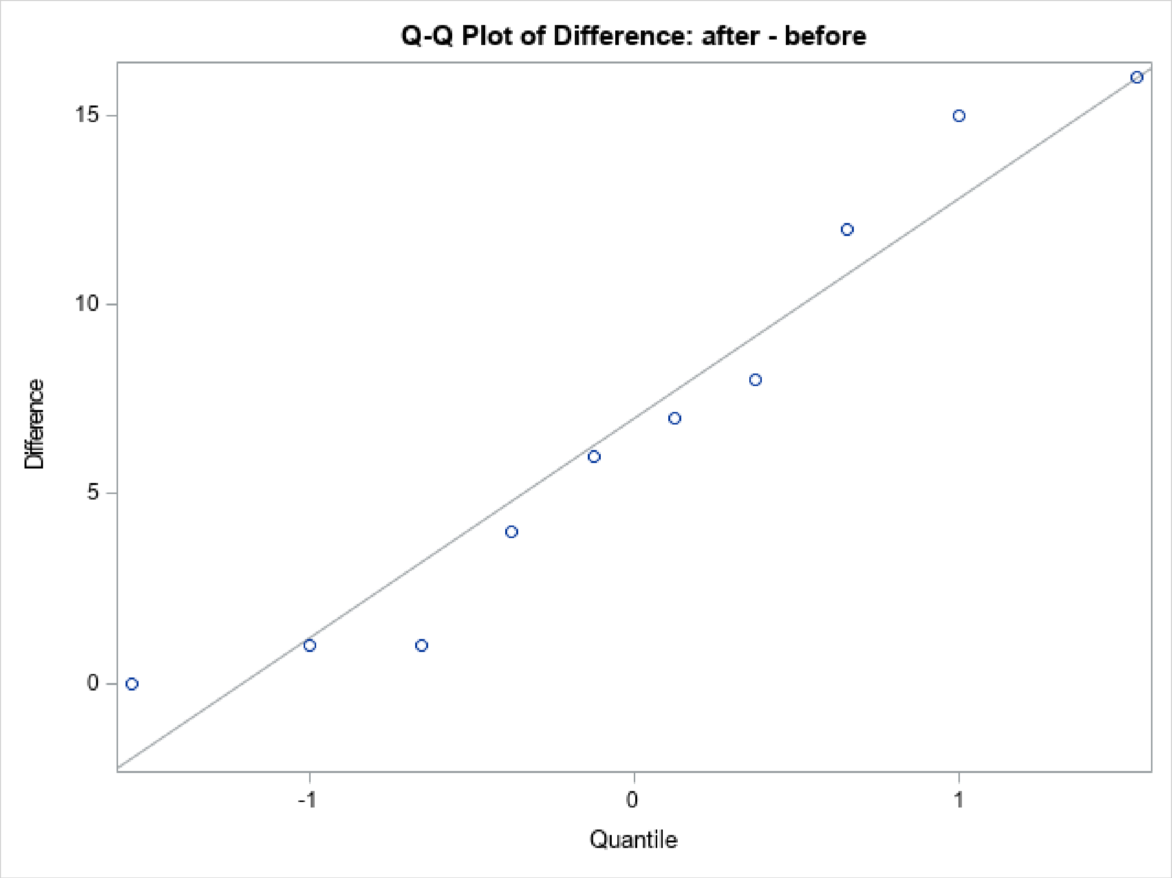
paired3


Paired Profiles for (after, before)
Mean
50
40
30
20
50
40
30
20
after
before

paired4


Agreement of before and after
Mean
before
50
40
30
20
after
20
30
40
50

paired5



Testing categorical data

For most of the analyses that you have learned about, all are analyzing quantitative data. But that leaves out a large portion of data, categorical data. Now we can see how to analyze things like:

(1) making sure a sample follows a specific distribution
(2) exploring whether or not two or more categories have a relationship
(3) analyzing data to see how one category is distributed over another

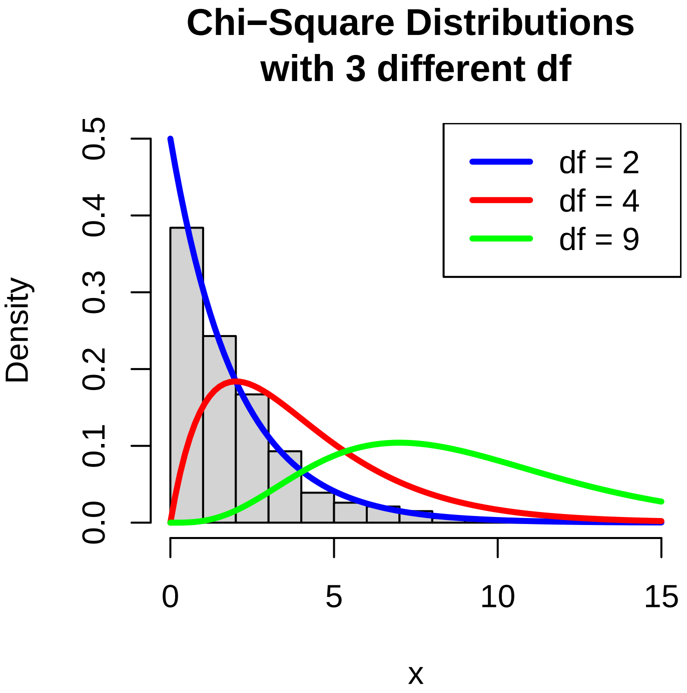
Chi-square distribution

While we have analyses for comparing more than 2 means, we cannot use them when trying to compare two or more proportions. However, there is a distribution that is related to the standard normal distribution (z) that works for comparing more than two proportions. Rather than a test statistic for each pair of proportions, we'd rather like to use just one to prevent the Type I error from inflating. What we do is measure the distance each sample value is from the average (from the "norm"). If we had a z-score for each pair, the sum of the squared z-scores would be a new (new to you) distribution called Chi-square (pronounced "ky" as in "sky"), denoted by χ^2. The distribution is a skewed distribution (skewed right) so it is not a symmetric distribution like z or t, until $df \to \infty$.

$$\chi^2 = \sum_{i=1}^{n} z_i^2 = z_1^2 + z_2^2 + \cdots + z_n^2$$

χ^2 with varying *df*

The following graph illustrates how the χ^2 distribution changes shape with increasing *df*.

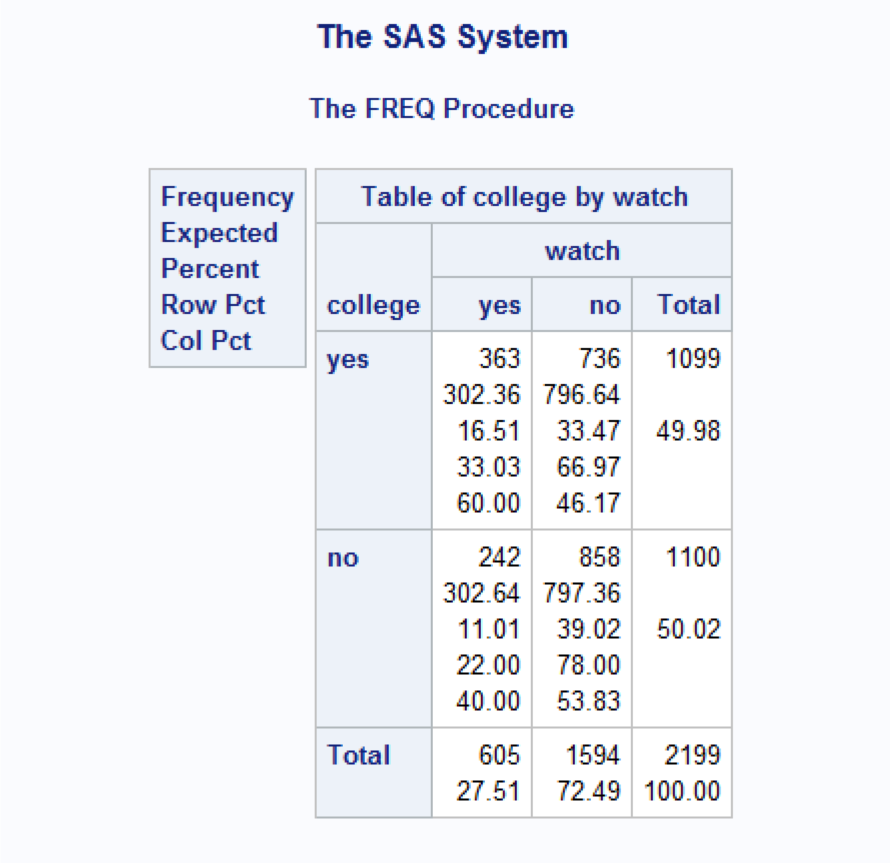


Independent proportions

A 2010 Pew Research foundation poll indicated that among 1099 college graduates, 33% watch the Daily Show. Meanwhile, 22% of the 1100 people with a high school degree (but no college degree) watch The Daily Show. Use of Fisher's exact test will work here since it is a 2X2 table

```
data dailyshow;
input college$ watch$ count;
cards;
yes yes 363
yes no 736
no yes 242
no no 858
;
proc freq data=dailyshow order=data;
tables college*watch / chisq expected;
weight count;
run;
```

2prop1



The SAS System

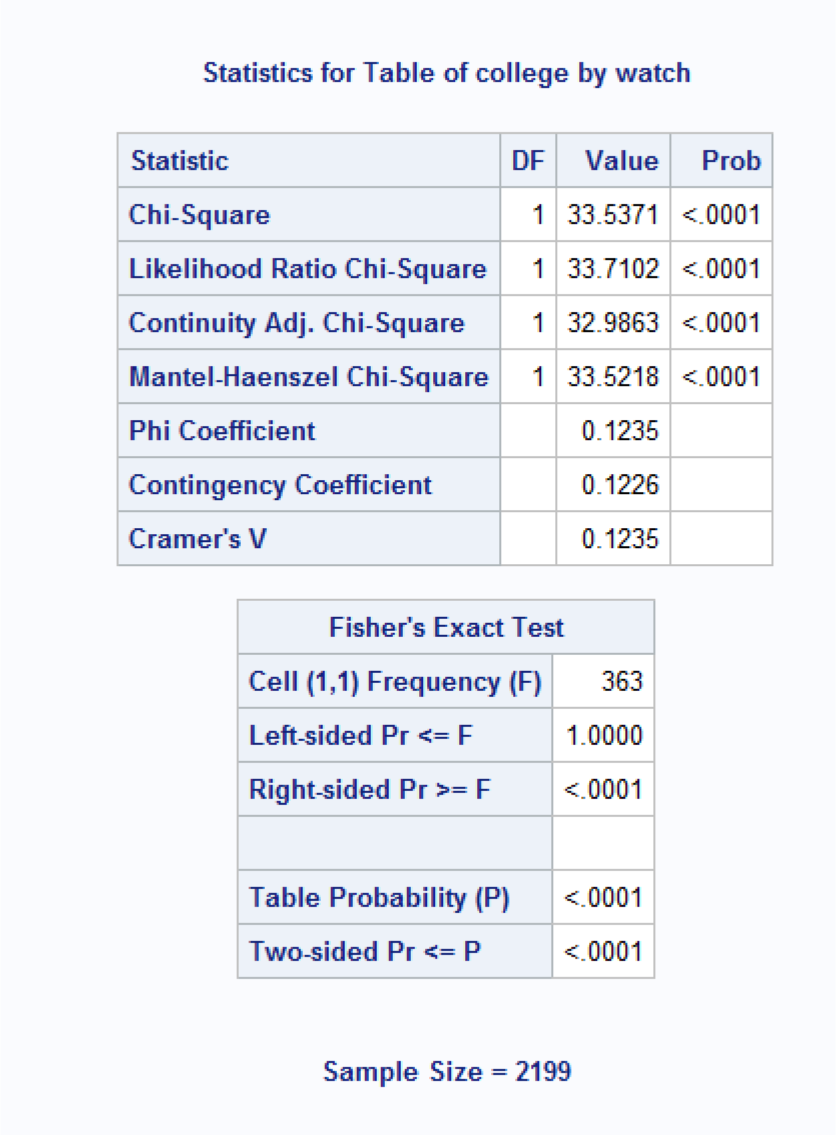
The FREQ Procedure

Frequency
Expected
Percent
Row Pct
Col Pct

Table of college by watch

	watch		
college	yes	no	Total
yes	363 302.36 16.51 33.03 60.00	736 796.64 33.47 66.97 46.17	1099 49.98
no	242 302.64 11.01 22.00 40.00	858 797.36 39.02 78.00 53.83	1100 50.02
Total	605 27.51	1594 72.49	2199 100.00

2prop2



Statistics for Table of college by watch

Statistic	DF	Value	Prob
Chi-Square	1	33.5371	<.0001
Likelihood Ratio Chi-Square	1	33.7102	<.0001
Continuity Adj. Chi-Square	1	32.9863	<.0001
Mantel-Haenszel Chi-Square	1	33.5218	<.0001
Phi Coefficient		0.1235	
Contingency Coefficient		0.1226	
Cramer's V		0.1235	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	363
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	<.0001
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Sample Size = 2199

Assumptions of Chi-square tests

(1) The data must be counts from categories
(2) Independence of observations
(3) $E_i \geq 5$; each individual expected value (E_i) must be at least 5

Test statistic (for all 3 tests), *df*

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected} = \sum \frac{(O-E)^2}{E}$$

df for GoF is $df = k - 1$, where k = number of categories

df for Independence and Homogeneity is $df = (r-1)(c-1)$

(r = number of rows, c = number of columns)

Goodness-of-Fit (GoF)

Chi-square for a one-way table (a table that has categories and counts for each category): In evaluating whether there is sufficient evidence that a set of observed counts, $O_1, O_2, \cdots, O_k$ in k categories are unusually different from what would be expected under a null hypothesis. The expected values under the null hypothesis, called $E_1, E_2, \ldots, E_k$.

GoF hypotheses

$H_0 : p_1 = p_2 = \cdots = p_k = p_0$ or

$$H_0 : The\ data\ follows\ <specified>\ distribution$$

H_a : At least one p_i differs or

$$H_a : H_0\ is\ not\ true\ (the\ data\ does\ not\ follow\ <specified>\ distribution))$$

GoF formulas

Expected value

$$E_i = np_i$$

You will need to find the probabilities associated with the null hypothesized distribution (given), then multiply the sample size (the sum of the observations) by each category probability to get the expected values.

GoF H_0 rejection

Rejection region

Reject H_0 iff $pvalue \leq \alpha$ where $pvalue = P(\chi^2 \geq \chi^2_{calc})$ (used for this course)

Conclusion (in context)
When the null hypothesis is rejected, in terms of the context of the data, it means that we think that the data does not follow the theoretical (specified) distribution. When we fail to reject the null hypothesis, we are maintaining that the data does follow the theoretical (specified) distribution

Test of Independence

The test of Independence explores whether two categorical random variables are independent or whether some level of dependency exists between them. Each dataset will be constructed into a table with I rows and J columns. Let n_{ij} denote the number of individuals in the sample falling in the $(i, j)^{th}$ cell (of row i, column j) of the table. The following is a prototype of a general table that displays the counts (n_{ij}) and is called a *two-way contingency table.* I and J (capital I,J) are the row and column totals, respectively.

Data organization

	1	2	$\ldots$	j	$\ldots$	J
1	n_{11}	n_{12}	$\ldots$	n_{1j}	$\ldots$	n_{1J}
2	n_{21}					$\vdots$
$\vdots$						
i	n_{i1}	$\ldots$		n_{ij}	$\ldots$	
$\vdots$						

	1	2	$\ldots$	j	$\ldots$	J
I	n_{I1}	$\ldots$				$n_{IJ} = n$

Independence test hypotheses

$H_0 : p_{ij} = (p_{i\cdot})(p_{\cdot j})$

Or H_0 : The row context and column are independent

$H_a : H_0$ is not true (meaning that rows and columns are dependent)

With $i = 1, 2, \ldots, I \ and \ j = 1, 2, \ldots, J$

Independence test formulas

Expected values

$$E_{ij} = \frac{n_i n_j}{n} = \frac{(rtotal)(ctotal)}{grandtotal}$$

Independence test rejection

Rejection region
Reject H_0 iff $pvalue \leq \alpha$ where $pvalue = P(\chi^2 \geq \chi^2_{calc})$

Conclusion (in context)
When the null hypothesis is rejected, in terms of the context of the data, it means that we think that the context of the rows and context of the columns are dependent (there is a dependency). When we fail to reject the null hypothesis, we are maintaining that the context of the rows and context of the columns are dependent (there is no relationship).

Homogeneous Test

We are assuming that each individual in every one of the I populations belongs in exactly one of J categories. An example would be to see if voting habits are the same over regions.

Homogeneous test hypotheses

$H_0 : p_{1j} = p_{2j} = \ldots = p_{Ij}$
OR
H_0 : The row is distributed the same over the column

$H_a : H_0$ is not true (the distribution is not the same for all categories)

With $i = 1, 2, \ldots, I \ and \ j = 1, 2, \ldots, J$

Homogeneous test formulas+

Test statistic
Same as Independence Test

Expected values
Same as Independence Test

Rejection region
Same as Independence Test

Conclusion (in context)
When the null hypothesis is rejected, in terms of the context of the data, it means that we think that the context of the rows are distributed differently across the context of the columns. When we fail to reject the null hypothesis, we are maintaining that the context of the rows are distributed similarly across the context of the columns.

PROC FREQ

```
proc freq data=SASdataset order=data;
tables row_var*col_var </ chisq expected>; weight numeric_var(s);
run;
```

`order=`: data (as is in dataset), formatted (sorts ascending), . . .
`weight`: numeric values of counts

There are other options available, mostly for suppressing row, column, or total percents, and other options

Independence and homogeneity

Data are the voting records; recorded are gender (M,F) and party affiliation (D,I,R).

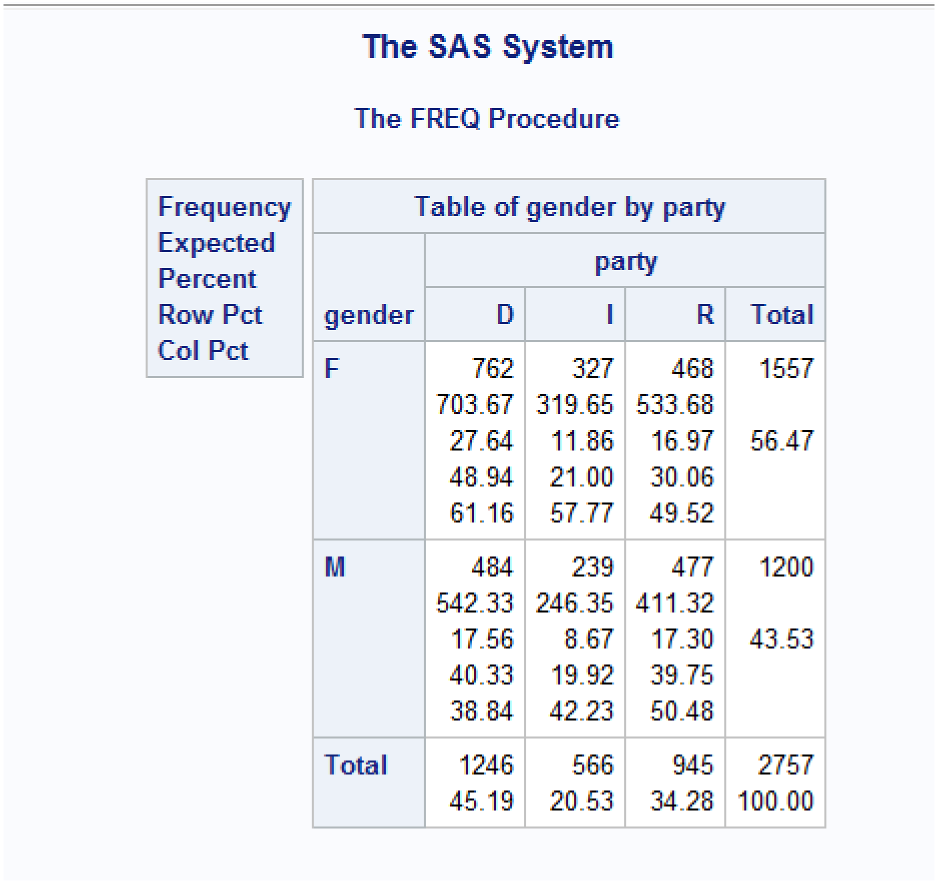
Independence: Is there evidence that there is an association between gender and party affiliation?

Homogeneous: Is there evidence that party affiliation is the same across genders?

Independence and homogeneity

```
data vote;
input gender$ party$ count;
cards;
F D 762
F I 327
F R 468
M D 484
M I 239
M R 477
;
proc freq data=vote order=data;
tables gender*party / chisq expected;
weight count;
run;
```

indep1



The SAS System

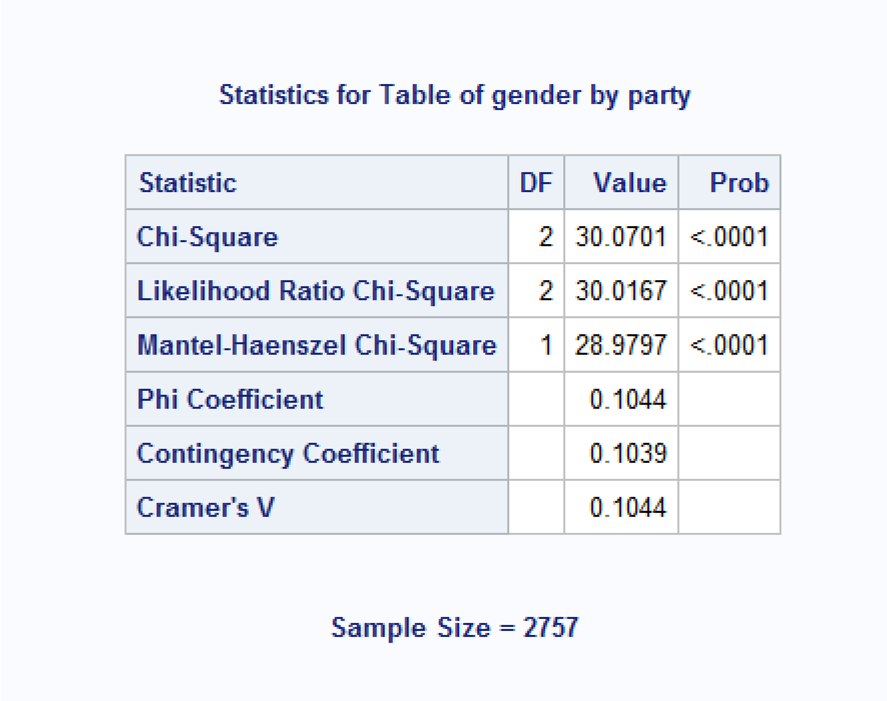
The FREQ Procedure

Frequency Expected Percent Row Pct Col Pct

Table of gender by party

	party			
gender	D	I	R	Total
F	762 703.67 27.64 48.94 61.16	327 319.65 11.86 21.00 57.77	468 533.68 16.97 30.06 49.52	1557 56.47
M	484 542.33 17.56 40.33 38.84	239 246.35 8.67 19.92 42.23	477 411.32 17.30 39.75 50.48	1200 43.53
Total	1246 45.19	566 20.53	945 34.28	2757 100.00

indep2



Statistics for Table of gender by party

Statistic	DF	Value	Prob
Chi-Square	2	30.0701	<.0001
Likelihood Ratio Chi-Square	2	30.0167	<.0001
Mantel-Haenszel Chi-Square	1	28.9797	<.0001
Phi Coefficient		0.1044	
Contingency Coefficient		0.1039	
Cramer's V		0.1044	

Sample Size = 2757

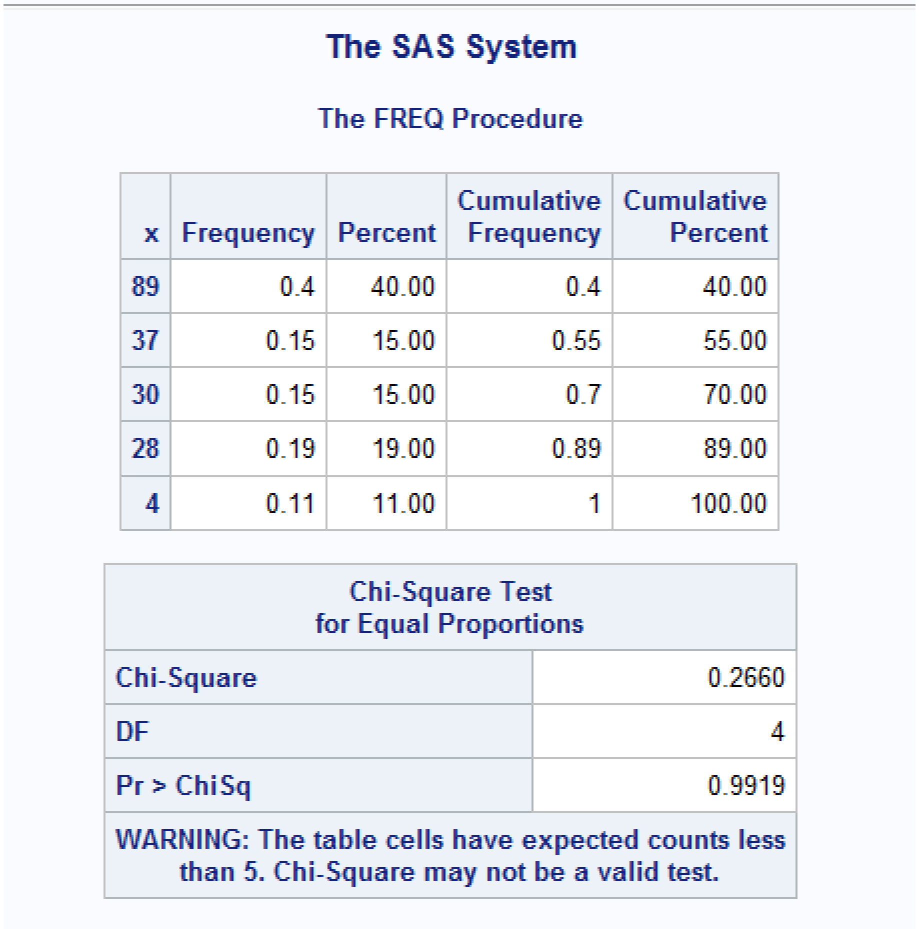
GoF

Chocolate! M&M population colors are specified by the Mars company. Does the distribution of candies match the popualtion proportions?

GoF

```
data gof;
input x p @@;
cards;
89 0.4 37 0.15 30 0.15 28 0.19 4 0.11
;
proc freq data=gof order=data;
tables x / chisq expected;
weight p;
run;
```

gof1



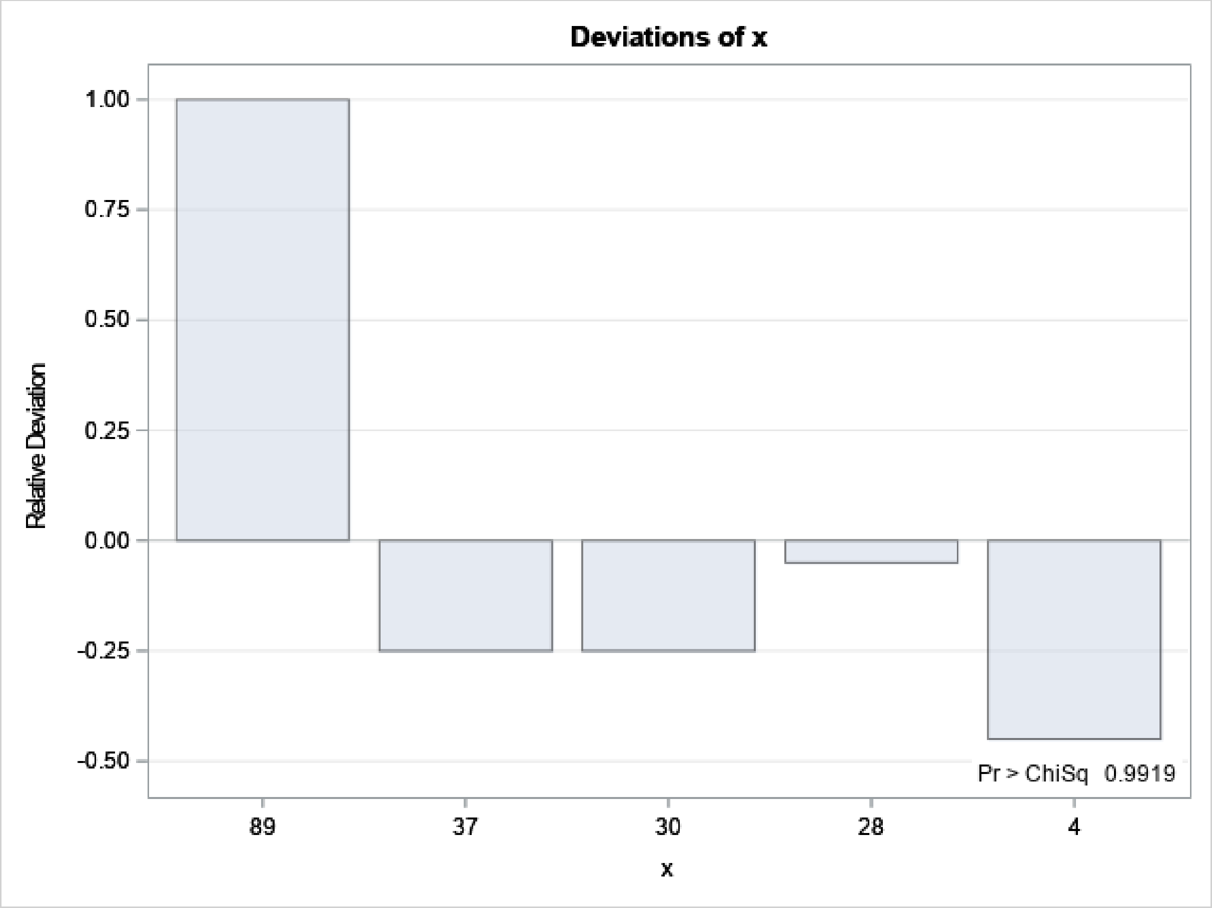
The SAS System

The FREQ Procedure

x	Frequency	Percent	Cumulative Frequency	Cumulative Percent
89	0.4	40.00	0.4	40.00
37	0.15	15.00	0.55	55.00
30	0.15	15.00	0.7	70.00
28	0.19	19.00	0.89	89.00
4	0.11	11.00	1	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	0.2660
DF	4
Pr > ChiSq	0.9919
WARNING: The table cells have expected counts less than 5. Chi-Square may not be a valid test.	

gof2



Sample Size = 1

Simple Linear Regression (slr)

- SLR analysis explores the linear association between an explanatory (independent) variable, usually denoted as x, and a response (dependent) variable, usually denoted as y
- This type of data is called bivariate data (data with two (bi) variables)
- The point is to see if we can use a mathematical linear model to describe the association (relationship) between the two variables
- Using one known value to estimate the other value, in addition to seeing how strong the relationship is
- You are familiar with $y = mx + b$ from algebra, where m is the slope and b is the y-intercept (value of y when $x = 0$), which is a mathematical linear equation, a *deterministic* equation.

The population regression model

Notice that it is basically the same as you have seen and used before ($y = mx + b$):

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Where:

- y_i: value of the response (dependent) variable
- β_0: the value of the y-intercept (when $x = 0$)
- β_1: the value of the slope (the change in y due to a one unit increase in x, **not** $\frac{rise}{run}$)
- ϵ_i: the residual (error) term

The sample regression model

Is used once there are estimated values from the data:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Where:

- $\hat{y}$: estimate of the value of the i^{th} response (dependent) variable

- $\hat{\beta}_0$: the estimate of the value of the y-intercept ($\hat{y}$ when $x = 0$)

- $\hat{\beta}_1$: the estimate of the value of the slope (the change in y due to a one unit increase in x (Not $\frac{rise}{run}$)

- Note that ϵ dropped off from the other model. This is because of the first assumption of regression, $E(\epsilon_i) = 0$: the mean of the residuals $= 0$.

The assumptions for SLR are the same as ANOVA.

Residuals

Residuals: ϵ_i are the population residuals and $\hat{\epsilon}_i = e_i$ are the sample residuals

$e_i = y_i - \hat{y}$. If $e_i > 0$, the model *underestimated* the response and if $e_i < 0$, the model *overestimated* the response.

Analysis tools: scatterplot graph

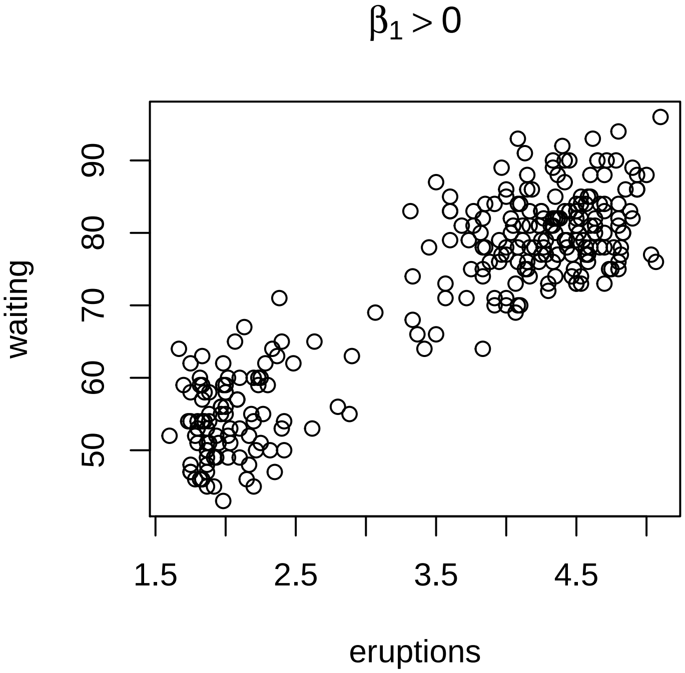
- First thing that is necessary is to look at a scatterplot of the two variables; it is a type of graph that you are familiar with from algebra
  - x is the explanatory (independent) variable and goes along the x-axis
  - y is the response (dependent) variable and goes along the y-axis
- The values of x and $\hat{y}$ are an ordered pair of data, $(x, \hat{y})$ that can be graphed on the Cartesian (rectangular) coordinate system
- The value of x that will be given is most often one that is an observed value of x so that an estimation of the residual, $e_i = y_i - \hat{y}_i$ can be calculated.

Analysis tools: scatterplot graph

- A scatterplot of the data shows if there is a linear association between the explanatory (independent) variable and the response (dependent) variable
  - When x and y both increase, the slope (relationship) is positive
  - When x increases while y decreases, the slope (relationship) is negative
- The point of visually checking the scatterplot **before** doing the regression analysis is decide if there is at least a fair linear relationship between x and y
  - If you do not have a linear relationship, then use of regression analysis is not recommended as the results cannot be used with the given dataset
- The regression line is also called a trend line.

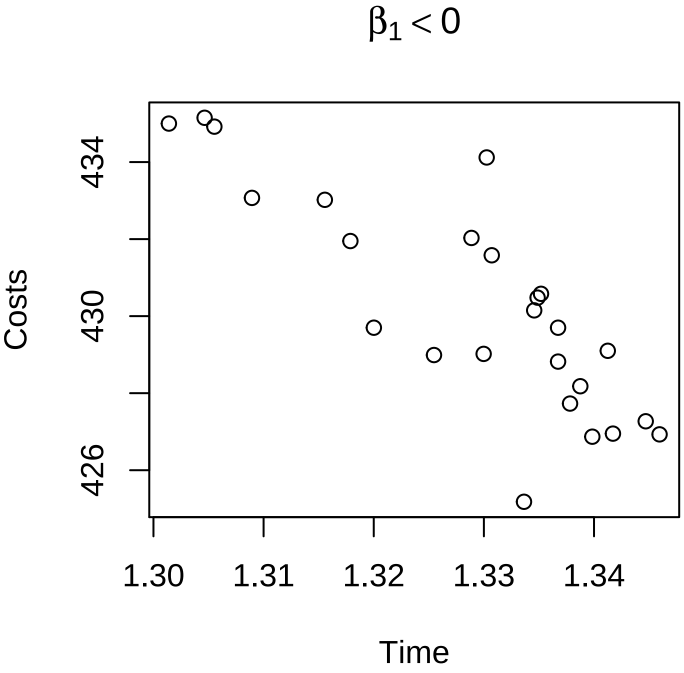
Analysis tools: scatterplot graph

This has positive slope (x increases and y increases)



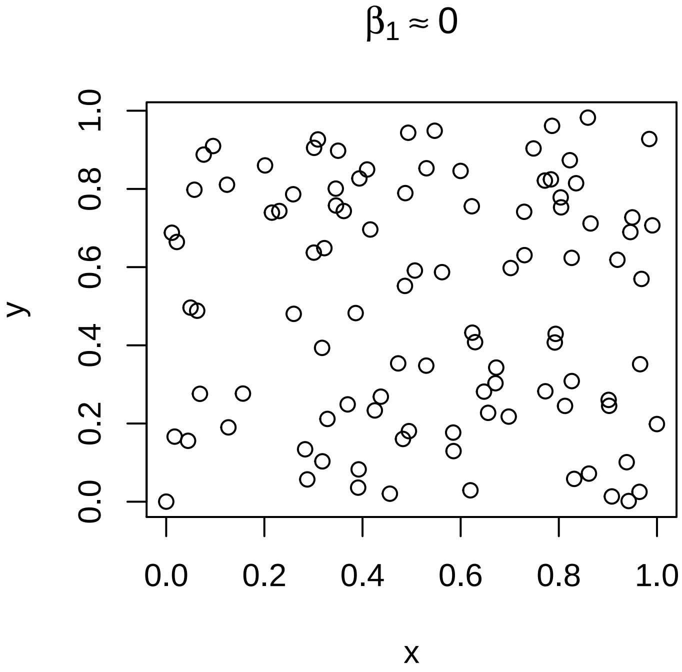
Analysis tools: scatterplot graph

This has negative slope (x increases and y decreases)



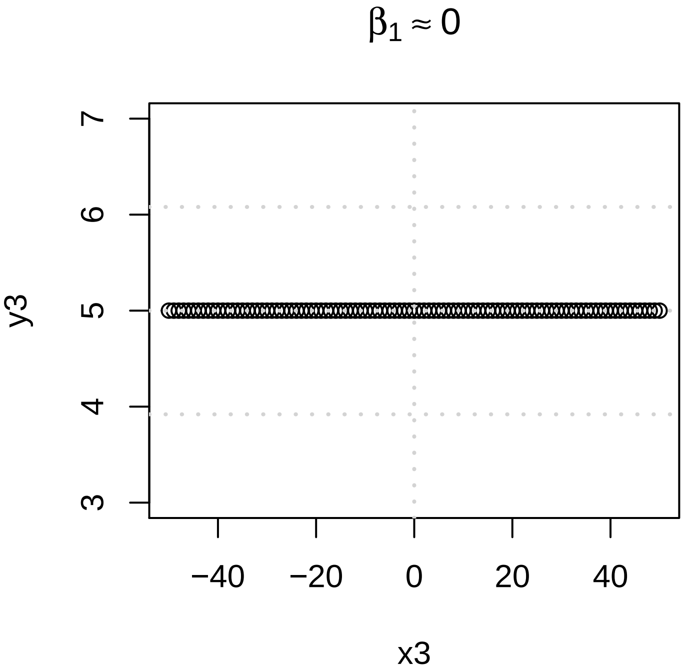
Analysis tools: scatterplot graph

This has 0 slope (and a lot of random scatter)



Analysis tools: scatterplot graph

This has 0 slope



Slope and intercept formulas

Slope:

$$\hat{\beta}_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{s_x^2(n-1)}$$

Intercept:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Correlation

To determine the strength of the relationship between two *quantitative* variables, we use a measure called *correlation*

Defn: Is a calculation that measures the strength and direction (positive or negative) of the *linear* relationship between 2 *quantitative* variables, x and y

Correlation $\neq$ causation

It is *extremely* important to note that just because two variables have a mathematical correlation **IT DOES NOT MEAN** X **CAUSES** Y**!!!**. To establish actual causation, repeatable experimentation must be done.

Correlation logistics

- It is bound between -1 and 1 ($-1 \leq r \leq 1$)
  - $r = -1$ and $r = 1$ are perfect linear relationships

  - $r = 0$ implies both no linear relationship and x, y are independent

- r makes no distinction between x and y

- r has no units of measurement
- if $r > 0$, then $\hat{\beta}_1 > 0$, $r < 0$, then $\hat{\beta}_1 < 0$
- Correlation is denoted as r for sample correlation and ρ for the population correlation.

$$r = \frac{1}{n-1} \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

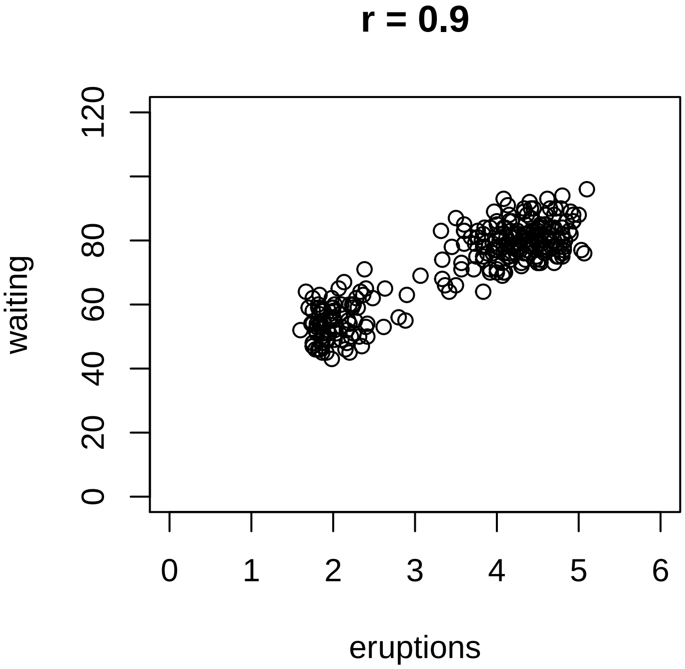
Coefficient of Determination, R^2

R^2 is called the *coefficient of determination*:

- It is the proportion (or ×100%) of observed variation that can be explained by the relationship between x and y
- $0 \leq R^2 \leq 1$: It is bound between 0 (0%) and 1 (100%)
  - The closer to 1 (100%), the more variation we can explain and also the stronger the linear relationship between x and y
    - An acceptable baseline for R^2 would be when $R^2 \geq 60\%$
- $R^2 = (r)^2 \therefore r = \pm\sqrt{R^2}$
  - if the slope is positive, then r is positive, if the slope is negative, then r is negative.

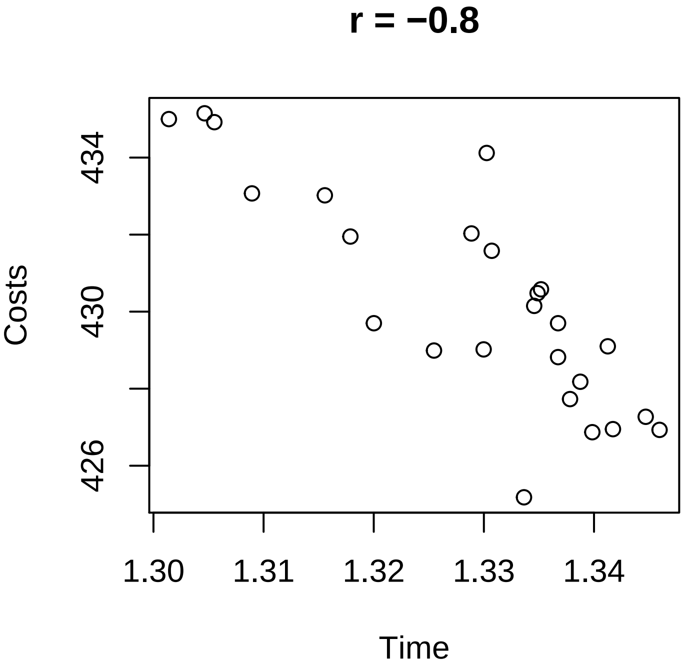
Analysis tools: scatterplot graph

Relatively strong, positive correlation



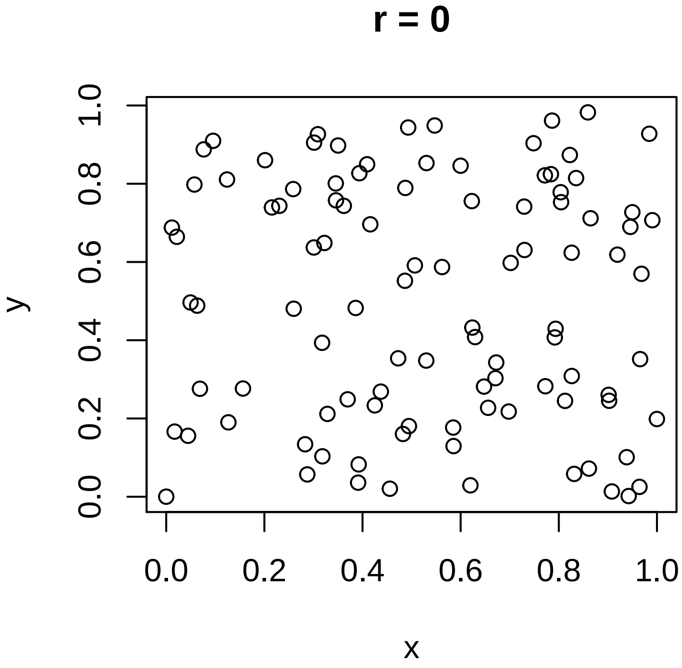
Analysis tools: scatterplot graph

Moderately strong, negative correlation



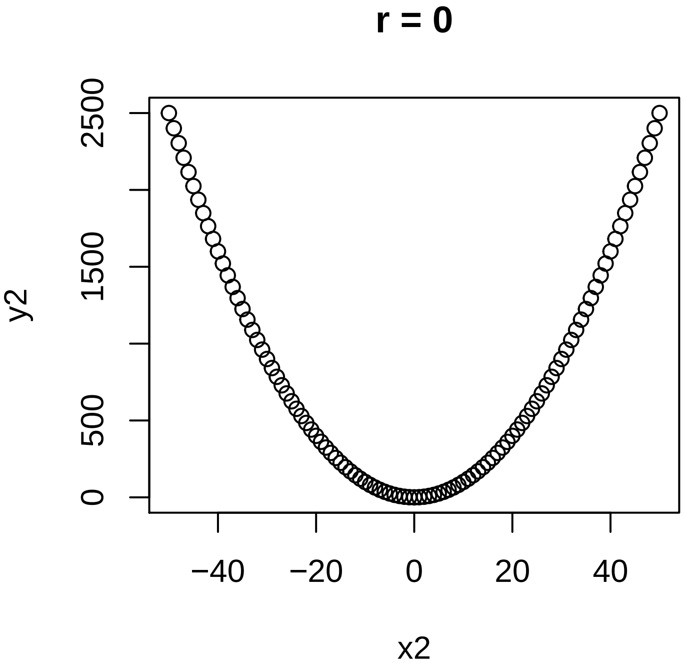
Analysis tools: scatterplot graph

No correlation



Analysis tools: scatterplot graph

No correlation but there *is* a relationship, it is not a linear relationship



PROC REG

General form of PROC REG:

```
PROC REG data=SAS-dataset <options>;
MODEL dependent(s)=regressor(s) </options>;
...;
RUN;
```

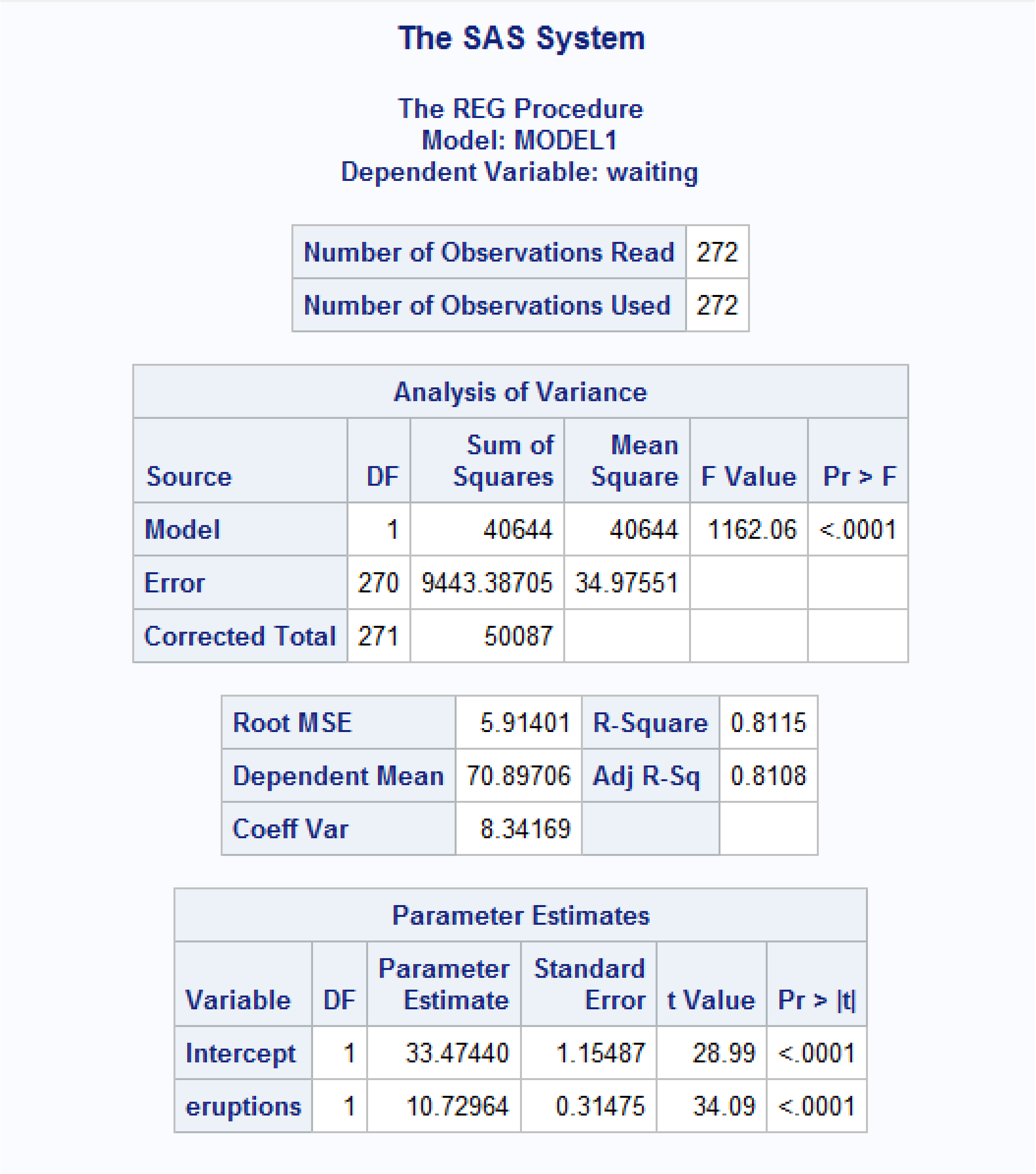
No PLOTS option is needed as PROC REG will automatically create diagnostic plots

MODEL statement options (part 1):
dw: for Durbin-Watson test

PROC REG

```
proc sgplot data=faithful;
scatter x=eruptions y=waiting;
run;
proc reg data=faithful;
model waiting=eruptions;
output out=new p=yhat r=res;
run; quit;
```

slr1



The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: waiting

Number of Observations Read	272
Number of Observations Used	272

Analysis of Variance

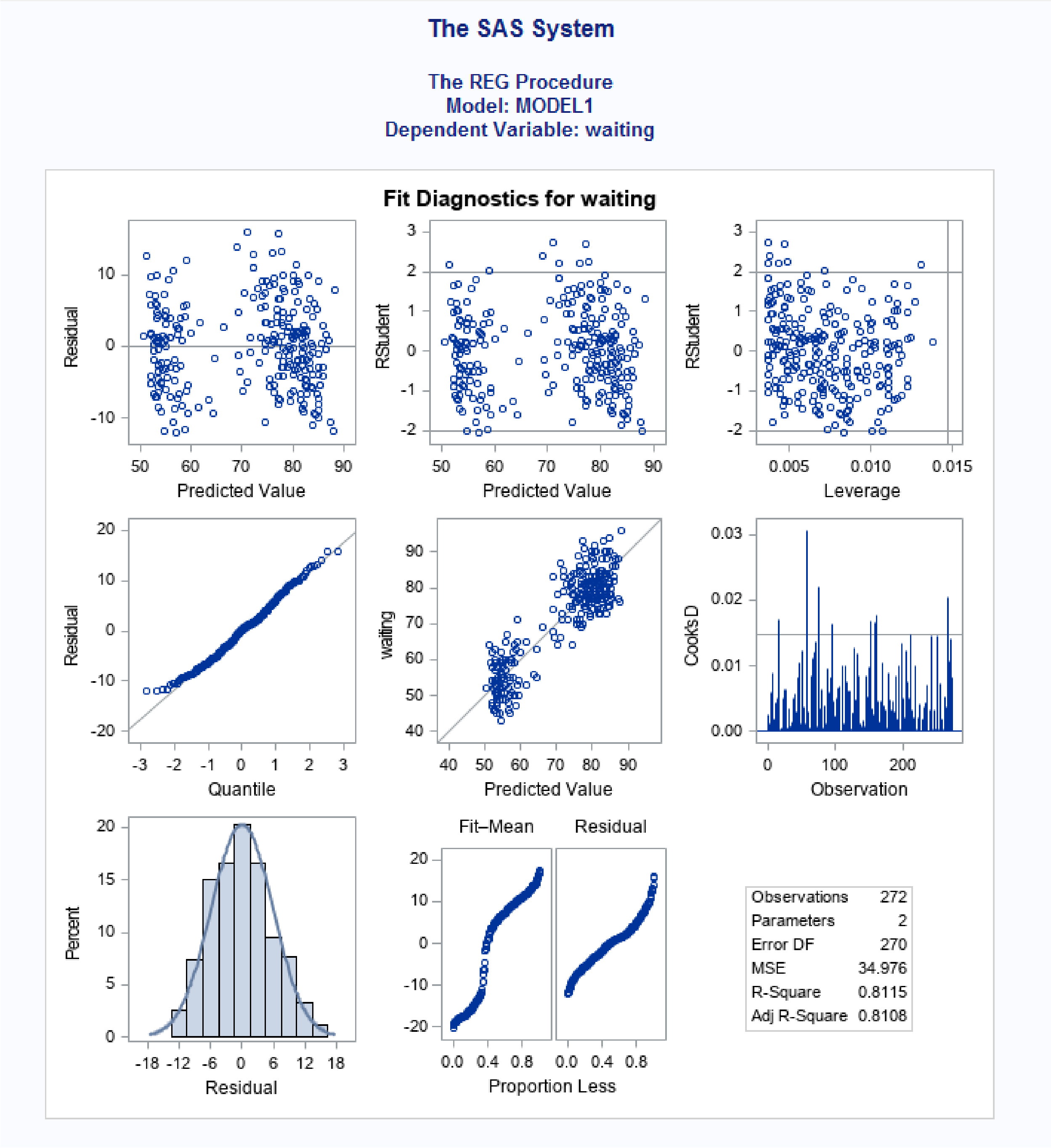
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	40644	40644	1162.06	<.0001
Error	270	9443.38705	34.97551		
Corrected Total	271	50087			

Root MSE	5.91401	R-Square	0.8115
Dependent Mean	70.89706	Adj R-Sq	0.8108
Coeff Var	8.34169		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	33.47440	1.15487	28.99	<.0001
eruptions	1	10.72964	0.31475	34.09	<.0001

slr2



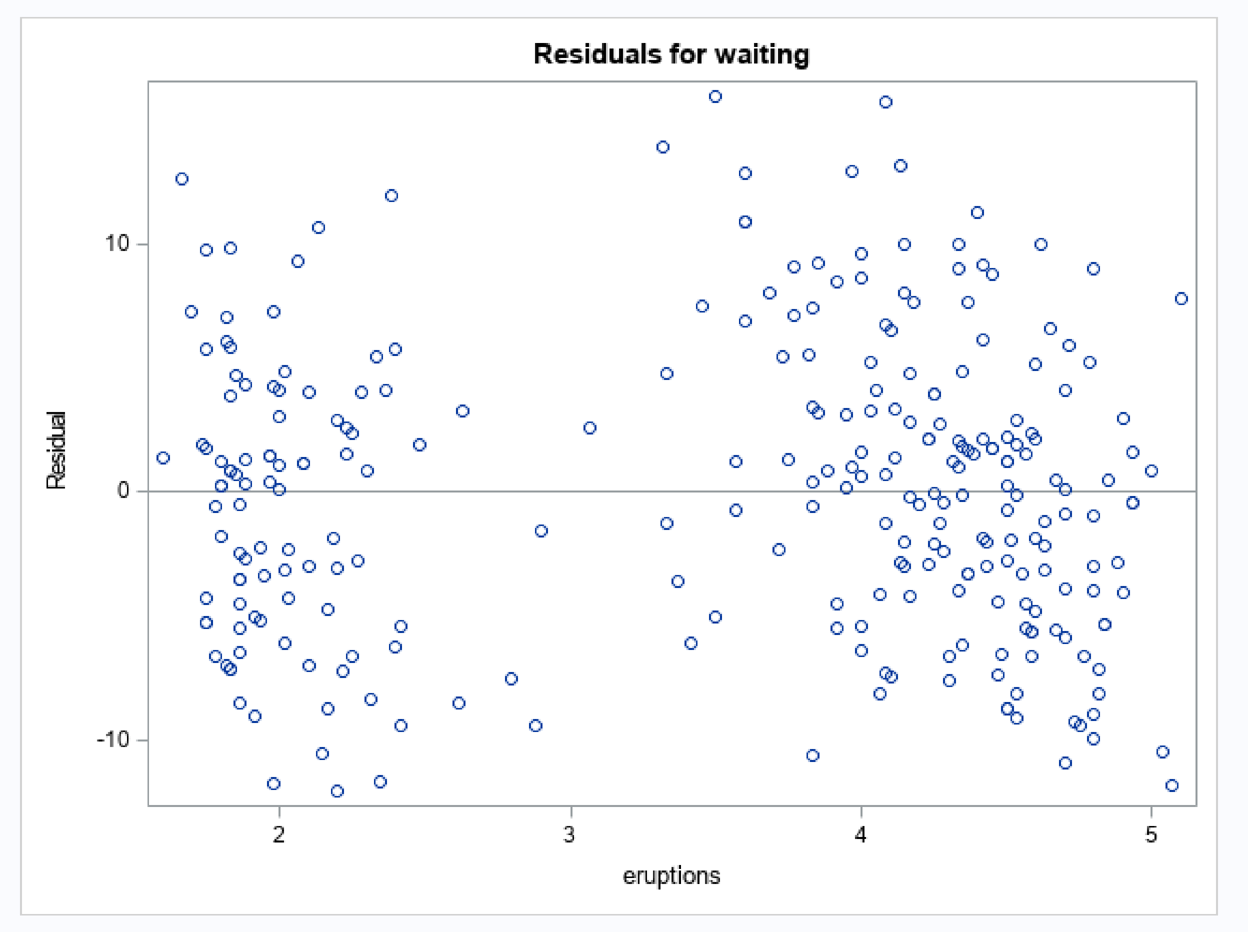
The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: waiting

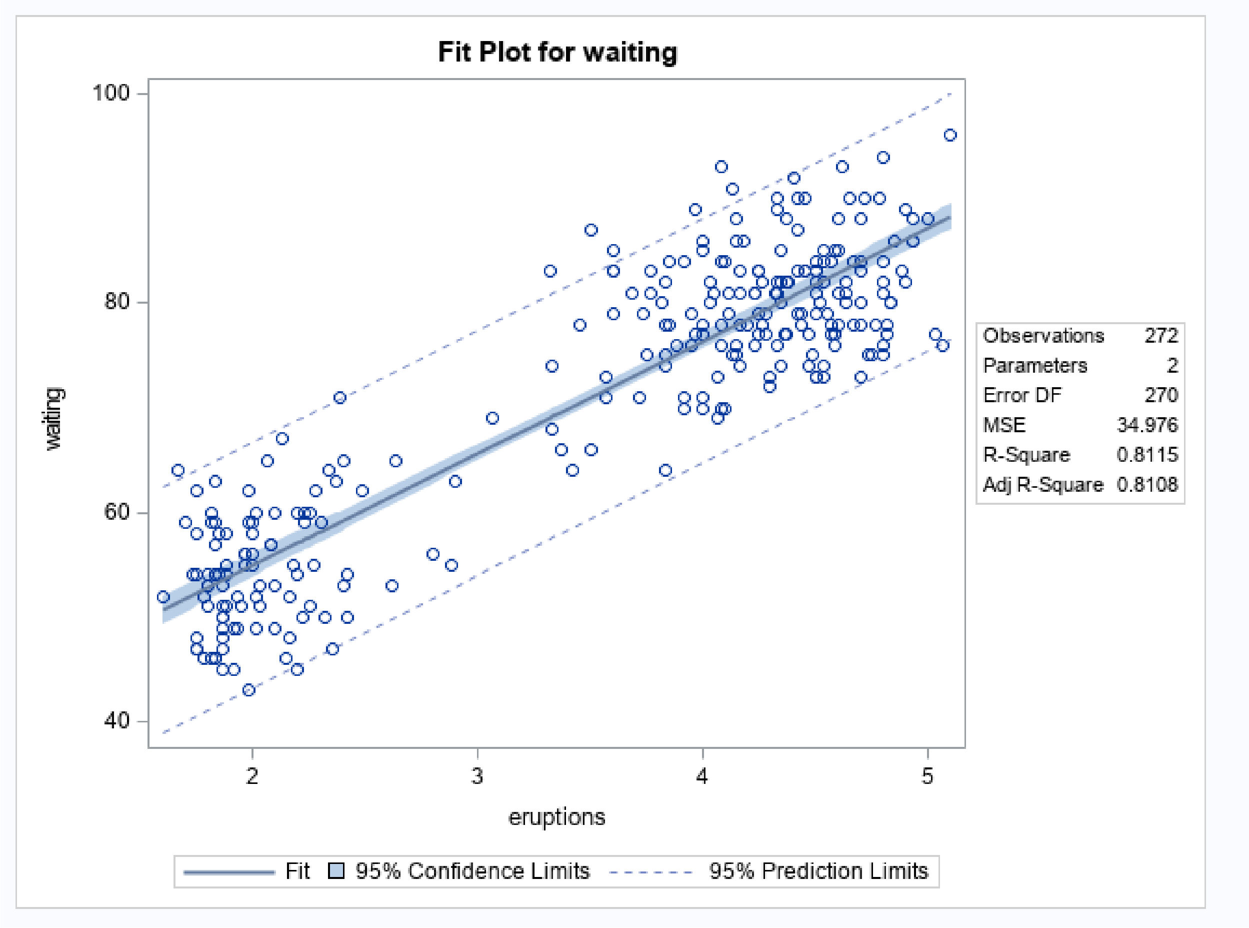
Fit Diagnostics for waiting

Observations	272
Parameters	2
Error DF	270
MSE	34.976
R-Square	0.8115
Adj R-Square	0.8108

slr3



slr4

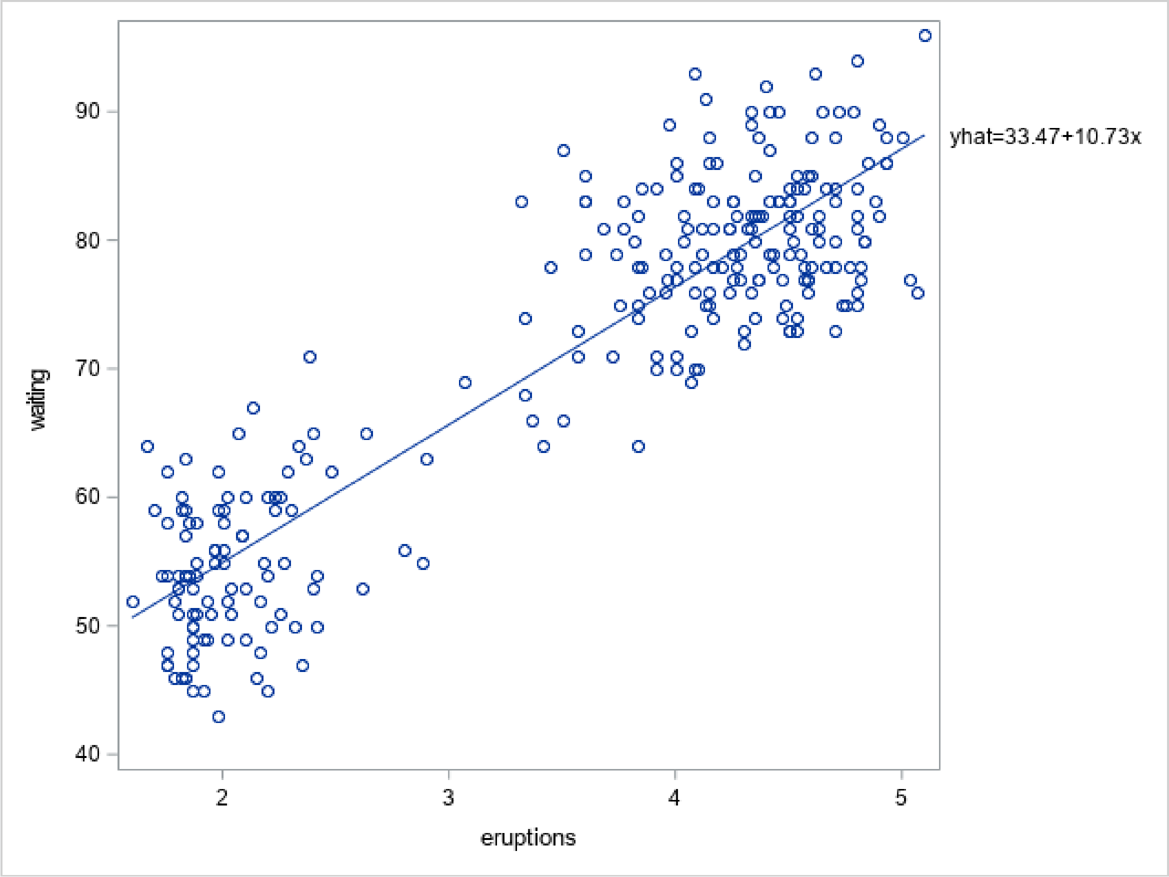


Observations	272
Parameters	2
Error DF	270
MSE	34.976
R-Square	0.8115
Adj R-Square	0.8108

sgplot scatter with regression line

```
proc sgplot data=f2;
scatter x=eruptions y=waiting;
series x=eruptions y=yhat / curvelabel="yhat=33.47+10.73x"
            curvelabelloc=outside;
run;
```

slr5



CIs for $\hat{\beta}_1$, $\hat{\beta}_0$

All the following standard errors are provided in the regression analysis output; SAS will provide the CI for β_k (the slope (or any partial slope in the case of multiple regression)) but not for β_0

$$\hat{\beta}_j \pm t^{\star}(se_{\hat{\beta}_j})$$

Where $\hat{\beta}_j$ is either $\hat{\beta}_0$ or $\hat{\beta}_1$; same goes for the se, $t^{\star} = t_{\alpha/2,df}$ and $df = n - 2$ for both cases.

$$se_{\hat{\beta}_0} = \sqrt{s_{\epsilon}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_x^2(n-1)}\right)}\ se_{\hat{\beta}_1} = \sqrt{\frac{s_{\epsilon}^2}{s_x^2(n-1)}}$$

$$s_{\epsilon}^2 = \frac{\sum(\hat{y}_i - y_i)^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

Hypothesis tests for the estimated slope (β_1) and intercept (β_0)

- Most often the slope $\hat{\beta}_1$ is the only real test of interest
- Many times the value of $x = 0$ is not in the dataset (or the fact that maybe $x = 0$ is not possible in the population the data was sampled from). Without $x = 0$ in the dataset (or even possible at all), the intercept does not make sense in context

- Additionally, the slope is what is driving the relationship whereas the intercept just represents the value where the regression line crosses through the y-axis

- There are some economic datasets and many others that utilize the intercept because it make sense both mathematically and realistically.

Hypothesis tests for the estimated slope (β_1) and intercept (β_0)

- The null hypothesis for the slope is to test if the slope is equal to zero
  - A slope of zero is a horizontal line, where any value of x has the same y value
- Most often of interest is whether or not it is significant, the alternative hypothesis is to see if the slope is different from zero
  - Realistically the hypothesized value could be something other than 0 if there is a need, like seeing if it has increased or decreased since the previous sample was taken and analyzed

Test for β_1, the slope

Hypotheses:

$$H_0 : \beta_1 = 0 \text{ vs. } H_a : \beta_1 \neq 0$$

Test Statistic:

$$t = \frac{\hat{\beta}_1 - \beta_1}{se_{\hat{\beta}_1}}$$

- The $se_{\hat{\beta}_1}$ and $df = n - 2$ are the same as for CIs

- Rejection criteria is the same as the t-tests learned in earlier modules (starting in module 9). Rejection of the null means the slope is significant; there is a significant relationship between x and y. Not rejecting the null means there is no significant relationship between x and y

Test for β_0, the intercept

Hypotheses:

$$H_0 : \beta_0 = 0 \text{ vs. } H_a : \beta_0 \neq 0$$

Test Statistic:

$$t = \frac{\hat{\beta}_0 - \beta_0}{se_{\hat{\beta}_0}}$$

- The $se_{\hat{\beta}_0}$ and $df = n - 2$ are the same as for CIs
- Rejection criteria is the same as the t-tests learned earlier; rejection of the null means the intercept is significant. Not rejecting the null just means the intercept is not significant (but has no impact on the significance of the slope)

CI for $\hat{\mu}$

This is referred to as a CI for μ, an average response, computed from the regression line for a given value of x, denoted as $x^\star$. Since it is an average response that is why it uses the notation of $\hat{\mu}$ and to distinguish it from a prediction interval (next slide).

$$\hat{\mu} \pm t^\star(se_{\hat{\mu}})$$

Where $\hat{\mu}_{|_{x=x^\star}} = \hat{\beta}_0 + \hat{\beta}_1 x^\star$, $t^\star = t_{\alpha/2,df}$ and $df = n-2$ for both CIs and PIs.

$$se_{\hat{\mu}} = \sqrt{s_\epsilon^2 \left(\frac{1}{n} + \frac{(x^\star - \overline{x})^2}{s_x^2(n-1)} \right)}$$

PIs (prediction intervals) for $\hat{y}$

This is referred to as a CI for $\hat{y}$, a single response, computed from the regression line for a given value of x, denoted as $x^\star$. Since it is a single response that is why it uses the notation of $\hat{y}$ and to distinguish it from a CI

$$\hat{y} \pm t^\star(se_{\hat{y}})$$

Where $\hat{y}_{|_{x=x^\star}} = \hat{\beta}_0 + \hat{\beta}_1 x^\star$, $t^\star = t_{\alpha/2,df}$ and $df = n-2$ for both CIs and PIs.

$$se_{\hat{y}} = \sqrt{s_\epsilon^2 \left(1 + \frac{1}{n} + \frac{(x^\star - \overline{x})^2}{s_x^2(n-1)} \right)}$$

CIs and PIs

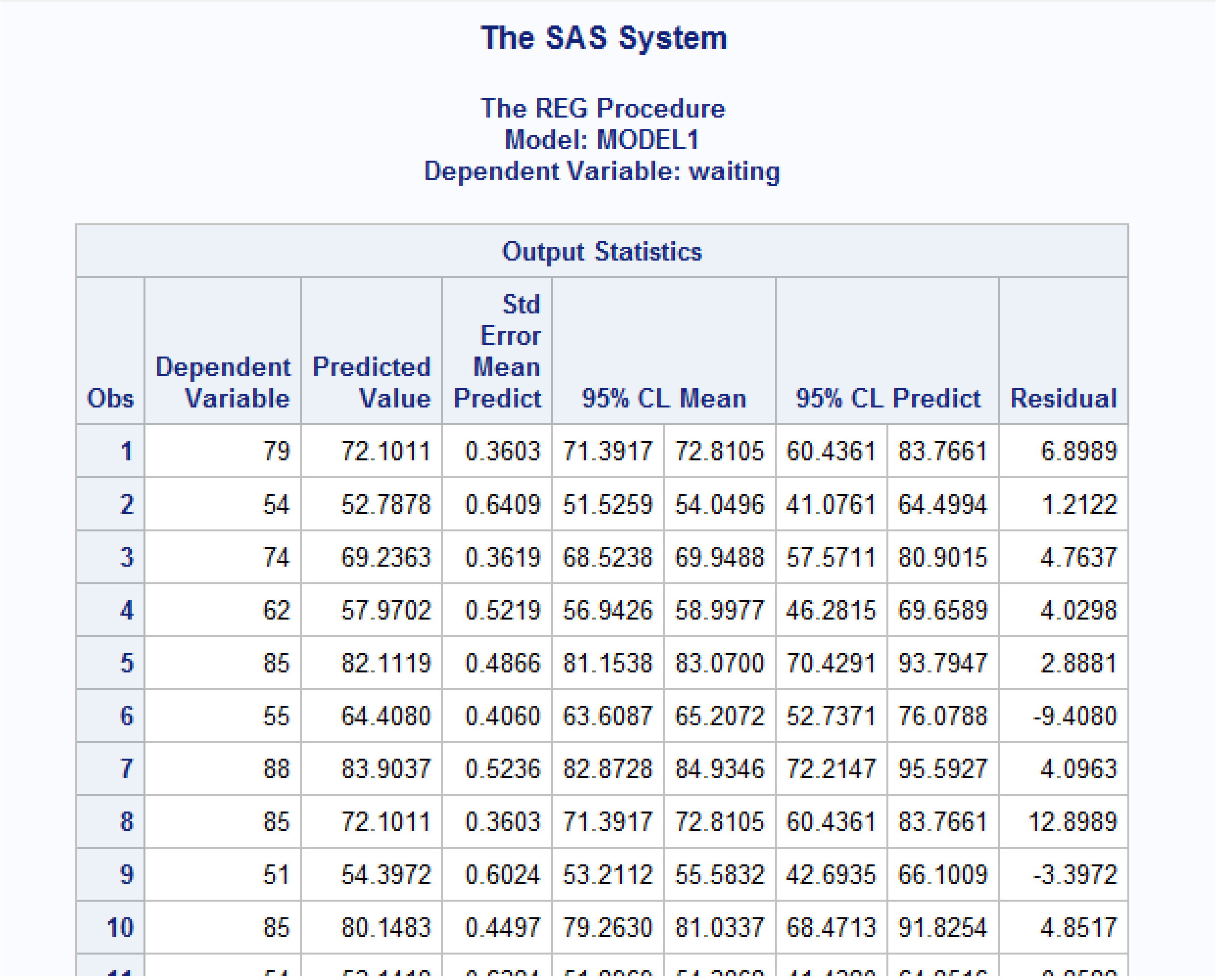
```
PROC REG data=SAS-dataset <options>;
MODEL dependent(s)=regressor(s) </options>;
...;
RUN;
```

MODEL statement options:
`clm`, `cli`, `clb`: for CIs (`clm`), PIs (`cli`), and CIs on β_k

`clm`, `cli` options

```
proc reg data=faithful;
model waiting=eruptions / clm cli clb;
run; quit;
```

slr6



The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: waiting

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	79	72.1011	0.3603	71.3917	72.8105	60.4361	83.7661	6.8989
2	54	52.7878	0.6409	51.5259	54.0496	41.0761	64.4994	1.2122
3	74	69.2363	0.3619	68.5238	69.9488	57.5711	80.9015	4.7637
4	62	57.9702	0.5219	56.9426	58.9977	46.2815	69.6589	4.0298
5	85	82.1119	0.4866	81.1538	83.0700	70.4291	93.7947	2.8881
6	55	64.4080	0.4060	63.6087	65.2072	52.7371	76.0788	-9.4080
7	88	83.9037	0.5236	82.8728	84.9346	72.2147	95.5927	4.0963
8	85	72.1011	0.3603	71.3917	72.8105	60.4361	83.7661	12.8989
9	51	54.3972	0.6024	53.2112	55.5832	42.6935	66.1009	-3.3972
10	85	80.1483	0.4497	79.2630	81.0337	68.4713	91.8254	4.8517

Assumption 1: $E(\epsilon_i) = 0$

Mean of the residuals is 0. For this, we look at a histogram of residuals to see if it is centered around zero (see if the histogram has the highest bar at zero)

Assumption 2: $V(\epsilon_i) = \sigma^2_\epsilon$

The variance of the residuals is constant (the same) for all values of $\hat{y}$. The plot of x=predicted and y=residuals and it should have no discernible pattern (random scatter)

Assumption 3: $Cov(\epsilon_i, \epsilon_i') = 0$

The covariance of any two residuals is equal to 0. Covariance of 0 implies that the two variables are independent. The Durbin-Watson (DW) test will find out if the residuals are independent. If $1.5 \leq DW \leq 2.5$ then the residuals are independent.

Assumption 4: $\epsilon_i \sim N(0, \sigma^2_\epsilon)$

Normality of residuals means that the histogram of residuals should be approximately symmetric/bell-shaped or that the QQplot (normal probability plot) shows that most points are along y=x line

Analysis of variance (ANOVA or AOV)

The methods learned for one- and two-sample only dealt with comparisons of two means or proportions. The question is, why not just do several 2-sample tests if we have at least two means? The reason is the Type I error, α, $\alpha = P(Reject\ H_0|H_0\ true)$ (rejecting a true null hypothesis). By doing several 2-sample t-tests simultaneously, since they would not be wholly independent, it increases the Type I error rate.

Analysis of variance

As an example, the number of 2-sample comparisons is the number of factor (treatment) groups choose 2 (as in a combination), $\binom{k}{2}$ where k is the number of factor groups and 2 because we are doing 2-sample comparisons. So if we had say $k = 4$ groups, then the number of comparisons to do in that case would be $\binom{4}{2} = \frac{4!}{2!(4-2)!} = 6$, *each having their own Type I error rate of 5%*, meaning that the overall Type I error rate for the entire experiment would be $6(0.05) = 0.3$. The ANOVA procedure protects the Type I error rate from inflating by doing multiple tests.

Hypotheses

The hypotheses for a (1-way) ANOVA for CRD (completely randomized design). The hypotheses only state that there are (or are not) differences among the factor group means **but does not indicate *where* the differences are, just if there are some**

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ or

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$$

$$H_a : H_0 \text{ not true (or at least one } \mu_i\ differs\ (or\ \alpha_i \neq 0))$$

The model

ANOVA uses a linear model to fit the data

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

y_{ij}: response (ith factor level, jth replicate (observation))
μ: the overall (grand) mean
α_i: the treatment (factor) effect (a slope); there is a different treatment effect (slope) for each factor level
ϵ_{ij}: residual error term

Residuals

The residuals are the errors; the difference between the observed value in the dataset and the estimated value from our model we fit (through the anova process). A residual is

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

e_{ij}: sample residual
y_{ij}: observed value of y
$\hat{y}_{ij}$: estimated value of y from estimated model

The residuals and estimated values are used in diagnostics for checking assumptions

Anova terms I

The results of the analysis is displayed in a table. Shown below is the generic version of the table and the following slides will define and give formulas for the values of the ANOVA output table.

Source	df	SS	MS	F	`Pr>F`
Treatment (Factor)	k-1	SST	MST	$=\frac{MST}{MSE}$	$P(F > F_{calc})$
Error (Residual)	n-k	SSE	MSE		
Total	n-1	TSS			

Anova terms II

Most of the calculations involve figuring out the variation (variances) between groups, within groups, and the total variation.

Sources of variation: (a) Factor (between), (b) Error (within, residuals), and (c) Total

Sums of squares (basically numerators of variances): (a) Factor or Treatment ($SS(Factor)$ or SST): sum of squared distances between each factor mean ($\overline{y}_i$) and the overall (grand) mean ($\overline{y}..$ or $\overline{\overline{y}}$), (b) Error ($SS(Error)$ or SSE): sum of squared distances between each individual observation (y_{ij}) and their corresponding factor mean ($\overline{y}_i$), and (c) Total($SS(Total)$ or TSS): sum of squared distances between each individual observation (y_{ij}) and the grand mean ($\overline{y}..$)

Anova terms III

Degrees of freedom (df): (a) Factor: $df_1 = k-1$ where k is the number of factor groups, (b) Error: $df_2 = n-k$ where n is the total number of observations in the experiment, and (c) Total: $df_{total} = n-1$

Mean squares (basically variances): (a) Factor ($MS(Factor)$): variance for factor is sum of squares for factor divided by the factor degrees of freedom (df_1), (b) Factor ($MS(Error)$): variance for error is sum of squares for error divided by the error degrees of freedom (df_2); also computed by the sum of each group variance multiplied by each group sample size minus 1,and (c) Total: could be calculated in the same manner but is not usually calculated nor used

Anova terms IV

The main goal of anova is to calculate the sums of squares (SS), mean square (MS), and the test statistic. The following are the calculations for all the values needed for the hypothesis test.

$$SS(Factor) = SST = \sum n_i(\overline{y}_i - \overline{y}..)^2$$

$$SS(Error) = SSE = \sum (y_{ij} - \overline{y}_i)^2 = \sum s_i^2(n_i - 1)$$

$$SS(Total) = TSS = \sum (y_{ij} - \overline{y}..)^2 = SST + SSE$$

Anova Terms V

Mean squares

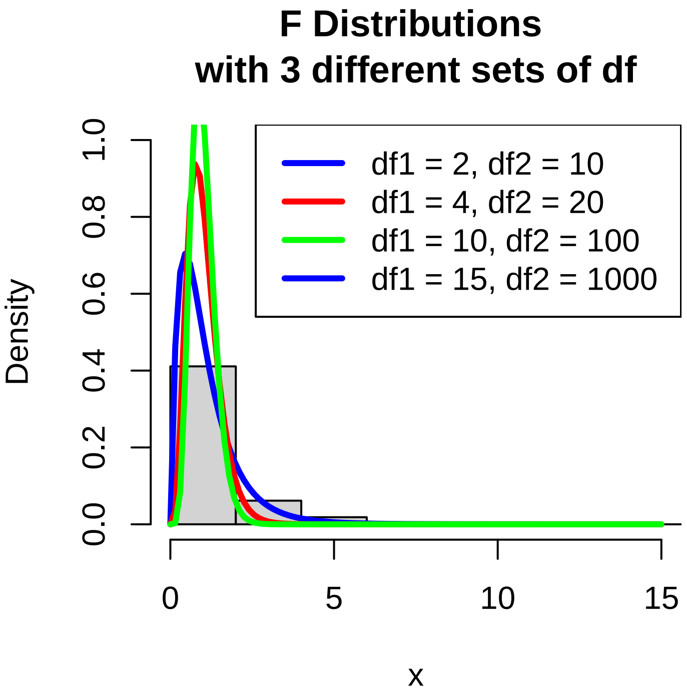
$$MST = \frac{SST}{df_1} = \frac{SST}{k-1}$$

$$MSE = \frac{SSE}{df_2} = \frac{SSE}{n-k}$$

There is no calculation of the Total mean square.

Anova Terms VI

Test Statistic: called an F statistic from the F probability distribution. Like the χ^2 distribution, F changes shape as *df* (2 of them) vary. The first *df* is df_1 and the second is df_2 from the ANOVA output table. The rows of the distribution table are df_1 and the columns are df_2.



Anova Terms VII

$$F = \frac{SST/(k-1)}{SSE/(n-k)} = \frac{MST}{MSE}$$

$$pvalue = P(F > F_{calc,df_1,df_2})$$

$$reject\ H_0\ if\ pvalue \leq \alpha$$

The F distribution has two degrees of freedom, df_1 and df_2. $df_1 = k-1$ and $df_2 = n-k$

Assumptions of ANOVA

(1) $E(\epsilon_{ij}) = 0$; the mean of the residuals should be approximately 0
(2) $V(\epsilon_{ij}) = \sigma^2_\epsilon$; the variance of the residuals should be constant for all values of the response
(3) $Cov(\epsilon_{ij}, \epsilon'_{ij}) = 0$; independence of residuals
(4) $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$; residuals should have an approximate normal distribution with mean 0 and constant variance

Example code will show how to check the assumptions graphically. Regardless of whether or not the null hypothesis is rejected in ANOVA, assumptions need to be checked to make sure the correct model was being used.

Diagnostics

(1) $E(\epsilon_{ij}) = 0$; hisotgram of residuals is centered around 0
(2) $V(\epsilon_{ij}) = \sigma^2_\epsilon$; residuals vs. predicted plot has no pattern
(3) $Cov(\epsilon_{ij}, \epsilon'_{ij}) = 0$; DW stat: $1.5 \leq DW \leq 2.5$
(4) $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$; QQplot has most points along $y = x$ line or histogram of residuals is approximately normal

anova

Multiple comparison procedures are executed along with the ANOVA

```
PROC GLM <options>;
CLASS variable(s);
MODEL dependent-variable(s) = independent-variable(s) </ options>;
BY variables;
CONTRAST 'label' effect values </ options>;
ESTIMATE 'label' effect values  </ options>;
MEANS effects </ options>;
OUTPUT out=newdatasetname;
RUN; QUIT;
```

1st line options: `PLOTS`: plots=(diagnostics residuals) will have the diagnostic plots necessary to check assumptions
MEANS statement options:
`TUKEY`, `LSD`, `BON`, `SHEFFE`: multiple comparison options

anova: handwashing

Washing hands is supposed to remove potentially harmful (and definitely gross) bacteria from your hands, thus minimizing the spread of illness and other random goobers (not the goofy kind). A completely randomized design was used to study different hand-washing methods to determine if there are differences in the amount of bacteria left on hands based on method. A total of 32 subjects were randomly assigned to one of 4 methods: water only (W), regular soap (S), antibacterial soap (ABS), and alcohol spray (AS). Is there sufficient evidence that at least one hand-washing method differs in the amount of bacteria left on the hand?

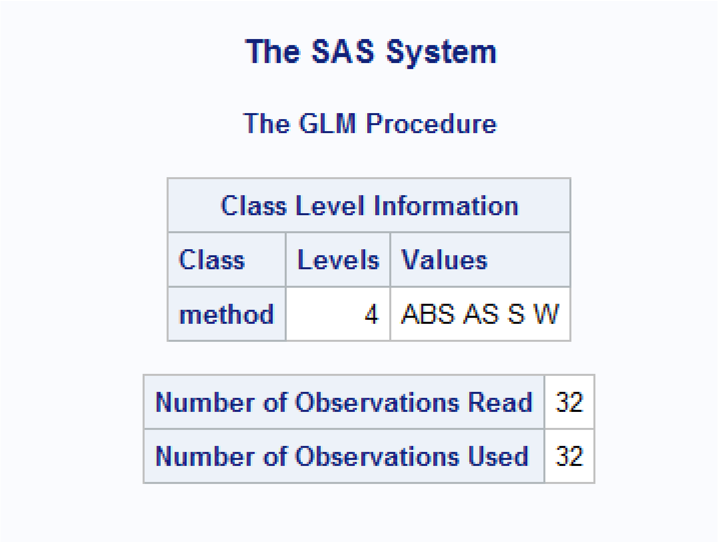
anova: handwashing

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ $(H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4)$

$H_a :$ H_0 not true

```
proc glm data=hands plots=(diagnostics residuals);
class method;
model bacteria=method;
means method / tukey lsd bon;
run; quit;
```

anova0



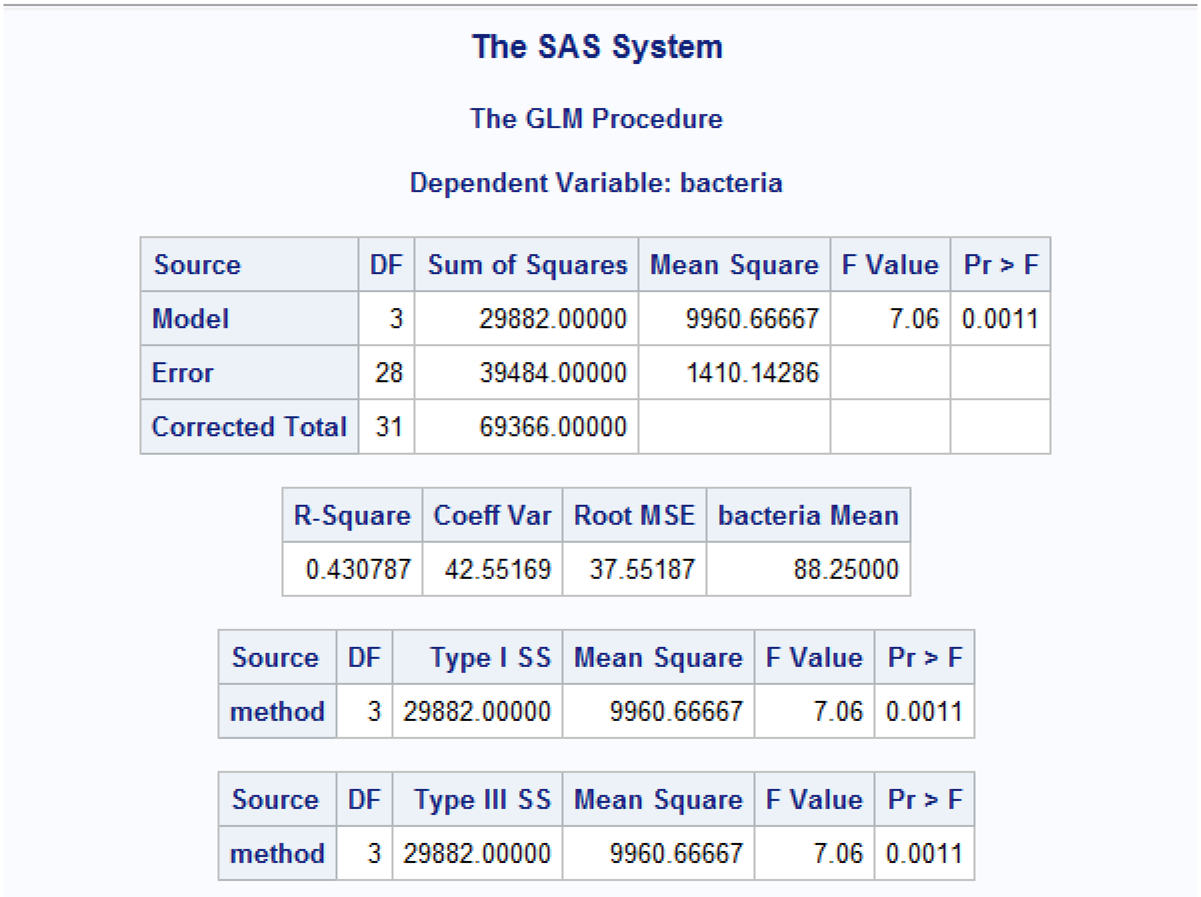
The SAS System

The GLM Procedure

Class Level Information		
Class	Levels	Values
method	4	ABS AS S W

Number of Observations Read	32
Number of Observations Used	32

anova1



The SAS System

The GLM Procedure

Dependent Variable: bacteria

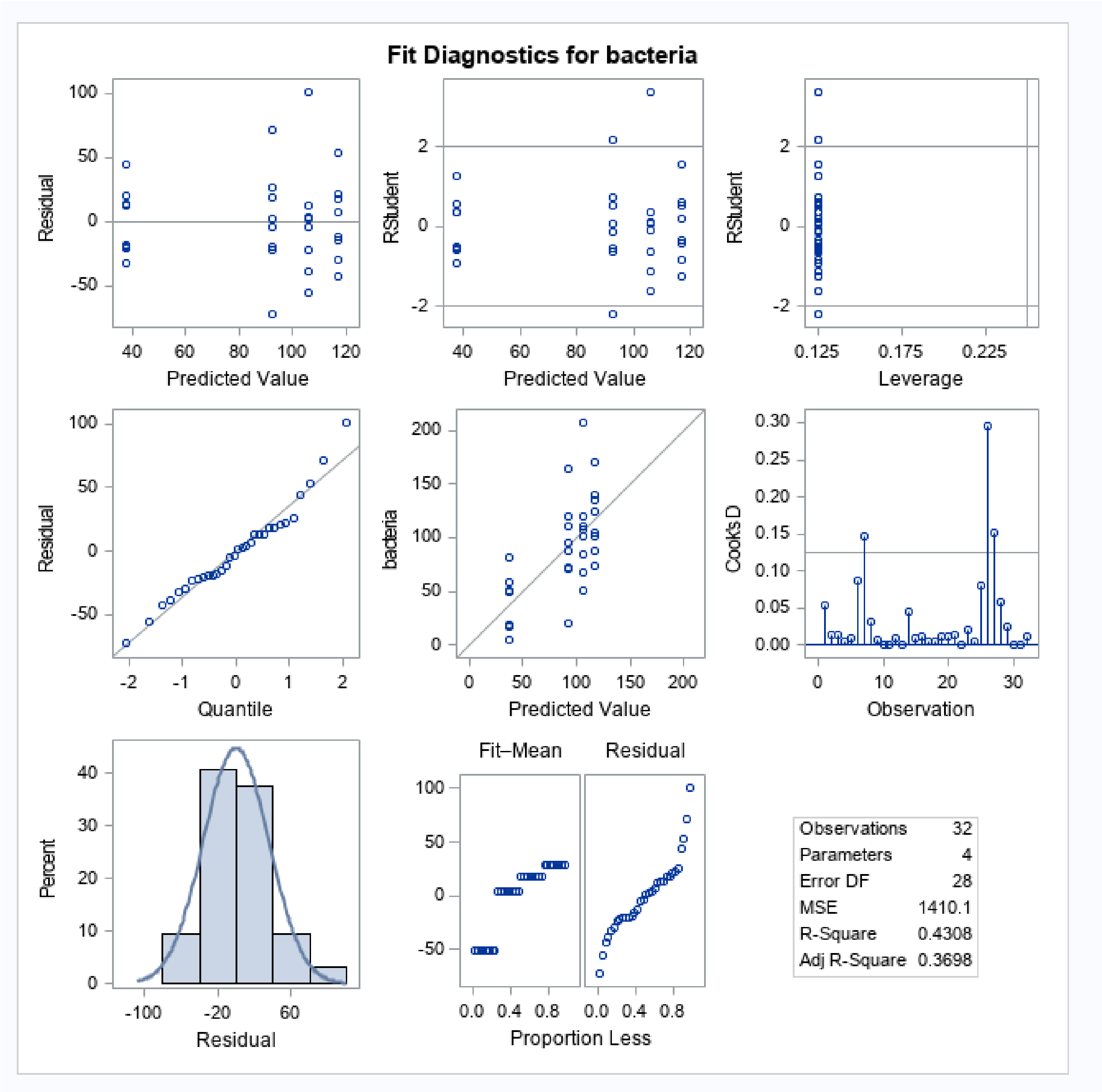
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	29882.00000	9960.66667	7.06	0.0011
Error	28	39484.00000	1410.14286		
Corrected Total	31	69366.00000			

R-Square	Coeff Var	Root MSE	bacteria Mean
0.430787	42.55169	37.55187	88.25000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
method	3	29882.00000	9960.66667	7.06	0.0011

Source	DF	Type III SS	Mean Square	F Value	Pr > F
method	3	29882.00000	9960.66667	7.06	0.0011

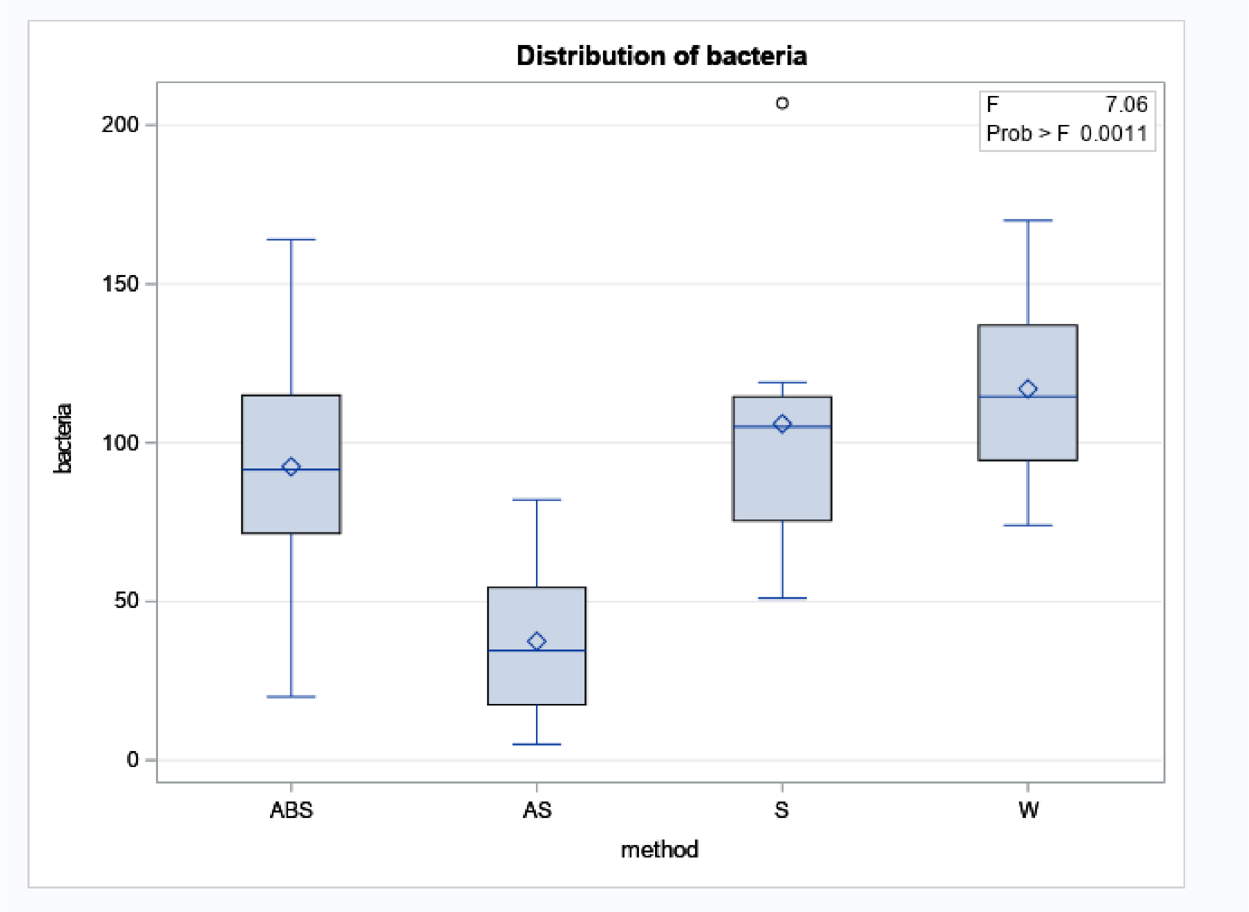
anova2



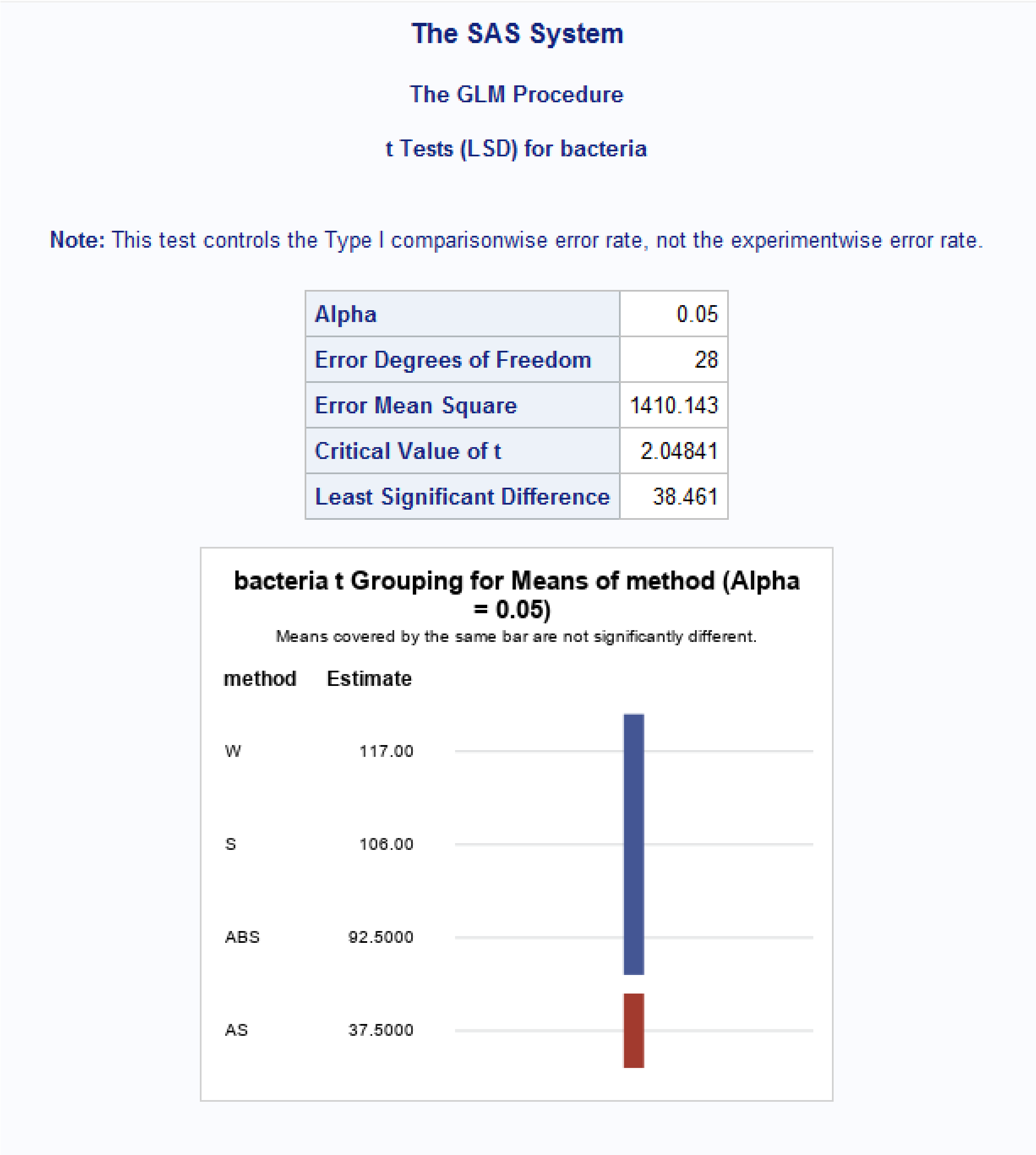
Fit Diagnostics for bacteria

```
Observations       32
Parameters          4
Error DF           28
MSE            1410.1
R-Square       0.4308
Adj R-Square   0.3698
```

anova3


Distribution of bacteria
F 7.06
Prob > F 0.0011
200
150
100
50
0
bacteria
ABS
AS
S
W
method

anova4



The SAS System

The GLM Procedure

t Tests (LSD) for bacteria

Note: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

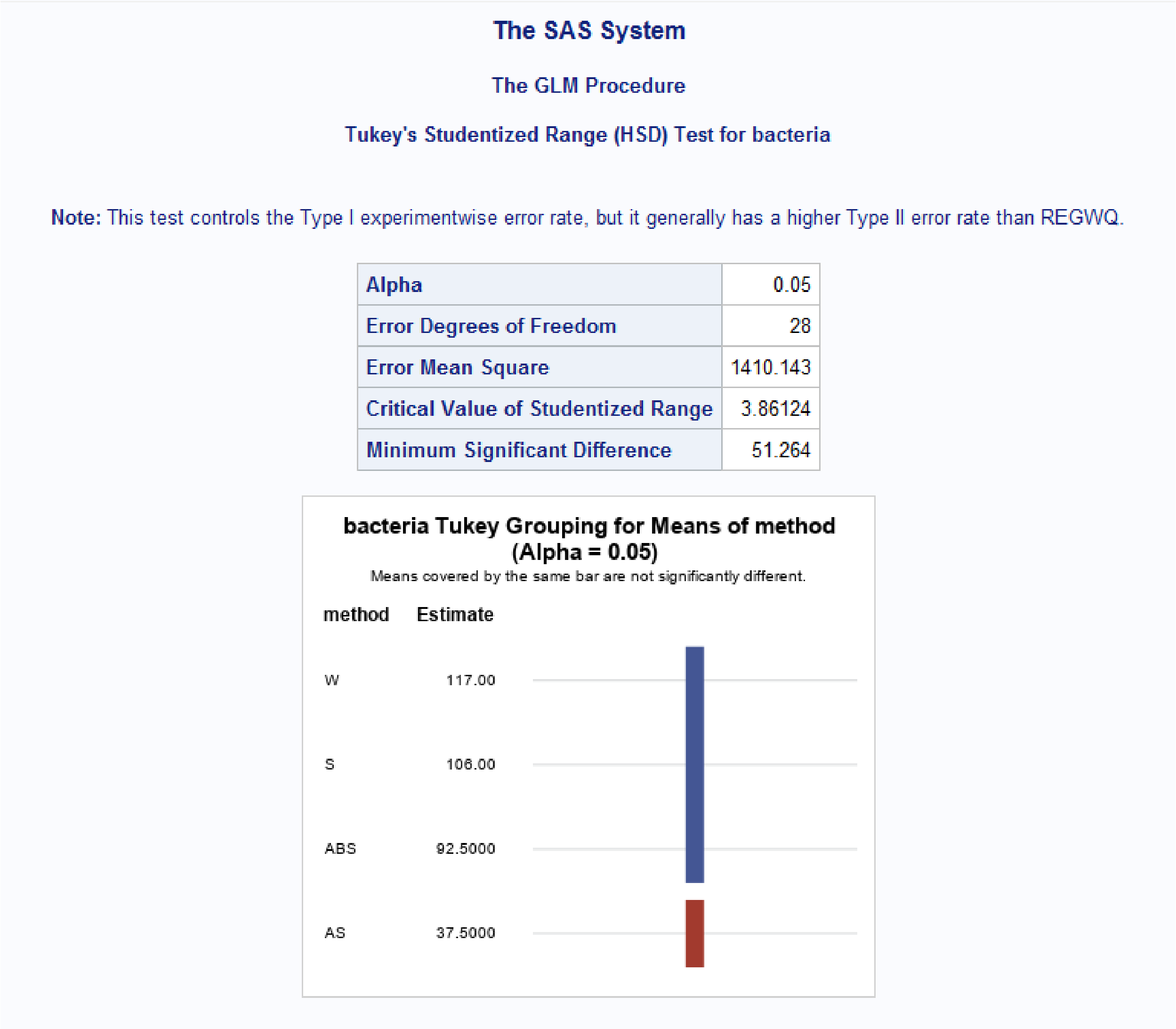
Alpha	0.05
Error Degrees of Freedom	28
Error Mean Square	1410.143
Critical Value of t	2.04841
Least Significant Difference	38.461

bacteria t Grouping for Means of method (Alpha = 0.05)

Means covered by the same bar are not significantly different.

method	Estimate
W	117.00
S	106.00
ABS	92.5000
AS	37.5000

anova5



The SAS System

The GLM Procedure

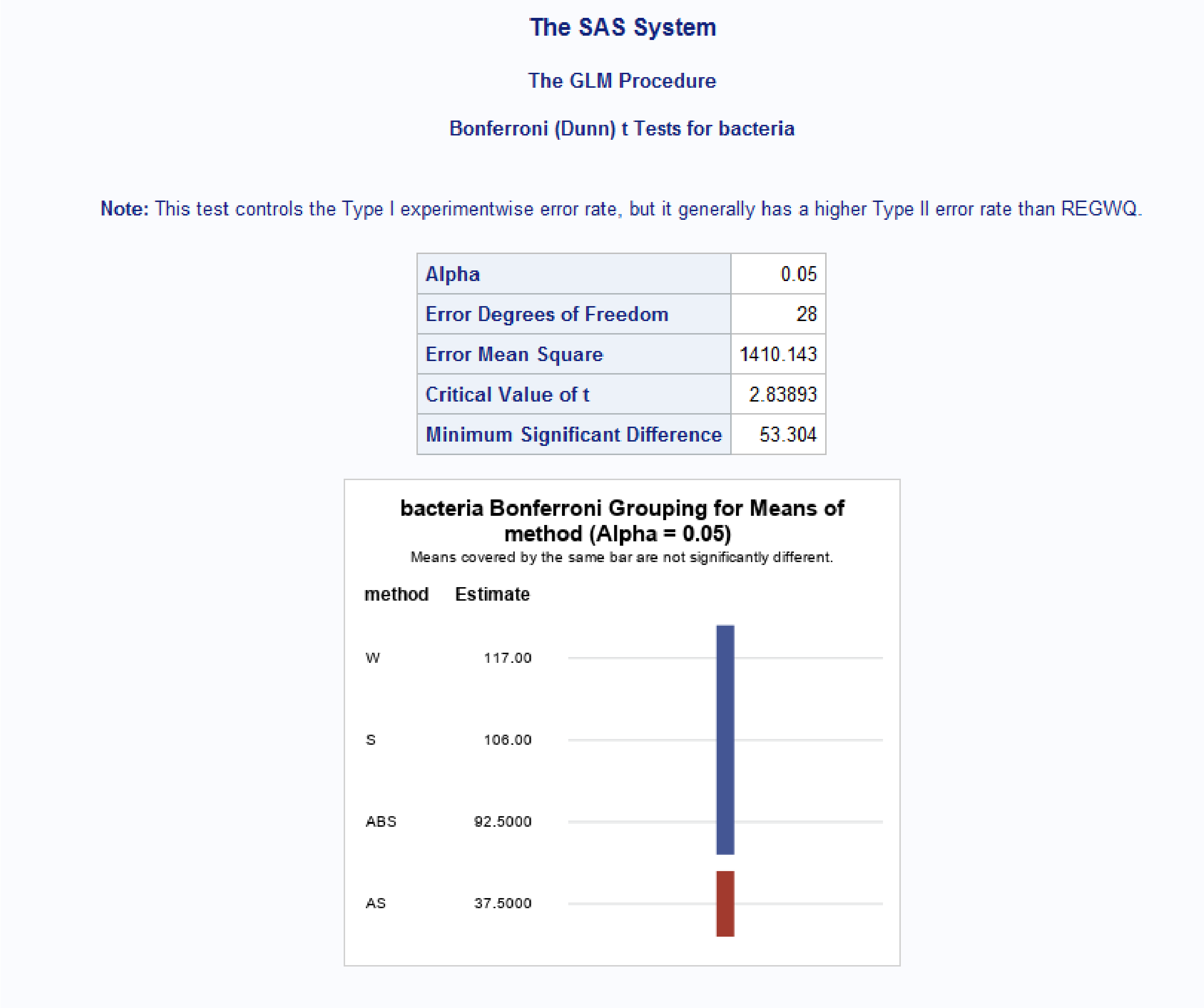
Tukey's Studentized Range (HSD) Test for bacteria

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	28
Error Mean Square	1410.143
Critical Value of Studentized Range	3.86124
Minimum Significant Difference	51.264

bacteria Tukey Grouping for Means of method
(Alpha = 0.05)
Means covered by the same bar are not significantly different.
method
Estimate
W
117.00
S
106.00
ABS
92.5000
AS
37.5000

anova6



The SAS System

The GLM Procedure

Bonferroni (Dunn) t Tests for bacteria

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	28
Error Mean Square	1410.143
Critical Value of t	2.83893
Minimum Significant Difference	53.304

bacteria Bonferroni Grouping for Means of method (Alpha = 0.05)

Means covered by the same bar are not significantly different.

method	Estimate
W	117.00
S	106.00
ABS	92.5000
AS	37.5000

anova: handwashing

Notice all the *df*, *SS*, *MS*, *F*, and *pvalue* is all input into a table of output. The main things of interest are the F value and *pvalue*. $F = 7.0636$ with $pvalue = 0.001111$. Since $pvalue = 0.001111 \leq \alpha(0.05)$, H_0 is rejected. There is at least one hand-washing method is better at removing bacteria from the hands. Another way to word it is that method of handwashing is significant.

Diagnostic graphs

The first graph should have the highest peak of the distribution be in the center around 0; it is met as long as the residual mean is close to 0. The example graph shows that the center of the distribution is centered right around 0, indicating that the mean of the residuals is approximately 0.

The second graph should show no pattern of increasing or decreasing variation or any other pattern that indicates the variance of the residuals is not constant (approximately equal for all values of y). The example graph shows that there is really no pattern (it kind of looks like there is a pattern but the vertical pattern is the grouping of the factor levels and is not a pattern to worry about).

Diagnostic graphs

The last graph is the normal probability plot of residuals, so it should indicate if the distribution is normal (approximately). The graph is created by plotting the sample quantiles of the data against the theoretical quantiles the data should have if the data is normal. If it is normal, most points should line up along the $y = x$ line without too many deviations or weird curves. The example graphs show most points are along the line so the residuals have an approximate normal distribution.

The assumptions are all met

Multiple Comparisons

Multiple comparisons are *only* to be done, if and only if (iff) the null hypothesis of ANOVA is rejected. (If the null is not rejected, you are saying there are no differences, so why would you try and find where the non-existent differences are?!?)

So now that we have seen an example of rejecting the null hypothesis of an ANOVA problem, we can just look and see if there are differences, right? Nope! That would be too easy, wouldn't it?

Multiple Comparisons

On an earlier slide from this lecture, the Type I error rate would increase, depending on how many 2-sample comparisons we do? That is why. The hand-washing example with $k = 4$ would require $\binom{k}{2} = \binom{4}{2} = 6$ 2-sample comparisons, and the larger k is, the more comparisons to do and the larger Type I error without a modified procedure to execute the comparisons.

There are many different multiple comparisons, we will learn one of the more commonly used ones called Tukey's Honest Significant Difference (Tukey's HSD).

Tukey's (not turkey) HSD

This is a modified 2-sample CI that uses a different statistical distribution called the Studentized Range distribution, denoted as $q_\alpha(k, df_2)$. You will not have to use the distribution, just interpret the output from the comparison.

Any pair of means will be determined to be significantly different if the magnitude of their difference is greater than the cutoff value, which is in essence a bound (margin of error). That is if,

$$|\overline{y}_i - \overline{y}_j| \geq HSD \ where \ HSD = q_\alpha(k, df_2)\sqrt{\frac{MSE}{n_i}}$$

Tukey's HSD

Let's wash some hands! Now that we rejected the null hypothesis, a multiple comparison, specifically Tukey's HSD, is appropriate.

Toward the bottom of the following output, there is an section with a header that reads `Treatments with the same letter are not significantly different.`, the treatment means are listed in order (largest to smallest) and there is a column called `groups`. The letters tell you which groups are statistically different. The groups that have the *same* groups letter are statistically the *same*. Different groups letters are statistically *different*.

There is also a value close to the groups that says `Minimum Significant Difference`. The value of the HSD is what the absolute value of the difference between any 2 means needs to be greater than if we wanted to look at the comparison in CI-type formatting (we will not here but something for future classes use of statistics)

Tukey's HSD

`Minimum Significant Difference: 51.26415`. The value 51.26415 is the HSD value that the absolute value of the difference between any 2 means needs to be greater than.

The groups lettering indicates that AS (alcohol spray) has the only different letter, `b`, and is significantly different than the other methods (all other methods share the letter `a` so they are all the same)