Introductory Inferential Methods

Statistics 427: R Programming

Module 15

2020

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)

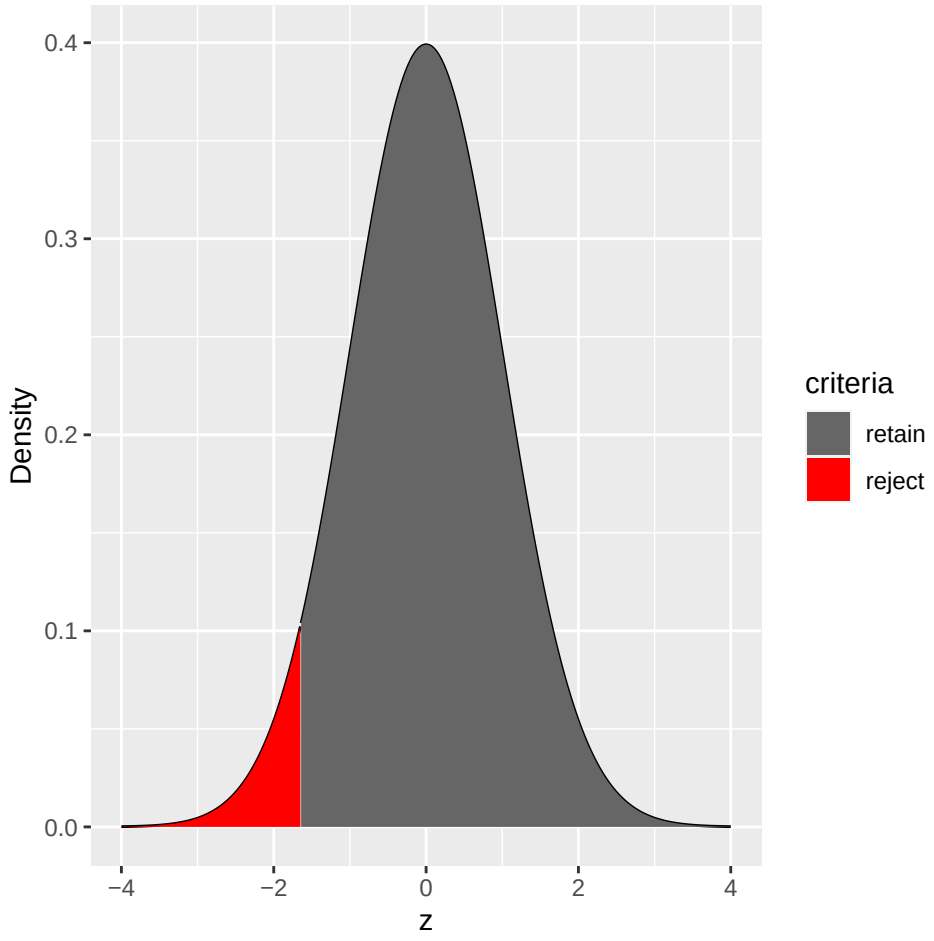
Two methods for rejection:

(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

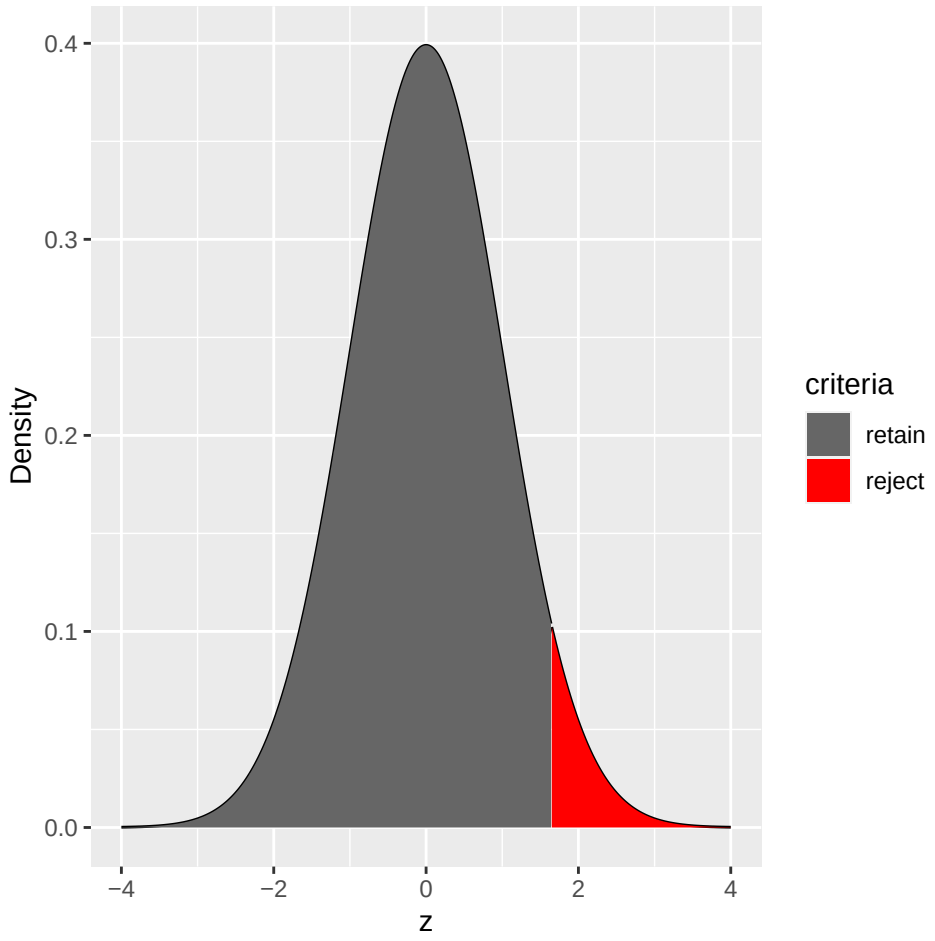
Critical Value Approach $H_a :<$

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 $H_a :>$

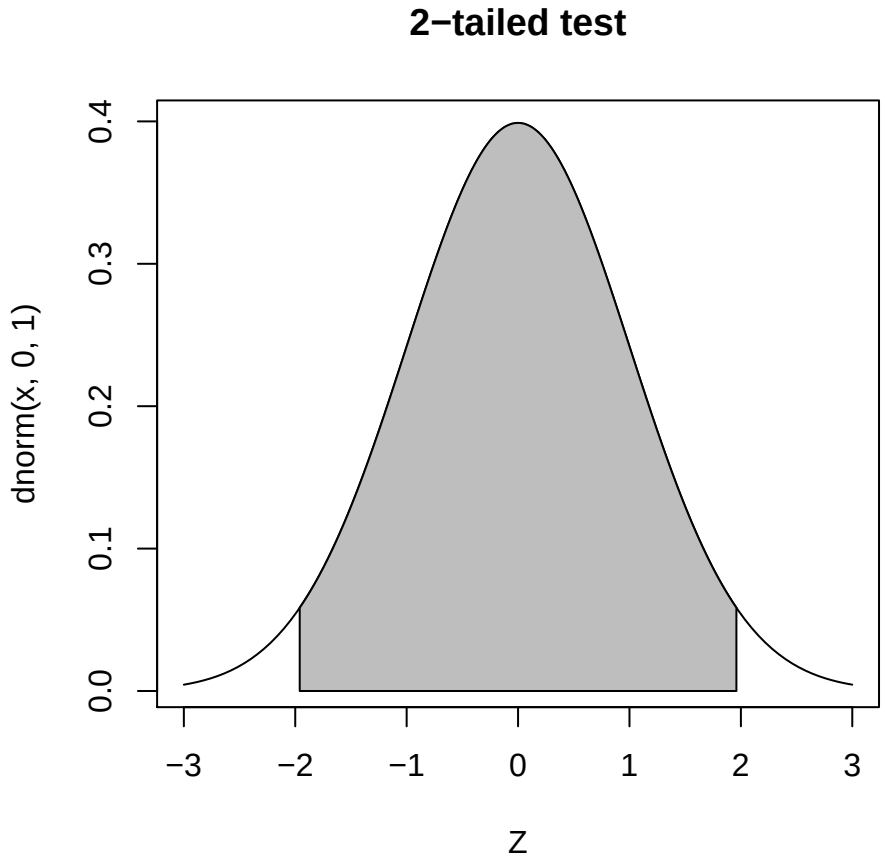
Reject H_0 iff $z_{calc} \geq z_\alpha$ (z_{calc} will most likely be a positive value and z_α *must* be positive)



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.

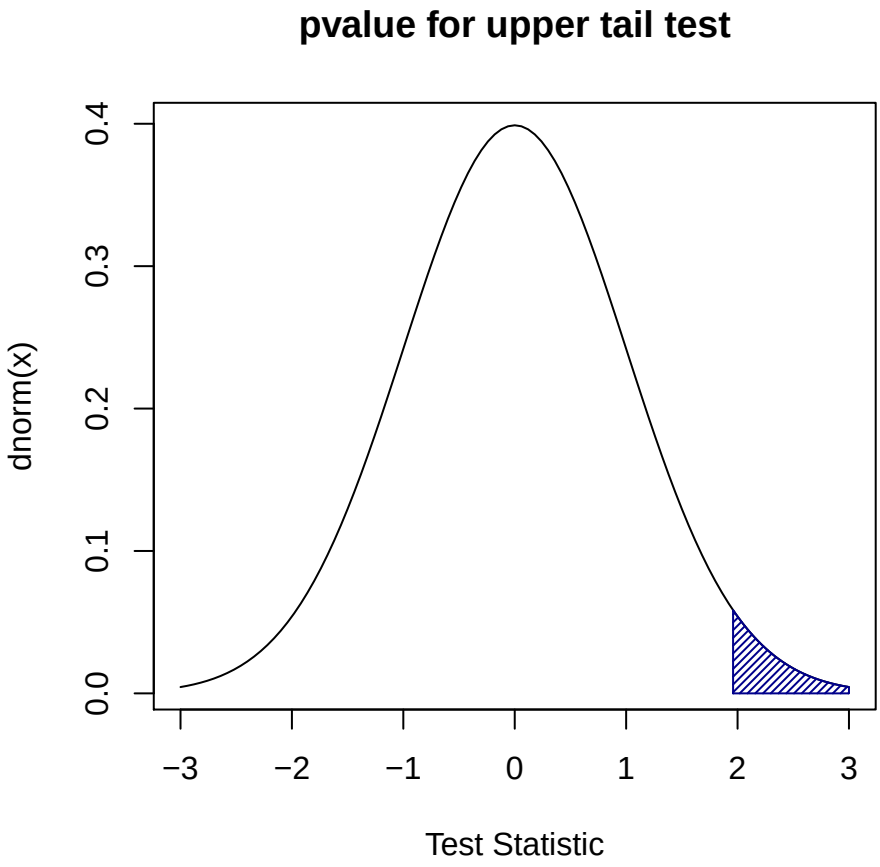
$$pvalue \leq \alpha \Rightarrow Reject\ H_0$$

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

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

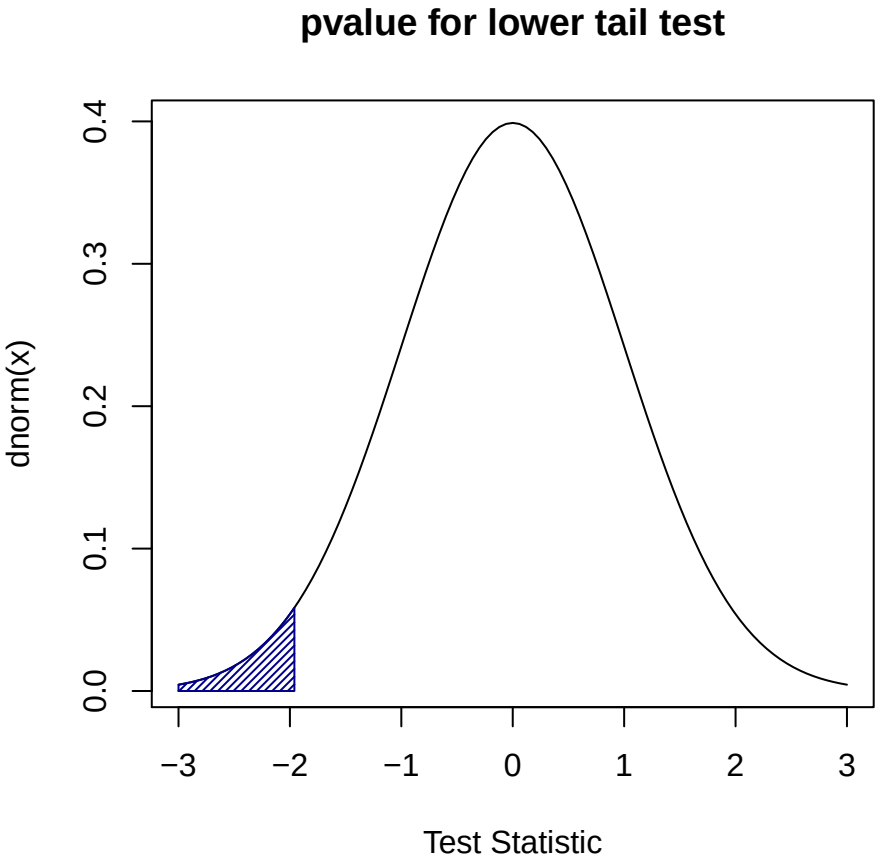
In this case, *pvalue* represents the rejection region in the right tail of the distribution.

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



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

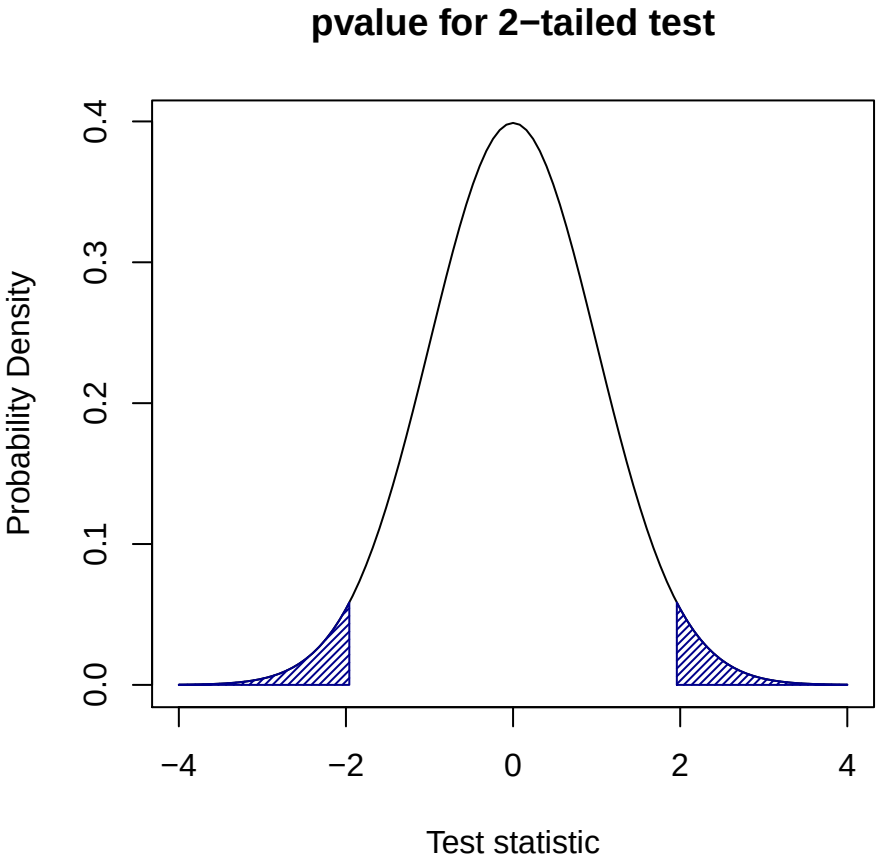
$$pvalue = P(Z \leq z_{calc})$$



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

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

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



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 `R`

The point of this course is to learn `R` 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 `R`.

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

Tests and CIs for one-sample

`t.test()` will give results for hypothesis tests and for CIs.

```
t.test(x,y,mu=,alternative=,conf.level=,...)
```

`x` and `y`: either 1 vector (`x`), 2 vectors of quantitative data, or a formula `y~x` with `y` numeric and `x` character
`mu`: `mu=0` is default; mu is hypothesized value (μ_0)
`alternative=`: `'two.sided'` (default), `'l'` (lowercase L) or `'less'`, `'g'` or `'greater'`
`conf.level=`: `0.95` is default; $CL = 1 - \alpha$
`...`: other options

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 $(lower, \infty)$

`t.test()` $H_a: \neq$

```
x=iris$Petal.Length
# test H0: mu=3.5 Ha: mu not= 3.5
t.test(x,mu=3.5)
```

```
    One Sample t-test

data:  x
t = 1.79, df = 149, p-value = 0.07549
alternative hypothesis: true mean is not equal to 3.5
95 percent confidence interval:
 3.473185 4.042815
sample estimates:
mean of x
    3.758
```

`t.test()` $H_a: >$

```
# test H0: mu=3.5 Ha: mu > 3.5 with alpha=10%
t.test(x,mu=3.5,alternative='g',conf.level=.9)
```

```
    One Sample t-test

data:  x
t = 1.79, df = 149, p-value = 0.03774
alternative hypothesis: true mean is greater than 3.5
90 percent confidence interval:
 3.57246     Inf
sample estimates:
mean of x
    3.758
```

`t.test()` $H_a: <$

```
# test H0: mu=3.5 Ha: mu < 3.5 with alpha=1%
t.test(x,mu=3.5,alternative='l',conf.level=.99)
```

```
    One Sample t-test
```

```
data:  x
t = 1.79, df = 149, p-value = 0.9623
alternative hypothesis: true mean is less than 3.5
99 percent confidence interval:
     -Inf 4.096955
sample estimates:
mean of x
    3.758
```

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 Welch test in R)
(2) Dependent means (mean difference)
(3) Two independent proportions (again, 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*)

Figuring out if $\sigma_1^2 \approx \sigma_2^2$ 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.

Variance test hypotheses

The hypotheses for this test are (always)

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

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 (Welch) 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 `R` 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

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

Variance test with `var.test()`

`var.test(x,y,alternative=,conf.level=,...)`
`x` and `y`: either 2 vectors of quantitative data or a formula `y~x` with `y` numeric and `x` character
`alternative=`: 'two.sided' (default), 'l' (lowercase L) or 'less', 'g' or 'greater'
`conf.level=`: `0.95` is default; can be modified
`...`: other options

`var.test()`

```
x=c(1,2,3,4,5); y=c(6,7,8,9,10)
var.test(x,y)
```

```
    F test to compare two variances

data:  x and y
F = 1, num df = 4, denom df = 4, p-value = 1
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1041175 9.6045299
sample estimates:
ratio of variances
                 1
```

var.test()

```
y=1:10; x=factor(c(rep('a',each=5),rep('b',each=5)))
var.test(y~x)
```

```
    F test to compare two variances

data:  y by x
F = 1, num df = 4, denom df = 4, p-value = 1
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1041175 9.6045299
sample estimates:
ratio of variances
                 1
```

Tests and CIs for independent means

`t.test()` will give results for hypothesis tests and for CIs.

`t.test(x,y,mu=,var.equal=F,paired=F,alternative=,`
`conf.level=,...)`
`x` and `y`: either 2 vectors of quantitative data or a formula `y~x` with `y` numeric and `x` character
`mu`: `mu=0` is default; mu is hypothesized value (mu-not)
`var.equal=F`: `F` is default for unpooled, `T` is for pooled
`paired=F`: `F` is default (independent means), `T` is for dependent means
`alternative=`: 'two.sided' (default), 'l' (lowercase L) or 'less', 'g' or 'greater'
`conf.level=`: `0.95` is default; can be modified
`...`: other options

t.test(): Pooled

```
x=c(1,2,3,4,5); y=c(6,7,8,9,10)
# pooled CI and test
t.test(x,y,var.equal=T)
```

```
    Two Sample t-test

data:  x and y
t = -5, df = 8, p-value = 0.001053
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.306004 -2.693996
sample estimates:
mean of x mean of y
        3         8
```

t.test() Unpooled

Also called a `Welch` test in `R`

```r
# unpooled CI and test
t.test(x,y)
```

```
    Welch Two Sample t-test

data:  x and y
t = -5, df = 8, p-value = 0.001053
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.306004 -2.693996
sample estimates:
mean of x mean of y
        3         8
```

`t.test()` Paired (dependent)

```r
# dependent means
x=c(1,2,3,4,5); y=c(2,10,4,-2,6)
t.test(x,y,paired=T)
```

```
    Paired t-test

data:  x and y
t = -0.45175, df = 4, p-value = 0.6749
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.145923  5.145923
sample estimates:
mean of the differences
                     -1
```

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

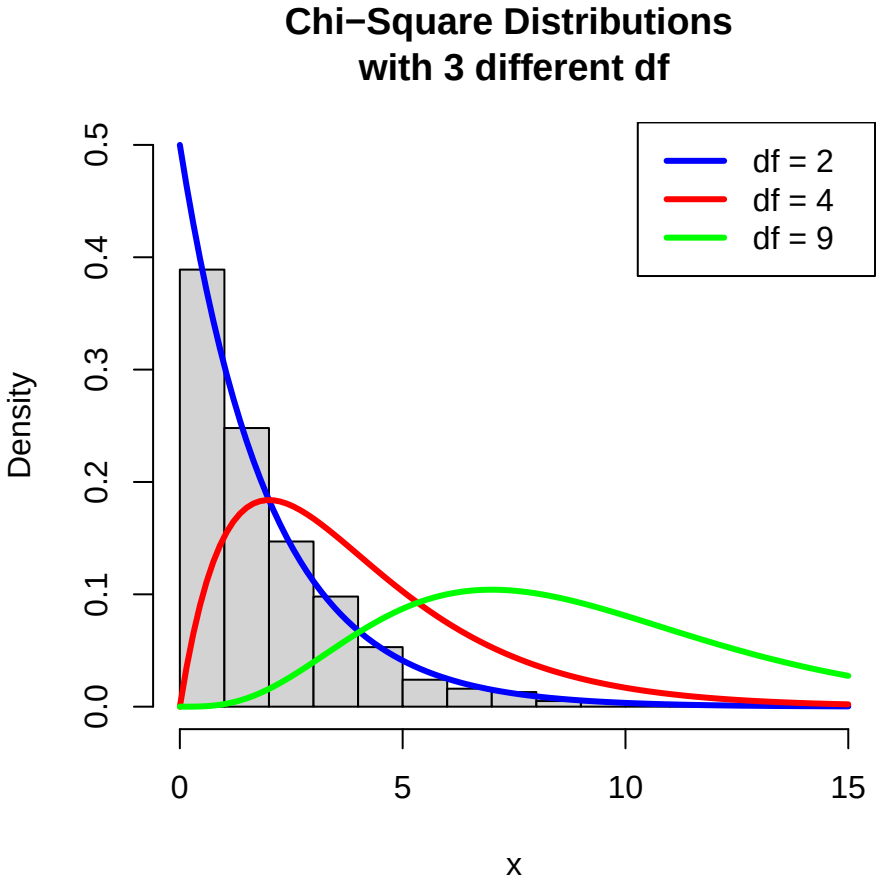
Chi-square distribution

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.



Assumptions of any Chi-square test

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

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

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.

`chisq.test()`

`chisq.test()` will give results for hypothesis tests for categorical data or proportions

`chisq.test(x,y,p,simulate.p.value=F,...)`
`x` and `y`: `x` is a numeric vector or a matrix, `y` is ignored if `x` is a matrix
`p`: a vector of probabilities as length of `x`, `p`≥ 0
`simulate.p.value=F`:`Fis default,Tif pvalue to be computed by Monte Carlo simulation if any $E_i<5$`...': other options

`chisq.test()` Independence and homogeneity

```
# independence and homogeneity
M=matrix(c(762,327,468,484,239,477),byrow=T,ncol=3,nrow=2)
dimnames(M)=list(gender=c("F","M"),party=c("Democrat","Independent","Republican"))
chisq.test(M)
```

```
    Pearson's Chi-squared test

data:  M
X-squared = 30.07, df = 2, p-value = 2.954e-07
```

`chisq.test()` GoF

```
# GoF
x=c(89,37,30,28,4)
p=c(0.40,0.15,0.15,0.19,0.11)
chisq.test(x,p=p)
```

```
    Chi-squared test for given probabilities

data:  x
X-squared = 20.516, df = 4, p-value = 0.000395
```

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

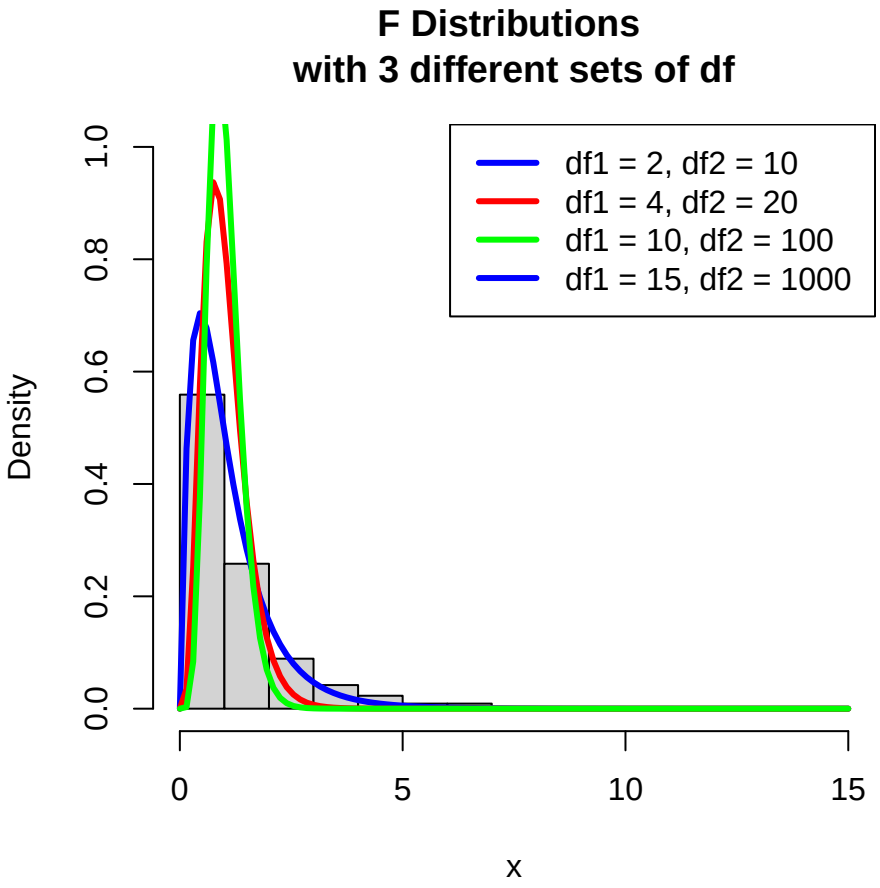
$$MST = \frac{SST}{df_1} = \frac{SST}{k-1}$$

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

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) $V(\epsilon_{ij}) = \sigma_\epsilon^2$; 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) $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$; 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 in R

There are a couple of different methods to carry out an ANOVA.

(1) `fit=lm(y~x,data=)` with `anova(fit)`
(2) `fit=aov(y~x,data=)` with `summary(fit)`

Both will give the same results. However, some multiple comparison procedures in R need the `aov()` version rather than the `lm()` version.

anova diagnostics in R

The diagnostics to check assumptions are three main graphs. The values needed to make the graphs are residuals and estimated values.

`pred=fitted(fit)`: fitted (estimated/predicted) value ($\hat{y}_{ij}$)
`res=rstudent(fit)`: residuals (standardized)
`hist(res)`: histogram of residuals to check if mean≈ 0 (and normality)
`plot(pred,res); abline(h=0)`: scatterplot of $x = pred$,$y = res$ with trendline $y = 0$ to check if variance of residuals is constant
`dwt(fit)`: DW test in `car` package; check only if time component in dataset (mostly used in regression)
`qqnorm(res); qqline(res)`: normal probability plot to check normality of residuals with $y = x$ line

anova: handwashing

Learning to read the output from a statistical software program. The following example will use output from the program R.

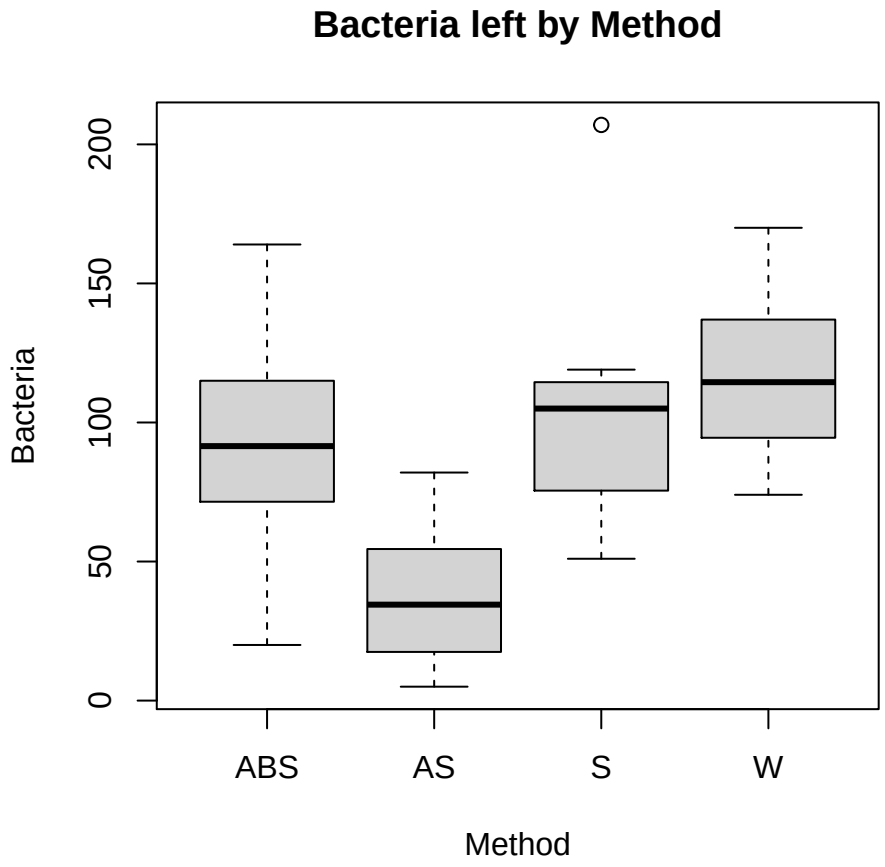
Washing hands is supposed to remove potentially harmful (and definitely gross) bacteria from your hands, thus minimizing the spread of illness and other random goobers (not the goofy kind). A completely randomized design was used to study different hand-washing methods to determine if there are differences 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

```
boxplot(Bacteria~Method,data=hands,main='Bacteria left by Method')
```



anova: handwashing

```
fit=lm(Bacteria~Method,data=hands)
anova(fit)
```

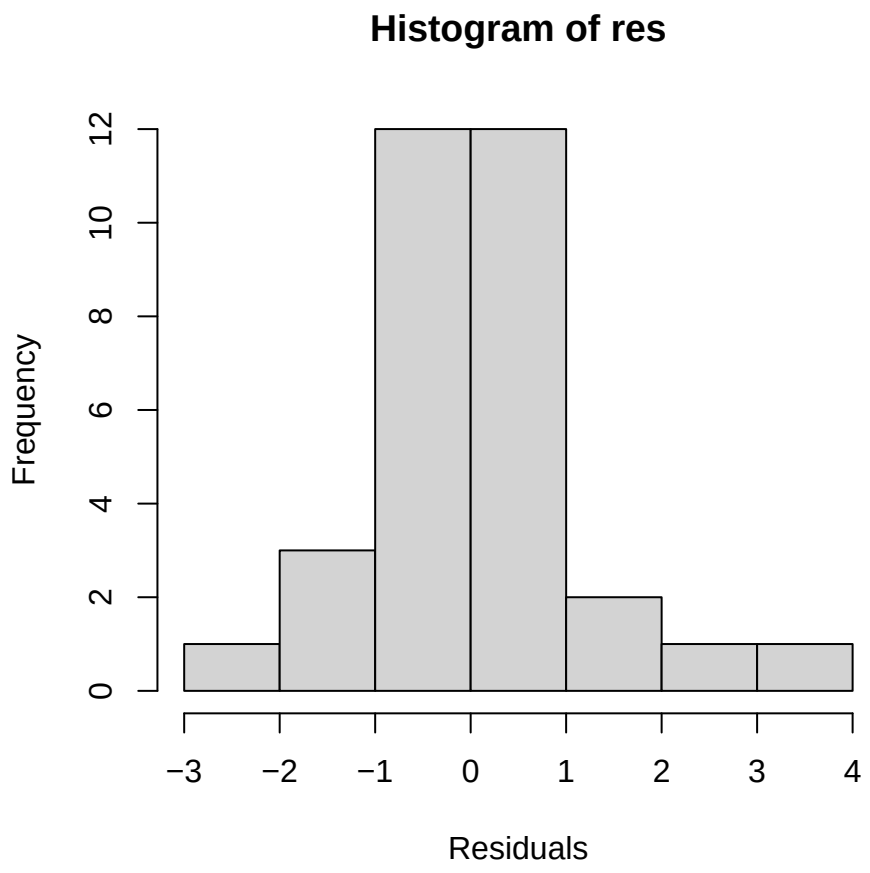
```
Analysis of Variance Table

Response: Bacteria
          Df Sum Sq Mean Sq F value   Pr(>F)
Method     3  29882  9960.7  7.0636 0.001111 **
Residuals 28  39484  1410.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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.

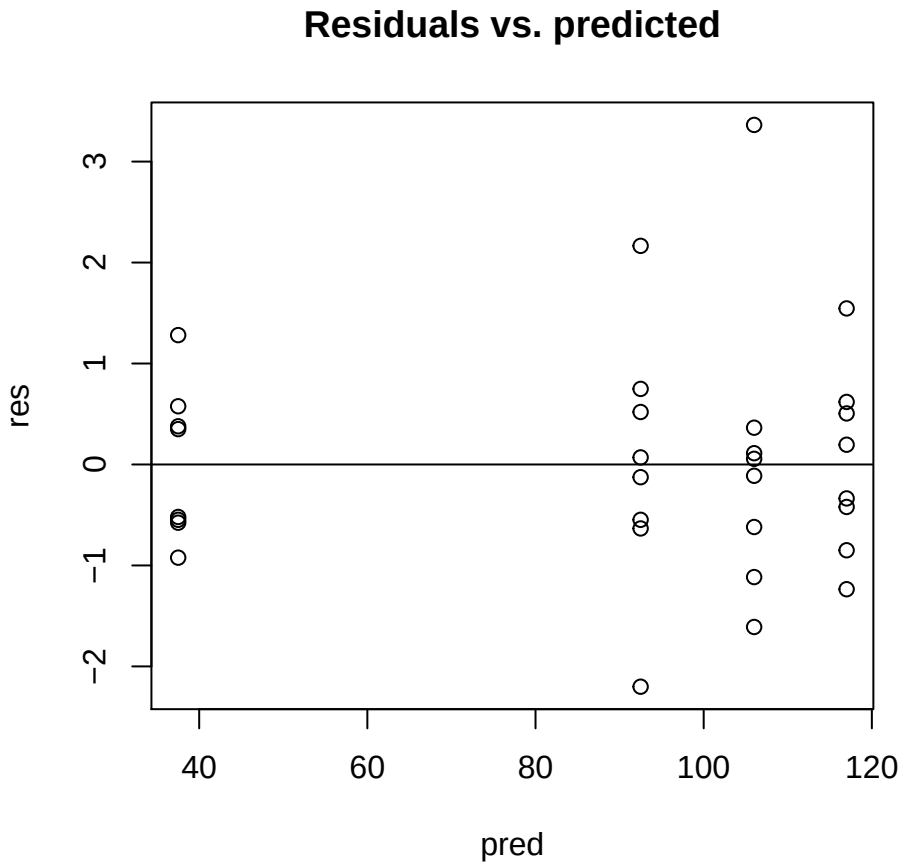
Diagnostic graphs

```
res=rstudent(fit); pred=fitted(fit)
hist(res,xlab='Residuals')
```



Diagnostic graphs

```
plot(pred,res,main='Residuals vs. predicted')
abline(h=0)
```



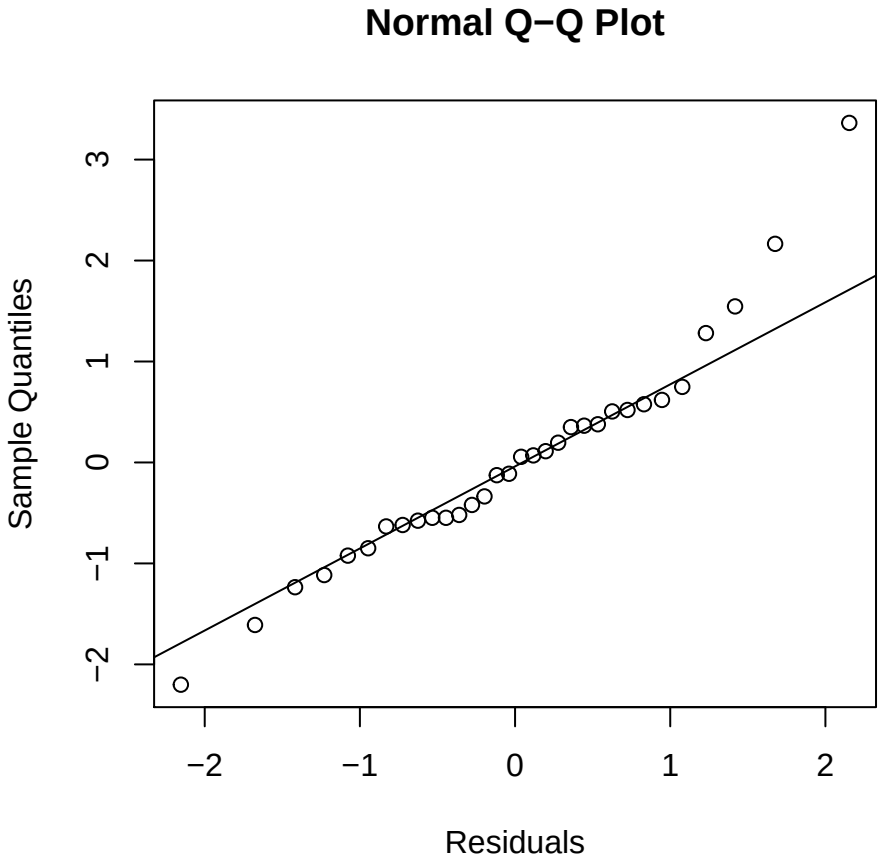
Diagnostic graphs

```
library(car)
dwt(fit)
```

```
 lag Autocorrelation D-W Statistic p-value
   1      -0.3131268      2.568781   0.058
 Alternative hypothesis: rho != 0
```

Diagnostic graphs

```
qqnorm(res,xlab='Residuals'); qqline(res)
```



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 DW test showed that the DW test statistic is within our acceptance range so the residuals are independent. Usually unless there is a time component in the dataset, this is not always necessary to check.

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,

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

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*.

Tukey's HSD

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 general form

`HSD.test(fit,'factor',group=T,console=T,...)`
`fit`: fit object of class `lm()`
`'factor'`: name of factor variable with quotes
`group=T`: use grouping letters for differences
`console=T`: display results in console

Tukey's HSD

```r
library(agricolae)
fit=aov(Bacteria~Method,data=hands) # HSD.test likes aov
HSD.test(fit,'Method',group=T,console=T)
```

```
Study: fit ~ "Method"

HSD Test for Bacteria

Mean Square Error:  1410.143

Method,  means

    Bacteria      std r Min Max
ABS     92.5 41.96257 8  20 164
AS      37.5 26.55991 8   5  82
S      106.0 46.95895 8  51 207
W      117.0 31.13106 8  74 170

Alpha: 0.05 ; DF Error: 28
Critical Value of Studentized Range: 3.861244

Minimun Significant Difference: 51.26415

Treatments with the same letter are not significantly different.

    Bacteria groups
W      117.0      a
S      106.0      a
ABS     92.5      a
AS      37.5      b
```

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).

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

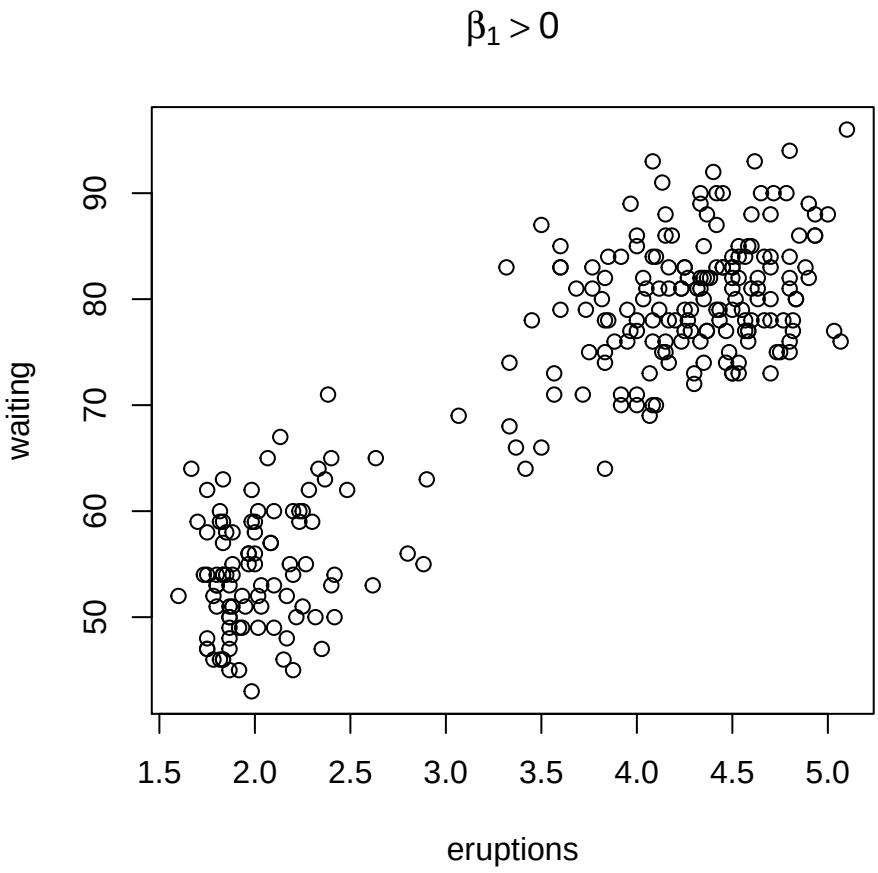
- A scatterplot of the data shows if there is a linear association between the explanatory (independent) variable and the response (dependent) variable
  - When x and y both increase, the slope (relationship) is positive
  - When x increases while y decreases, the slope (relationship) is negative

- 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)

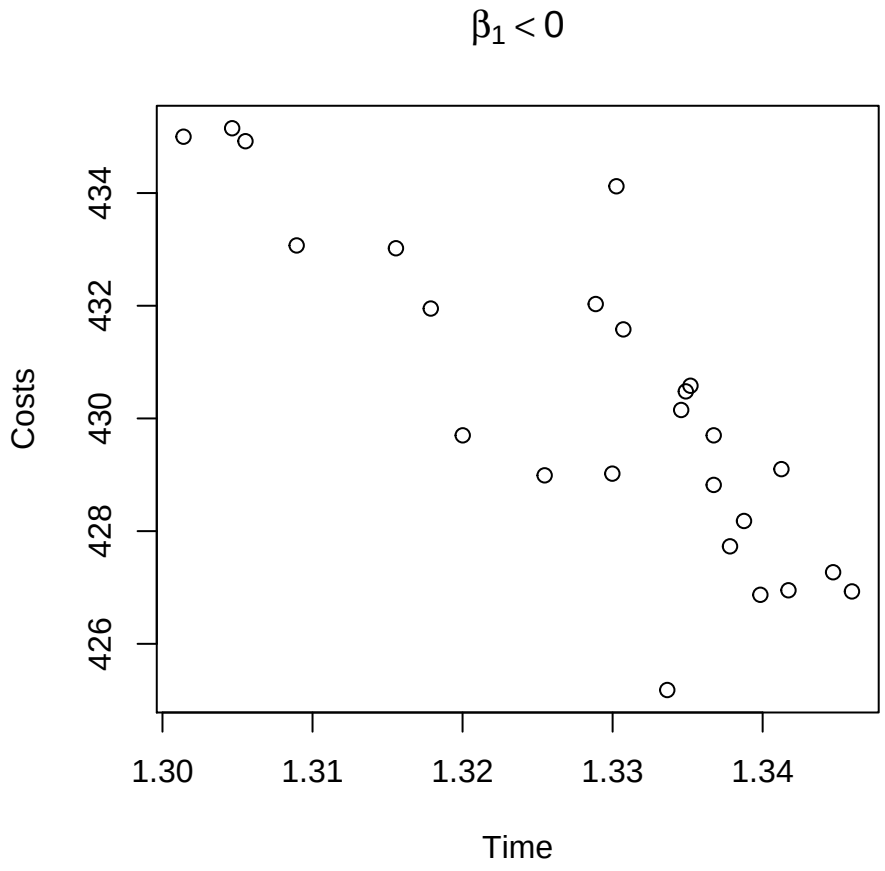
```
plot(waiting~eruptions,main=bquote(beta[1]>0),data=faithful)
```



Analysis tools: scatterplot graph

This has negative slope (x increases and y decreases)

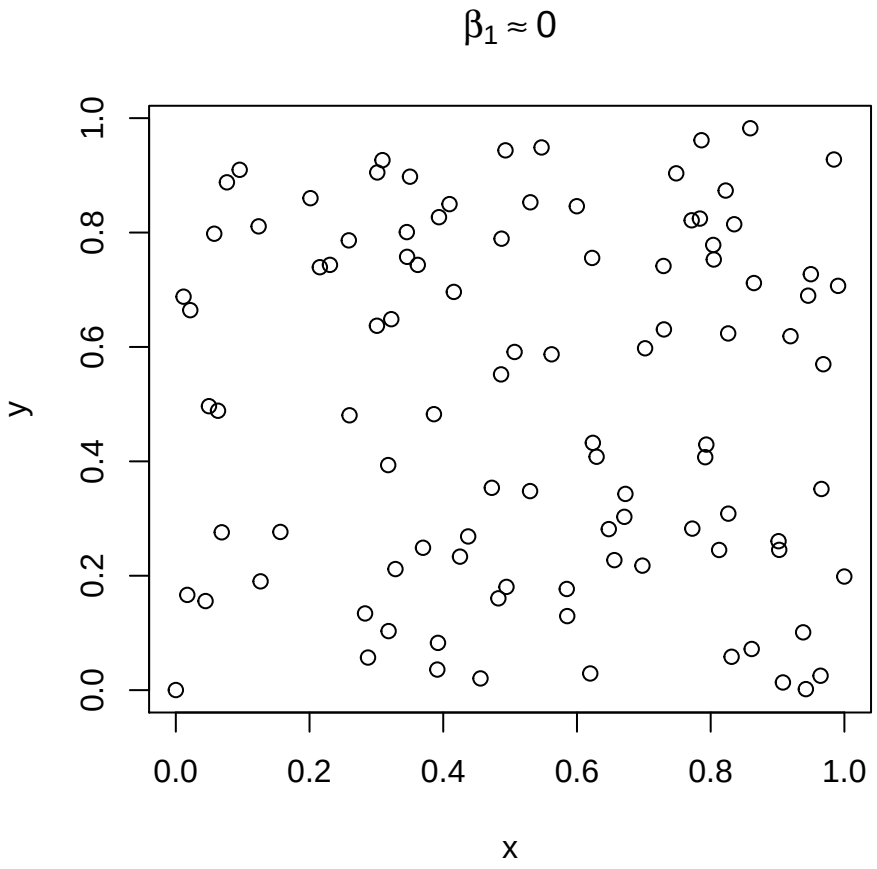
```
with(decagon,plot(Time,Costs,main=bquote(beta[1]< 0)))
```



Analysis tools: scatterplot graph

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

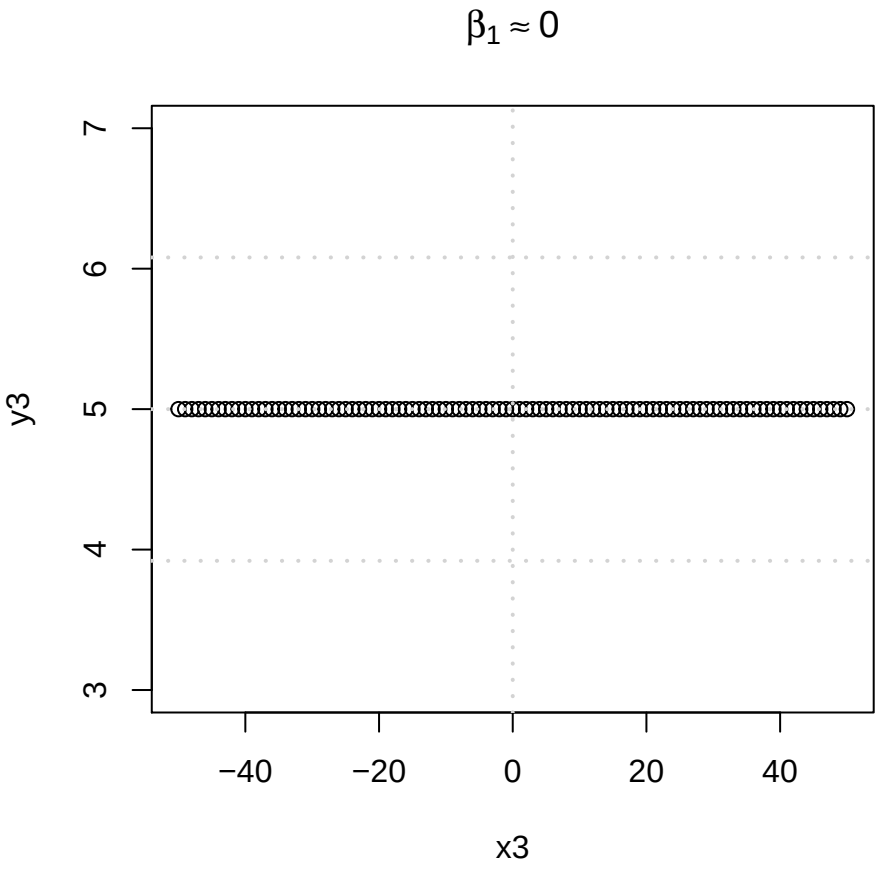
```
with(randu[1:100,],plot(x,y,main=bquote(beta[1]%~~%0)))
```



Analysis tools: scatterplot graph

This has 0 slope

```
plot(x3,y3,main=bquote(beta[1]%~~%0)); grid(2,4,lwd=2)
```

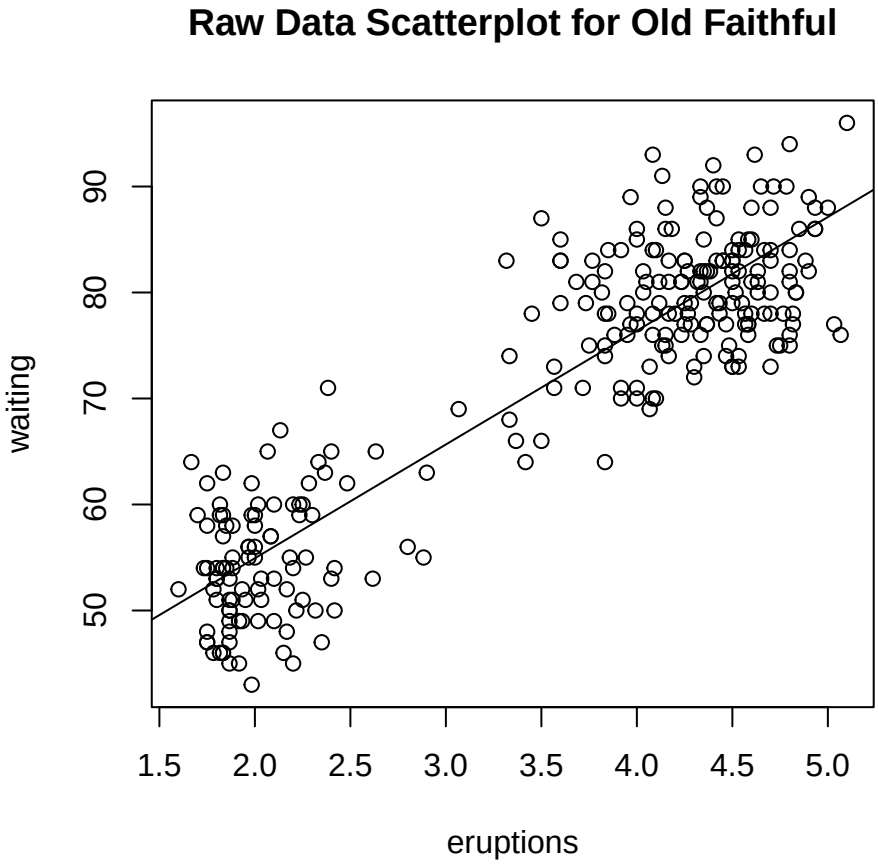


Analysis tools: scatterplot graph with regression line

Many times in regression, we want to see what the line of the regression equation will look like on the scatterplot of the raw data. It is not strictly necessary but the point of this analysis is to explore and understand the linear relationship between two variables. If you do not have a linear relationship, then use of this 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 with regression line

```
plot(waiting~eruptions,main="Raw Data Scatterplot for Old Faithful",data=faithful)
abline(lm(waiting~eruptions,data=faithful))
```



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

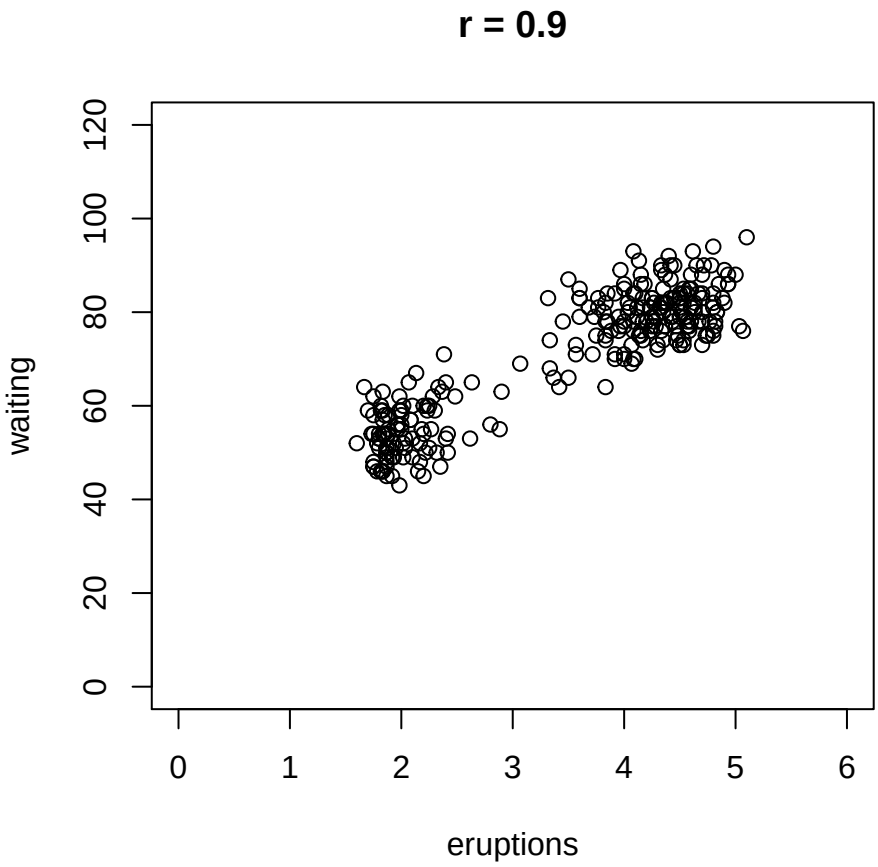
R^2 is called the *coefficient of determination*:

- It is the proportion (or $\times 100\%$) 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

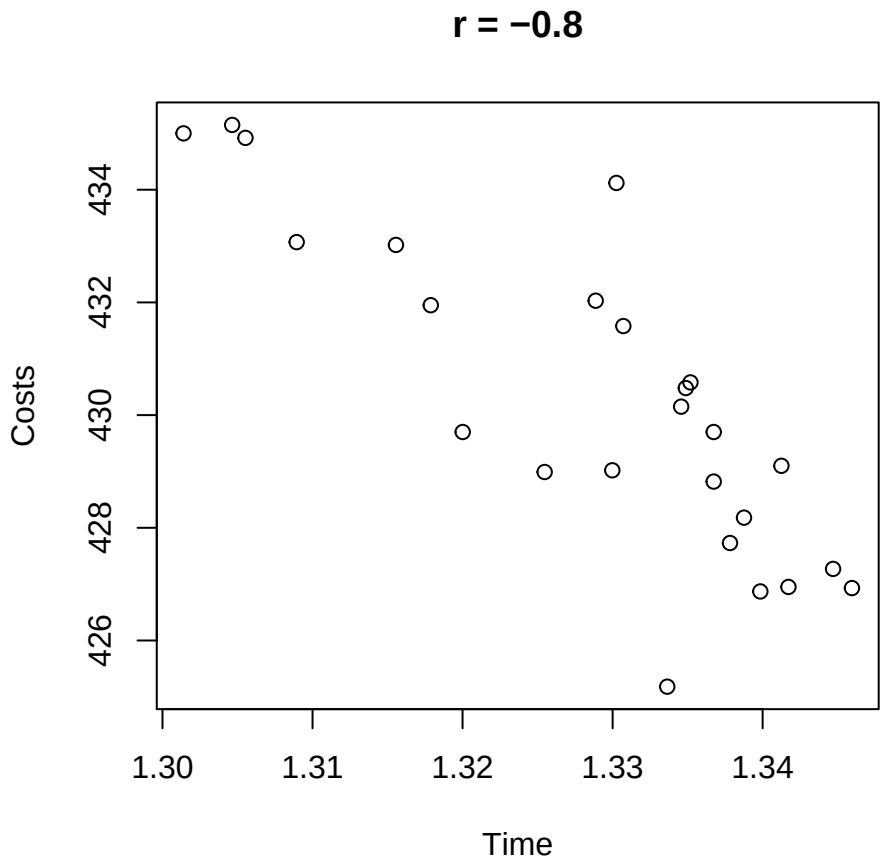
```
plot(waiting~eruptions,main="r = 0.9",data=faithful,xlim=c(0,6),ylim=c(0,120))
```



Analysis tools: scatterplot graph

Moderately strong, negative correlation

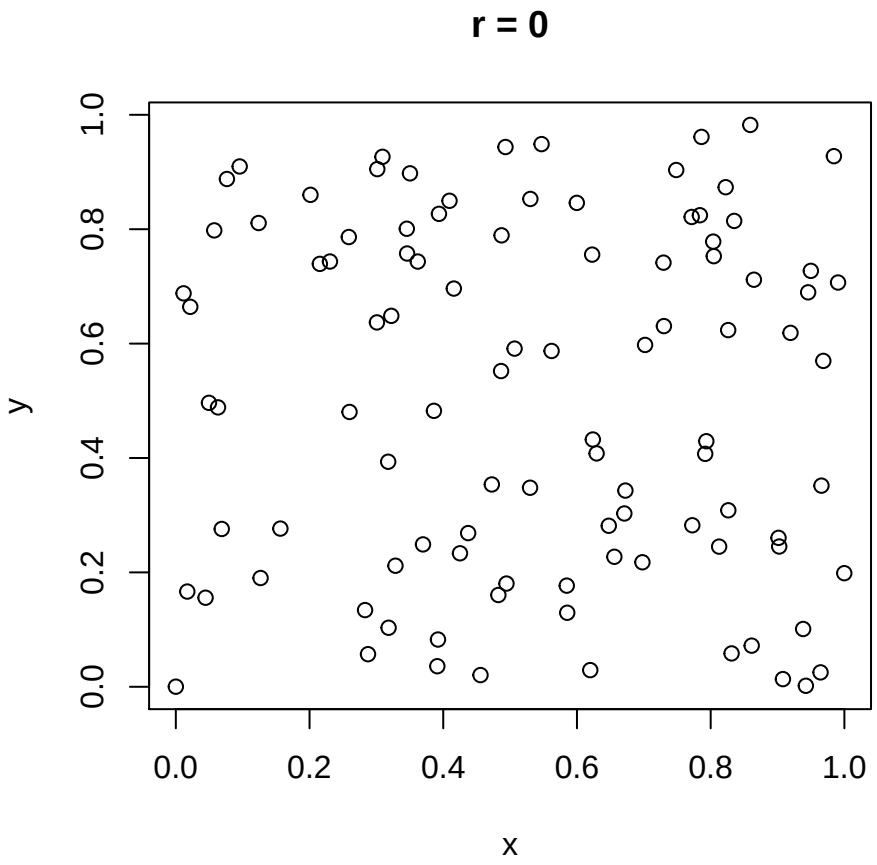
```
with(decagon,plot(Time,Costs,main="r = -0.8"))
```



Analysis tools: scatterplot graph

No correlation

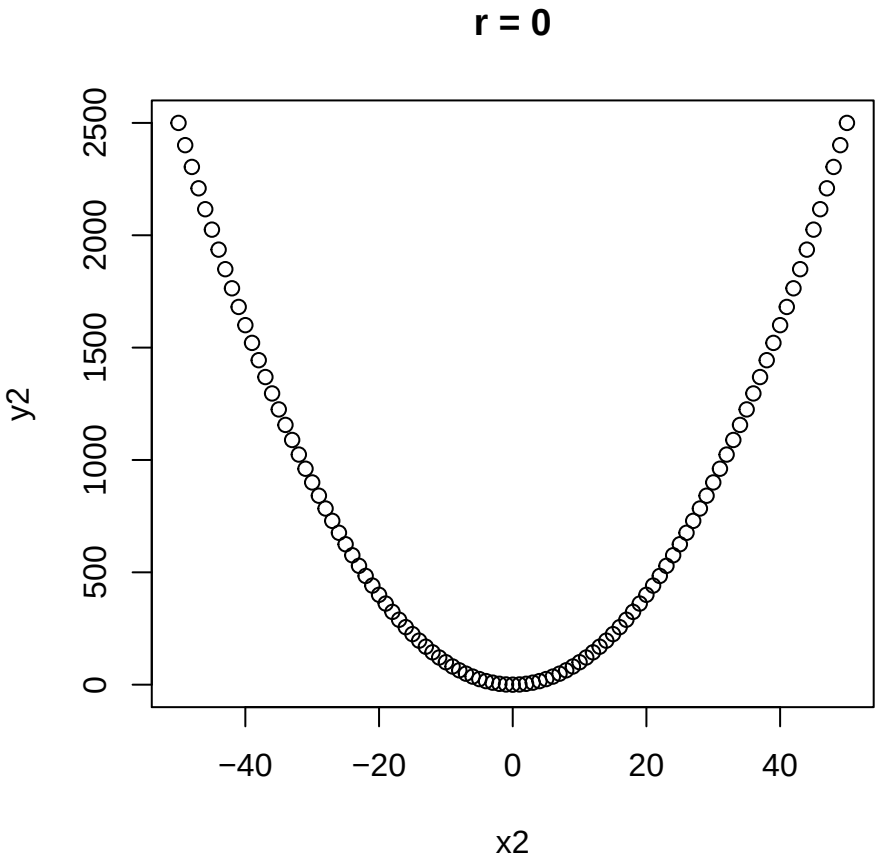
```
with(randu[1:100,],plot(x,y,main="r = 0"))
```



Analysis tools: scatterplot graph

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

```
plot(x2,y2,main="r = 0")
```



General form of `lm()` for model

First a linear model must be created. Since other objects from the model will be needed for diagnostics, use the assignment statement to create an object of the `lm()` model.

`lm(formula,data,...)`

`formula`: usually `y~x` or some derivation thereof (`y~x` for SLR, `y~x1+x2+...+` for multiple regression, etc.)
`data`: dataset object name

SLR analysis output displayed with `summary()`

The next thing once the model is run and stored as an object, to display the analysis results. `summary()` will display parameter estimates (slope and intercept) and other regression statistics. There are sums of squares calculations that are only displayed with `anova()`.

Using `lm()` and `summary()`

```
fit=lm(waiting~eruptions,data=faithful); summary(fit)
```

```
Call:
lm(formula = waiting ~ eruptions, data = faithful)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0796  -4.4831   0.2122   3.9246  15.9719
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.4744     1.1549   28.98   <2e-16 ***
eruptions    10.7296     0.3148   34.09   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.914 on 270 degrees of freedom
Multiple R-squared:  0.8115,    Adjusted R-squared:  0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

Objects in the `fit` model

There are many objects that are a part of the fit model calculations and they can be extracted for use in other calculations. An object of class "lm" is a list containing at least the values found on the following table. To access these objects, `fit$objectname` where `objectname` is one of the objects from the table on the next slide.

List of `fit` model objects

Object	Defn
coefficients	named vector of coefficients
residuals	residuals (y-yhat)
fitted.values	fitted mean values (yhat)
weights	only for weighted fits
df.residual	residual df
contrasts	only if contrasts used
...	more options

`coefficients` object

The `coefficients` are useful once extracted for use of the regression equation for estimations. The `coefficients` object is a named vector, meaning that even when using the index number method, the name can still be displayed. The use of `[[]]` (double square brackets will eliminate the name).

As an example, to create an object called `slope`, it is the 2nd element in the `coefficients` vector, intercept is the first element. A single set of square brackets around 1 or 2 will keep the name of the `coefficients` vector, while double square brackets will eliminate the name.

`slope=fit$coefficients[2]` or `slope=fit$coefficients[[2]]`

Extraction of equation coefficients

```
intercept=fit$coefficients[1]; b0=fit$coefficients[[1]]
slope=fit$coefficients[2]; b1=fit$coefficients[[2]]
intercept; slope
```

```
(Intercept)
    33.4744
```

```
eruptions
 10.72964
```

```
b0; b1
```

```
[1] 33.4744
```

```
[1] 10.72964
```

Using the regression equation

Use of the equation works just like you are used to; given a specified value of x, solve the equation for the estimated y value called $\hat{y}$ (y-hat)

Find the values of $\hat{y}$ and e_i for each of the following values: $(2.283, 62), (5.1, 96)$

Estimations ($\hat{y}_i$ and e_i)

```
x1=2.283; y1=62; x2=5.1; y2=96
yhat1=b0+b1*x1; yhat2=b0+b1*x2
e1=y1-yhat1; e2=y2-yhat2
yhat1; e1
```

```
[1] 57.97017
```

```
[1] 4.029832
```

```
yhat2; e2
```

```
[1] 88.19557
```

```
[1] 7.804432
```

Extracting R^2 to calculate r

R^2 is on the `R` output and is found under `Multiple R-squared`. $R^2 = 0.8115$. R^2 is the 8th element of the summary of the fit model. `summary(fit)[[8]]`

```
rsq=summary(fit)[8]; rsq; R2=summary(fit)[[8]]; R2
```

```
$r.squared
[1] 0.8114608
```

```
[1] 0.8114608
```

```
r=sqrt(R2); r; cor(faithful)
```

```
[1] 0.9008112
```

```
           eruptions   waiting
eruptions 1.0000000 0.9008112
waiting   0.9008112 1.0000000
```

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

All the following standard errors are provided in the regression analysis output.

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

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.

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

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

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

`predict.lm(object,newdata,interval=,level=,...)`

`object`: fit model; object of class "lm"
`newdata`: (optional) data frame for specified variable values
`interval`: 'confidence' or 'prediction'
`level`: confidence level; 0.95 is default
`...`: more options

The first thing to do is to create a data frame with the specific observations you want use for CIs or PIs (x^*) and then use it with the fit object to find intervals. If you do not, the function will calculate the intervals using *all* of the data points.

CIs PIs with `predict.lm()`

```
newfaith=data.frame(faithful[c(4,149),]) # same ones from before
predict.lm(fit,newfaith,interval='confidence')
```

```
         fit      lwr      upr
4   57.97017 56.94264 58.99769
149 88.19557 86.97223 89.41890
```

```
predict.lm(fit,newfaith,interval='prediction')
```

```
         fit      lwr      upr
4   57.97017 46.28148 69.65886
149 88.19557 76.48804 99.90309
```

Diagnostic plots used to check assumptions of slr

For checking assumptions, we need 3 graphs and one test:
- Histogram of the residuals (#1,4)
- Scatterplot of residuals vs. predicted (#2)
- Independence of residuals are checked with a DW test (#3)
- A normal probability plot, also called a QQ plot (#4)

Extracting residuals and fitted values

There are functions as well as the residuals and fitted values can be extracted from the model. Less typing when using functions.

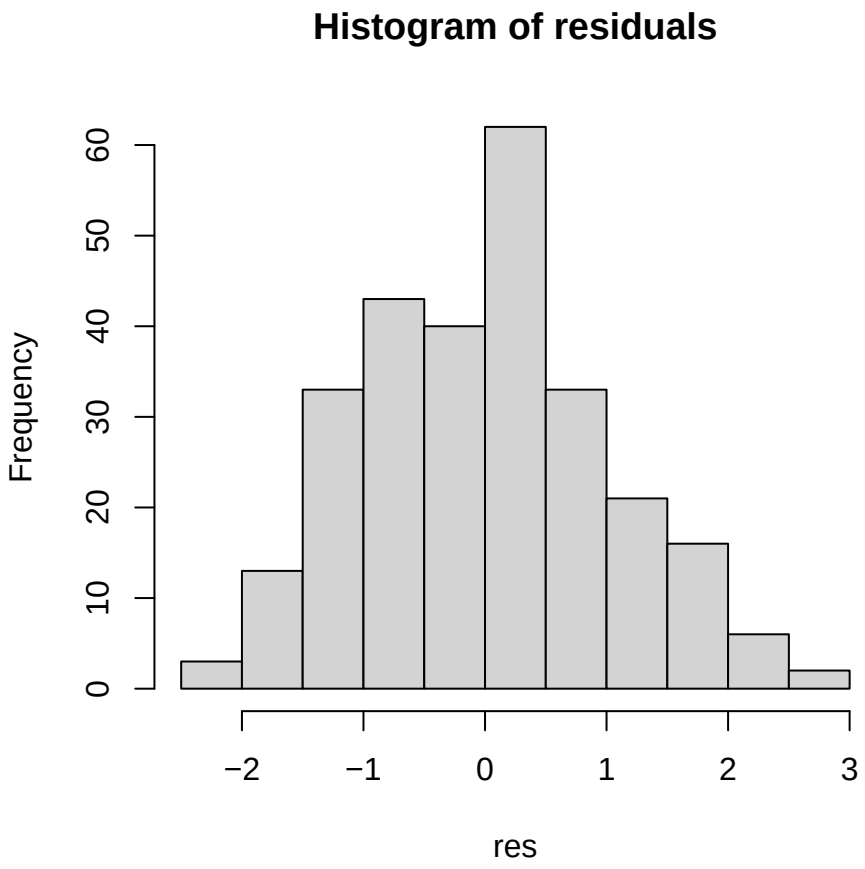
`res=rstudent(fit)` for residuals (these are standardized residuals; like z-scores for every residual).

`pred=fitted(fit)` for estimated values. The name is unimportant, they can be called yhat, fits, est, ... whatever.

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)

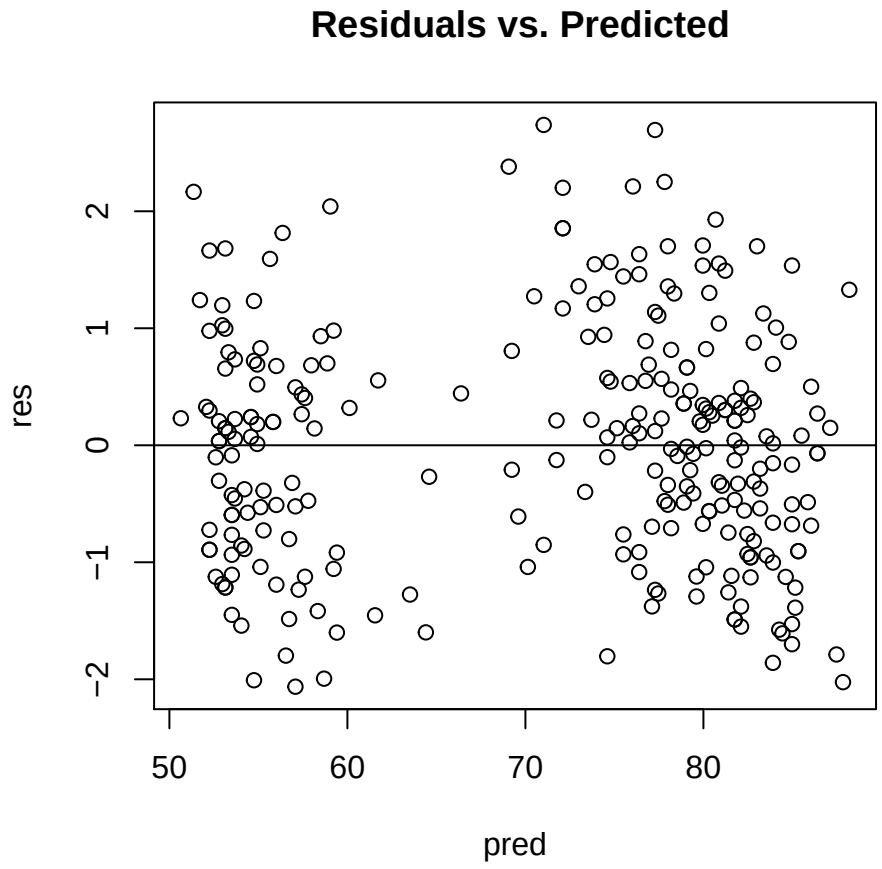
```
res=rstudent(fit); pred=fitted(fit)
hist(res,main='Histogram of residuals')
```



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)

```
plot(pred,res,main=' Residuals vs. Predicted'); abline(h=0)
```



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.

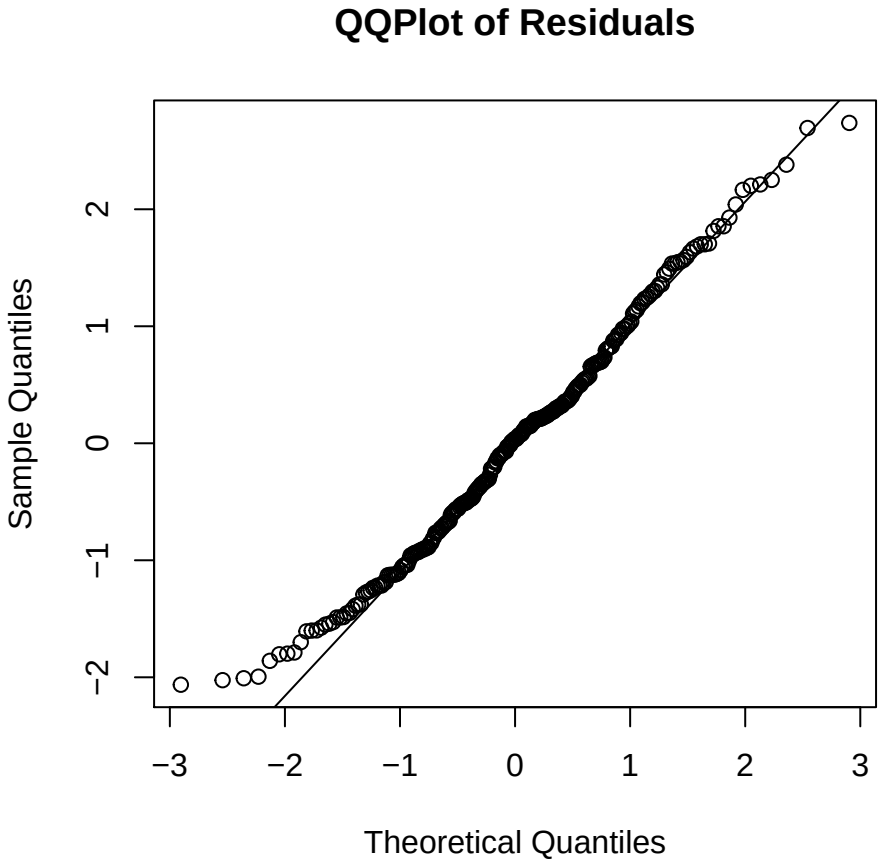
```
library(car)
dwt(fit)
```

```
 lag Autocorrelation D-W Statistic p-value
   1      -0.2767457      2.542647       0
 Alternative hypothesis: rho != 0
```

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

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

```
qqnorm(res,main='QQPlot of Residuals'); qqline(res)
```


QQPlot of Residuals
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
2
3
−2
−1
0
1
2