STAT 210
Applied Statistics and Data Analysis
Week 8 - Summary

Joaquin Ortega

King Abdullah University of Science and Technology

Video 25: Experimental Design II

Hypothesis Test of No Treatment Effect

The null hypothesis is

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_k \quad vs. \quad H_1 : \text{ at least two of the } \tau_i \text{ differ.}$$

Compare the sums of squares for the complete model with the *reduced model*:

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

with the same assumptions as before for the noise.

The test is based on

$$\frac{SST - SSE}{SSE} = \frac{SSA}{SSE}.$$

Hypothesis Test of No Treatment Effect

It can be shown that

- SSE/σ^2 has a χ^2_{n-k} distribution
- SSA/σ^2 has a χ^2_{k-1} distribution under H_0,
- these variables are independent.

In consequence, under H_0 the quotient

$$\frac{SSA/\sigma^2(k-1)}{SSE/\sigma^2(n-k)} = \frac{(n-k)SSA}{(k-1)SSE} \sim F_{k-1,n-k} \tag{1}$$

and we can use this relation to test H_0.

Hypothesis Test of No Treatment Effect

Define

$$MSA = \frac{SSA}{k-1} \qquad \text{and} \qquad MSE = \frac{SSE}{n-k},$$

then (1) becomes

$$\frac{MSA}{MSE} \sim F_{k-1,n-k}$$

and if *msE* and *msA* represent the observed values of these variables, the decision rule for testing H_0 at level of significance α is

$$\text{reject } H_0 \text{ if} \quad \frac{msA}{msE} > F_{1-\alpha,k-1,n-k} \tag{2}$$

Hypothesis Test of No Treatment Effect

The values for the sums of squares, degrees of freedom, mean squares, and F test are usually written in an Analysis of Variance table.

Table 3: Anova table for the one-way analysis of variance

Source	SS	d.f.	MS	F_{obs}	Critical F
Treatment	SSA	$k-1$	$MSA = \frac{SSA}{k-1}$	$F = \frac{MSA}{MSE}$	`qf(1-`α`, k-1, n-k)`
Error	SSE	$n-k$	$MSE = \frac{SSE}{n-k}$		
Total	SST	$n-1$			

Computational Formulae

$SSA = \sum_i r_i \bar{y}_{i\bullet}^2 - n\bar{y}_{\bullet\bullet}^2$

$SSE = \sum_i \sum_j y_{ij}^2 - \sum_i r_i \bar{y}_{i\bullet}^2$

$SST = \sum_i \sum_j y_{ij}^2 - n\bar{y}_{\bullet\bullet}^2$

Hypothesis Test of No Treatment Effect

In R, use the aov and summary functions.

For the tire tread example, the code is

```
library(PASWR)
mod0 <- lm(StopDist ~ tire, data = Tire)
mod1 <- aov(StopDist ~ tire, data = Tire)
summary(mod1)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## tire         3   5673  1891.0   5.328 0.00732 **
## Residuals   20   7099   354.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Hypothesis Test of No Treatment Effect

The same table is produced using the function `anova()` on the model we obtained with the `lm` function (`mod0`).

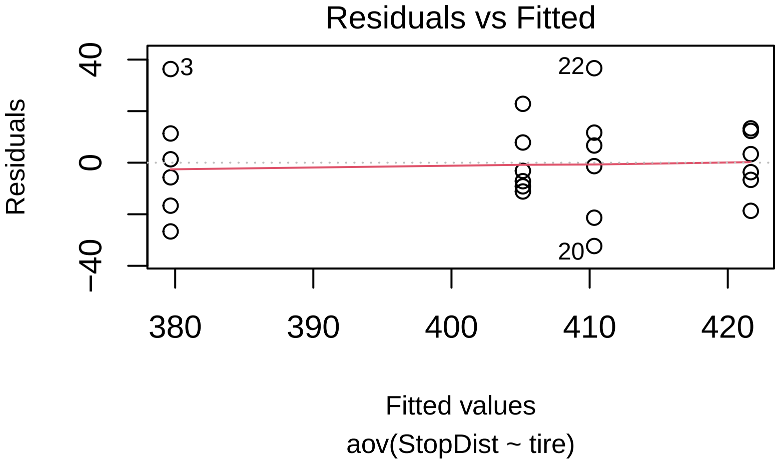
```
anova(mod0)
```

```
## Analysis of Variance Table
##
## Response: StopDist
##           Df Sum Sq Mean Sq F value   Pr(>F)
## tire       3 5673.1 1891.04  5.3278 0.007316 **
## Residuals 20 7098.8  354.94
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Hypothesis Test of No Treatment Effect

Look at diagnostic plots to check the assumptions of the model.

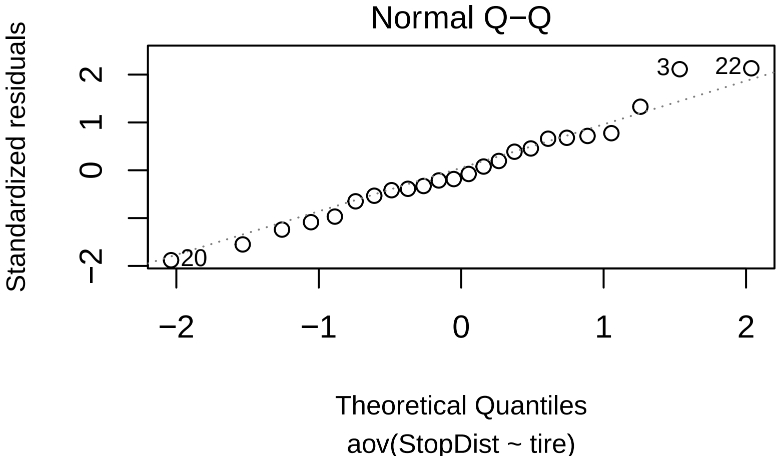
```
plot(mod1, which=1, cex.lab=0.8, cex.sub=0.8)
```



Hypothesis Test of No Treatment Effect

We look for departures from normality in the standardized residuals. Considering the fit we observe, the normality assumption seems justified.

```
plot(mod1, which=2, cex.lab=0.8, cex.sub=0.8)
```



Hypothesis Test of No Treatment Effect

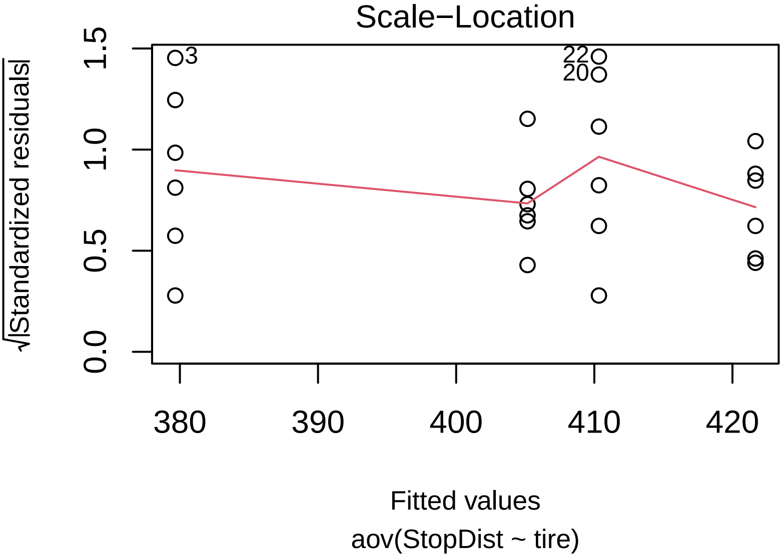
The third graph plots the square root of standardized residuals versus fitted values, and again we look for changes in the variance.

The red line joins the average values of the square root of standardized residuals, and therefore is a indication of the average distance to the origin.

As in the first graph, we see that some points seem to have more spread than others, perhaps pointing to heteroscedasticity. However, there are very few points to be certain.

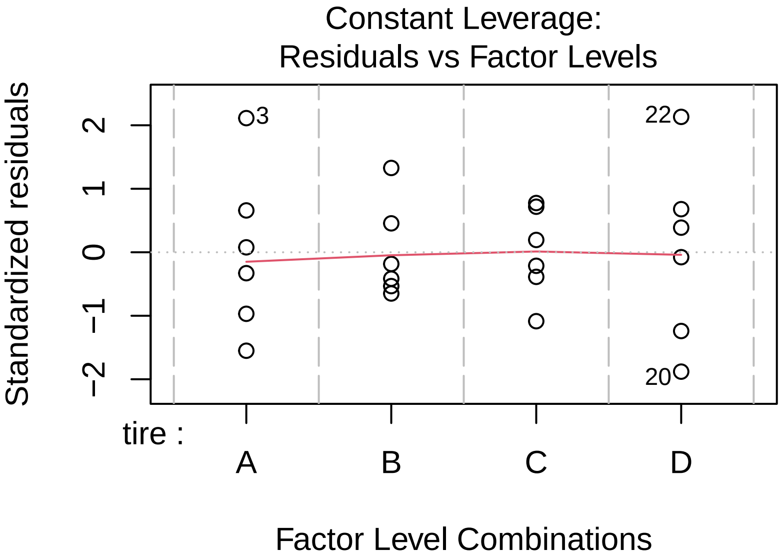
Hypothesis Test of No Treatment Effect

```
plot(mod1, which=3, cex.lab=0.8, cex.sub=0.8)
```



Hypothesis Test of No Treatment Effect

```
plot(mod1, which=5)
```



Effect Sizes

To see the effect sizes in tabular form use `model.tables`

```
model.tables(mod1, se=T)
```

```
## Tables of effects
##
##  tire
## tire
##       A       B       C       D
## -24.542   0.958  17.458   6.125
##
## Standard errors of effects
##          tire
##         7.691
## replic.     6
```

Effect Sizes

Specifying means you get

```
model.tables(mod1,'means',  se=T)
```

```
## Tables of means
## Grand mean
##
## 404.2083
##
##  tire
## tire
##     A     B     C     D
## 379.7 405.2 421.7 410.3
##
## Standard errors for differences of means
##          tire
##         10.88
## replic.     6
```

Pairwise comparisons

Pairwise comparisons

If the result of the F test is to reject the null hypothesis of no treatment effects, one is naturally interested in determining where the difference lies. For this, it becomes necessary to compare the individual groups.

This can be (partly) done looking at the regression coefficients using `summary` and `lm`.

Pairwise comparisons

```
summary(mod0)
```

```
##
## Call:
## lm(formula = StopDist ~ tire, data = Tire)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.333  -9.667  -2.250  11.417  36.667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  379.667      7.691  49.363  < 2e-16 ***
## tireB         25.500     10.877   2.344 0.029497 *
## tireC         42.000     10.877   3.861 0.000973 ***
## tireD         30.667     10.877   2.819 0.010594 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.84 on 20 degrees of freedom
## Multiple R-squared:  0.4442, Adjusted R-squared:  0.3608
## F-statistic: 5.328 on 3 and 20 DF,  p-value: 0.007316
```

Pairwise comparisons

The problem is that if we perform many tests, the probability of finding one of them to be significant by chance alone increases.

Consider one hundred statistical tests at the 5% level and assume all null hypotheses are true. We expect to reject 5 of them by chance alone. This is the expected number of Type I errors.

If the tests are independent, the probability of rejecting at least one null hypothesis can be calculated using the binomial distribution:

```
1 - dbinom(0,100,0.05); 1 - dbinom(0,100,0.01)
```

```
## [1] 0.9940795
```

```
## [1] 0.6339677
```

Pairwise comparisons

The Bonferroni correction is based on the Bonferroni inequalities:

$$P(\cup_1^n B_i) \leq \sum_1^n P(B_i)$$

Dividing the significance level by the number of tests, we get a test with a true significance level smaller or equal to the nominal significance level.

This is equivalent to multiplying the p-values by the number of tests.

Pairwise comparisons

```r
with(Tire, pairwise.t.test(StopDist, tire,
                p.adjust.method = 'bonferroni'))
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  StopDist and tire
##
##   A      B      C
## B 0.1770 -      -
## C 0.0058 0.8696 -
## D 0.0636 1.0000 1.0000
##
## P value adjustment method: bonferroni
```

Pairwise comparisons

Tukey's method for pairwise comparisons

The procedure simultaneously considers all pairs of effects and adjusts the critical region by using the studentized range statistic instead of student's t-distribution.
The test is

$$\text{reject } H_0 \text{ if} \quad |\hat{\tau}_u - \hat{\tau}_s| > \sqrt{2} q_{I,n-k,1-\alpha/2} \hat{\sigma}(\bar{y}_{s\bullet} - \bar{y}_{u\bullet})$$

where $q_{I,n-k,1-\alpha}$ is the $1-\alpha$ percentile of the studentized range and $\hat{\sigma}(\bar{y}_{s\bullet} - \bar{y}_{u\bullet})$ is the estimated standard error for the difference between the averages.

If $X_1, \ldots X_I$ are independent random variables with a $N(\mu, \sigma^2)$ distribution and

$$R = \max_i X_i - \min_i X_i$$

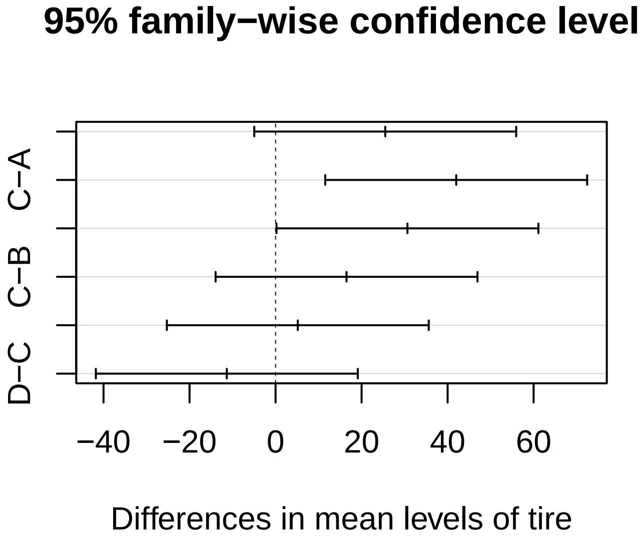
then $R/\hat{\sigma}$ follows the studentized range distribution.

Pairwise comparisons

```r
(mod1.tky <-TukeyHSD(mod1))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = StopDist ~ tire, data = Tire)
##
## $tire
##           diff         lwr      upr     p adj
## B-A  25.500000  -4.9446409 55.94464 0.1213153
## C-A  42.000000  11.5553591 72.44464 0.0049515
## D-A  30.666667   0.2220258 61.11131 0.0479540
## C-B  16.500000 -13.9446409 46.94464 0.4464584
## D-B   5.166667 -25.2779742 35.61131 0.9637307
## D-C -11.333333 -41.7779742 19.11131 0.7273681
```

Pairwise comparisons



Video 26: Experimental Design III

Power and the number of replicates

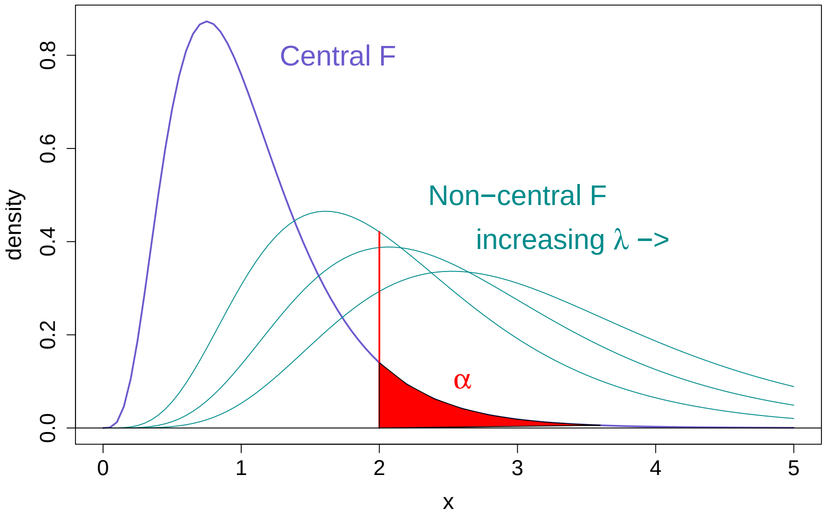
The test statistic msA/msE follows the F-distribution when the null hypothesis is true, but when the null hypothesis is false it follows the noncentral F-distribution.

The noncentral F-distribution has a wider spread than the central F-distribution.

The probability of exceeding the critical limit from the central F-distribution is an increasing function of the noncentrality parameter, λ.

When the sampling distribution is the noncentral F, the probability of exceeding the critical limit from the central F-distribution is the power of the test.

Power and the number of replicates



Power and the number of replicates

If there is a difference among the cell means, the power is given by

$$Power(\lambda) = \int_{F_{k-1,k(r-1),\alpha}}^{\infty} f(x, k-1, k(r-1), \lambda)\, dx$$

where

- $F_{k-1,k(r-1),\alpha}$ is the α-th percentile of the central F distribution, with $k-1$ and $k(r-1)$ degrees of freedom
- $f(x, k-1, k(r-1), \lambda)$ is the non-central F density with non-centrality parameter λ and
- $\lambda = \frac{r}{\sigma^2} \sum_{i=1}^{k} (\mu_i - \bar{\mu}_\bullet)^2$

For a fixed value of $\frac{1}{\sigma^2} \sum_{i=1}^{k} (\mu_i - \bar{\mu}_\bullet)^2$ the power increases with r.

These computations can be carried out with the `Fpower1` function in the `daewr` package.

Power and the number of replicates

Example

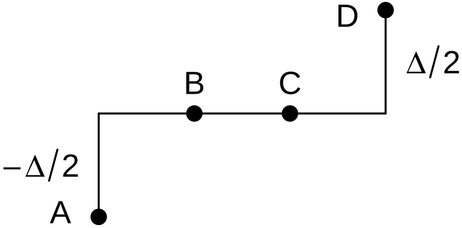
In the tire example, suppose that the standard tread is `D`, and a difference of fewer than 30 feet in the braking distance is of no interest to the manufacturer, but a difference larger than this value would be of interest.

In this case, we regard $\Delta = 30$ as a practical difference in cell means.

We need a value, or at least a lower bound, for the sum $\sum_{i=1}^{k}(\mu_i - \bar{\mu}_\bullet)^2$ under the condition that at least one of the values for the stopping distance for treads `A`, `B` or `C` differ from `D` by at least 30 feet.

Power and the number of replicates

The minimum value is attained when the result for one of the cell means for `A`,`B` or `C` is lower than the average by $\Delta/2$, the braking distance for `D` is higher than the average by $\Delta/2$, and the other two values are equal to the average.



Power and the number of replicates

This results in

$$\sum_{i=1}^{k}(\mu_i - \bar{\mu}_\bullet)^2 = \left(\frac{\Delta}{2}\right)^2 + 0 + 0 + \left(\frac{\Delta}{2}\right)^2 = \frac{\Delta^2}{2} = 450.$$

By previous experience, the manufacturer knows that a reasonable estimate for the variance of the braking distance is 225 ft^2 (standard deviation = 15).

The noncentrality parameter can be calculated as

$$\lambda = \frac{r}{225}450.$$

The power is calculated for $r = 2, \ldots, 10$ using the `Fpower1` function in the `daewr` package.

Power and the number of replicates

```r
library(daewr)
rmin <-2 #smallest number of replicates considered
rmax <-10 # largest number of replicates considered
alpha <- rep(0.05, rmax - rmin +1)
sigma <- 15; nlev <- 4; nreps <- rmin:rmax; Delta <- 30
(power <- Fpower1(alpha,nlev,nreps,Delta,sigma))
```

```
##       alpha nlev nreps Delta sigma     power
##  [1,]  0.05    4     2    30    15 0.1698028
##  [2,]  0.05    4     3    30    15 0.3390584
##  [3,]  0.05    4     4    30    15 0.5037050
##  [4,]  0.05    4     5    30    15 0.6442332
##  [5,]  0.05    4     6    30    15 0.7545861
##  [6,]  0.05    4     7    30    15 0.8361289
##  [7,]  0.05    4     8    30    15 0.8935978
##  [8,]  0.05    4     9    30    15 0.9325774
##  [9,]  0.05    4    10    30    15 0.9581855
```

Non-Parametric Tests

Comparing Two Populations: The Rank Sum Test

The rank sums test was proposed by Wilcoxon. Later, Mann and Whitney proposed an equivalent test. Therefore, any combination of these three surnames may appear to refer to this test.

In this test, we are interested in differences in means or medians. Assume that we have samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ from two continuous symmetric distributions with the same variance but (possibly) different mean. If the distributions are not symmetric, a test of medians is possible.

If the distributions are the same, we can consider that the pooled sample $x_1, \ldots, x_n, y_1, \ldots, y_m$ is a random sample of size $N = n + m$ from the common distribution.

Hence, if these observations are ordered according to magnitude, we would expect to see the xs and ys well mixed.

Comparing Two Populations: The Rank Sum Test

Let μ_X be the (population) mean for the X distribution and similarly for μ_Y. Our test is

$$H_0 : \mu_X = \mu_Y \qquad \text{vs.} \qquad H_A : \mu_X \neq \mu_Y.$$

The (Wilcoxon) rank sums test proceeds as follows:

- The two samples are joined together, giving a sample of size $N = n + m$.
- This sample is ordered, and the ranks (positions in the ordered sample) for the elements of the x sample are added up.

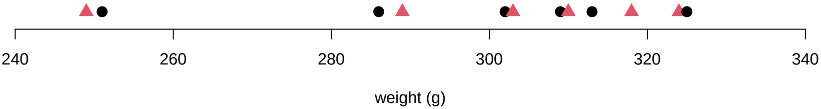
This is the test statistic W.

If the null hypothesis of equal distributions is true, all $\binom{N}{n}$ possible assignments of ranks for the x sample are equally likely, each having probability $1/\binom{N}{n}$.

Example

Two samples of fish were drawn from different lakes and the fish were weighted. The weights in grams are

```
sampl1 <- c(286, 251, 325, 313, 309, 302)
sampl2 <- c(249, 324, 289, 303, 310, 318)
sample <- c(rep(1,6), rep(2,6))
wcx <- data.frame(weight=c(sampl1,sampl2), sample)
ord <- order(wcx[,1])
wcx.ord <- cbind(wcx[ord,],rank=1:12)
```



Comparing Two Populations: The Rank Sum Test

The command `wilcox.test()` in R performs the rank sums test:

```
wilcox.test(sampl1,sampl2)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  sampl1 and sampl2
## W = 17, p-value = 0.9372
## alternative hypothesis: true location shift is not equal to 0
```

The Kruskal-Wallis test

The Kruskal-Wallis test is an extension of the rank-sum test to the case of d multiple samples.

The null hypothesis is that all the samples come from the same distribution, while the alternative is that at least two of the samples come from different distributions.

The only requisite is that the population distributions have to be continuous.

Because the underlying distributions of the d populations are assumed to be identical under the null hypothesis, this test can be applied to means, medians, or any other quantile.

The Kruskal-Wallis test

The hypotheses expressed in terms of the means are

$$H_0 : \mu_1 = \cdots = \mu_d \quad \text{vs.} \quad H_A : \mu_i \neq \mu_j \text{ for at least one pair of } i, j.$$

The $n_1, n_2, \ldots, n_d$ observations are pooled together and ordered from 1 to $N = n_1 + \cdots + n_d$ to obtain the ranks.

The standardized test statistic used by `R` is

$$H = \frac{12 \sum_{i=1}^{d} n_i (\bar{R}_i - \bar{R})}{N(N+1)},$$

where

- n_i is the number of observations for the i-th treatment,
- $\bar{R}_i$ is the average of the ranks in the i-th treatment and
- $\bar{R}$ is the overall average of the ranks.

The Kruskal-Wallis test

When there are ties in the average ranks for the groups, adjustments in the test statistic must be made.

As the size of the smallest group goes to infinity, the test statistic converges in distribution to a χ^2 distribution with $d - 1$ degrees of freedom.

To use this approximation, it is usually required that the minimum group size be at least five.

In `R`, `kruskal.test()` performs this test with the corresponding corrections when ties are present.

Let us compare the results with this test to those obtained before for the `yields` example.

The Kruskal-Wallis test

```
results <- read.table('yields.txt',header=T)
frame <- stack(results)
names(frame) <- c('yield','soil')
with(frame,summary(aov(yield~soil)))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)  
## soil         2   99.2   49.60   4.245  0.025 *
## Residuals   27  315.5   11.69                 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
with(frame, kruskal.test(yield~soil))
```

```
## 
##  Kruskal-Wallis rank sum test
## 
## data:  yield by soil
## Kruskal-Wallis chi-squared = 7.5813, df = 2, p-value = 0.02258
```

The Kruskal-Wallis test

The other example we considered was the tire experiment:

```
anova(mod0)
```

```
## Analysis of Variance Table
##
## Response: StopDist
##           Df Sum Sq Mean Sq F value   Pr(>F)
## tire       3 5673.1 1891.04  5.3278 0.007316 **
## Residuals 20 7098.8  354.94
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
with(Tire, kruskal.test(StopDist ~ tire))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  StopDist by tire
## Kruskal-Wallis chi-squared = 9.9133, df = 3, p-value = 0.01932
```

Video 27: Linear Regression I

Linear Regression

Suppose we have a joint sample

$$(X_1, Y_1), (X_2, Y_2) \ldots, (X_n, Y_n)$$

and we want to determine whether there exists a relationship between them.

The simplest relation is a linear model such as

$$Y = \beta_0 + \beta_1 X. \tag{3}$$

In this model, Y is the **response** or dependent variable and X is a (continuous) **explanatory** or independent variable, also known as a **regressor**.

There are two **parameters** in the model, the slope β_1 and the intercept β_0.

Simple Linear Regression

Since we have a sample of values from both variables $(X_i, Y_i), i = 1, \ldots, n$, the model is usually written as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n. \tag{4}$$

This model is

- **simple**, because it has only one regressor or independent variable,
- **linear**, because it is linear in the parameters: none of the parameters appears as an exponent or raised to a power or multiplied by another parameter,

and is also linear on the variables because the predictor variable only appears raised to the power 1.

Simple Linear Regression

We usually assume that the ϵ_i are centered, $E[\epsilon_i] = 0$ and have equal variance $Var(\epsilon_i) = \sigma^2, i = 1, \ldots, n$. We will also assume that they follow a Gaussian distribution and are independent.

The expected value of Y given X is

$$\begin{aligned} E[Y|X] &= E[\beta_0 + \beta_1 X + \epsilon_i] \\ &= \beta_0 + \beta_1 E[X] + E[\epsilon_i] \\ &= \beta_0 + \beta_1 X. \end{aligned}$$

The distribution of Y **when X is known** is Gaussian with mean $\beta_0 + \beta_1 X$ and variance σ^2.

The slope β_1 represents the expected change in Y when X changes one unit.

When $\beta_1 = 0$, the response Y is independent of the explanatory variable X.

Simple Linear Regression

The problem we want to solve is the estimation of the parameter for the models from a sample of values $(x_1, y_1), \ldots, (x_n, y_n)$.

We can write the relation as a system of linear equations

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\
\vdots \quad & \quad \vdots \qquad \vdots \qquad \vdots \\
y_n &= \beta_0 + \beta_1 x_n + \epsilon_n.
\end{aligned}$$

Simple Linear Regression

In matrix notation this can be written as

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$$

where

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}; \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}; \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

$\mathbf{X}$ is known as the **design matrix**, while β is the vector of parameters.

Estimation

Estimation

The model we want to fit is

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

where (x_i, y_i) are the observed values.

The errors are the differences between the observed values and the values that the model predicts: $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$.

We adopt the least-squares criterion for choosing the parameter values. We want

$$(\hat{\beta}_0, \hat{\beta}_1) = \text{argmin}_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{5}$$

Estimation

The sum on the right of (5) is known as the **error sum of squares** *SSE*:

$$SSE = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

and in matrix notation

$$SSE = \boldsymbol{\epsilon}' \boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta).$$

The fitted values are given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

and the residuals are given by

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

Estimation

Derivation of the estimated values for the βs

We want to minimize SSE. For this, we take partial derivatives wrt the parameters and set them to zero:

$$
\begin{aligned}
\frac{\partial SSE}{\partial \beta_0} &= -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\
\frac{\partial SSE}{\partial \beta_1} &= -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0
\end{aligned}
\tag{6}
$$

These are known as the **normal equations**.

The least squares estimators for the model parameters are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{7}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}. \tag{8}$$

Back to Example 1

Let's go back to example 1 and see what information we can get from the `lm` function:

```
summary(lm1)
```

```
## 
## Call:
## lm(formula = FL ~ CL)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.86395 -0.51746 -0.02826  0.50456  1.77009 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.15316    0.23477   0.652    0.515    
## CL           0.48060    0.00714  67.313   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.717 on 198 degrees of freedom
## Multiple R-squared:  0.9581, Adjusted R-squared:  0.9579 
## F-statistic:  4531 on 1 and 198 DF,  p-value: < 2.2e-16
```

Residuals and Properties of the Regression Line

Residuals

The difference between the observed values y_i, and the fitted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, \ldots, n$ are the residuals:

$$\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

The model we have fitted is the one that minimizes the sum of squares of these residuals. Observe that

$$\begin{aligned} \sum_{i=1}^{n} \hat{\varepsilon}_i &= \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i \\ &= n\bar{y} - n(\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 n\bar{x} = 0. \end{aligned} \tag{9}$$

Properties of the Regression Line

We have already seen in (9) that $\sum_i \hat{\epsilon}_i = 0$. Other properties are:

1. $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$.
2. $\sum_{i=1}^n x_i \hat{\epsilon}_i = 0$.
3. $\sum_{i=1}^n \hat{y}_i \hat{\epsilon}_i = 0$.
4. The regression line always goes through $(\bar{x}, \bar{y})$.

Video 28: Simple Linear Regression II

Ordinary Least Squares in Matrix Notation

Recall that the error sum of squares in matrix notation is given by

$$SSE = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta). \tag{10}$$

Multiplying out the terms in this expression we have

$$SSE = \mathbf{Y}'\mathbf{Y} - \beta'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta. \tag{11}$$

Since $\beta'\mathbf{X}'\mathbf{Y}$ is a scalar, it is equal to its transpose so

$$\beta'\mathbf{X}'\mathbf{Y} = (\beta'\mathbf{X}'\mathbf{Y})' = \mathbf{Y}'\mathbf{X}\beta$$

and we get

$$SSE = \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta. \tag{12}$$

The derivative of SSE is given by

$$\frac{\partial SSE}{\partial \beta} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\beta$$

Ordinary Least Squares in Matrix Notation

Setting this expression equal to zero and solving for β we obtain

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{13}$$

which is the matrix version of the normal equations.

$$\begin{aligned}
\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} &= \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \\
&= \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix} \\
&= \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i \\ -\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i + n\sum_{i=1}^{n} x_i y_i \end{pmatrix}
\end{aligned} \tag{14}$$

Sampling Distribution of $\hat{\beta}$.

Sampling Distribution of $\hat{\beta}$.

We have assumed that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ and therefore

$$\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}).$$

Thus,

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}). \tag{15}$$

Consequently, $\hat{\boldsymbol{\beta}}$ is a linear unbiased estimator for $\boldsymbol{\beta}$.

The unbiased estimator for the error variance σ^2 is

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2. \tag{16}$$

Hypothesis tests

Hypothesis tests

A test statistic for $H_0 : \beta_i = \beta_{i,0}$ versus $H_1 : \beta_i \neq \beta_{i,0}$ can be obtained from the pivotal quantity

$$\frac{\hat{\beta}_i - \beta_{i,0}}{s(\hat{\beta}_i)} \tag{17}$$

for $i = 0, 1$, which has a t_{n-2} distribution under H_0.

For the test $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$, the function `summary` applied to a linear model object will provide

$$t_{obs} = \hat{\beta}_i / s(\hat{\beta})$$

and the corresponding p-value:

$$p - \text{value} = 2P(t_{n-2} \geq |t_{obs}|).$$

Back to Example 1

Let us review again the results for the initial model

```
summary(lm1)
```

```
##
## Call:
## lm(formula = FL ~ CL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.86395 -0.51746 -0.02826  0.50456  1.77009
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.15316    0.23477   0.652    0.515
## CL           0.48060    0.00714  67.313   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.717 on 198 degrees of freedom
## Multiple R-squared:  0.9581, Adjusted R-squared:  0.9579
## F-statistic:  4531 on 1 and 198 DF,  p-value: < 2.2e-16
```

Confidence Intervals

Confidence Intervals

Using the pivotal quantity (17) we can get confidence intervals for the parameters of the model.

A $100(1-\alpha)\%$ confidence interval for $\beta_i, i = 0, 1$ is

$$CI_{1-\alpha}(\beta_i) = \left(\hat{\beta}_i - t_{n-2,1-\alpha/2}, \hat{\beta}_i + t_{n-2,1-\alpha/2}\right)$$