Statistics for linguists

2023-11-29

Linear models, ANOVA, Chi-squared test

From last week

- P-values
- Confidence intervals (see video on moodle)
- Models in general
- Linear models

Concepts you mentioned

- Standard deviation and mean
- Cross-tabulation with a chi-square test
- Binomial test
- Rank tests
- Bivariate analysis
- Multiple logistic regression analysis
- Regression model
- Inter-rater reliability
- Cluster-based permutation test
- One-tail test
- bootstrapping
- null distribution
- surrogate distribution
- baseline distribution
- mixed-effect models
- z-score

Concepts you mentioned

- **Standard deviation and mean**
- Cross-tabulation with a chi-square test
- Binomial test
- **Rank tests**
- Bivariate analysis
- Multiple logistic regression analysis
- **Regression model**
- Inter-rater reliability
- Cluster-based permutation test
- **One-tail test**
- bootstrapping
- **null distribution**
- surrogate distribution
- baseline distribution
- mixed-effect models
- **z-score**

Concepts you mentioned

- ~~Standard deviation and mean~~
- **Cross-tabulation with a chi-square test**
- Binomial test
- ~~Rank tests~~
- Bivariate analysis
- **Multiple logistic regression analysis**
- ~~Regression model~~
- Inter-rater reliability
- Cluster-based permutation test
- ~~One-tail test~~
- **bootstrapping**
- ~~null distribution~~
- surrogate distribution
- **baseline distribution**
- **mixed-effect models**
- ~~z-score~~

For this week

- Linear models (continued)
- Analysis of variance
- Chi-squared test

For this week

- Packages to load

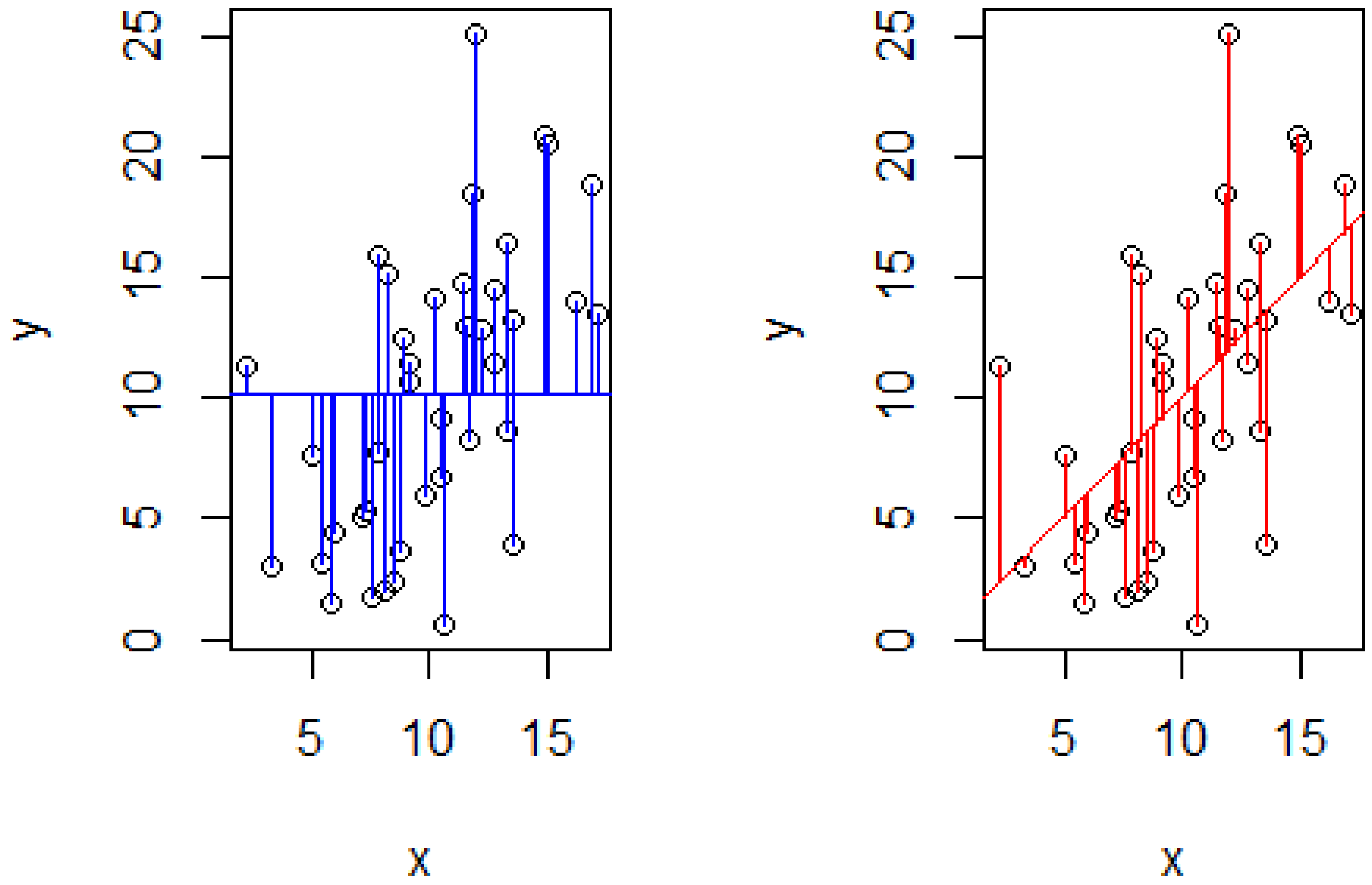
```
library(tidyverse)
library(lattice)
library(Rling)
library(languageR)
library(nhstplot)
library(reshape)
elp.df <- read.csv("YourPath/ELP_full_length_frequency.csv")
senses <- read.csv("YourPath/winter_2016_senses_valence.csv")
data(ldt)
data(sharedref)
```

Analysis of Variance

- What is typically referred to as ANOVA (Analysis of variance) often refers to a type of linear model, where all the predictors are categorical.

- Here’s a way to visualize the difference:

  - In a regression (linear model), you add a line that has an intercept and a slope
  - In an ANOVA (linear model), you add more than one horizontal line with separate intercepts but 0 slope each



Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

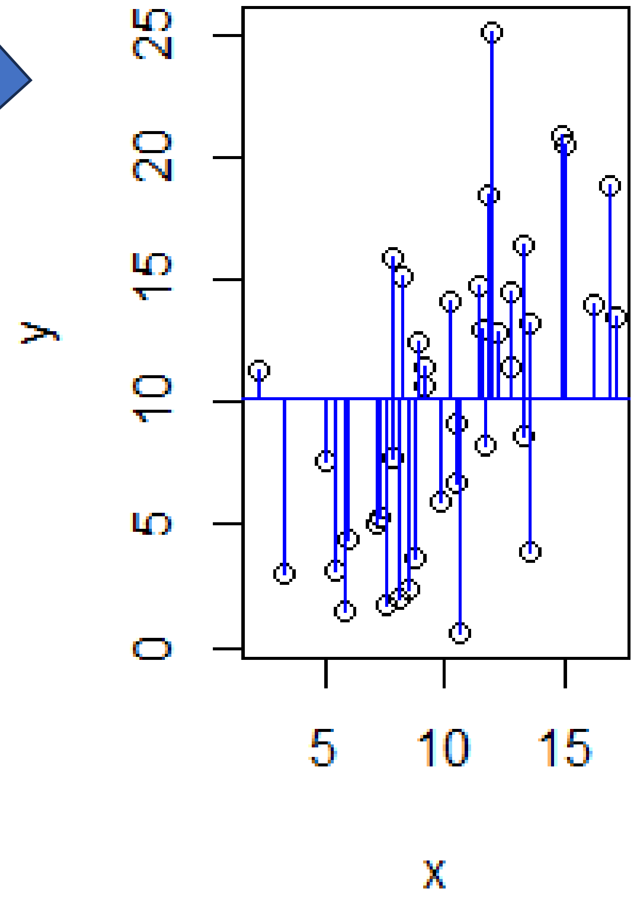
Regression model

- **Null model**: For Response latencies – you can develop a 'model' of response latencies based *only* on response latencies – it just says "assume the mean"

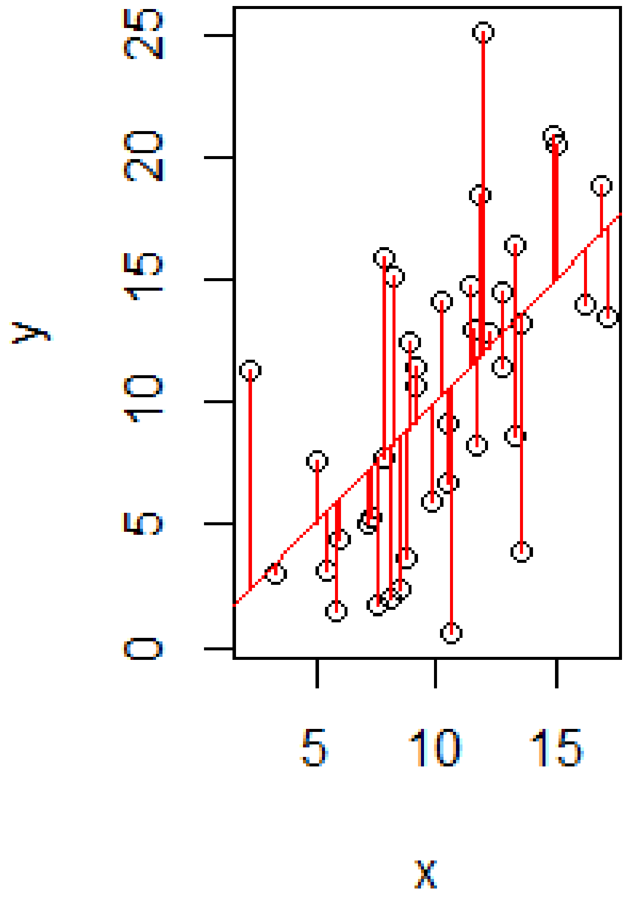
- **Regression model**: Or you can develop a model of response latencies based on the length of words – this model says "assume a response latency *i* for a given length of word *j* according to the following line"

- Your statistics are asking which one is better

Null model – just add a horizontal line through y datapoints corresponds to the mean of *y* and makes no reference to *x*



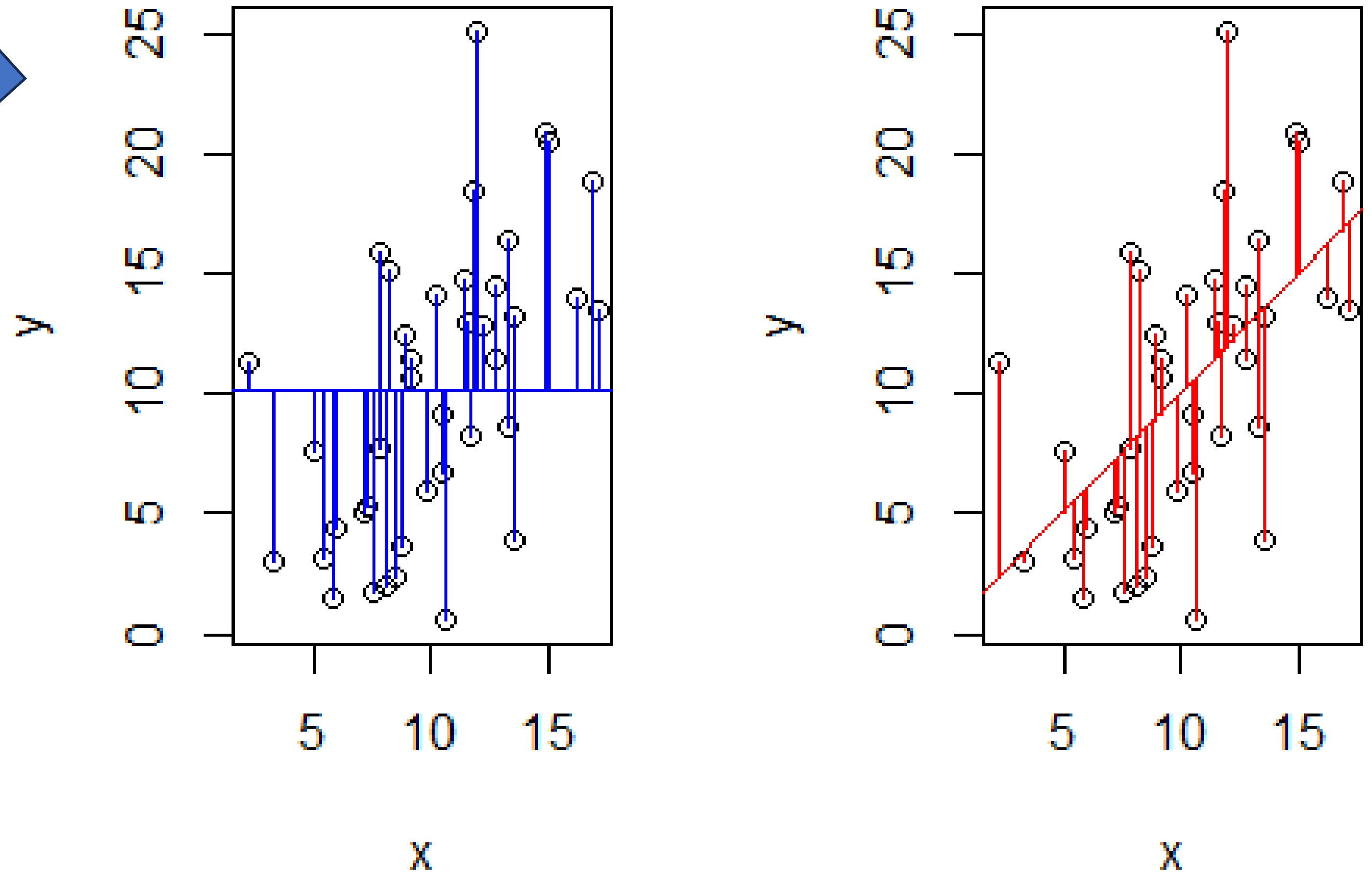
Regression model (linear model): add a line that corresponds to y changing as x changes: x informs *y*



Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

How do I test whether a y-variable only line (null) is better than an x predicts y linear model?

Add up all the residuals (distances from the means) and see if there is a big enough difference given your sample size.



Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

Linear model

$$y = a + \beta x$$

$$y = a + \beta x + \epsilon$$

$$\epsilon \sim N(0, \sigma)$$

Linear model

- What makes a model a ***statistical* model** is that it has some *stochastic component*
- In a classical linear model, this is the error term
- The error term is supposed to follow a normal distribution with 0 mean

$$y = a + \beta x$$

Formula for a straight line

$$y = a + \beta x + \boxed{\epsilon}$$

Error

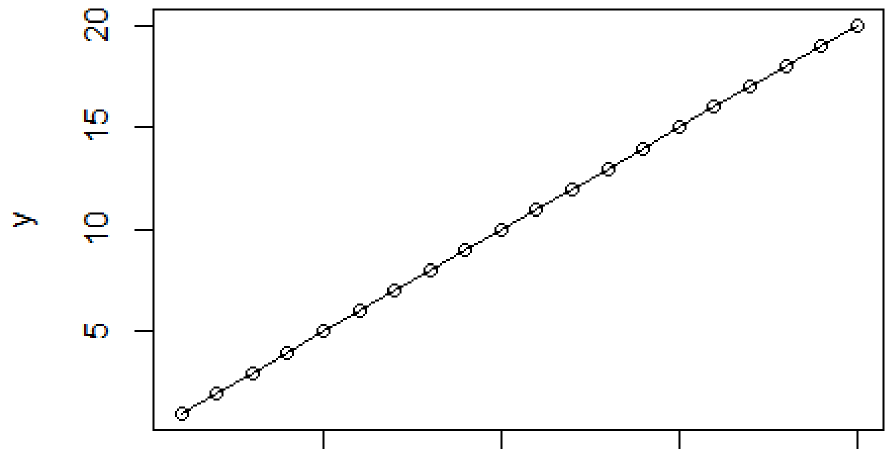
$$\epsilon \sim N(0, \sigma)$$

Normally distributed

0 Mean

Standard deviation

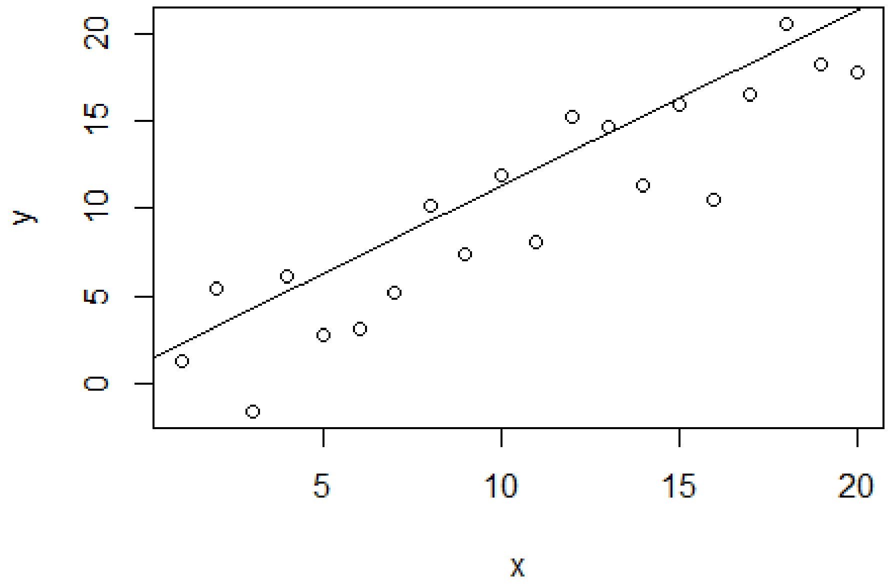
```r
##Deductive model
x <- seq(from=1, to=20)
b <- 1
a <- 0
y <- a + b*x
plot(y~x)+lines(y,x)
```




```r
##Statistical model

x <- seq(from=1, to=20)
b <- 1
a <- 0
e <- rnorm(m=0,sd=3, n=20)
y <- a + b*x + e
plot(y~x)+abline(y,x)
```

$\epsilon \sim N(0,3)$

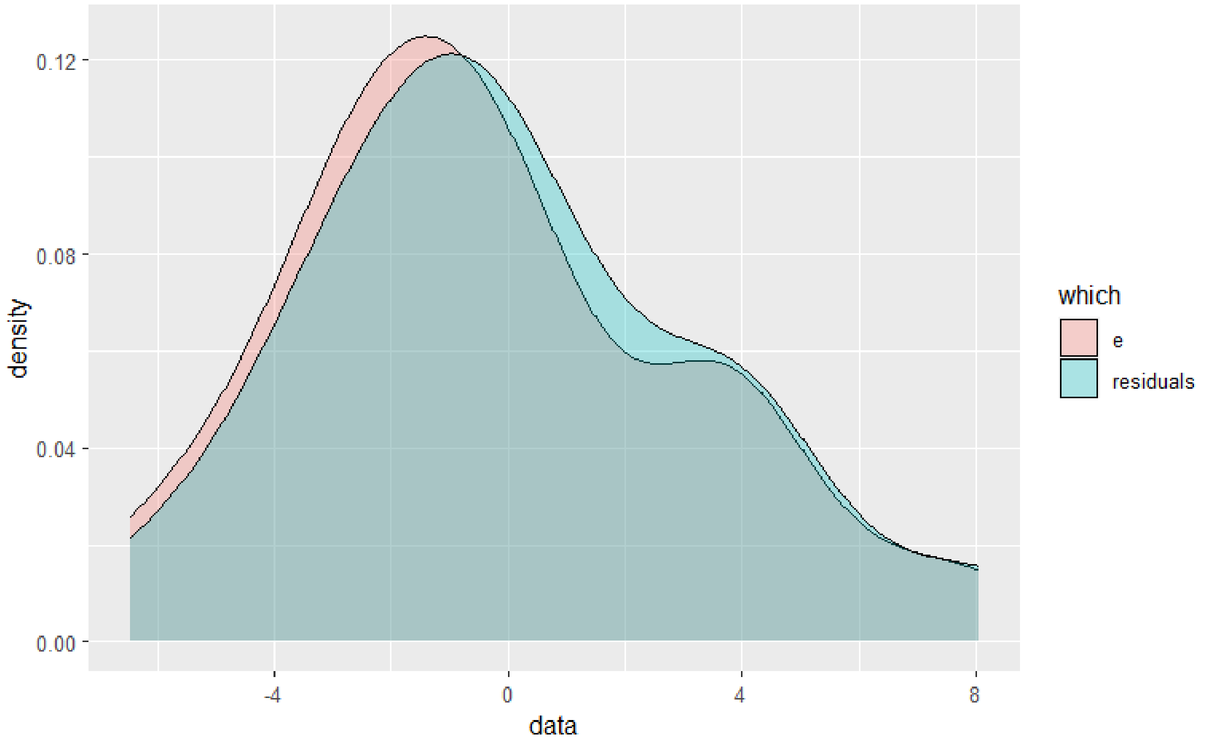


```r
y_hat <- predict(lm(y~x))
residuals <- y - y_hat

error_residuals <- make.groups(e, residuals)

ggplot(error_residuals, aes(x=data, fill=which))+
  geom_density(alpha=0.3)
```

$$\boldsymbol{d} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$$



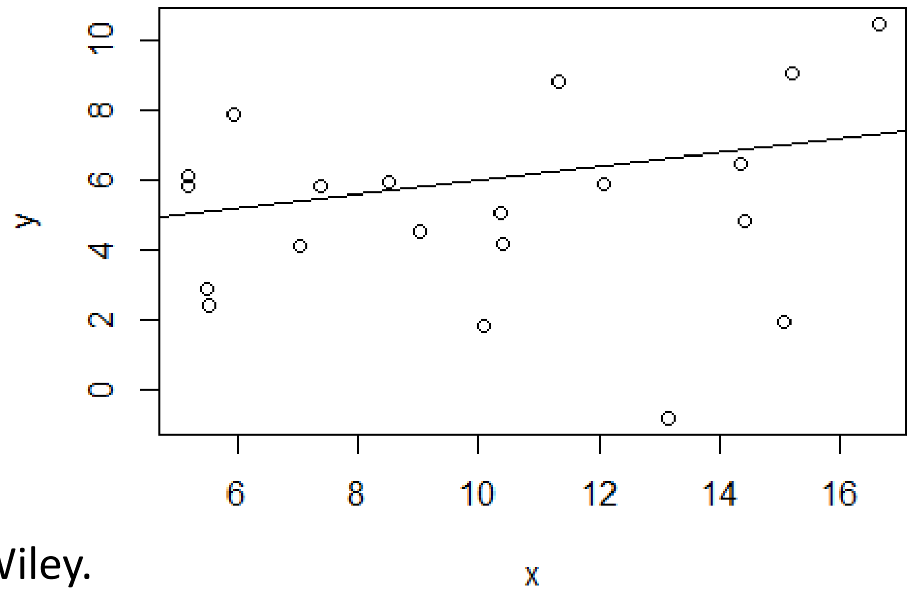
- Our generated errors are almost the same as the residuals

Based on
Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

Linear models and inference

- How do I make inferences about my line?
- Is my line explaining any of the variability?
- Note: you can draw a line through two vectors that are unrelated to one another - this does not mean that they are related.

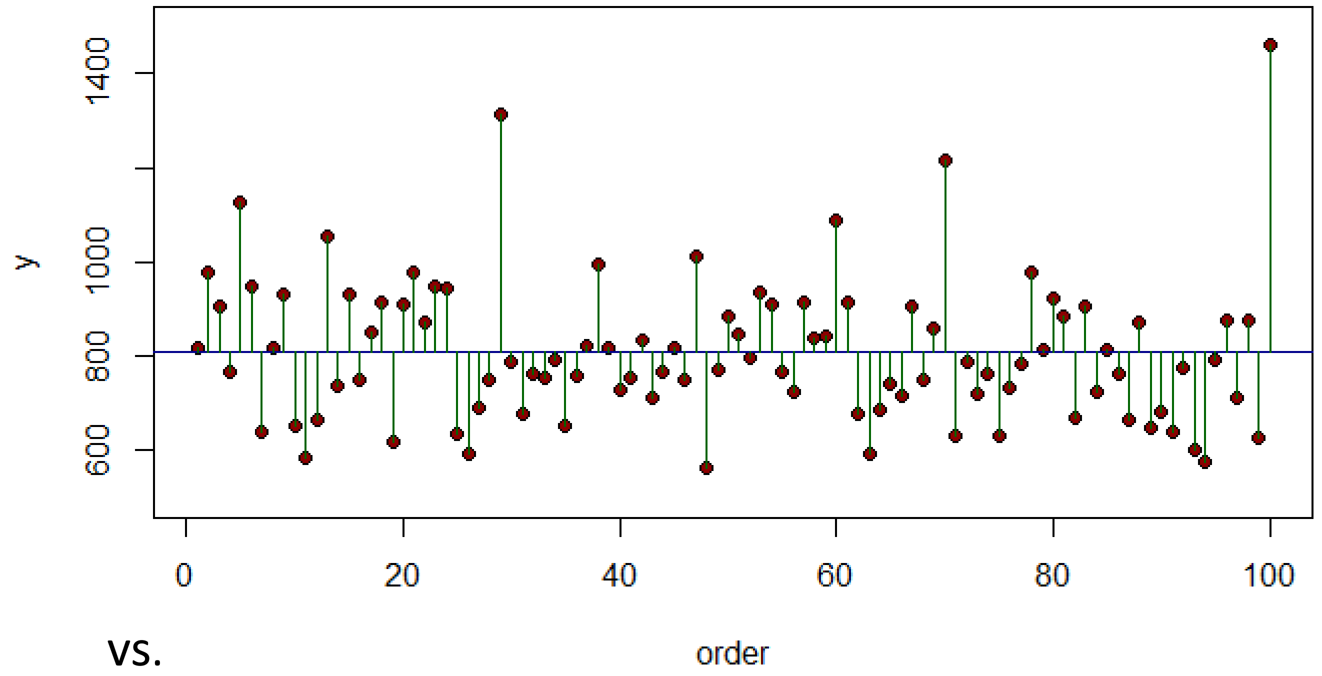
```
##Line through unrelated vectors
x <- rnorm(20, 10, 4)
y <- rnorm(20, 5, 3)
plot(y~x)+abline(a=4, b=0.2)
```



Based on
Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

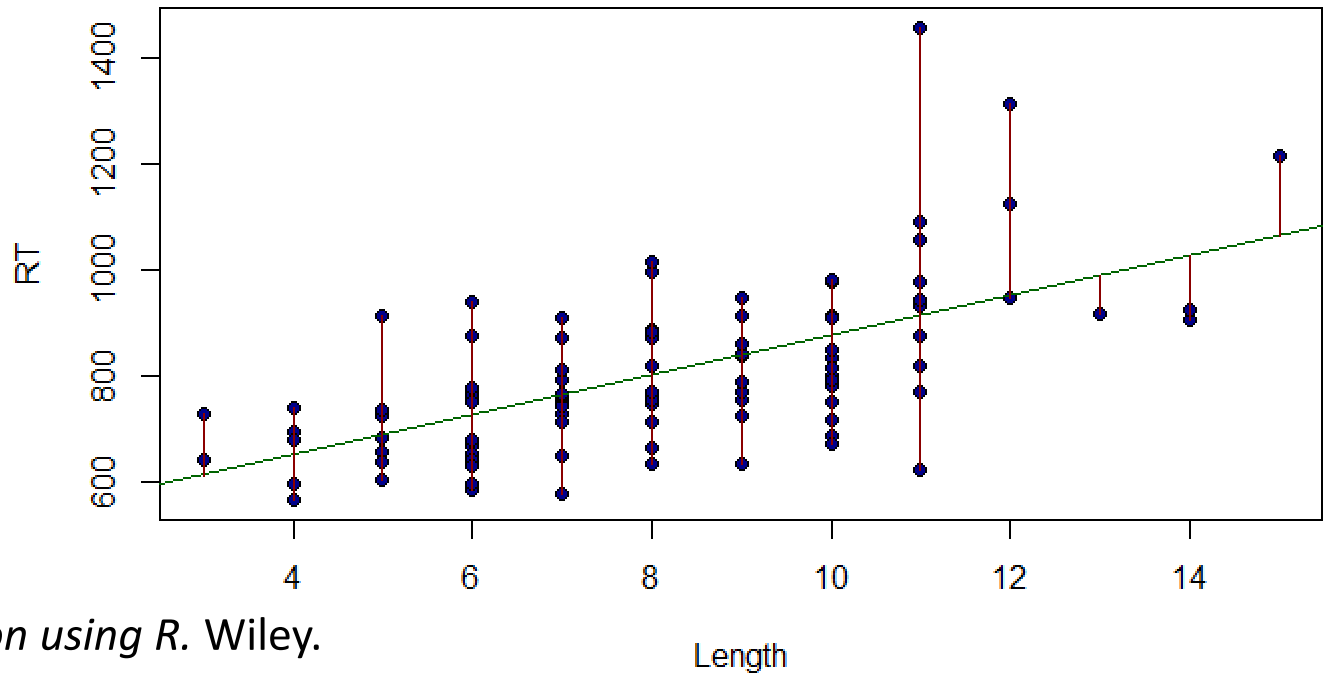
Linear models and inference

- Compare the variance accounted for by just predicting the mean of *y* (23472)



vs.

- ... to the variance accounted for by varying *y* according to *x* in a linear model (1202)



Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

```r
model_1 <- lm(Mean_RT~Length, data=ldt)
summary(model_1)
```

```
##
## Call:
## lm(formula = Mean_RT ~ Length, data = ldt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -291.74  -77.81   -3.69   47.92  546.22
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  498.443     41.949  11.882  < 2e-16 ***
## Length        37.644      4.879   7.716 1.02e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 121.5 on 98 degrees of freedom
## Multiple R-squared:  0.3779, Adjusted R-squared:  0.3716
## F-statistic: 59.53 on 1 and 98 DF,  p-value: 1.019e-11
```

```r
model_1 <- lm(Mean_RT~Length, data=ldt)
summary(model_1)
```

```
##
## Call:
## lm(formula = Mean_RT ~ Length, data = ldt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -291.74  -77.81   -3.69   47.92   546.22
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  498.443     41.949  11.882  < 2e-16 ***
## Length        37.644      4.879   7.716 1.02e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 121.5 on 98 degrees of freedom
## Multiple R-squared:  0.3779, Adjusted R-squared:  0.3716
## F-statistic: 59.53 on 1 and 98 DF,  p-value: 1.019e-11
```

$$y = a + \beta x + \epsilon$$

```r
model_1 <- lm(Mean_RT~Length, data=ldt)
summary(model_1)
```

```
##
## Call:
## lm(formula = Mean_RT ~ Length, data = ldt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -291.74  -77.81   -3.69   47.92  546.22
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  498.443     41.949  11.882  < 2e-16 ***
## Length        37.644      4.879   7.716 1.02e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 121.5 on 98 degrees of freedom
## Multiple R-squared:  0.3779, Adjusted R-squared:  0.3716
## F-statistic: 59.53 on 1 and 98 DF,  p-value: 1.019e-11
```

'R-squared': How much variabililty the model explains (0 = no variability, 1 = all the variability)

Exercise

- Exercise:
- Load the data(ELP_Frequency)
  - Winter 2019

- Run lm() models with reaction time as dependent variable and once with frequency as predictor and once with length as a predictor.

- Which predictor is better?

Winter, Bodo. 2019. *Statistics for Linguists: An Introduction using R.* Routledge.
https://osf.io/34mq9/

Analysis of Variance

- ANOVA (Analysis of Variance) does the same type of calculation but its better to think (initially) in terms of multiple lines rather than one line with a slope and intercept.

- Null model (H0): (same as for linear model) variance with all the data pooled
- Alternative model (H1): total variance with all the data put into groups

Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

Analysis of Variance

- Let's look at the senses data from Winter (2016)

```
head(senses)
```

```
##         Word Modality      Val
## 1   abrasive    Touch 5.398113
## 2 absorbent    Sight 5.876667
## 3     aching    Touch 5.233370
## 4     acidic    Taste 5.539592
## 5      acrid    Smell 5.173947
## 6   adhesive    Touch 5.240000
```

Winter, Bodo. 2016. Taste and smell words form an affectively loaded and emotionally flexible part of the English lexicon. *Language, Cognition and Neuroscience* http://dx.doi.org/10.1080/23273798.2016.1193619
https://osf.io/34mq9/

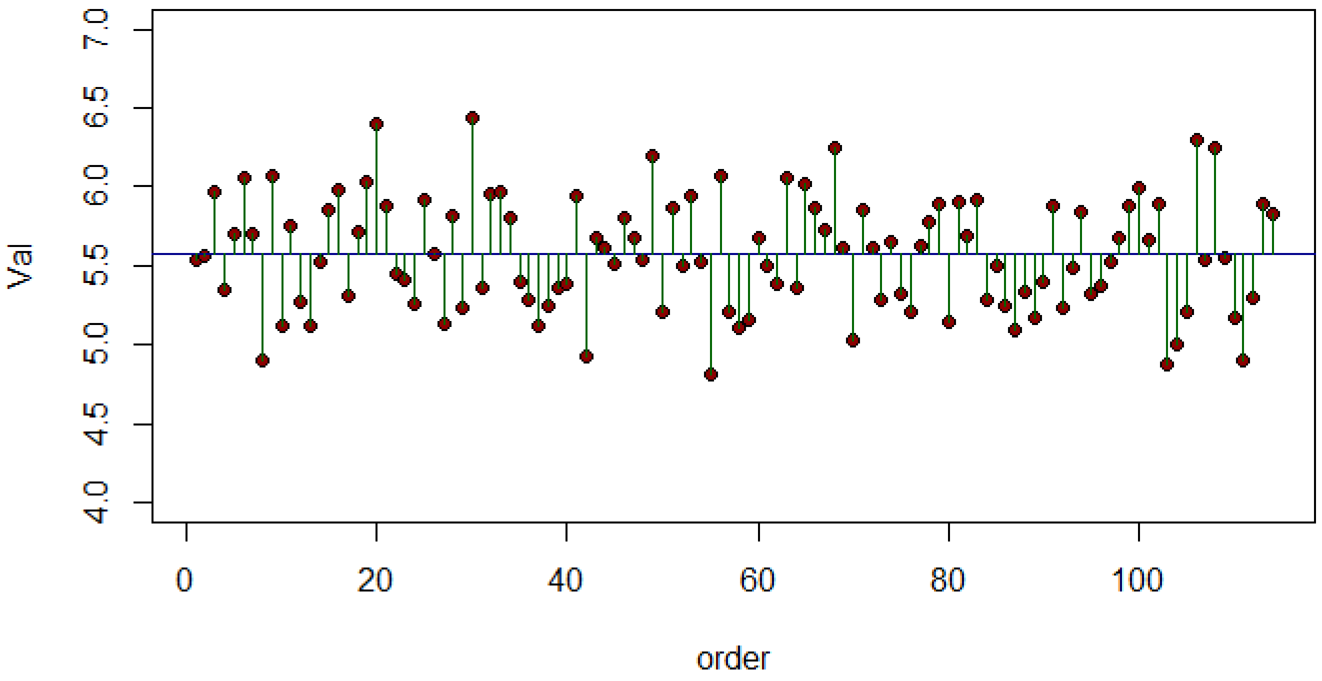
Analysis of Variance

- Filter for **Taste** and **Sound**
- This will simplify our analysis

```
senses_01 <- filter(senses, Modality =="Taste" | Modality =="Sound")
modality <- senses_01$Modality
Val <- senses_01$Val
```

Analysis of Variance

- This is how we visualize our *null* model of the **valence metric**



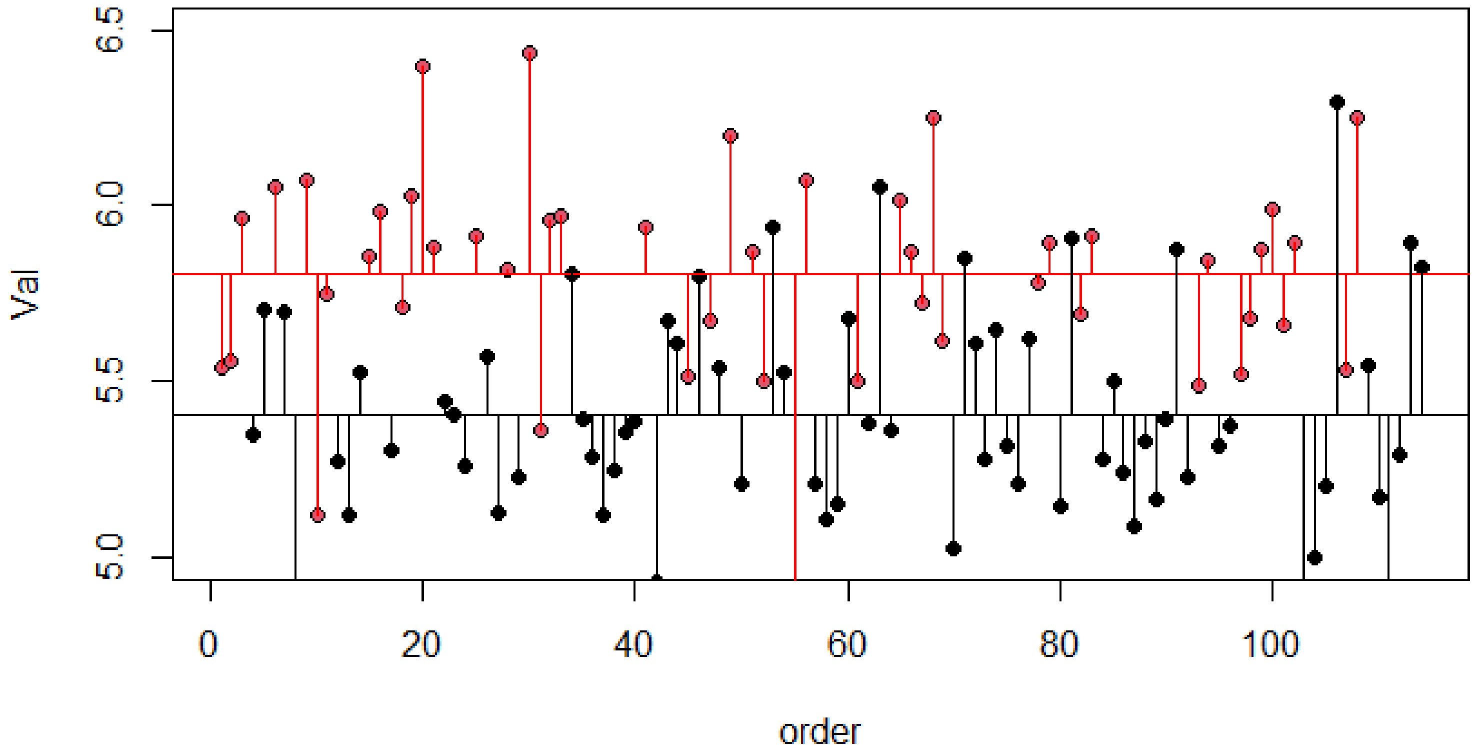
Winter, Bodo. 2016. Taste and smell words form an affectively loaded and emotionally flexible part of the English lexicon. *Language, Cognition and Neuroscience* http://dx.doi.org/10.1080/23273798.2016.1193619
https://osf.io/34mq9/

```
plot(1:114,Val,ylim=c(4,7),ylab="y",xlab="order",pch=21,bg="darkred")
abline(h=mean(Val),col="darkblue")
for(i in 1:114)
  lines(c(i,i),c(mean(Val),Val[i]),col="darkgreen")
```

Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

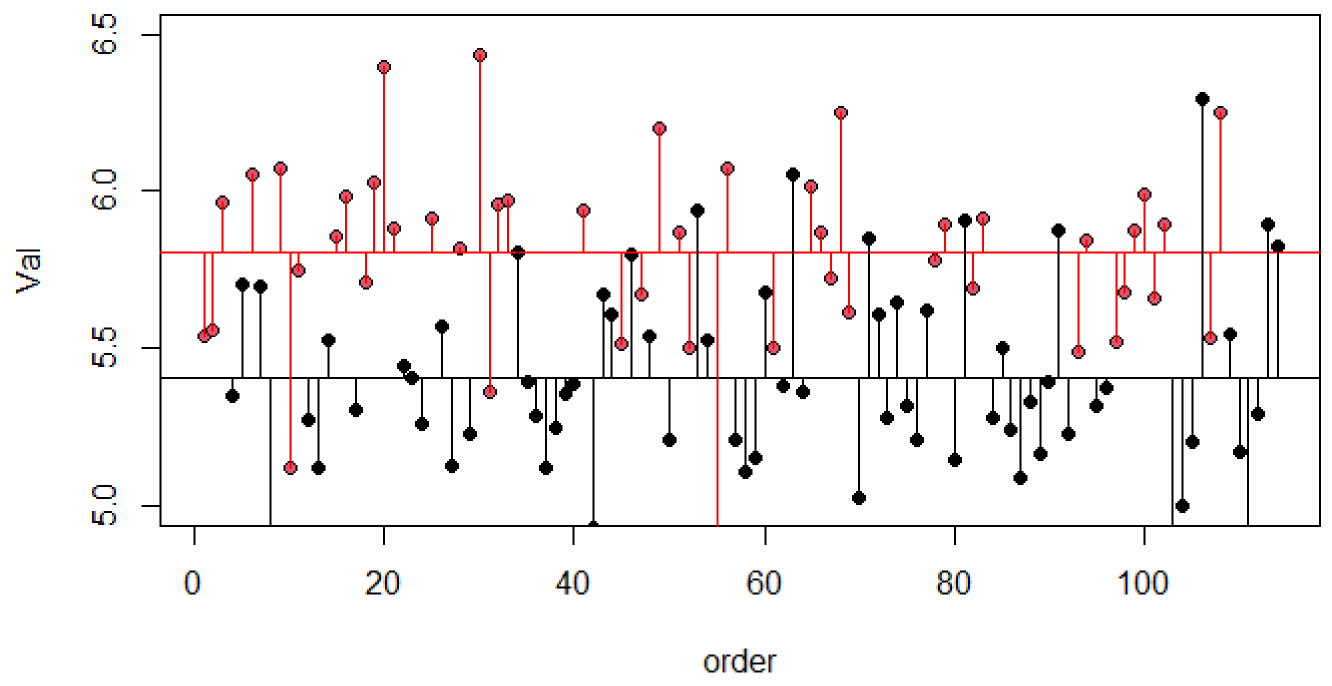
Analysis of Variance

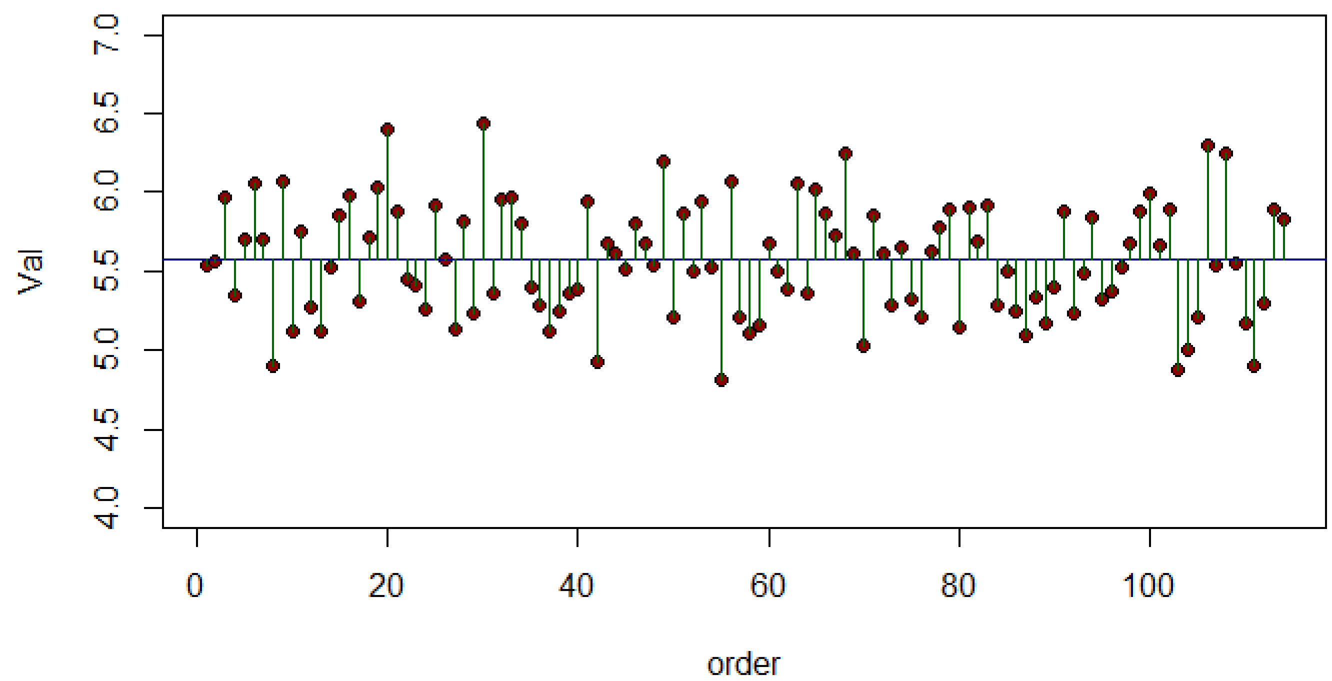
- Our alternative hypothesis is that the variance in VAL can be explained by splitting the data into two groups



- Quiz:
- If the means between the senses were not different, where wowuld the lines be?

Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.





- Quiz:

If the mean difference is different, would the **residual lines** be larger or smaller than when we compute them from the **residual lines from the groups pooled**?

Error sum of squares (analyzing variances)

- **Total sum of squares**: The sum of squares of the residuals of all the pooled.

- **Error sum of squares**: The combined sum of squares of the residuals of the data split up.

- **Treatment effect**: Total of squares minus the error sum of squares.

Error sum of squares (analyzing variances)

- **Total sum of squares**: The sum of squares of the residuals of all the pooled.

- **Error sum of squares**: The combined sum of squares of the residuals of the data split up.

$$SSE = \sum_{j=1}^{k} \sum \left(y - \overline{y}_j\right)^2$$

Error sum of squares

```
sound <- senses_01[senses_01$Modality=="Sound",]
taste <- senses_01[senses_01$Modality=="Taste",]
residuals_Sound <- sound$Val - mean(sound$Val)
residuals_Taste <- taste$Val - mean(taste$Val)
error_sum_of_squares <- sum(residuals_Sound^2) + sum(residuals_Taste^2)
error_sum_of_squares
```

```
## [1] 10.13909
```

Analysis of variance

```
total_sum_of_squares <- sum((senses_01$Val - mean(senses_01$Val))^2)
```

Total sum of squares

Error sum of squares

$$SSE = \sum_{j=1}^{k} \Sigma \left(y - \overline{y}_j\right)^2$$

```
sound <- senses_01[senses_01$Modality=="Sound",]
taste <- senses_01[senses_01$Modality=="Taste",]
residuals_Sound <- sound$Val - mean(sound$Val)
residuals_Taste <- taste$Val - mean(taste$Val)
error_sum_of_squares <- sum(residuals_Sound^2) +
sum(residuals_Taste^2)
```

Treatment effect

```
treatment_sum_of_squares <- total_sum_of_squares -
error_sum_of_squares
```

F table

	Sum of squares	degrees of freedom	Mean square	F ratio
Sense	4.48	1	4.48	49.78
Error	10.13	114 – 2	0.09	
Total	14.62	113		

$$\text{F ratio} = \frac{\frac{SSA\ (treatment)}{df}}{\frac{SSE\ (error\ sum\ of\ squares)}{df}}$$

F table

Degrees of freedom are the maximum number of logically independent values, which may vary in a data sample. Degrees of freedom are calculated by subtracting one from the number of items within the data sample.

https://www.investopedia.com/terms/d/degrees-of-freedom.asp

	Sum of squares	Degrees of freedom	Mean square	F ratio
Sense	4.48	**1**	4.48	49.78
Error	**10.13**	**n – 2**	0.09	
Total	14.62	n -1		

$$\text{F ratio} = \frac{\frac{SSA\ (treatment)}{df}}{\frac{SSE\ (error\ sum\ of\ squares)}{df}}$$

```
F_ratio <- treatment_sum_of_squares / (error_sum_of_squares/112)
```

Analysis of variance

```
summary(aov(Val~Modality, data=senses_01))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Modality      1  4.485   4.485   49.54 1.65e-10 ***
## Residuals   112 10.139   0.091
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Quiz

- Quiz: ceritus paribus ..
  - What happens to F if the error sum of squares increases?
  - What happens to F if the treatment increases?
  - What happens to F if the sample size increases?
  - What happens to F if there are more groups compared?

Exercise

- Load the **sharedref** data from lingR (Ch. 8 Levshina)
- Conduct and interpret an ANOVA that uses **cohort** as the predictor and **mod** as the dependent variable
- Conduct and interpret an ANOVA that uses **age** as the predictor and **mod** as the dependent variable

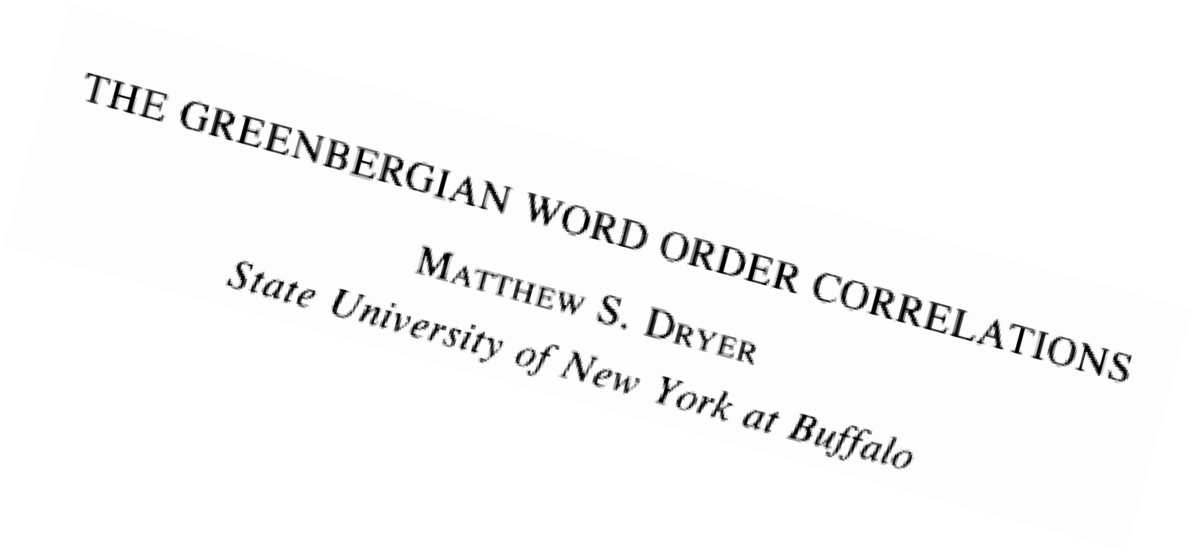
Levshina, Natalia. 2018. *How to do Linguistics with R: Data exploration and statistical analysis.* John Benjamins Publishing.

Counts & contingency tables

- A lot of statistical information comes from counts (e.g. frequency of words in different texts)
- The data are usually presented in a contingency table.
- Data from Matthew Dryer (1992)
- OV = Object-Verb words order / VO = Verb-Object word order
- Postp = positions / Prep = prepositions
- H1:postpositions are associated with OV word order and prepositions are associated with VO order

Dryer, Matthew S. 1992. The Greenbergian Word Order Correlations. *Languages* 68:1, 81-138.

Contingency table (word order associations)


THE GREENBERGIAN WORD ORDER CORRELATIONS
MATTHEW S. DRYER
State University of New York at Buffalo

	OV	VO
Postp	107	12
Prep	7	70

```
adpos <- matrix(c(107,12,7,70),ncol=2,byrow=TRUE)
rownames(adpos)<-c("PostP","Prep")
colnames(adpos)<-c("OV","VO")
adpos
```

Dryer, Matthew S. 1992. The Greenbergian Word Order Correlations. *Languages* 68:1, 81-138.

Counts and probabilities

- These are **observed frequencies**
- We now need a model that predicts the **expected frequencies**
- Using these data, what is the probability of a random language from this sample having OV?

Dryer, Matthew S. 1992. The Greenbergian Word Order Correlations. *Languages* 68:1, 81-138.

Counts and probability

```
wordorder <- cbind(c(107, 7), c(12, 70))
rownames(wordorder) <- c("Postp", "Prep")
colnames(wordorder) <- c("OV", "VO")
wordorder <- rbind(wordorder, c(114,82))
wordorder <- cbind(wordorder, c(119,77,196))
rownames(wordorder) <- c("PostP", "Prep", "Column Total")
colnames(wordorder) <- c("OV", "VO", "Row total")
wordorder
```

Expected Frequency

- Raw total (Postp = 119, Prep = 77)
- Column total (OV = 114, VO = 82)
- Grand total = 196
- Expected = (Raw total * Column total) / Grand total

Dryer, Matthew S. 1992. The Greenbergian Word Order Correlations. *Languages* 68:1, 81-138.

Expected frequency

- The expected frequency refers to what the values would be if VO/OV and Postp/Prep were independent.

	OV	VO
Postp	(114*119)/196	(82*119)/196
Prep	(114*77)/196	(82*77)/196

```
E <- cbind(c((114*119)/196, (114*77)/196),
           c(82*119/196,(82*77)/196))
rownames(E) <- c("Postp", "Prep")
colnames(E) <- c("OV", "VO")
E
```

Expected frequency

- What we’ve done is created a hypothetical “null distribution” against which we can measure how surprising our actual data are.

	OV	VO
Postp	69.21429	49.78571
Prep	44.78571	32.21429

Expected frequency vs. real frequencies

- What we’ve done is created a hypothetical “null distribution” against which we can measure how surprising our actual data are.

	OV	VO
Postp	69.21429	49.78571
Prep	44.78571	32.21429

	OV	VO
Postp	107	12
Prep	7	70

Expected frequency vs. real frequencies

- Its clear that the expected and the observed are different
- But because of errors in sampling there is always some variation, so we are interested in whether the expected frequencies are significantly different

	OV	VO
Postp	69.21429	49.78571
Prep	44.78571	32.21429

	OV	VO
Postp	107	12
Prep	7	70

Chi-squared test

- The classical way of doing this is Karl Pearson's chi-squared test.

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

https://upload.wikimedia.org/wikipedia/commons/2/21/Karl_Pearson_2.jpg

```r
E <- cbind(c((114*119)/196, (114*77)/196),
           c(82*119/196,(82*77)/196))
rownames(E) <- c("Postp", "Prep")
colnames(E) <- c("OV", "VO")
E
```

$$Expected = \frac{Rowtotal * Columntotal}{Grandtotal}$$

```r
E.df <- melt(E)
colnames(E.df)<-c("Adposition", "Verb.Object", "
Expected.Frequency")
E.df$Observed.Frequency <- c(107,7,12,70)
E.df
```

Putting expected and observed data in the same data set

```r
E.df$oe <- ((E.df$Observed.Frequency - E.df$Expected.Frequency)^2) / E.df$Expected.Frequency
sum(E.df$oe)
```

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Chi-squared test and p values

- Is this number big?
- What is the critical value of the chi-squared test?
- To calculate this, we need the degrees of freedom and the cut off area you want.
- R = number of rows
- C = number of columns

Chi-squared test and p-value

- We have another hypothetical distribution based like the t distribution.