GLM for t-tests and ANOVAs

PSM 2

Bennett Kleinberg

29 Jan 2019

Welcome

Probability, Statistics & Modeling II

Lecture 4

The GLM for group mean comparisons

About the module

- 7.5 ECTS (0.5 UCL credits)
- = 150 learning hours
- = 11 weeks with *14 hours per week*
- 3 contact-hours per week
- –> 11 hours self-study per week

Expected self-study

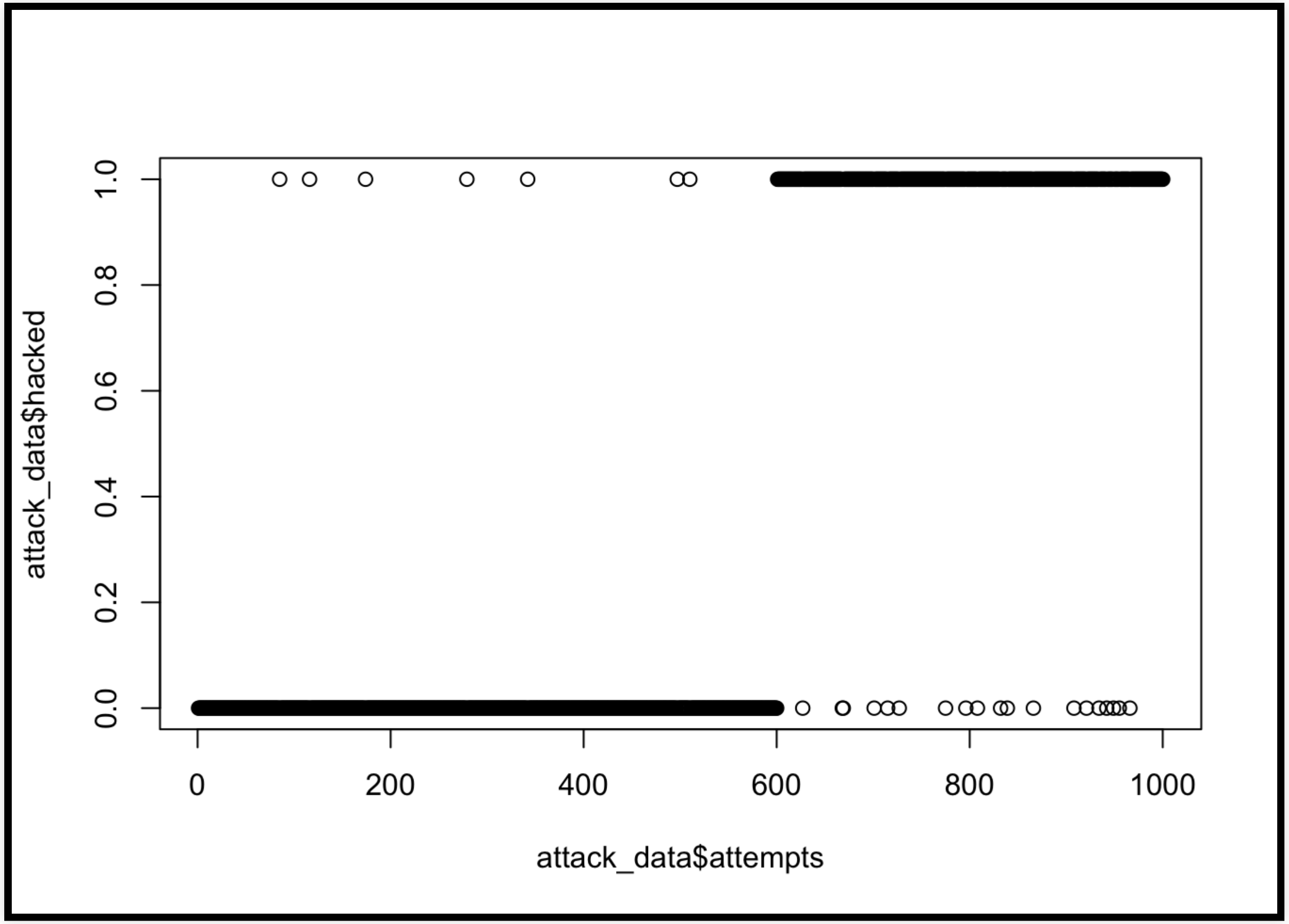
- Revise the lecture (your responsibility)
- Replicate the code/examples
- Read the required literature (read, annotate, summarise)
- Read additional literature if necessary
- Design own code examples to understand the concept
- If still unclear: post it on Moodle or ask us

What question do you have?

Brief recap logistic regression

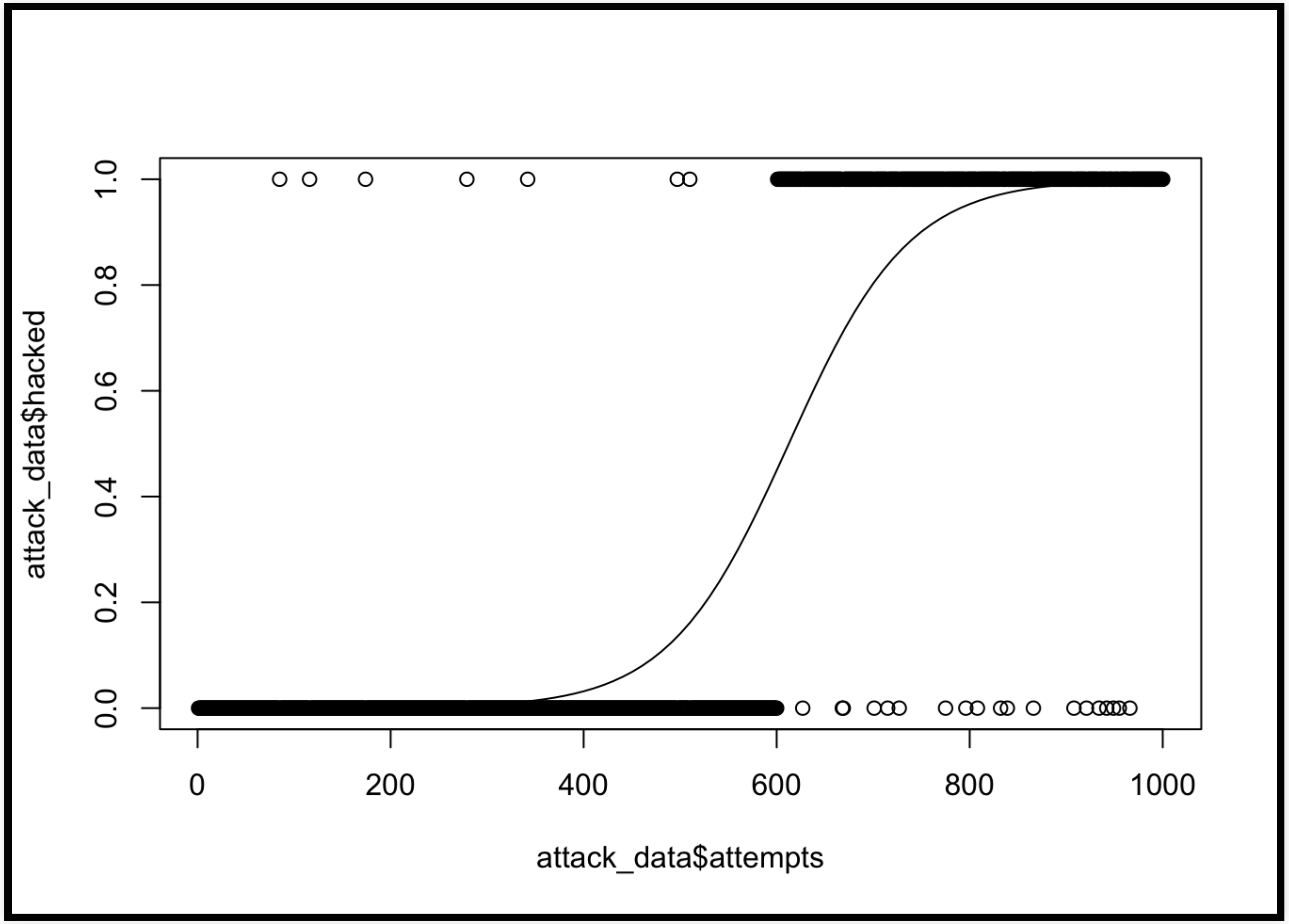
- problem of linear models for discrete/binary outcome variables
- needed: transformation of the outcome variable

From ...



attack_data$hacked
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
attack_data$attempts

… to:



attack_data$hacked
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
attack_data$attempts

Re-transform for the interpretation!

- remember: we model the log-odds
- so: un-log the natural logarithm
- then: use the odds, or transform to probabilities

Your turn

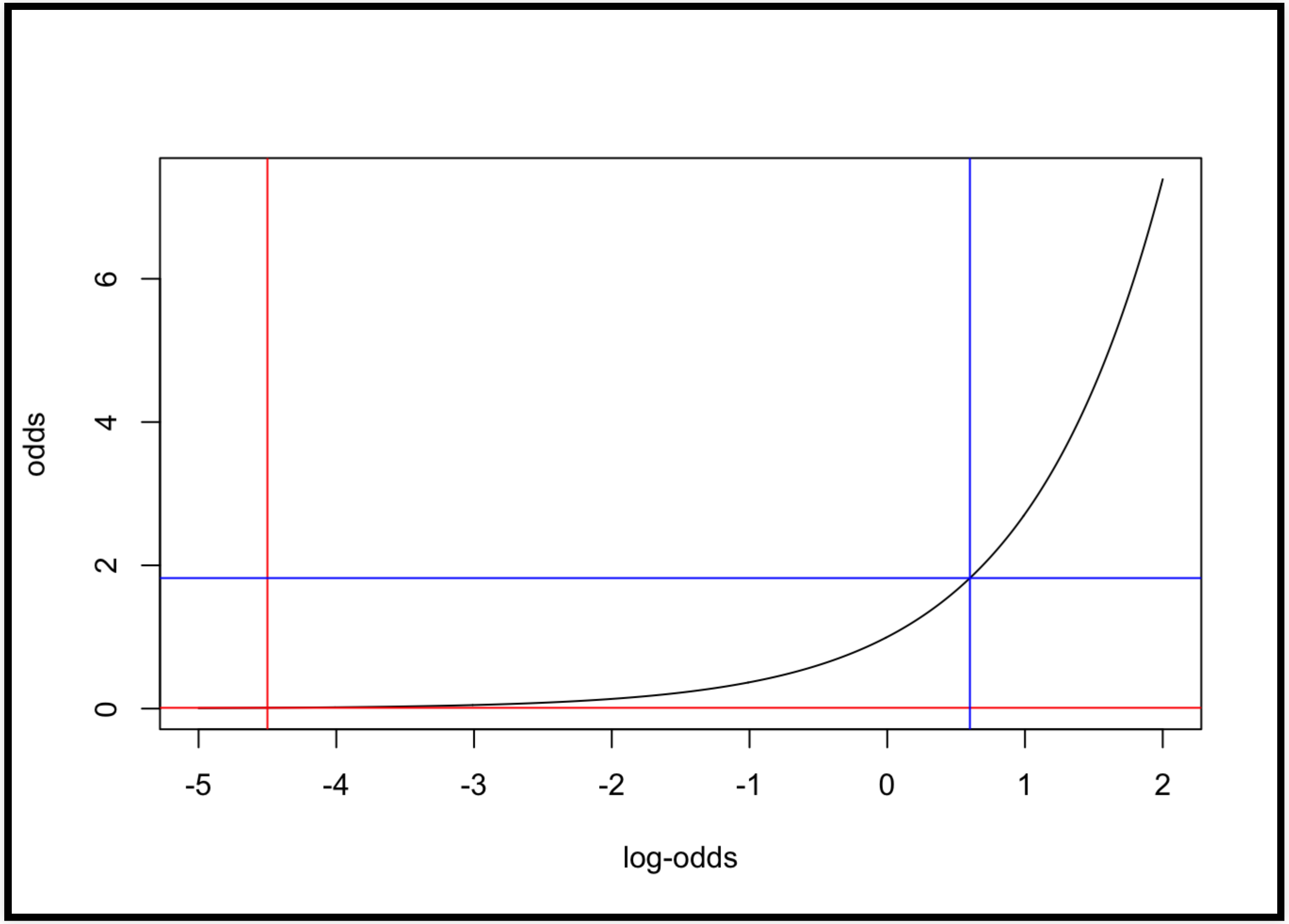
- Linear model
  - `income ~ age`
  - intercept = 10,0000
  - b1 = 5,000

Your turn

- Linear model
  - `income ~ age*gender`
  - intercept = 10,0000
  - b1 (age) = 4,000
  - b2 (gender, 0 = female, 1 = male) = 12,000
  - b3 (age*gender) = 2,000

Your turn

- Logistic model
  - `failed ~ hours_spent`
  - intercept = -10.00
  - b1 = 0.60

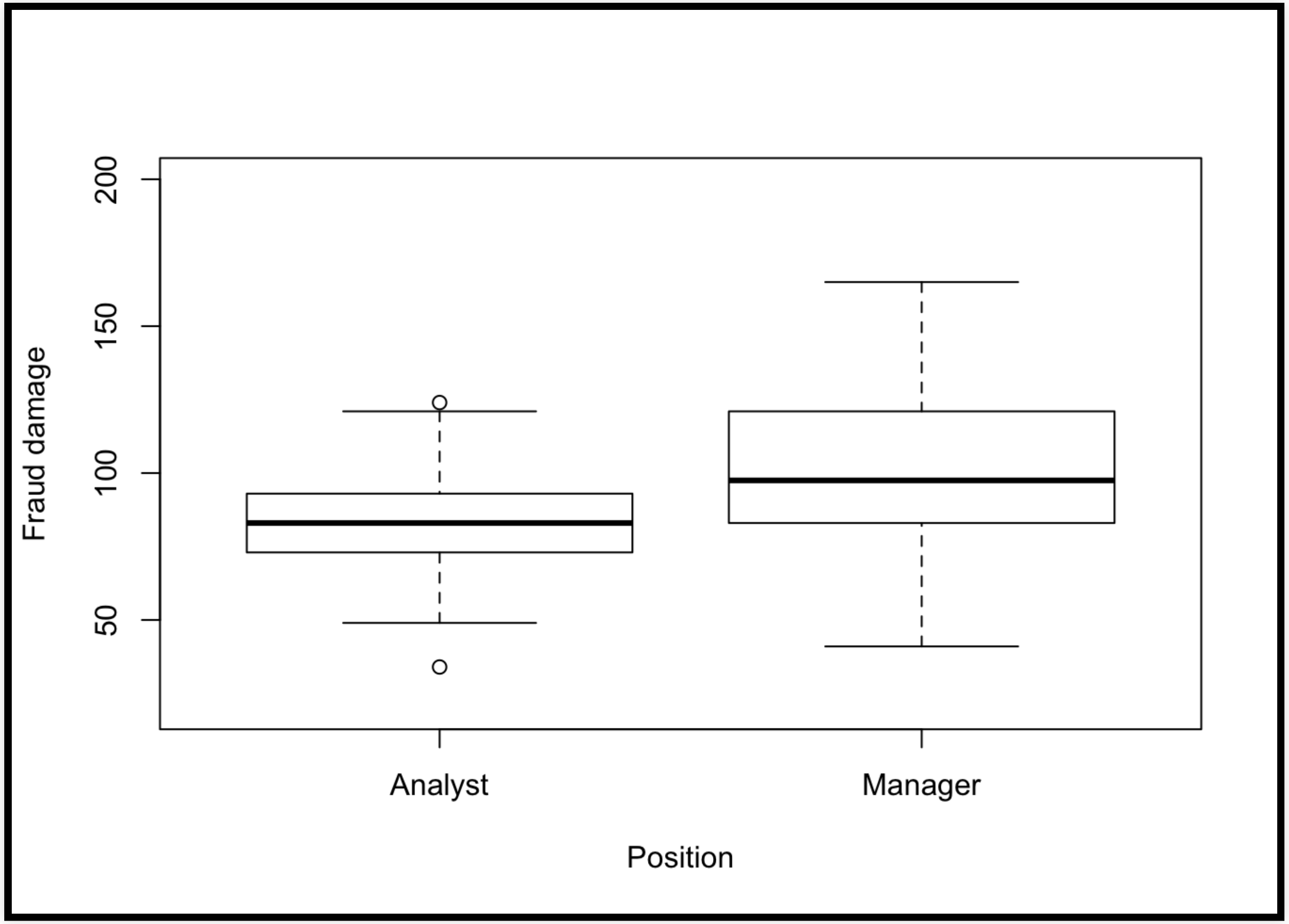


odds
0
2
4
6
-5
-4
-3
-2
-1
0
1
2
log-odds

Today

- Mean comparison
  - t-tests
  - ANOVA
  - GLM implementation

Mean comparisons


Fraud damage
200
150
100
50
Analyst
Manager
Position

Numerical values

Mean:

```
## Analyst Manager
##   82.92  101.02
```

SD:

```
##  Analyst  Manager
## 18.11735 27.84028
```

Is fraud by managers more damaging than fraud by analysts?

Non-statistical answer

Yes, because:

```
## Analyst Manager
##   82.92  101.02
```

And: 101.50 > 78.66

Why is this problematic?

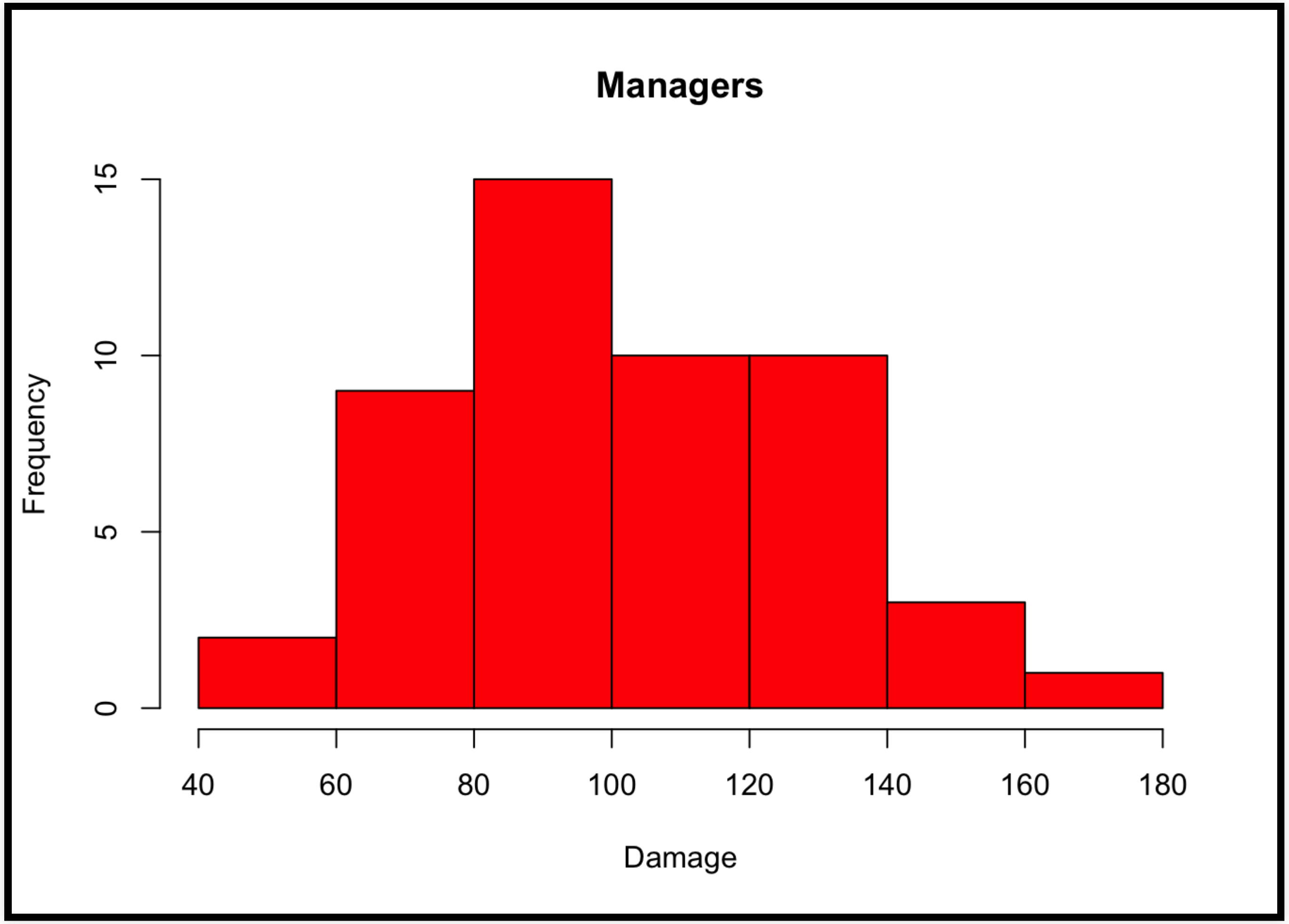
Sample –> Population

```
# damage_ = c(round(rnorm(10000, 100, 30), 0), round(rnorm(10000, 80, 20
# df_ = data.frame(damage = damage, role = rep(c('Manager', 'Analyst'),
# plot(df_$damage, col=df_$role)
```

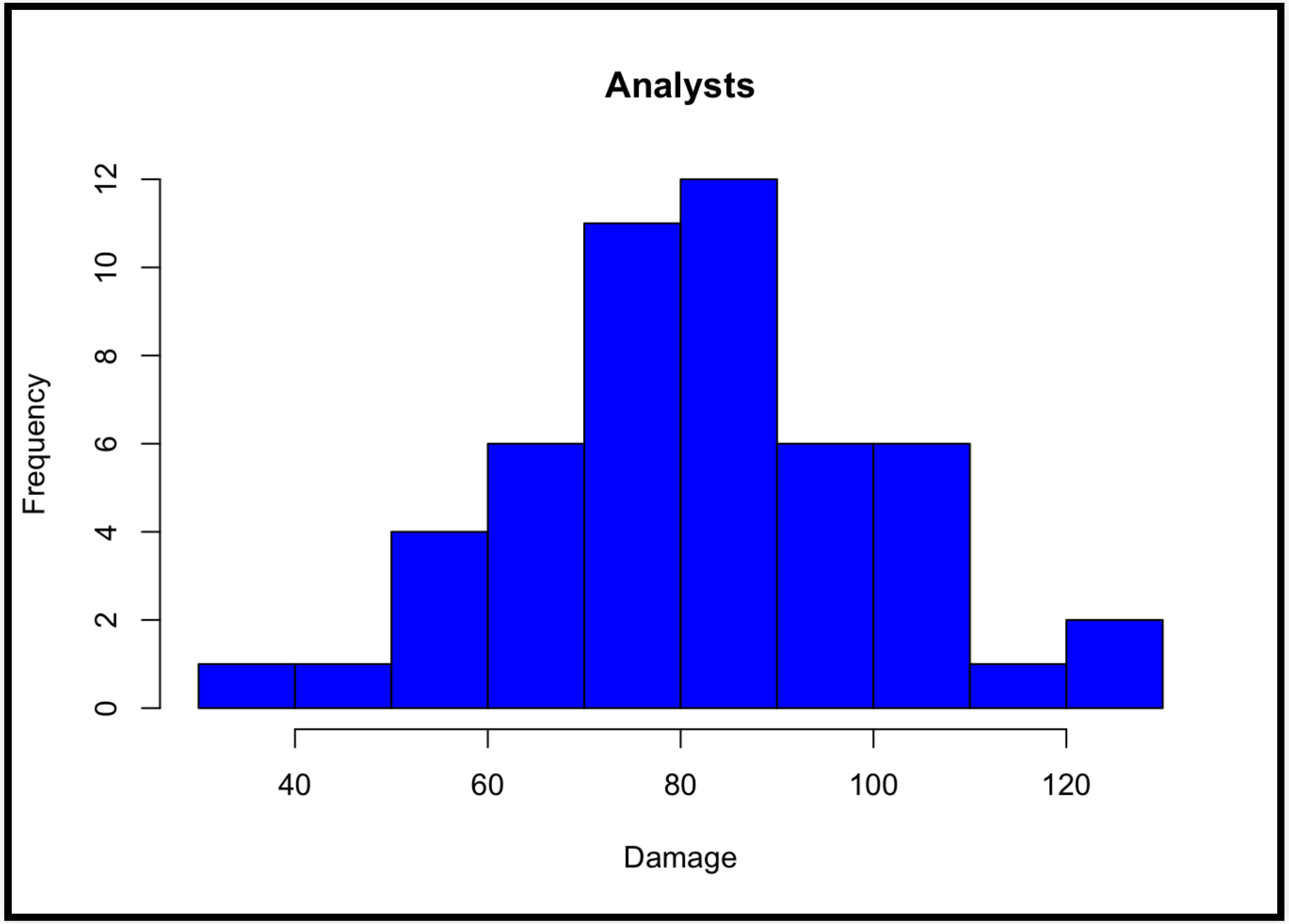
Inferential statistics

- infere parameters of the population
- from a sample (of that population)

All data stem from distributions

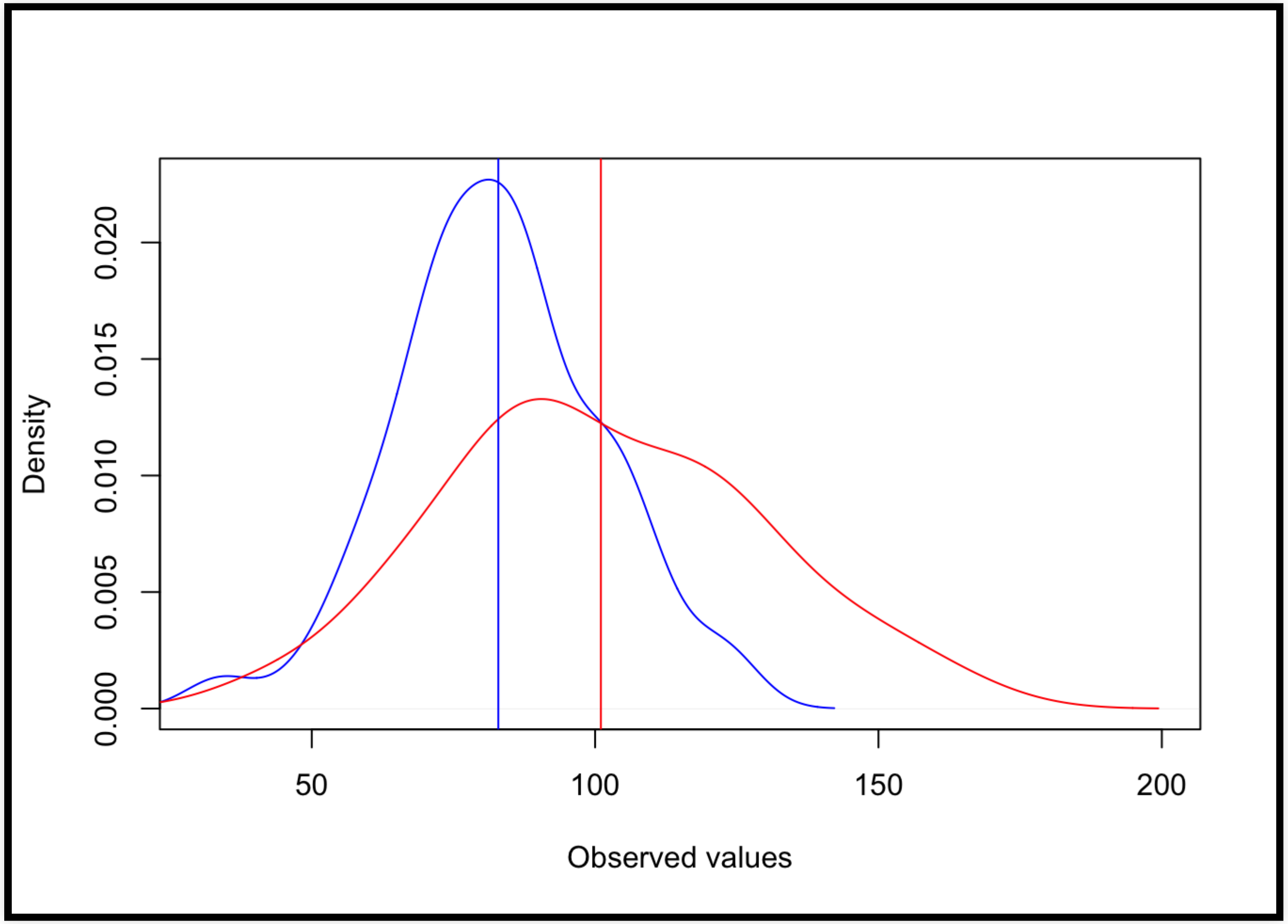


All data stem from distributions



Comparison of 2 groups

- Are the means of the groups statistically significantly different?
- Different phrasing: do the samples stem from different distributions?


Density
0.000
0.005
0.010
0.015
0.020
50
100
150
200
Observed values

Do the samples stem from different distributions?

- decision criterion needed
- problem: there is always some overlap
- wanted: a threshold that we deem practically irrelevant in overlap

Hypothesis testing

- Null hypothesis: there is no differece (i.e. Managers == Analysts)
- Alternative hypothesis: there is a difference (e.g. Managers > Analysts)

```
mean(df$damage[df$role == 'Manager'] - df$damage[df$role == 'Analyst'])
```

```
## [1] 18.1
```

> *Wanted: a value that expresses the frequentist probability of observing the mean difference of 18.10 (or more extreme) if the null hypothesis were true.*

–> **called the p-value**

Thresholds for the p-value

- aim: test whether you can reject the null hypothesis
- wanted: a threshold that we deem practically irrelevant in overlap
- threshold is called the significance level (alpha level)
- analogous to: a threshold that we deem acceptable in making a Type I error
- used to be: `p < .05`
- changes under way:
  - redefine to `p < .005`
  - justify your own threshold
  - rejection of p-values altogether

Calculating the p-value

For two groups: t-test

- Assumes:
  - normality of outcome variable
  - independence of observations
  - equality of variance (corrected by default)

(non-parametric tests –> next week)

t-test

```
t.test(df$damage ~ df$role)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  df$damage by df$role
## t = -3.8531, df = 84.191, p-value = 0.0002269
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -27.441161  -8.758839
## sample estimates:
## mean in group Analyst mean in group Manager 
##                 82.92                101.02
```

t-test reporting

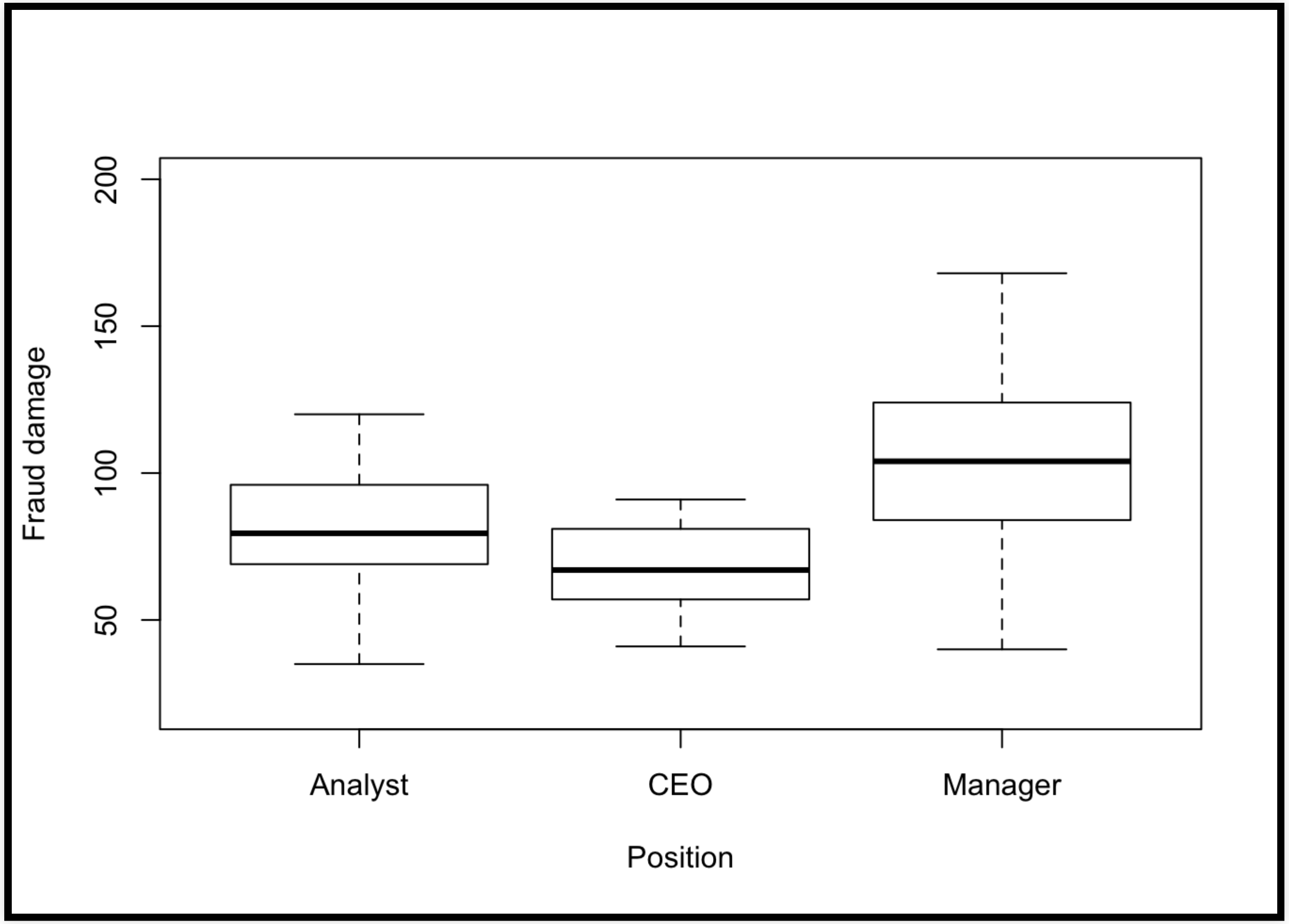
```
t = -3.8531, df = 84.191, p-value = 0.0002269
```

> *The damage in \$ lost was higher for managers (M = 101.92, SD = 27.83) than for analysts (M = 82.92, SD = 18.12), $t(84.19) = -3.85$, $p < .001$.*

Note: always three decimals for the p-value, unless $p < .001$.

(more in week 7)

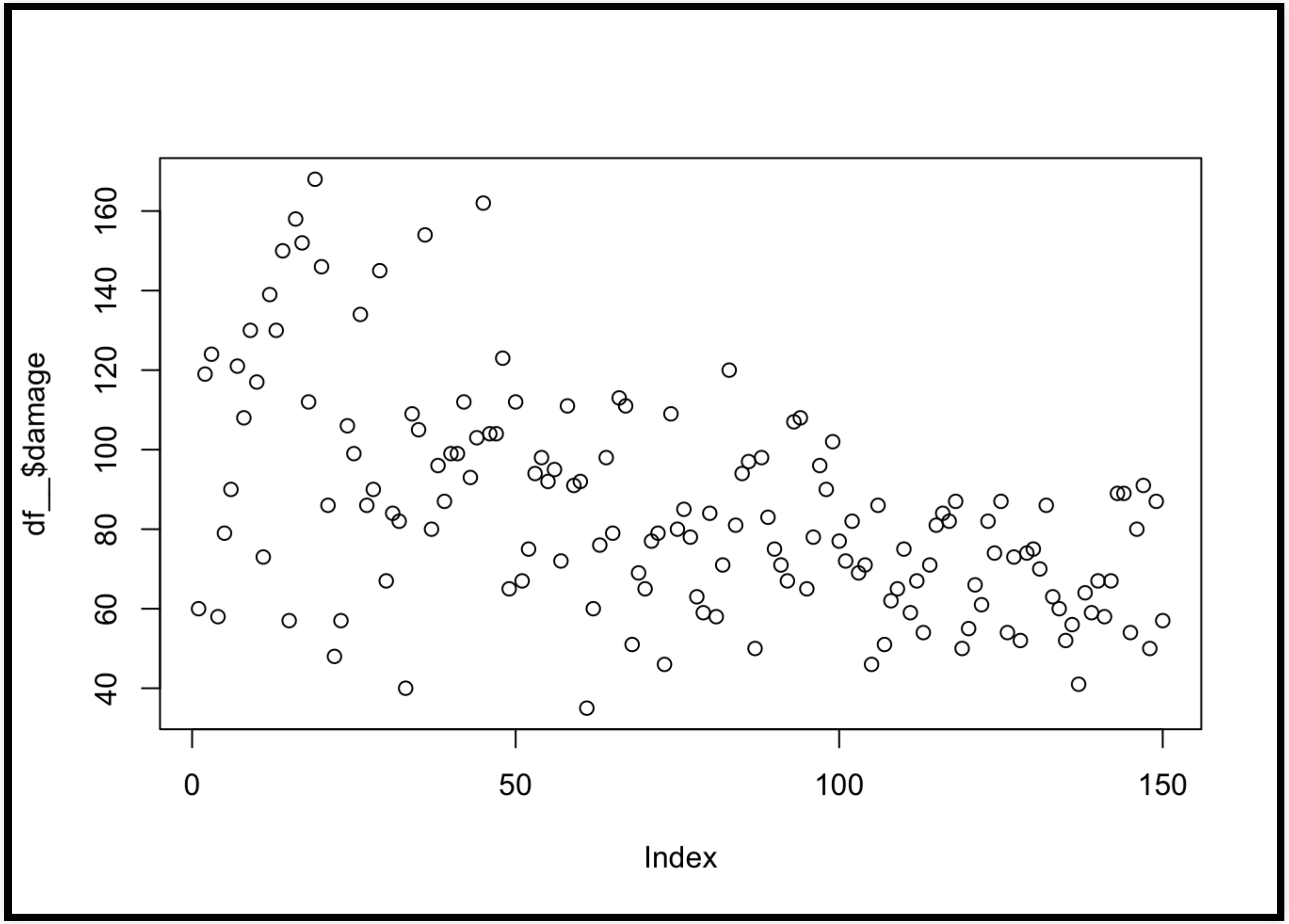
What if there are three groups?


Fraud damage
200
150
100
50
Analyst
CEO
Manager
Position

Idea

- Test whether all three means are the same
- Could use 3 t-tests:
  - Analysts vs CEOs
  - CEOs vs Managers
  - Managers vs Analysts
- problem: multiple t-tests increase Type 1 error
- Null hypothesis: Analyst = CEO = Manager
- Alt. hypothesis: the means are affected by the factor Position (3 levels)

Variance

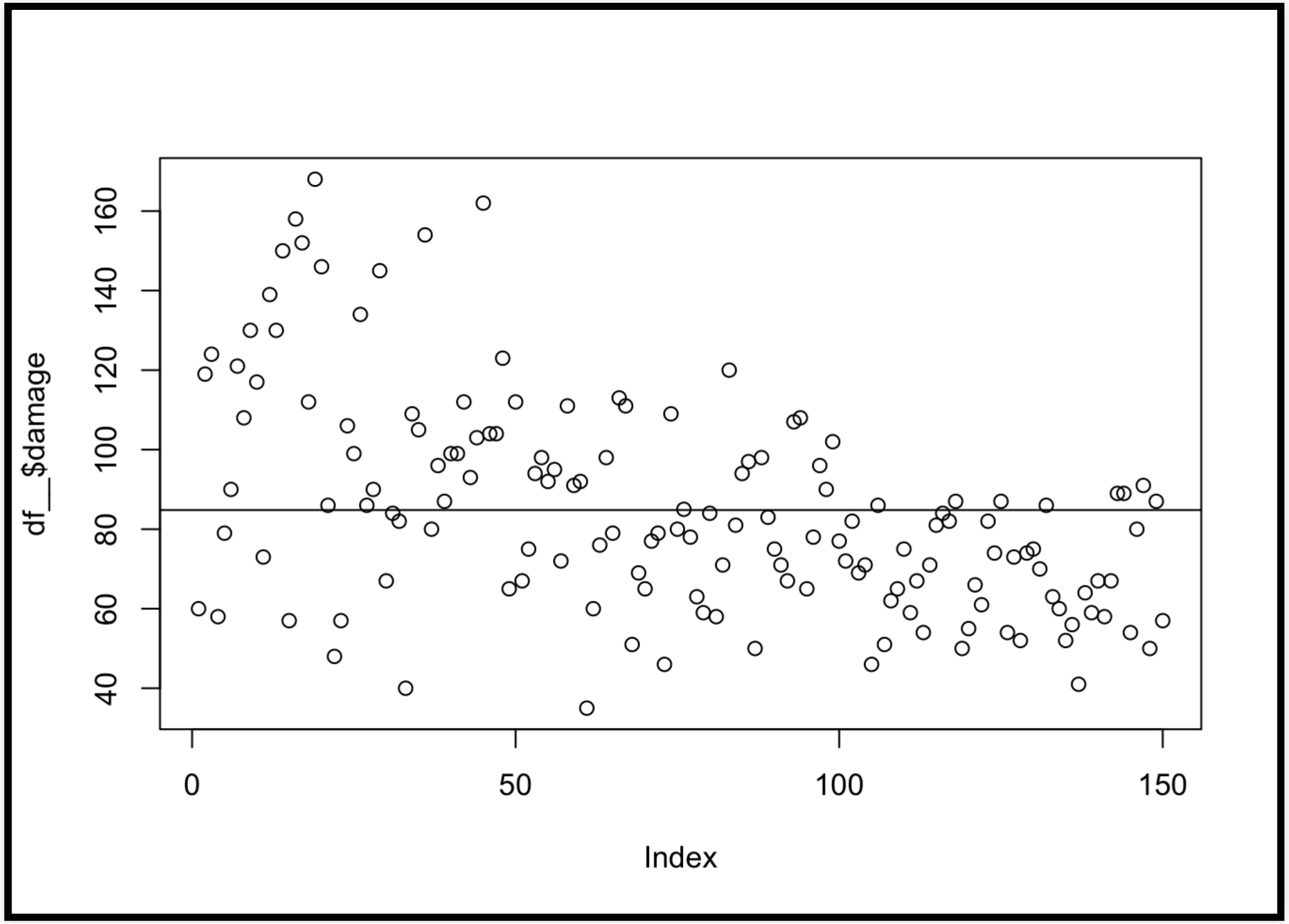


Analysis of VARIANCE

Total variance = explained variance + unexplained variance

- **explained variance**: variation that is attributable to the factor Position
- **unexplained variance**: variation not attributable to the factor Position

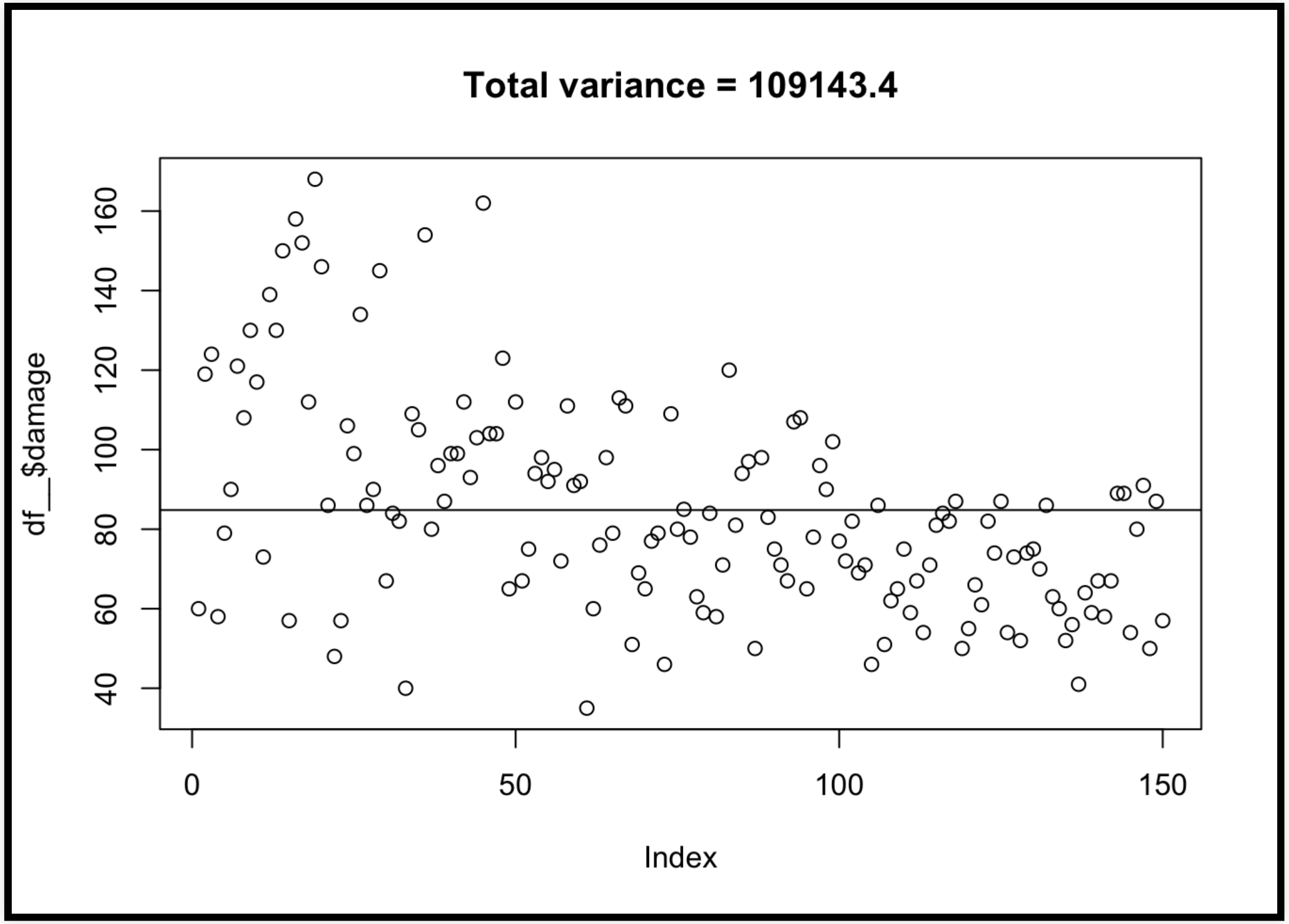
Total variance



Total variance

damage	grandmean	squared_diff
60	84.80667	615.37071
119	84.80667	1169.18404
124	84.80667	1536.11738
58	84.80667	718.59738
79	84.80667	33.71738
90	84.80667	26.97071
121	84.80667	1309.95738
108	84.80667	537.93071
130	84.80667	2042.43738
117	84.80667	1036.41071

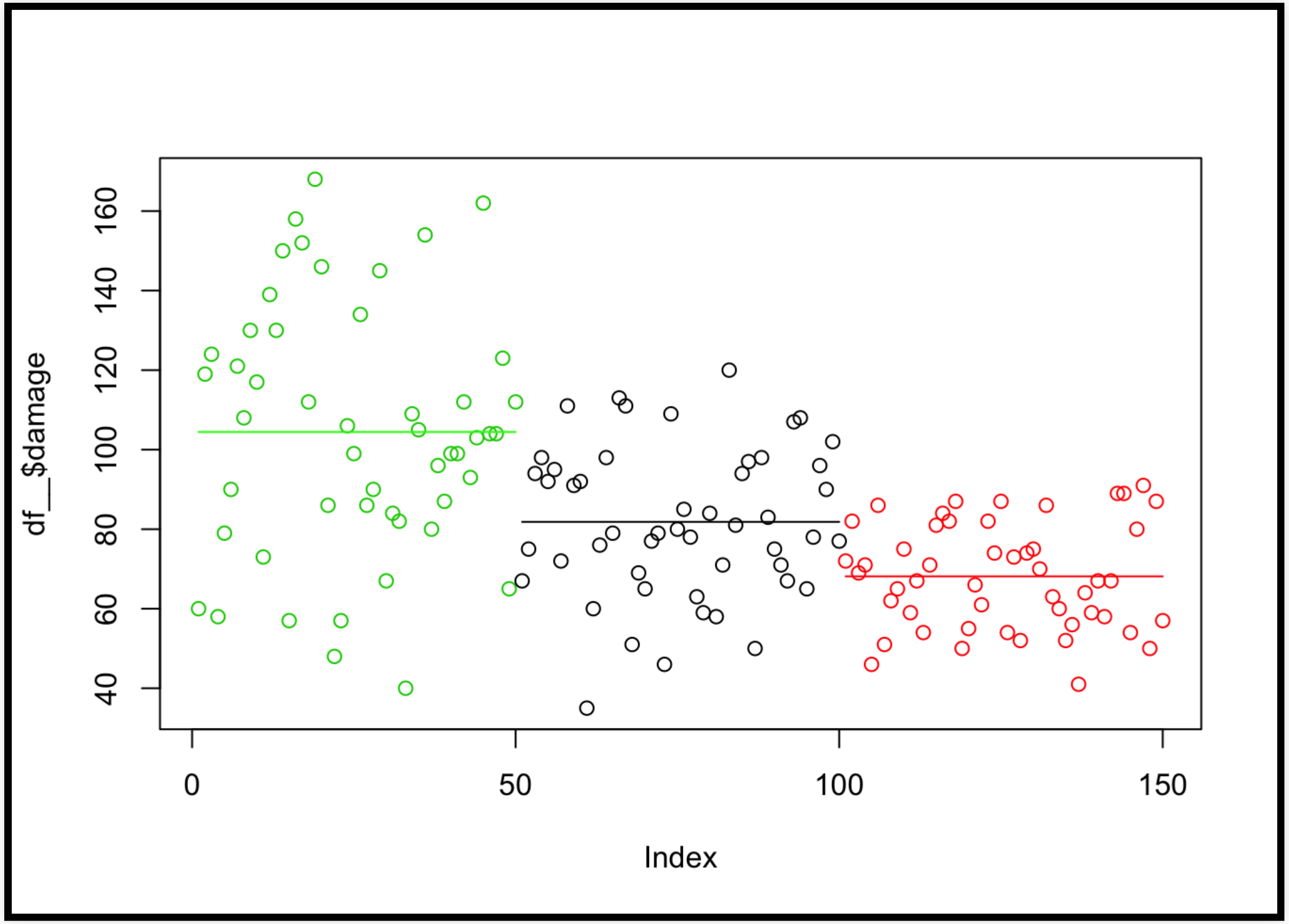
Total variance



Partitioning of variance

Total variance = explained variance + unexplained variance

109143.40 = **explained variance** + unexplained variance


df__$damage
40
60
80
100
120
140
160
0
50
100
150
Index

Partitioning of variance

109143.40 = **explained variance** + unexplained variance

explained variance

- (variance group 1 * size of group 1)
- (variance group 2 * size of group 2)
- (variance group 3 * size of group 3)

Shortcut:

- (mean group 1 - grand mean)^2
- (mean group 2 - grand mean)^2
- (mean group 3 - grand mean)^2
- sum these and multiply by *n* per group

Explained variance

group	groupmean	grandmean	squared_diff
Manager	104.44	84.81	385.47
Analyst	81.84	84.81	8.80
CEO	68.14	84.81	277.78
SUM	-	-	672.05

```
672.05 * 50
```

```
## [1] 33602.5
```

Partitioning of variance

109143.40 = **33602.50** + unexplained variance

–>

unexplained variance = 75540.90

We want to know: how much more variance is explained compared to non-explained.

Degrees of freedom

source	variance
explained (factor Position)	33602.50
unexplained	75540.90
total	109143.40

But: different number of values used for calculation!

Degrees of freedom (df)

df = number of values that are free to vary

source	variance	df
explained (factor Position)	33602.50	2
unexplained	75540.90	147
total	109143.40	149

- total_df = explained_df + unexplained_df
- total_df = *n* - 1
- unexplained_df = *n* - k
- k = number of levels of the factor (here: Position)

Corrected table of variance

source	variance	df	mean SSq
explained (factor Position)	33602.50	2	16801
unexplained	75540.90	147	514
total	109143.40	149	-

How much more variance is explained compared to non-explained?

The F-test

F-statistic = mean SSq explained / mean SSq unexplained

```
16801 / 514
```

```
## [1] 32.68677
```

> *The explained variance (due to the factor Position) is 32.69 times higher than the unexplained variance.*

Is this significant?

ANOVA in R

```
summary(aov(df__$damage ~ df__$role))
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)    
## df__$role       2  33602   16801   32.69 1.79e-12 ***
## Residuals     147  75541     514                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The one-way ANOVA revealed that there was a significant main effect of Position (CEO, Manager, Analyst) on the damage in USD, $F(2, 147) = 32.69$, $p < .001$.

ANOVA as omnibus test

Now we know whether there is an overall effect …

Important: only now do you have statistical justification proceed with *follow-up contrasts*.

If ANOVA *ns* –> analysis stops here!!!!!!!

ANOVA interpretation

Step 1: The one-way ANOVA revealed that there was a significant main effect of *Position* (CEO, Manager, Analyst) on the damage in USD, $F(2, 147) = 32.69$, $p < .001$.

Step 2: follow-up contrasts

- t-test: CEO vs Manager
- t-test: CEO vs Analysts
- t-test: Manager vs Analysts

Follow-up contrasts

```
t.test(df__$damage[df__$role != 'Analyst'] ~ df__$role[df__$role != 'Ana
t.test(df__$damage[df__$role != 'Manager'] ~ df__$role[df__$role != 'Mana
t.test(df__$damage[df__$role != 'CEO'] ~ df__$role[df__$role != 'CEO'])
```

comparison	t	df	p
CEO vs Manager	-7.4734	65.80	< .001
CEO vs Analysts	4.1717	87.70	< .001
Manager vs Analysts	-4.33	80.31	< .001

ANOVA interpretation

Step 1: The one-way ANOVA revealed that there was a significant main effect of *Position* (CEO, Manager, Analyst) on the damage in USD, $F(2, 147) = 32.69$, $p < .001$.

Step 2: follow-up contrasts

```
knitr::kable(tapply(df__$damage, df__$role, mean), col.names = c('mean'))
```

	mean
Analyst	81.84
CEO	68.14
Manager	104.44

Follow-up contrasts revealed that the damaga (in \$) was smaller when caused by CEOs (M = 68.14, SD = 13.31) than when caused by Managers (M = 104.44, SD = 31.67), $t(65.80) = -7.47$, $p < .001$. …

Types of ANOVAs

- One predictor (factor)
  - One-way ANOVA
  - 2+ levels
- Two predictors (factors)
  - Two-way ANOVA
  - 2+ levels by 2+ levels
  - e.g. `role*gender` –> 2 by 3 ANOVA
- Always note whether the factor is within-subjects or between-subjects
  - All factors within: within-subjects ANOVA
  - All facors between: between-subjects ANOVA
  - Both: mixed ANOVA

ANOVA & GLM

- both resemble each other (main effects, interaction)
- more so: they are the same for categorical predictors
- predictors in ANOVA: means vs grandmean
- predictors in linear regression: dummy coded levels

ANOVA & GLM

```
lm(damage ~ role, data=df__)
```

	beta	SE	t-statistic	p-value
(Intercept)	81.84	3.205884	25.528057	0.0000000
roleCEO	-13.70	4.533805	-3.021744	0.0029654
roleManager	22.60	4.533805	4.984775	0.0000017

Beta coefficients are the group means respective to the reference group!

ANOVA & GLM

	beta
(Intercept)	81.84
roleCEO	-13.70
roleManager	22.60

```
knitr::kable(tapply(df__$damage, df__$role, mean), col.names = c('mean'))
```

	mean
Analyst	81.84
CEO	68.14
Manager	104.44

ANOVA & GLM

Look at the output:

Linear Model:

```
F-statistic: 32.69 on 2 and 147 DF,  p-value: 1.793e-12
```

ANOVA:

```
              Df Sum Sq Mean Sq F value   Pr(>F)    
df__$role      2  33602   16801   32.69 1.79e-12 ***
Residuals    147  75541     514                     
```

ANOVA & GLM

Same omnibus logic:

- if the F-statistic in the regression is not significant
- … then you cannot conclude an overall effect of the factor

t-test & ANOVA

- ANOVA = regression with categorical predictor(s) with 2+ levels
- t-test = one-way ANOVA with 2 two levels.
- –> t-test = regression with one categorical predictors with 2 levels

t-test & GLM

- t-test

```
#Managers and Analysts only
t = -3.8531, df = 84.191, p-value = 0.0002269
```

- as ANOVA

```
summary(aov(df$damage ~ df$role))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## df$role       1   8190    8190   14.85 0.000208 ***
## Residuals    98  54063     552                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- ANOVA and t-test: $F = t^2$ -> $t = \sqrt{F}$

```
(-3.8531)^2
```

```
## [1] 14.84638
```

```
sqrt(14.84)
```

```
## [1] 3.852272
```

t-test & GLM

If the ANOVA is a linear regression,
so is the t-test:

```
lm(damage ~ role, data=df)
```

```
knitr::kable(coefficients(summary(lm(damage ~ role, data=df))))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	82.92	3.321626	24.963681	0.0000000
roleManager	18.10	4.697488	3.853123	0.0002083

RECAP

- comparing 2 groups
- t-test
- t-test as GLM
- comparing multiple groups
- ANOVA as GLM

Outlook

Tutorial

- logistic regression
- ANOVA

Next week

- Non-parametric methods
- discrete data analysis

END