Behavioural Security Group

Usable Security and Privacy

Crash Course:
Inferential testing

Inferential Statistics – Theory

- Drawing conclusions from sample data
  - we usually only have a sample
- Considering the observed sample,
  - is a known population value different?
  - are two observed values different from each other?
  - are two variables related to each other?

- Hypothesis testing
  - p-values
  - effect size
  - power / sample size

Hypothesis Testing (two tailed)

- Hypothesis:
  - The SUS score of condition A is different from condition B

- Null-hypothesis:
  - The SUS score of conditions A and B are equal

- Statistical test to reject the null-hypothesis (two tailed)
  - If $p<0.05$ then reject null-hypothesis
    - i.e. ~~The SUS score of conditions A and B are equal~~
    - Ergo: The SUS score of condition A is different from condition B
  - if $p>=0.05$ then fail to reject the null-hypothesis
    - i.e. no statement can be made
    - NOT: The SUS score of conditions A and B are equal

Hypothesis Testing (one tailed)

- Hypothesis:
  - The SUS score of condition A is greater than condition B
- Null-hypothesis:
  - The SUS score of conditions A is not greater than condition B
- Statistical test to reject the null-hypothesis (one tailed)
  - Test is twice as powerful
  - Should only be done if there is good theoretical grounds why other tail can be ruled out
  - Rarely the case for us

Hypothesis Nomenclature

- H0: null hypothesis
- H1: alternative hypothesis
- A statistical tests will or will not reject H0
  - implies that H1 will be accepted
  - you cannot accept H0

The p-value

The probability of obtaining a difference as large as the observed one in the sample

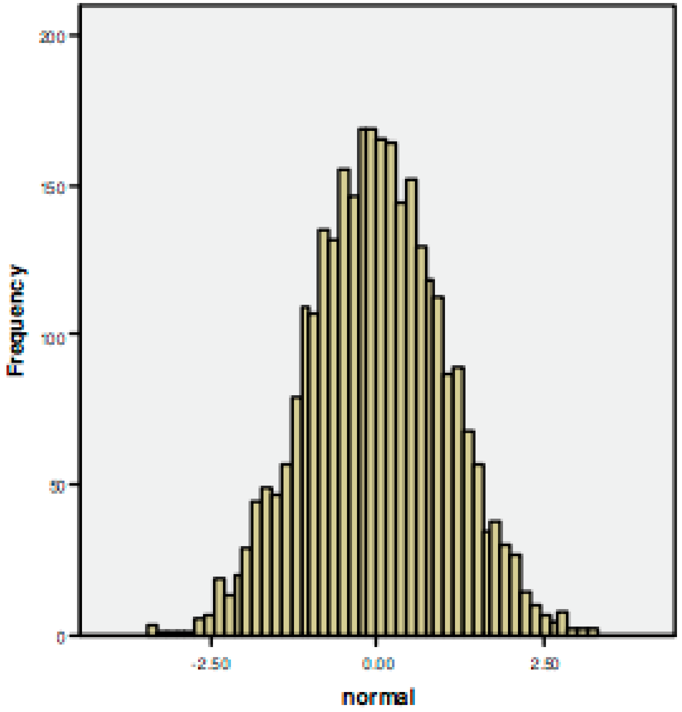
- if the null hypothesis was true.
- i.e. in truth there is no difference and the observed difference is down to chance

The p-value

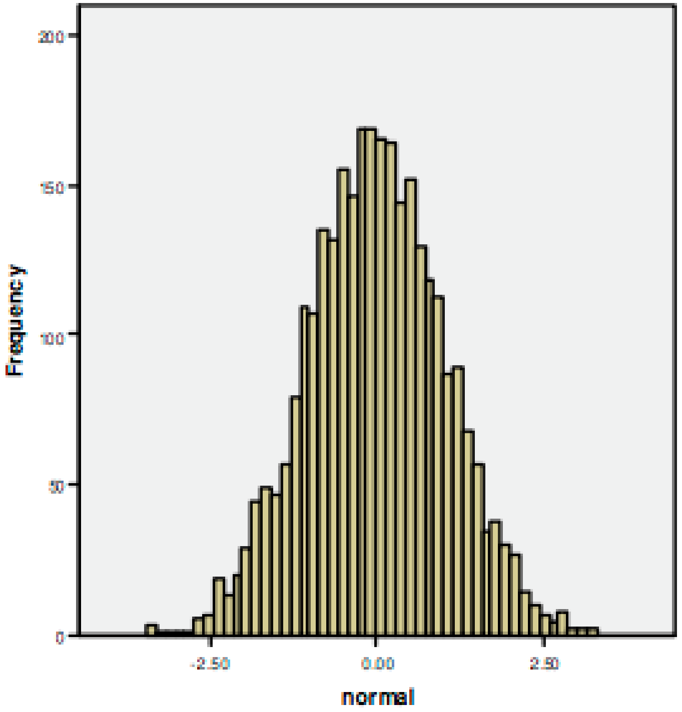
The probability of obtaining a difference as large as the observed one in the sample

- if the null hypothesis was true.
- i.e. in truth there is no difference

null hypothesis is true – both populations are equal



Student Population



Developer Population

The p-value

The probability of obtaining a difference as large as the observed one in the sample

- if the null hypothesis was true.
- i.e. in truth there is no difference

	H0 is true	**H0 is false**
Reject H0	Type I error *false positive* α	*correct* 1-β: Power
Fail to reject H0	*correct* 1-α	Type II error *false negative* β

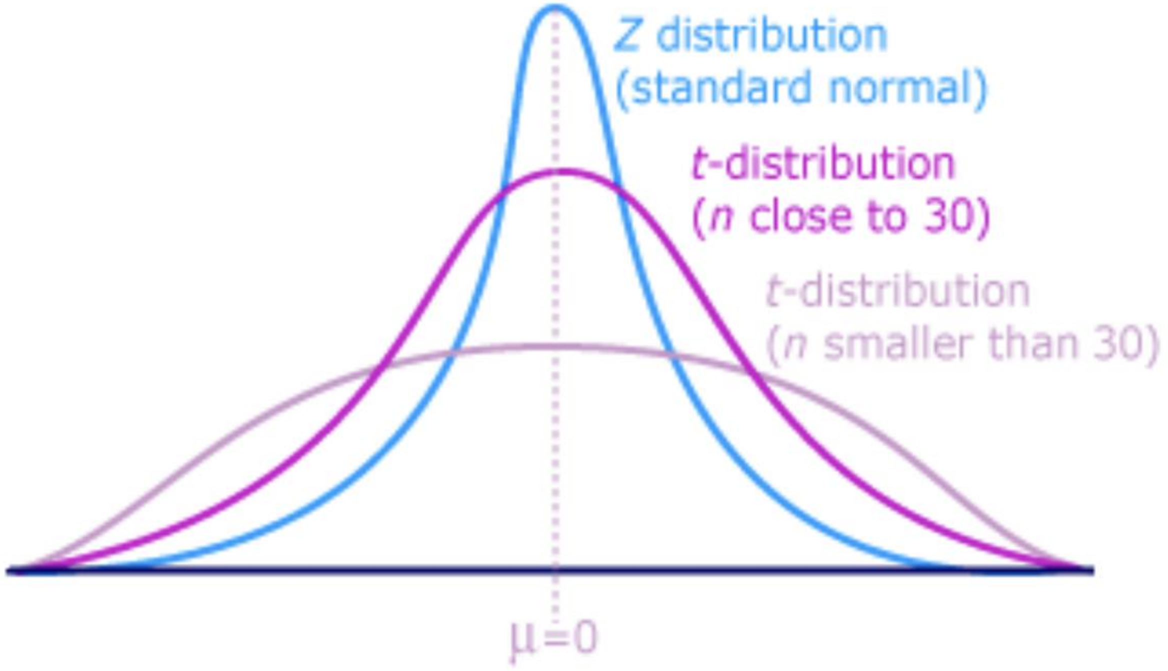
- p depends on effect size, variance and sample size
- we usually set α to .05, i.e. we look at least for $p < .05$
  - if $p < .05$ a result is “statistically significant” or “significant at the .05 level”.

Why 0.5?

Guinness is good for your statistics

Guinness - a data driven company

- William S. Gosset
  - Worked for Guinness from 1899 to 1937 as an experimental brewer
  - Needed to know how many samples to take to make an inference on batches of beer
  - Invented the “Student’s t-distribution”



Why 0.05?

- Book: Statistical Methods for Research Workers (1925)
  - Ronald Fisher proposed the level p = 0.05,
  - i.e. a 1 in 20 chance of being a chance occurrence
    - when drawing samples from two populations where the null hypothesis is true

Hypothesis Testing 0.05?

on average
one in twenty studies will meet the .05 threshold by random chance

It’s time to talk about ditching statistical significance,
Nature Editorial 567, 283 (2019)

Moving to a World Beyond “$p < 0.05$”
Wasserstein et al. American Statistical Association (2019)

Hypothesis Testing

- Everything is significant with a large enough sample!
  - presentation of results should address why this is meaningful.
- Effect size
  - the degree of relationship or extent of relationship
  - e.g. correlation measure (e.g. Pearson’s r), odds ratio, etc.

Power

- The probability of correctly rejecting the null hypothesis when it is false in the population
  - i.e. is our test even able to reliably detect an effect
  - “A study with insufficient power might not be worth doing.”

- Power is influenced by
  - effect size (in the population)
  - variance (in the population)
  - sample size
  - desired alpha

- Desired power can be used to estimate necessary sample size
  - commonly power is desired to be ≈ .8

Power Table t-Test / Cohen’s d

Power Tables for Effect Size d
(from Cohen 1988, pg. 55)

two-tailed α = .05 or one-tailed α = .025

	d										
Power	**.10**	**.20**	**.30**	**.40**	**.50**	**.60**	**.70**	**.80**	**1.0**	**1.20**	**1.40**
.25	332	84	38	22	14	10	8	6	5	4	3
.50	769	193	86	49	32	22	17	13	9	7	5
.60	981	246	110	62	40	28	21	16	11	8	6
2/3	1144	287	128	73	47	33	24	19	12	9	7
.70	1235	310	138	78	50	35	26	20	13	10	7
.75	1389	348	155	88	57	40	29	23	15	11	8
.80	1571	393	175	99	64	45	33	26	17	12	9
.85	1797	450	201	113	73	51	38	29	19	14	10
.90	2102	526	234	132	85	59	44	34	22	16	12
.95	2600	651	290	163	105	73	54	42	37	19	14
.99	3675	920	409	231	148	103	76	58	38	27	20

- 0.2 – 0.5 small effect
- 0.5 – 0.8 medium effect
- > 0.8 large effect

TESTING IN R

R Pitfalls

- Data types
  - factor / nominal
  - numerical / interval / ratio

```
df$var
[1] 1 3 5 4
Levels: 1 3 4 5
is.factor(df$var) true
is.numeric(df$var) false

df$var = as.factor(df$var) #converts to factor
df$var = as.numeric(df$var)   #converts to numeric representation of levels
[1] 1 2 4 3
df$var=as.numeric(as.character(df$var)) #converts to numeric
[1] 1 3 5 4
```

R ~

- ~ is part of a formula
  - thing on the left of the ~ is the dependent variable and the things on the right are the indpendent variables

Tests of Proportion / Counts

Samples	Level	Tests
1	2	One-sample χ^2 test, binomial test
1	3+	One-sample χ^2 test, multinomial test
2+	2+	N-sample χ^2 test, Fisher's exact test

```
                                                                        Please.state.your.gender
Have.you.every.written.code.that.has.been.used.outside.of.the.university.context.. female male
                                                                     No       1    1
                                                                     Yes      7   23
```

A Multitude of Possible Analyses

- One Sample
  - Test one variable against a known value
- Two Samples
  - compare one variable in two samples
  - independent samples
    - between subjects conditions
    - e.g. male/female
  - dependent samples
    - within subjects conditions
    - repeated measures

Tests of Proportion / Counts

Samples	Levels	Tests
1	2	One-sample χ^2 test, binomial test
1	3+	One-sample χ^2 test, multinomial test
2+	2+	N-sample χ^2 test, Fisher's exact test

Example Proportions Test

- Hypothesis / Alternative Hypothesis / H1
  - It is not equally likely that students have had their code used outside a university context or not.
- Null Hypothesis / H0
  - It is equally likely that students have had their code used outside of a university context or not.
- Data:

	Have.you.every.written.code.that.has.been.used.outside.of.the.university.context..	Freq
1	No	2
2	Yes	30

Create Table for Testing

- p contains questionnaire data
- factor: convert variable to nominal factor
- xtabs: creates a contingency table
  - https://stat.ethz.ch/R-manual/R-devel/library/stats/html/xtabs.html

````
```{r}
#convert to nominal factor
p$Have.you.every.written.code.that.has.been.used.outside.of.the.university.context.. =
factor(p$Have.you.every.written.code.that.has.been.used.outside.of.the.university.context..)

#create table
codeTabs = xtabs( ~ Have.you.every.written.code.that.has.been.used.outside.of.the.university.context.., data=p)
codeTabs # show counts
```
````

```
Have.you.every.written.code.that.has.been.used.outside.of.the.university.context..
 No Yes
  2  30
```

One Sample Chi Squared Test

```{r}
#one sample chi squared test
chisq.test(codeTabs)

```

```
        Chi-squared test for given probabilities

data:  codeTabs
X-squared = 24.5, df = 1, p-value = 7.431e-07
```

Binomial Test

```{r}
#binomial test
binom.test(codeTabs)
```

```
        Exact binomial test

data:  codeTabs
number of successes = 2, number of trials = 32, p-value = 2.463e-07
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.007660736 0.208069430
sample estimates:
probability of success 
                0.0625 
```

Tests of Proportion / Counts

Samples	Levels	Tests
1	2	One-sample χ^2 test, binomial test
1	3+	One-sample χ^2 test, multinomial test
2+	2+	N-sample χ^2 test, Fisher's exact test

Example Proportions Test

- Hypothesis / Alternative Hypothesis / H1

- Null Hypothesis / H0

- Data:

```{r}
pSkill = factor(p$How.would.you.judge.yourself.as.a.programmer.)
skillTabs = xtabs( ~ pSkill)
skillTabs
```

```
pSkill
    Beginner       Expert Intermediate
           5            4           23
```

Example Proportions Test

- Hypothesis / Alternative Hypothesis / H1
  - There are unequal counts of beginners, intermediates and experts
- Null Hypothesis / H0
  - There is no difference in counts of beginners, intermediates and experts
- Data:

````
```{r}
pSkill = factor(p$How.would.you.judge.yourself.as.a.programmer.)
skillTabs = xtabs( ~ pSkill)
skillTabs
```
````

```
pSkill
    Beginner       Expert Intermediate
           5            4           23
```

Multinomial Test

````
```{r}
library(XNomial)
xmulti(skillTabs, c(1/3, 1/3, 1/3), statName="Prob")
```
````

```
P value (Prob) = 6.008e-05
```

Post-hoc test with correction

````
```{r}
# post hoc binomial tests with correction for multiple comparisons
beginner = binom.test(sum(p$How.would.you.judge.yourself.as.a.programmer. =="Beginner"), nrow(p), p=1/3)
intermediate = binom.test(sum(p$How.would.you.judge.yourself.as.a.programmer. == "Intermediate"), nrow(p), p=1/3)
expert = binom.test(sum(p$How.would.you.judge.yourself.as.a.programmer. =="Expert"), nrow(p), p=1/3)

p.adjust(c(beginner$p.value, intermediate$p.value, expert$p.value), method="holm")
```
````

```
[1] 0.03795405248 0.00003536582 0.02672803498
```

What’s the deal with multiple testing?

Tests of Proportion / Counts

Samples	Levels	Tests
1	2	One-sample χ^2 test, binomial test
1	3+	One-sample χ^2 test, multinomial test
2+	2+	N-sample χ^2 test, Fisher's exact test

Example Proportions Test

- Hypothesis / Alternative Hypothesis / H1
  - It is more likely that one of the genders have had their code used outside a university context than the other (two tailed)
- Null Hypothesis / H0
  - There is no difference in likelihood between the two genders wrt. having had their code used outside of a university context.
- Data:

```
                                                                     Please.state.your.gender
Have.you.every.written.code.that.has.been.used.outside.of.the.university.context.. female male
                                                                   No          1    1
                                                                   Yes         7   23
```

Create Contingency Table

- factor: convert variable to nominal factor
- xtabs: creates a contingency table

```
```{r}

#convert to ordinal factor
p$Please.state.your.gender=factor(p$Please.state.your.gender)

codeTabsGender = xtabs( ~ Have.you.every.written.code.that.has.been.used.outside.of.the.university.context.. +
Please.state.your.gender, data=p) # the '+' sign indicates two vars

codeTabsGender

```
```

```
                                                                      Please.state.your.gender
Have.you.every.written.code.that.has.been.used.outside.of.the.university.context.. female male
                                                                               No       1    1
                                                                               Yes      7   23
```

Two sample Chi Squared Test

- Requirements
  - Independence of observations
  - All fields must have a count >5

```{r}
chisq.test(codeTabsGender) #since not all values are greater than 5 chisq might be inacuarte.
Use fisher's test instead.

```

```
Chi-squared approximation may be incorrect
	Pearson's Chi-squared test with Yates' continuity correction

data:  codeTabsGender
X-squared = 0, df = 1, p-value = 1
```

Fisher’s exact test

````
```{r}

fisher.test(codeTabsGender)

```
````

```
          Fisher's Exact Test for Count Data

 data:  codeTabsGender
 p-value = 0.4435
 alternative hypothesis: true odds ratio is not equal to 1
 95 percent confidence interval:
    0.03666286 268.20815022
 sample estimates:
 odds ratio
   3.139603
```

Comparing central tendency

Variables	Levels	Within / Between	Parametric	Non-parametric
1	2	B	Independent-samples t-test	Mann-Whitney U test aka. Wilcoxon rank sum test aka. Wilcoxon-Mann-Whitney-Test
1	3+	B	One-way ANOVA	Kruskal-Wallis test
1	2	W	Paired-samples t-test	Wilcoxon signed-rank test
1	3+	W	One-way repeated measures ANOVA	Friedman test
2+	3+	B	Factorial ANOVA Linear Models (LM)	Aligned Rank Transform (ART) Generalised Linear Models (GLM)
2+	3+	W	Factorial repeated measures ANOVA Linear Mixed Models (LMM)	Aligned Rank Transform (ART) Generalised Linear Mixed Models (GLMM)

Parametric vs Non-Parametric

- Parametric makes assumptions about distribution of data
- Analysis of Variance (ANOVA)
- 3 ANOVA assumptions
  - Independence
    - Subjects sampled independently of each other
    - Measures on subjects are independent of other subjects
    - Snowball sampling violates this assumption
  - Normality
    - Residuals are normally distributed
  - Homoscedasticity / Homogeneity of Variance
- Non-parametric tests usually have less power (when assumptions are met)

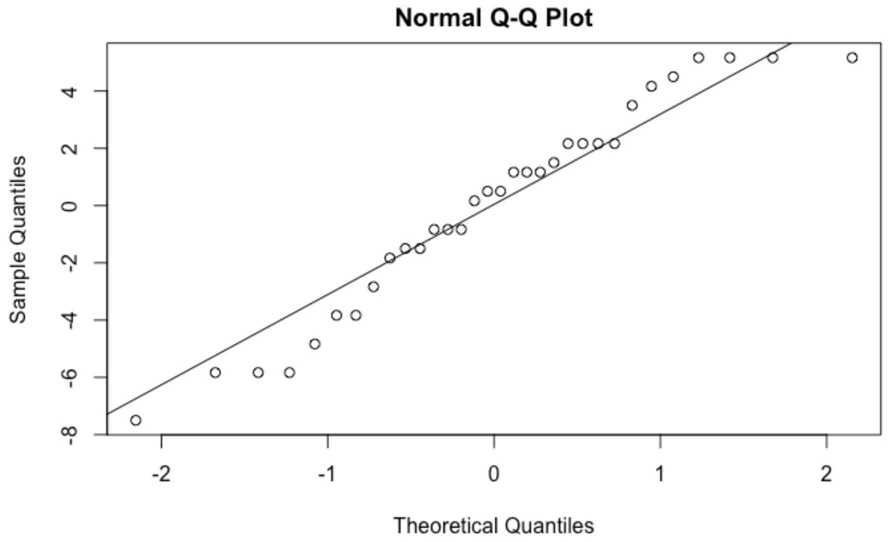
Test normality of residuals

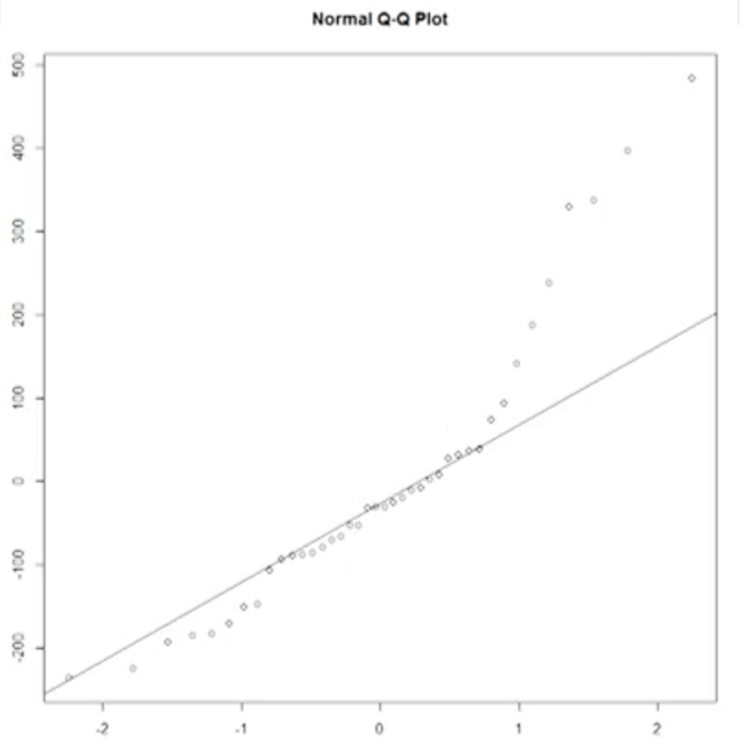
- Shapiro-Wilk-Test:
  - Null hypothesis: Data is distributed normally.
  - Significant test means null hypothesis has to be rejected (data likely not normal)

```
m = aov(How.many.years.of.programming.experience.do.you.have. ~ Please.state.your.gender, data=p) # fit model
shapiro.test(residuals(m)) # test residuals
qqnorm(residuals(m)); qqline(residuals(m)) # plot residuals
```

```
Shapiro-Wilk normality test

data:  residuals(m)
W = 0.95073, p-value = 0.1512
```





Test for homogeneity of variance

- Levene’s test
- Brown-Forsythe test
  - uses median, thus is more robust against outliers
- Null hypothesis: Variances are equal
  - Significant test means assumptions are violated

```
```{r}
# tests for homoscedasticity (homogeneity of variance)
library(car)
# Levene's test
leveneTest(How.many.years.of.programming.experience.do.you.have. ~ Please.state.your.gender, data=p, center=mean)
# Brown-Forsythe test
leveneTest(How.many.years.of.programming.experience.do.you.have. ~ Please.state.your.gender, data=p, center=median)
```
```

```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  1  2.7335 0.1091
      29
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  1  2.9251 0.09789 .
      29
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Comparing central tendency

Variables	Levels	Within / Between	Parametric	Non-parametric
1	2	B	Independent-samples t-test	Mann-Whitney U test aka. Wilcoxon rank sum test aka. Wilcoxon-Mann-Whitney-Test
1	3+	B	One-way ANOVA	Kruskal-Wallis test
1	2	W	Paired-samples t-test	Wilcoxon signed-rank test
1	3+	W	One-way repeated measures ANOVA	Friedman test
2+	3+	B	Factorial ANOVA Linear Models (LM)	Aligned Rank Transform (ART) Generalised Linear Models (GLM)
2+	3+	W	Factorial repeated measures ANOVA Linear Mixed Models (LMM)	Aligned Rank Transform (ART) Generalised Linear Mixed Models (GLMM)

One Sample

- One-sample t-test
  - “Is the mean of variable different from a known value?”
  - rarely used
  - if we know the population mean of a measure, e.g. age

```
t.test(age, mu = 45, alternative = "two.sided")

data:  age
t = -1.2026, df = 19, p-value = 0.2439
alternative hypothesis: true mean is not equal to 45
95 percent confidence interval:
 27.0501 49.8499
sample estimates:
mean of x
    38.45
```

Two Samples

- Between-subjects effects / independent samples
  - “Does the mean differ with respect to subgroups?”
- Independent Samples t-Test
  - Assumptions
    - Independence of observations
    - Homogeneity of variance
    - Normality of the dependent variable within groups / normality of residuals

Example t-test

- Hypothesis / Alternative Hypothesis / H1
  - The mean years of programming experience of one gender is higher than of the other (two tailed)
- Null Hypothesis / H0
  - No difference in mean between the two genders
- Data:

```
> summary(p[p$Please.state.your.gender == "female",]$How.many.years.of.programming.experience.do.you.have.)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   3.000   4.000   4.375   5.500   8.000 
> summary(p[p$Please.state.your.gender == "male",]$How.many.years.of.programming.experience.do.you.have.)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.000   4.000   7.000   7.217  10.000  19.000 
```

R example

- [off-screen]: Testing of assumptions

```
> t.test(How.many.years.of.programming.experience.do.you.have. ~ Please.state.your.gender, data=p, var.equal=TRUE)

	Two Sample t-test

data:  How.many.years.of.programming.experience.do.you.have. by Please.state.your.gender
t = -1.884, df = 29, p-value = 0.06962
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.9279983  0.2432156
sample estimates:
mean in group female   mean in group male 
            4.375000             7.217391 
```

- If homogeneity of variances is violated, but distribution is normal
  - Welch-Test
  - `t.test(dependent.variable ~ independent.variable, data=df)`

`var.equal = TRUE` is left out

Comparing central tendency

Variables	Levels	Within / Between	Parametric	Non-parametric
1	2	B	Independent-samples t-test	Mann-Whitney U test aka. Wilcoxon rank sum test aka. Wilcoxon-Mann-Whitney-Test
1	3+	B	One-way ANOVA	Kruskal-Wallis test
1	2	W	Paired-samples t-test	Wilcoxon signed-rank test
1	3+	W	One-way repeated measuress ANOVA	Friedman test
2+	3+	B	Facotrial ANOVA Linear Models (LM)	Aligned Rank Transform (ART) Generalised Linear Models (GLM)
2+	3+	W	Factorial repeated measures ANOVA Linear Mixed Models (LMM)	Aligned Rank Transform (ART) Generalised Linear Mixed Models (GLMM)

Wilcoxon Rank Sum Test

- Does not require normal distribution

```
```{r}
# Mann-Whitney U test / Wilcoxon Rank Sum test
library(coin)
wilcox_test(How.many.years.of.programming.experience.do.you.have. ~ Please.state.your.gender, data=p, distribution="exact")
```
```

```
	Exact Wilcoxon-Mann-Whitney Test

data:  How.many.years.of.programming.experience.do.you.have. by Please.state.your.gender (female, male)
Z = -1.89, p-value = 0.05896
alternative hypothesis: true mu is not equal to 0
```

Comparing central tendency

Variables	Levels	Within / Between	Parametric	Non-parametric
1	2	B	Independent-samples t-test	Mann-Whitney U test aka. Wilcoxon rank sum test aka. Wilcoxon-Mann-Whitney-Test
1	3+	B	One-way ANOVA	Kruskal-Wallis test
1	2	W	Paired-samples t-test	Wilcoxon signed-rank test
1	3+	W	One-way repeated measures ANOVA	Friedman test
2+	3+	B	Factorial ANOVA Linear Models (LM)	Aligned Rank Transform (ART) Generalised Linear Models (GLM)
2+	3+	W	Factorial repeated measures ANOVA Linear Mixed Models (LMM)	Aligned Rank Transform (ART) Generalised Linear Mixed Models (GLMM)

Remember last week?

- Which statistical test can be used to test the following hypothesis?

Interface A differs from Interface B concerning usability. *(Usability is measured with a question on a 7-point likert scale from 1=not at all usable to 7=very usable.)*

- ○ N-Sample Chi-Square Test
- ○ N-Sample Fishers Exact Test
- ○ One-sample Chi-Square-Test (Binomial test)
- ○ 1-Sample Chi-Square-Test (Multinomial test)
- ○ Independent samples t-test
- ○ Mann-Whitney U test / Wilcoxon Rank Sum Test / Wilcoxon-Mann-Whitney-Test
- ○ I don't know

https://pingo.coactum.de/426954

Comparing central tendency

Variables	Levels	Within / Between	Parametric	Non-parametric
1	2	B	Independent-samples t-test	Mann-Whitney U test aka. Wilcoxon rank sum test aka. Wilcoxon-Mann-Whitney-Test
1	3+	B	One-way ANOVA	Kruskal-Wallis test
1	2	W	Paired-samples t-test	Wilcoxon signed-rank test
1	3+	W	One-way repeated measures ANOVA	Friedman test
2+	3+	B	Factorial ANOVA Linear Models (LM)	Aligned Rank Transform (ART) Generalised Linear Models (GLM)
2+	3+	W	Factorial repeated measures ANOVA Linear Mixed Models (LMM)	Aligned Rank Transform (ART) Generalised Linear Mixed Models (GLMM)

More Factors

- If the factor has multiple levels (i.e. is not binary), we use ANOVA.
  - one-way ANOVA if there is one factor with more than 2 levels
  - two/three/…-way ANOVA if there are more factors
  - MANOVA if there is more than one factor variable as the dependent variable

- Result: $p < .05$ indicates that not all group means are the same
  - it does not say which is different!
  - ➔ post-hoc analysis

Example ANOVA

- Hypothesis / Alternative Hypothesis / H1

- Null Hypothesis / H0

- Data:

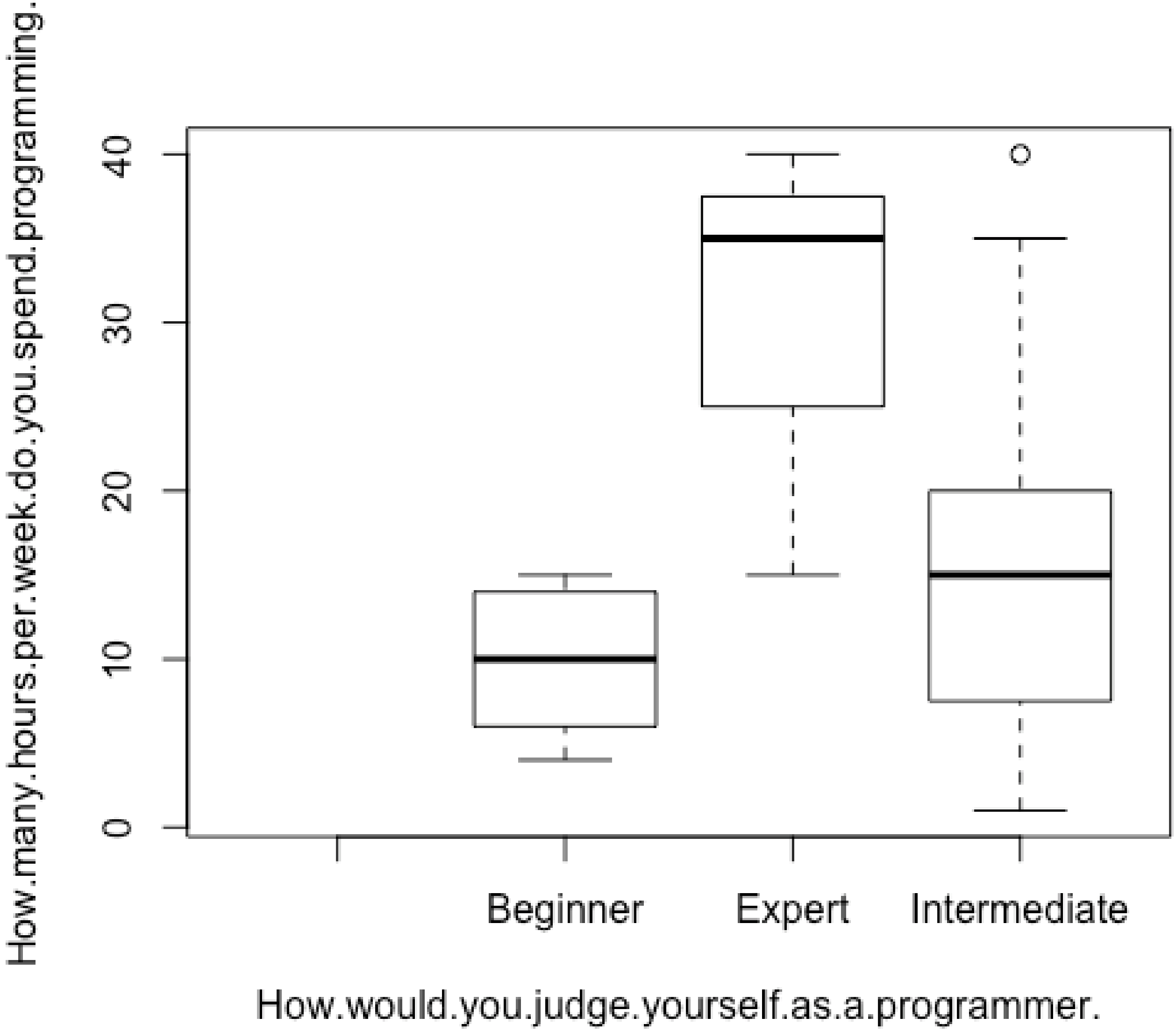
```
> ddply(p, ~ How.would.you.judge.yourself.as.a.programmer., function(data) summary(data$How
.many.hours.per.week.do.you.spend.programming.))
  How.would.you.judge.yourself.as.a.programmer. Min. 1st Qu. Median     Mean 3rd Qu. Max.
1                                      Beginner    4     6.0     10  9.80000    14.0   15
2                                        Expert   15    25.0     35 30.00000    37.5   40
3                                  Intermediate    1     7.5     15 15.52174    20.0   40
```

Example ANOVA

- Hypothesis / Alternative Hypothesis / H1
  - The mean hours per week programming differs between the three levels of experience
- Null Hypothesis / H0
  - No difference in mean between the three levels of experience
- Data:

```
> ddply(p, ~ How.would.you.judge.yourself.as.a.programmer., function(data) summary(data$How
.many.hours.per.week.do.you.spend.programming.))
  How.would.you.judge.yourself.as.a.programmer. Min. 1st Qu. Median     Mean 3rd Qu. Max.
1                                      Beginner    4     6.0     10  9.80000    14.0   15
2                                        Expert   15    25.0     35 30.00000    37.5   40
3                                  Intermediate    1     7.5     15 15.52174    20.0   40
```

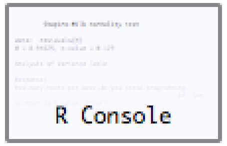
Visualize!

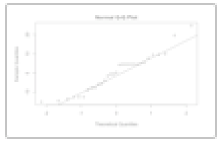

How.many.hours.per.week.do.you.spend.programming.
0
10
20
30
40
Beginner
Expert
Intermediate
How.would.you.judge.yourself.as.a.programmer.

One-way ANOVA

- Omnibus test

````
```{r}
# fit model
m = aov(How.many.hours.per.week.do.you.spend.programming. ~ How.would.you.judge.yourself.as.a.programmer., data=p)
# test residuals
shapiro.test(residuals(m))
# plot residuals
qqnorm(residuals(m)); qqline(residuals(m))
# report anova
anova(m)
```
````





```
	Shapiro-Wilk normality test

data:  residuals(m)
W = 0.94625, p-value = 0.123

Analysis of Variance Table

Response: How.many.hours.per.week.do.you.spend.programming.
                                              Df  Sum Sq Mean Sq F value  Pr(>F)  
How.would.you.judge.yourself.as.a.programmer.  2  785.46  392.73  3.8442 0.03348 *
Residuals                                     28 2860.54  102.16                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Post-hoc testing

- Pairwise comparison – corrected for multiple testing

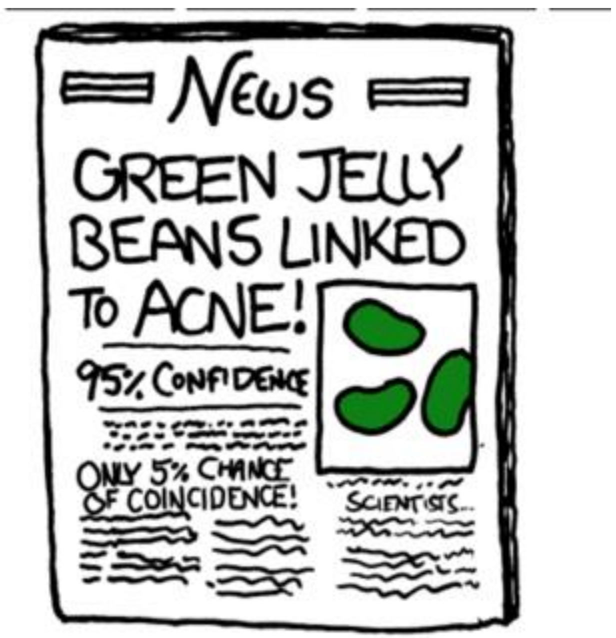
```{r}
library(multcomp)
summary(glht(m, mcp(How.would.you.judge.yourself.as.a.programmer.="Tukey")), test=adjusted(type="holm"))
# Tukey means compare all pairs
```

```
	 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = How.many.hours.per.week.do.you.spend.programming. ~ 
    How.would.you.judge.yourself.as.a.programmer., data = p)

Linear Hypotheses:
                              Estimate Std. Error t value Pr(>|t|)  
Expert - Beginner == 0          20.200      7.381   2.737   0.0320 *
Intermediate - Beginner == 0     5.722      4.987   1.147   0.2610  
Intermediate - Expert == 0     -14.478      6.205  -2.334   0.0541 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- holm method)
```



Comparing central tendency

Variables	Levels	Within / Between	Parametric	Non-parametric
1	2	B	Independent-samples t-test	Mann-Whitney U test aka. Wilcoxon rank sum test aka. Wilcoxon-Mann-Whitney-Test
1	3+	B	One-way ANOVA	Kruskal-Wallis test
1	2	W	Paired-samples t-test	Wilcoxon signed-rank test
1	3+	W	One-way repeated measures ANOVA	Friedman test
2+	3+	B	Factorial ANOVA Linear Models (LM)	Aligned Rank Transform (ART) Generalised Linear Models (GLM)
2+	3+	W	Factorial repeated measures ANOVA Linear Mixed Models (LMM)	Aligned Rank Transform (ART) Generalised Linear Mixed Models (GLMM)

Kruskal-Wallis Test

```
```{r}
library(coin)
kruskal_test(How.many.hours.per.week.do.you.spend.programming. ~ How.would.you.judge.yourself.as.a.programmer., data=p,
distribution="asymptotic") # can't do exact with 3 levels
```
```

```
	Asymptotic Kruskal-Wallis Test

data:  How.many.hours.per.week.do.you.spend.programming. by
	 How.would.you.judge.yourself.as.a.programmer. (Beginner, Expert, Intermediate)
chi-squared = 4.6765, df = 2, p-value = 0.09649
```

- If omnibus test significant
  - follow up with pairwise comparison
  - Mann-Whitney-U test
  - Correct for multiple testing

Comparing central tendency

Variables	Levels	Within / Between	Parametric	Non-parametric
1	2	B	Independent-samples t-test	Mann-Whitney U test aka. Wilcoxon rank sum test aka. Wilcoxon-Mann-Whitney-Test
1	3+	B	One-way ANOVA	Kruskal-Wallis test
1	2	W	Paired-samples t-test	Wilcoxon signed-rank test
1	3+	W	One-way repeated measures ANOVA	Friedman test
2+	3+	B	Factorial ANOVA Linear Models (LM)	Aligned Rank Transform (ART) Generalised Linear Models (GLM)
2+	3+	W	Factorial repeated measures ANOVA Linear Mixed Models (LMM)	Aligned Rank Transform (ART) Generalised Linear Mixed Models (GLMM)

Two Related Samples

- Within-subjects effects / related samples / repeated measures
  - “Does the mean of two or more variables differ?”
  - for example: pre- and post-treatment
  - generally more powerful than between-subjects designs, because we are able to remove the effect of individual differences.
- Assumptions
  - Independence of observations (subject a’s responses are independent of subject b’s responses)
  - Normality of differences

Example

- Programming Self-Assessment in Systemnahe Programmierung
  - Test: in first two weeks
  - Re-test: before exam
- Hypothesis / Alternative Hypothesis / H1
- Null Hypothesis / H0
- Data:

```{r}
summary(sysprog$programming.skill.x)
summary(sysprog$programming.skill.y)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

Example

- Programming Self-Assessment in Systemnahe Programmierung
  - Test: in first two weeks
  - Re-test: before exam
- Hypothesis / Alternative Hypothesis / H1
  - The mean self-assessed programming skill differs between pre- and post-test
- Null Hypothesis / H0
  - No difference in mean between pre- and post-test
- Data:

````
```{r}
summary(sysprog$programming.skill.x)
summary(sysprog$programming.skill.y)
```
````

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.000   4.000   5.000   4.609   6.000   6.000 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  3.000   4.000   5.000   4.957   6.000   6.000 
```

R Paired t-test

```
```{r}
#within-subjects t-test
t.test(x=sysprog$programming.skill.x, y=sysprog$programming.skill.y, paired = TRUE)
```
```

```

	Paired t-test

data:  sysprog$programming.skill.x and sysprog$programming.skill.y
t = -2.0057, df = 22, p-value = 0.05734
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.70747356  0.01182139
sample estimates:
mean of the differences 
             -0.3478261 
```

R Non-Paired t-test

```
```{r}
#between-subjects t-test
t.test(x=sysprog$programming.skill.x, y=sysprog$programming.skill.y, paired = FALSE)
```
```

```
	Welch Two Sample t-test

data:  sysprog$programming.skill.x and sysprog$programming.skill.y
t = -1.1242, df = 40.371, p-value = 0.2676
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9729436  0.2772914
sample estimates:
mean of x mean of y 
 4.608696  4.956522 
```

Comparing central tendency

Variables	Levels	Within / Between	Parametric	Non-parametric
1	2	B	Independent-samples t-test	Mann-Whitney U test aka. Wilcoxon rank sum test aka. Wilcoxon-Mann-Whitney-Test
1	3+	B	One-way ANOVA	Kruskal-Wallis test
1	2	W	Paired-samples t-test	Wilcoxon signed-rank test
1	3+	W	One-way repeated measures ANOVA	Friedman test
2+	3+	B	Factorial ANOVA Linear Models (LM)	Aligned Rank Transform (ART) Generalised Linear Models (GLM)
2+	3+	W	Factorial repeated measures ANOVA Linear Mixed Models (LMM)	Aligned Rank Transform (ART) Generalised Linear Mixed Models (GLMM)

Wilcoxon Signed Rank Test

```{r}
wilcox.test(x=sysprog$programming.skill.x, y=sysprog$programming.skill.y, paired = TRUE,
alternative = "two.sided", exact=FALSE)
```

```
	Wilcoxon signed rank test with continuity correction

data:  sysprog$programming.skill.x and sysprog$programming.skill.y
V = 16.5, p-value = 0.06463
alternative hypothesis: true location shift is not equal to 0
```