
university of
groningen

university of groningen

Basic statistical tests

Martijn Wieling
University of Groningen

This lecture

- Dataset for this lecture
- Comparing one or two groups: t-test
  - Non-parametric alternatives: Mann-Whitney U and Wilcoxon signed rank
- Assessing the dependency between two categorical variables: χ^2 test
- Comparing more than two groups: ANOVA

Some basic points

- This lecture focuses on *how-to-use* and *when-to-use*, rather than on the underlying calculations
    - If you want more information about the tests and concepts illustrated in this lecture, I recommend the books from Levshina, Winter or (free) Navarro
- Make sure to report **effect size** as significance is dependent on sample size

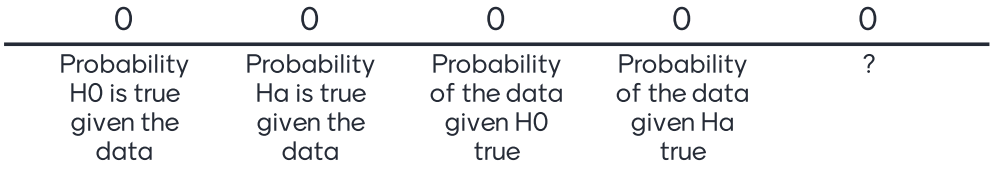
DIFFERENCE (IN s)	n	p
0.01	40,000	0.05
0.10	400	0.05
0.25	64	0.05
0.54	16	0.05

Question 1

Go to **www.menti.com/8a981a**

What is a p-value?

Mentimeter



Press ENTER to show correct

Dataset for this lecture

```
load("dat.rda")
head(dat)
```

```
#    Speaker Language  PronDist PronDistCat LangDist LangDistAlt Age Sex AEO LR NrLang
# 1  arabic1   arabic  0.185727   Different  0.63699     0.44864  38   F  12  4      0
# 2 arabic10   arabic -0.172175     Similar  0.63699     0.44864  26   M   5  2      2
# 3 arabic13   arabic -0.035423     Similar  0.63699     0.44864  25   M  15  1      2
# 4 arabic12   arabic  0.372547   Different  0.63699     0.44864  32   M  11  8      0
# 5 arabic17   arabic -0.175237     Similar  0.63699     0.44864  35   M  15  0      1
# 6 arabic18   arabic  0.168120   Different  0.63699     0.44864  18   M   6  0      1
```

Dataset structure

```
str(dat)
```

```
# 'data.frame': 712 obs. of  11 variables:
#  $ Speaker    : Factor w/ 712 levels "afrikaans1","afrikaans2",..: 21 22 25 24 27 28 26 30 31 23 ...
#  $ Language   : Factor w/ 159 levels "afrikaans","agni",..: 7 7 7 7 7 7 7 7 7 7 ...
#  $ PronDist   : num  0.1857 -0.1722 -0.0354 0.3725 -0.1752 ...
#  $ PronDistCat: Factor w/ 2 levels "Different","Similar": 1 2 2 1 2 1 1 2 2 2 ...
#  $ LangDist   : num  0.637 0.637 0.637 0.637 0.637 ...
#  $ LangDistAlt: num  0.449 0.449 0.449 0.449 0.449 ...
#  $ Age        : num  38 26 25 32 35 18 22 36 23 30 ...
#  $ Sex        : Factor w/ 2 levels "F","M": 1 2 2 2 2 2 2 2 1 1 ...
#  $ AEO        : num  12 5 15 11 15 6 16 12 10 14 ...
#  $ LR         : num  4 2 1 8 0 0 0 1 0 4 ...
#  $ NrLang     : int  0 2 2 0 1 1 2 2 2 1 ...
```

Comparing one or two groups: t-test

- Values between two groups (or vs. value) can be compared using the t-test
- Assumptions:
  - Randomly selected sample(s)
  - Independent observations (except for paired data)
  - Data has interval scale (difference between two values is meaningful) or ratio scale (meaningful difference and true 0)
    - E.g., interval scale: temperature in C; ratio scale: length in cm.
  - Data in sample(s) normally distributed (for $N \leq 30$)
  - Variances in samples homogeneous (Welch's adjustment, default in R, corrects for this)
  - Note: *Likert scale* is ordinal data, so t-test in principle not adequate
    - But in practice not problematic (De Winter & Dodou, 2011)
- **Visualize** the data if possible (facilitates interpretation)

Question 2

Go to **www.menti.com/8a981a**

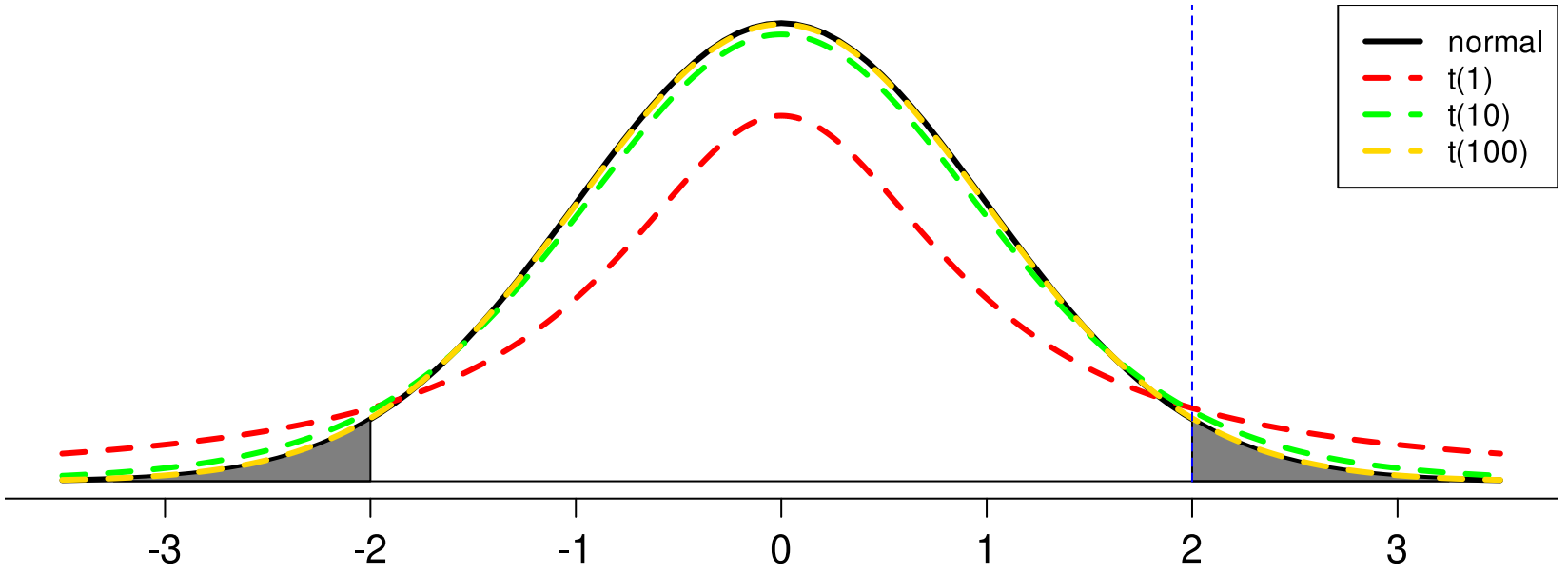
What is a good way to visualize the values of two groups?

Mentimeter

Press ENTER to show correct

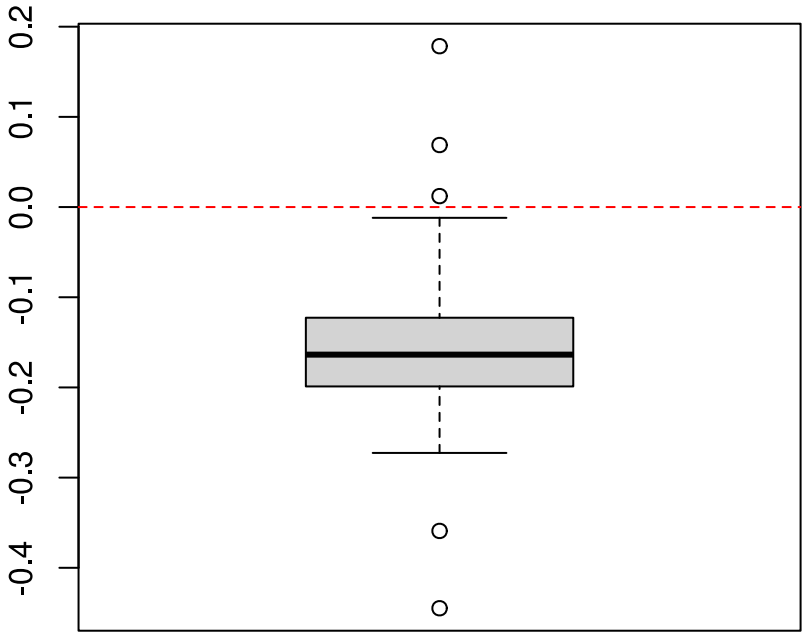
t-test

- Result of t-test is a t-value, which is compared to the appropriate t-distribution
- t-distribution depends on degrees of freedom (therefore: report dF!)



Group mean vs. value: visualization

```
german <- droplevels(dat[dat$Language == "german", ])
boxplot(german$PronDist)
abline(h = 0, col = "red", lty = 2)
```



Group mean vs. value: one sample t-test

```
t.test(german$PronDist, mu = 0)
```

```
#
#   One Sample t-test
#
# data:  german$PronDist
# t = -5.33, df = 21, p-value = 2.7e-05
# alternative hypothesis: true mean is not equal to 0
# 95 percent confidence interval:
#  -0.208787 -0.091657
# sample estimates:
# mean of x
#  -0.15022
```

One sample t-test: effect size

```
library(lsr)
cohensD(german$PronDist, mu = 0)
```

```
# [1] 1.1373
```

- Cohen's d measures the difference in terms of the number of standard deviations
  - Rough guideline: Cohen's d < 0.3: small effect size; 0.3 - 0.8: medium; > 0.8: large

Try it yourself!

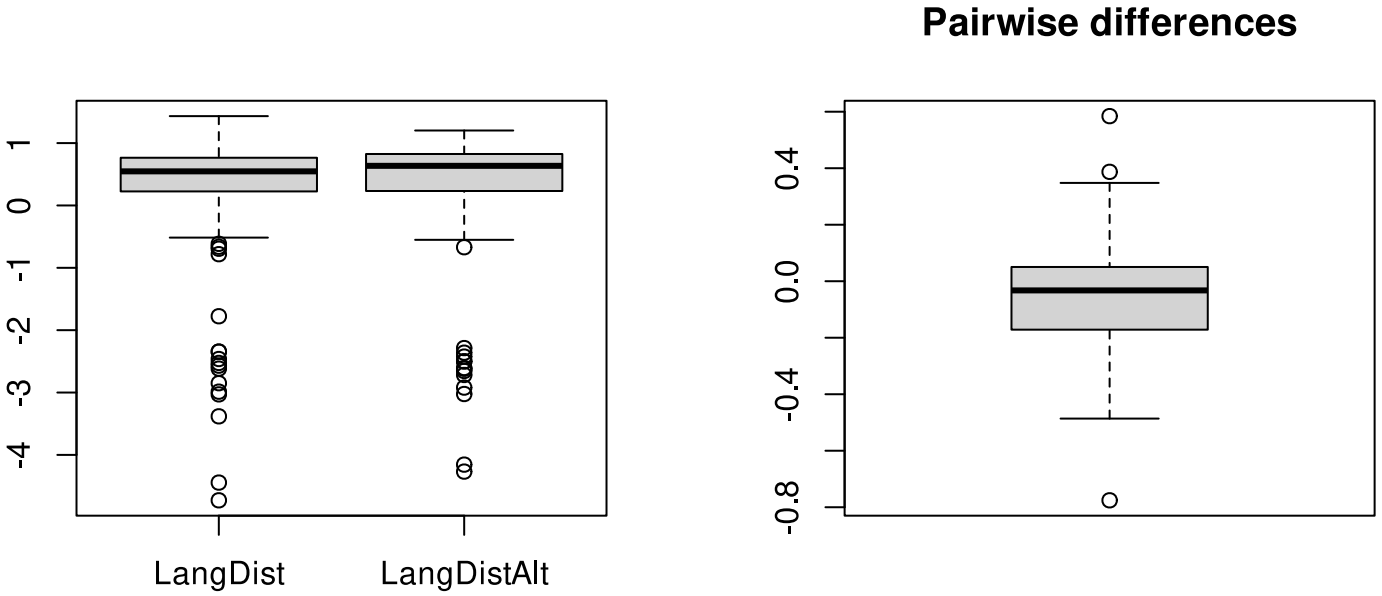
- Install the *Mathematical Biostatistics Boot Camp* swirl course:

```
library(swirl)
install_from_swirl("Mathematical_Biostatistics_Boot_Camp")
```

- Run `swirl()` in RStudio and finish the following lesson of the *Mathematical Biostatistics Boot Camp* course:
    - *Lesson 1*: One Sample t-test

Comparing paired data: visualization

```
# aggregate data per language (159 languages)
lang <- aggregate(cbind(LangDist, LangDistAlt) ~ Language, data = dat, FUN = mean)
par(mfrow = c(1, 2))
boxplot(lang[, c("LangDist", "LangDistAlt")])
boxplot(lang$LangDist - lang$LangDistAlt, main = "Pairwise differences")
```



Paired samples t-test

```
t.test(lang$LangDist, lang$LangDistAlt, paired = T)
```

```
#
#   Paired t-test
#
# data:  lang$LangDist and lang$LangDistAlt
# t = -3.73, df = 158, p-value = 0.00027
# alternative hypothesis: true mean difference is not equal to 0
# 95 percent confidence interval:
#  -0.085703 -0.026367
# sample estimates:
# mean difference
#       -0.056035
```

Question 3

Go to **www.menti.com/8a981a**

Which statement is true for the paired t-test applied to a small dataset (N=10)?

Mentimeter

0	0	0
both series of values need to be norm. distributed	the differences need to be normally distributed	?

Press ENTER to show correct

Paired samples t-test = one sample t-test

```
t.test(lang$LangDist - lang$LangDistAlt, mu = 0)  # identical to one-sample test of differences
```

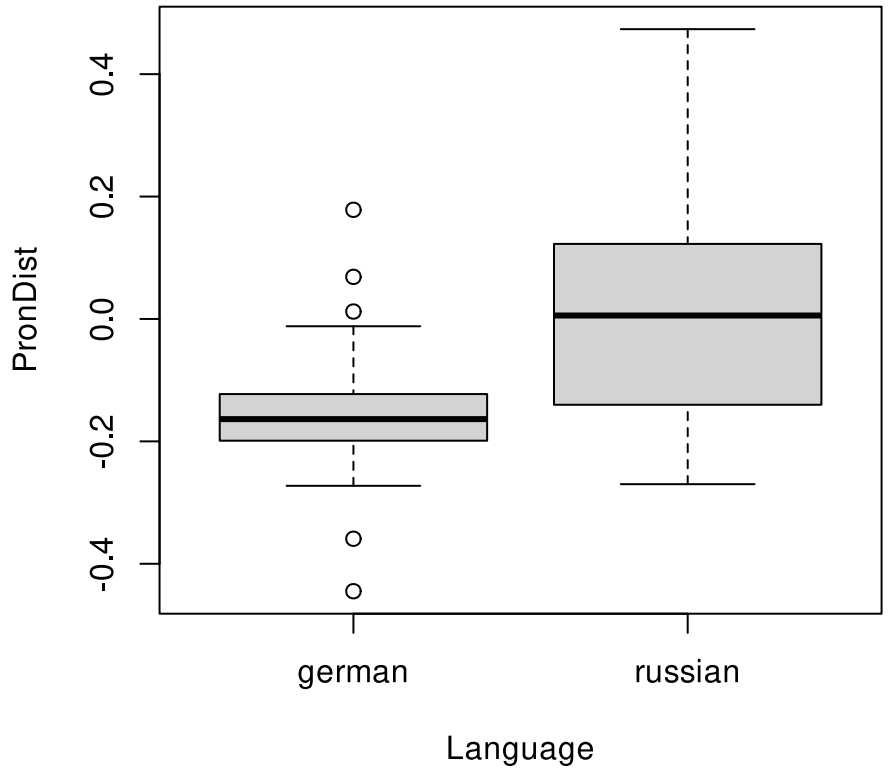
```
#
#   One Sample t-test
#
# data:  lang$LangDist - lang$LangDistAlt
# t = -3.73, df = 158, p-value = 0.00027
# alternative hypothesis: true mean is not equal to 0
# 95 percent confidence interval:
#  -0.085703 -0.026367
# sample estimates:
# mean of x
# -0.056035
```

```
cohensD(lang$LangDist, lang$LangDistAlt, method = "paired")  # effect size
```

```
# [1] 0.29585
```

Comparing two groups: visualization

```
rusger <- droplevels(dat[dat$Language %in% c("russian", "german"), ])
boxplot(PronDist ~ Language, data = rusger)
```



Comparing two groups: independent samples t-test

```
t.test(PronDist ~ Language, data = rusger, alternative = "two.sided")
```

```
#
#   Welch Two Sample t-test
#
# data:  PronDist by Language
# t = -3.56, df = 42.5, p-value = 0.00092
# alternative hypothesis: true difference in means between group german and group russian is not equal to 0
# 95 percent confidence interval:
#  -0.267719 -0.074108
# sample estimates:
#  mean in group german mean in group russian
#             -0.150222              0.020691
```

```
cohensD(PronDist ~ Language, data = rusger)
```

```
# [1] 1.0166
```

Reporting results of a t-test

- Pronunciation difference from native English was smaller for the German speakers (mean: -0.15, sd: 0.132) than for the Russian speakers (mean: 0.02, sd: 0.194). The difference was -0.17 (Cohen's d: 1.02, large effect) and reached significance using an independent samples Welch's unequal variances t-test at an α-level of 0.05, $t(42.5) = -3.56, p < 0.001$.

Assumptions met?

- ✓ Randomly selected sample(s)
- ✓ Independent observations (except for pairs)
- ✓ Data has interval or ratio scale
- ? Variance in samples homogeneous (corrected with Welch's adjustment)
- ? Data in compared samples are **normally distributed** (for $N \leq 30$)

Testing if variances are equal (homoscedasticity)

- Testing homoscedasticity using Levene's test

```
library(car)
leveneTest(PronDist ~ Language, data = rusger)
```

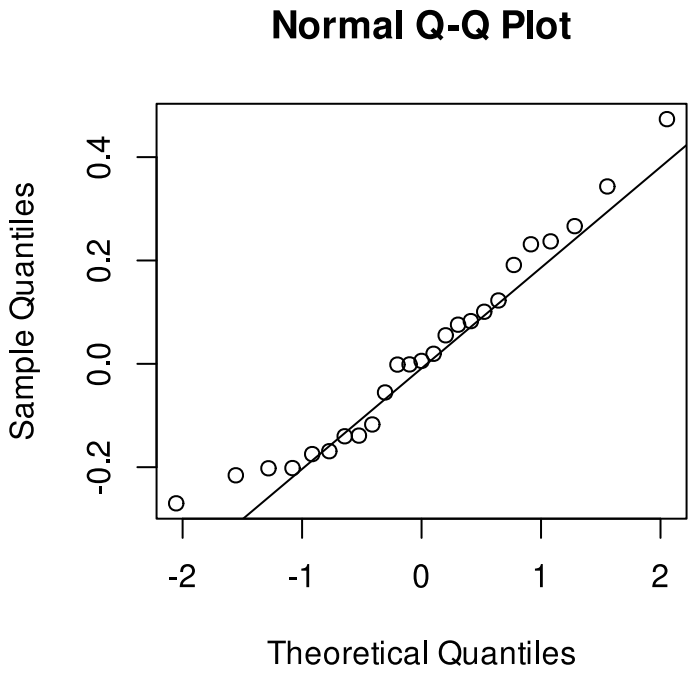
```
# Levene's Test for Homogeneity of Variance (center = median)
#       Df F value Pr(>F)
# group  1       5   0.03 *
#       45
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Levene's test shows that the variances are different and the default Welch's adjustment is warranted
  - But note that the Welch's t-test can always be used as it is more robust and power is comparable to that of the normal t-test

Assessing normality: Russian data (1)

- For investigating normality, a normal quantile plot can be used

```
russian <- droplevels(dat[dat$Language == "russian", ])
qqnorm(russian$PronDist)  # plot actual values vs. theoretical quantiles
qqline(russian$PronDist)  # plot reference line of normal distribution
```



Assessing normality: Russian data (2)

- Alternatively, one can use the Shapiro-Wilk test of normality

```
shapiro.test(russian$PronDist)
```

```
#
#   Shapiro-Wilk normality test
#
# data:  russian$PronDist
# W = 0.958, p-value = 0.38
```

Question 4

Go to **www.menti.com/8a981a**

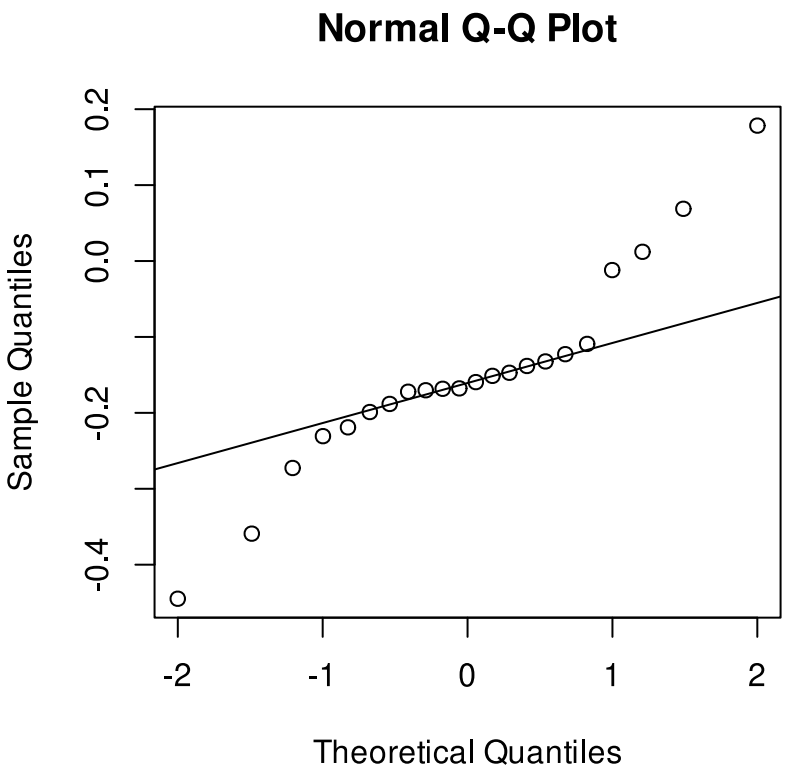
Which approach is better to assess normality?

Mentimeter

Press ENTER to show correct

Assessing normality: German data (1)

```
qqnorm(german$PronDist)
qqline(german$PronDist)
```



Assessing normality: German data (2)

```
shapiro.test(german$PronDist)
```

```
#
#   Shapiro-Wilk normality test
#
# data:  german$PronDist
# W = 0.929, p-value = 0.12
```

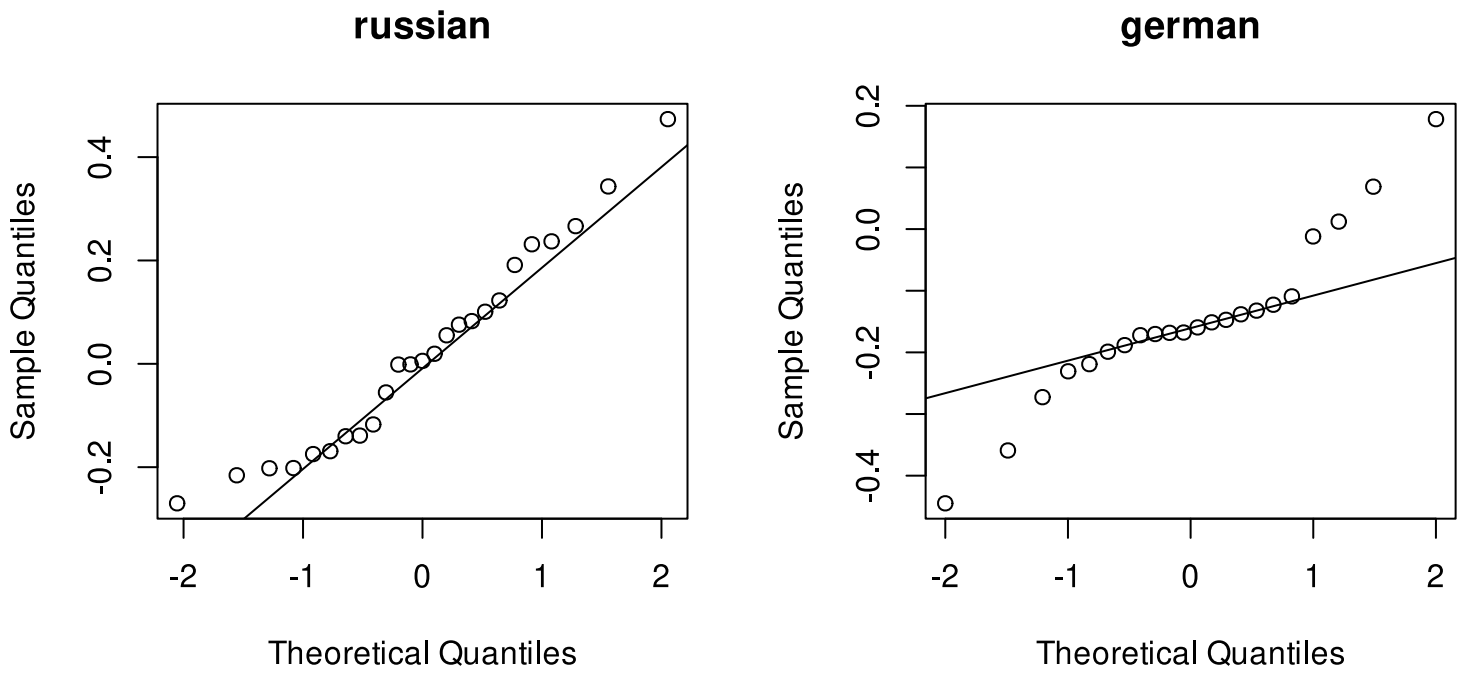
- Sensitivity to sample size of the Shapiro-Wilk test is clear: I would judge the data as *non-normal* on the basis of the normal quantile plot
- Given the small size of the sample (N = 22$), a non-parametric alternative is needed

Non-parametric alternatives

- Non-parametric fallbacks
  - One sample t-test and paired t-test: Wilcoxon signed rank test
  - Independent samples t-test: Mann-Whitney U test (= Wilcoxon rank sum test)
  - In both cases: `wilcox.test` (similar to `t.test`)

Comparing two groups: Mann-Whitney U test (1)

```
par(mfrow = c(1, 2))  # visualization indicates non-parametric test necessary
qqnorm(russian$PronDist, main = "russian")
qqline(russian$PronDist)
qqnorm(german$PronDist, main = "german")
qqline(german$PronDist)
```



Comparing two groups: Mann-Whitney U test (2)

```
(model <- wilcox.test(PronDist ~ Language, data = rusger))  # default 2-tailed
```

```
#
#   Wilcoxon rank sum exact test
#
# data:  PronDist by Language
# W = 140, p-value = 0.0035
# alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.effsize <- function(pval2tailed, N) {
    (r <- abs(qnorm(pval2tailed/2)/sqrt(N)))  # r = z / sqrt(N)
}

# rough guideline: r around 0.1 (small), > 0.3 (medium), > 0.5: large
wilcox.effsize(model$p.value, nrow(rusger))
```

```
# [1] 0.42616
```

Reporting results of Mann-Whitney U (or Wilcoxon)

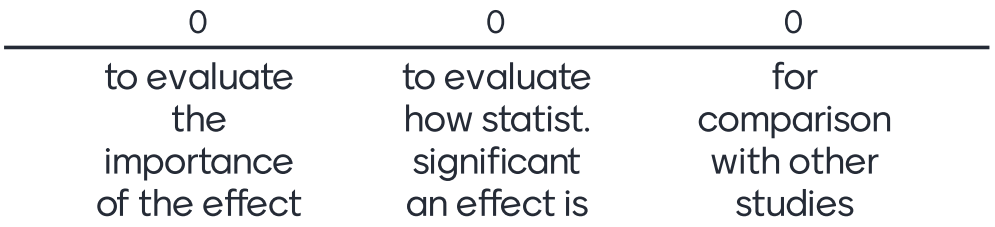
- Pronunciation difference from native English was smaller for the German speakers (median value: -0.16) than for the Russian speakers (median value: 0.006). The effect size r of the difference was 0.43 (medium) and reached significance using a Mann-Whitney U test ($U = 140$, with $n_g = 22$ and $n_r = 25$) at an α-level of 0.05 ($p = 0.003$).

Question 5

Go to **www.menti.com/8a981a**

Why report effect size?

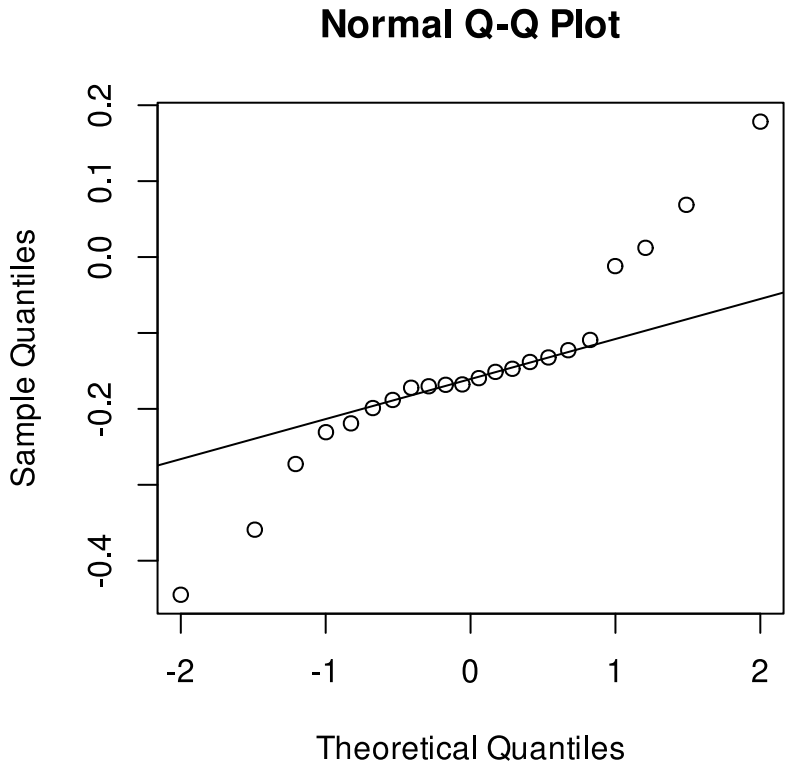
Mentimeter



Press ENTER to show correct

Group mean vs. value: Wilcoxon signed rank (1)

```
# visualization indicates non-parametric necessary
qqnorm(german$PronDist)
qqline(german$PronDist)
```



Group mean vs. value: Wilcoxon signed rank (2)

```
(model <- wilcox.test(german$PronDist, mu = 0))
```

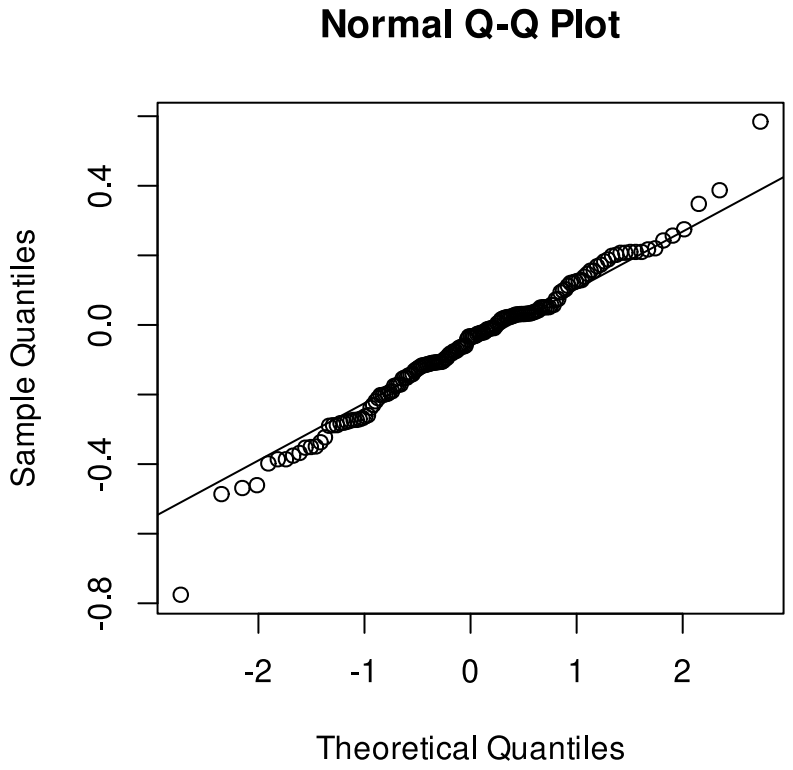
```
#
#   Wilcoxon signed rank exact test
#
# data:  german$PronDist
# V = 20, p-value = 0.00018
# alternative hypothesis: true location is not equal to 0
```

```
wilcox.effsize(model$p.value, nrow(german))
```

```
# [1] 0.79948
```

Comparing paired data: Wilcoxon signed rank

```
# No non-parametric test necessary
qqnorm(lang$LangDist - lang$LangDistAlt)
qqline(lang$LangDist - lang$LangDistAlt)
```



Comparing paired data: Wilcoxon signed rank

- Using a Wilcoxon signed rank test is not necessary, given the size of the dataset (159 languages) and the normal distribution, but it is included for completeness

```
(model <- wilcox.test(lang$LangDist, lang$LangDistAlt, paired = TRUE))
```

```
#
#   Wilcoxon signed rank test with continuity correction
#
# data:  lang$LangDist and lang$LangDistAlt
# V = 4362, p-value = 0.00059
# alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.effsize(model$p.value, nrow(lang))
```

```
# [1] 0.27242
```

Dependency between two cat. variables: χ^2 test

- Requirements:
  - Sample randomly selected from the population of interest
  - Independent observations
  - Every observation can be classified into exactly one category
  - Expected frequency for each combination at least 5 (or: Fisher's exact test)
- Intuition: compare expected frequencies with observed frequencies
  - Larger differences between expected and observed: more likely two categorical variables dependent

χ^2 test

```
languages <- c("farsi", "swedish", "polish")
dat3 <- droplevels(dat[dat$Language %in% languages, ])
(tab <- table(dat3$PronDistCat, dat3$Language))
```

```
#
#             farsi polish swedish
#   Different     6      6       1
#   Similar       4      5       9
```

```
chisq.test(tab)
```

```
#
#   Pearson's Chi-squared test
#
# data:  tab
# X-squared = 6.25, df = 2, p-value = 0.044
```

χ^2 test: effect size

```
cramersV(tab)  # from library(lsr)
```

```
# [1] 0.4489
```

- Rough guidelines for effect size:
  - Small effect: $w = 0.1$
  - Medium effect: $w = 0.3$
  - Large effect: $w = 0.5$
  - With $w = V \times \sqrt{min(R, C) - 1}$
    - With more rows ® and columns ©, a lower Cramer's V can still be the same size of effect

χ^2 test: reporting results

- Fisher's exact test of independence was performed to examine the relation between Language and Pronunciation Distance Category. The relation between the two variables was significant in a sample size of 31 at an α-level of 0.05, $\chi^2(2) = 6.25, p = 0.04$. The effect size was medium, with Cramer's V: 0.45.

However, χ^2 test not appropriate: Fisher's exact test

```
chisq.test(tab)$expected  # warning as not all expected values >= 5
```

```
# Warning in chisq.test(tab): Chi-squared approximation may be incorrect
```

```
#
#              farsi polish swedish
#   Different 4.1935 4.6129  4.1935
#   Similar   5.8065 6.3871  5.8065
```

```
fisher.test(tab)  # solution: use Fisher's exact test as the appropriate alternative
```

```
#
#   Fisher's Exact Test for Count Data
#
# data:  tab
# p-value = 0.053
# alternative hypothesis: two.sided
```

ANOVA for differences between 3 or more groups

- Intuition of ANOVA: compare between-group variation and within-group variation
  - If between-group variation (SS_b: sum of squares) is large relative to within-group variation (SS_w) the difference is more likely to be significant
  - See this freely downloadable, well-written statistics book

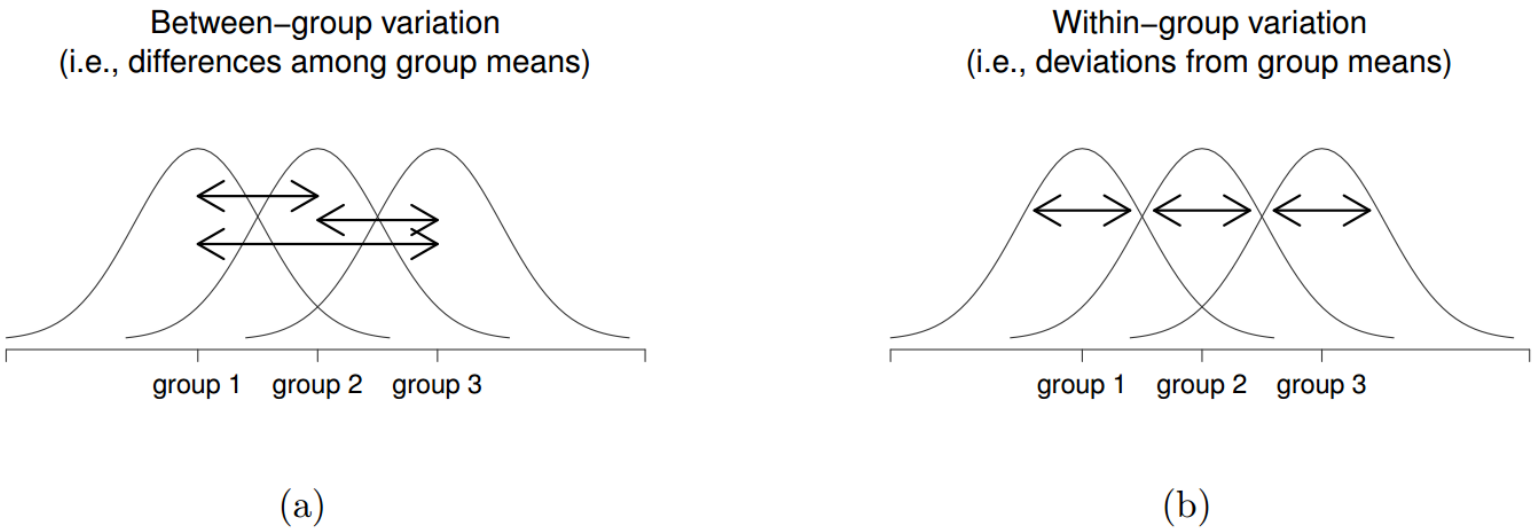


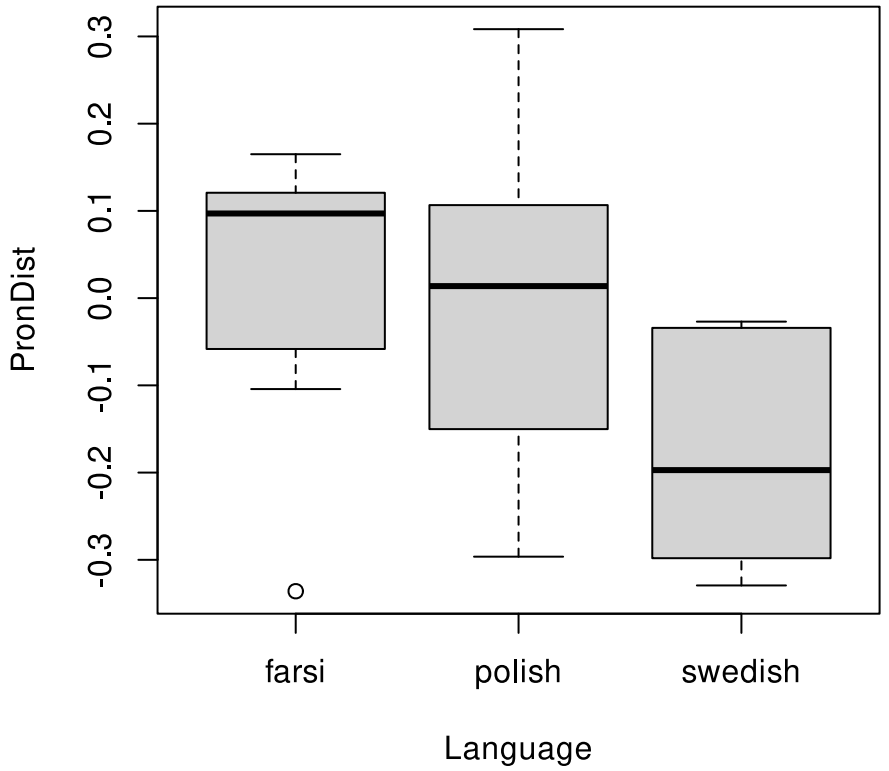
Figure 14.2: Graphical illustration of "between groups" variation (panel a) and "within groups" variation (panel b). On the left, the arrows show the differences in the group means; on the right, the arrows highlight the variability within each group.

Assumptions for ANOVA

- Randomly selected sample(s)
- Independent observations in the groups
- Data has interval scale or ratio scale
- Data in each sample is normally distributed and/or equal sample sizes
- Variance in samples homogeneous

Differences between 3+ groups: one-way ANOVA (1)

```
# start with visualization
boxplot(PronDist ~ Language, data = dat3)
```



Differences between 3+ groups: one-way ANOVA (2)

```
result <- aov(PronDist ~ Language, data = dat3)
# alternative if variances are not equal: oneway.test(), alternative if
# non-normal distribution: kruskal.test()

summary(result)  # is the ANOVA significant?
```

```
#             Df Sum Sq Mean Sq F value Pr(>F)
# Language     2  0.213  0.1067    4.16  0.026 *
# Residuals   28  0.718  0.0256
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
etaSquared(result)  # from library(lsr); small: 0.02, medium: 0.13, large: 0.26
```

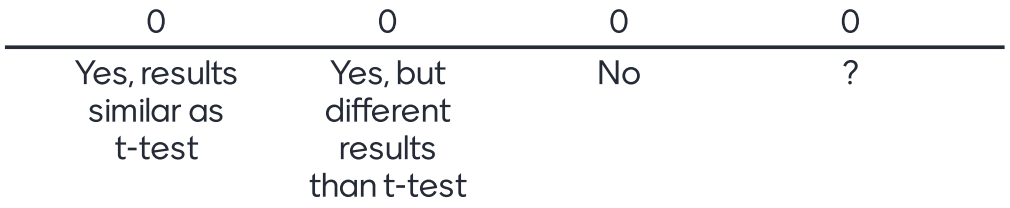
```
#           eta.sq eta.sq.part
# Language 0.22908     0.22908
```

Question 6

Go to **www.menti.com/8a981a**

Can an ANOVA be used to compare 2 groups?

Mentimeter



Press ENTER to show correct

ANOVA: reporting results

- At an α-level of 0.05, a one-way ANOVA showed a significant effect of Language on Pronunciation Difference from English: $F(2, 28) = 4.16$ ($p = 0.03$). The effect size of Language, partial eta squared η_p^2, was equal to 0.23 (medium).
  - As the F-distribution depends on two values (dF1 and dF2), both values need to be reported
    - dF1: number of levels of the categorical variable - 1
    - dF2: number of observations - number of levels of the categorical variable

ANOVA post-hoc test

```
posthocPairwiseT(result)  # from library(lsr)
```

```
#
#   Pairwise comparisons using t tests with pooled SD
#
# data:  PronDist and Language
#
#         farsi polish
# polish  0.63  -
# swedish 0.04  0.06
#
# P value adjustment method: holm
```

```
# alternative: TukeyHSD(result)
```

ANOVA post-hoc: reporting results

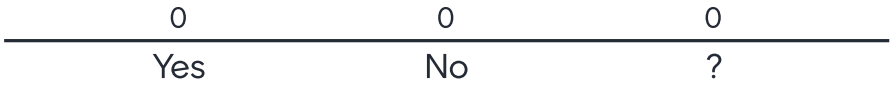
- Post-hoc comparisons were conducted using pairwise t-tests using the Holm method to correct for multiple comparisons. The post-hoc comparison (using an α-level of 0.05) revealed that Swedish had a lower Pronunciation Difference from English (mean: -0.172, sd: 0.126) than Farsi (mean: 0.021, sd: 0.156, $p = 0.04$), but not Polish (mean: -0.013, sd: 0.188, $p = 0.06$). Furthermore, Farsi and Polish did not differ significantly ($p = 0.63$).

Question 7

Go to **www.menti.com/8a981a**

If an ANOVA test is significant, does at least one pair differ significantly?

Mentimeter



Press ENTER to show correct

Testing assumptions: variances equal?

- Testing homoscedasticity using Levene's test

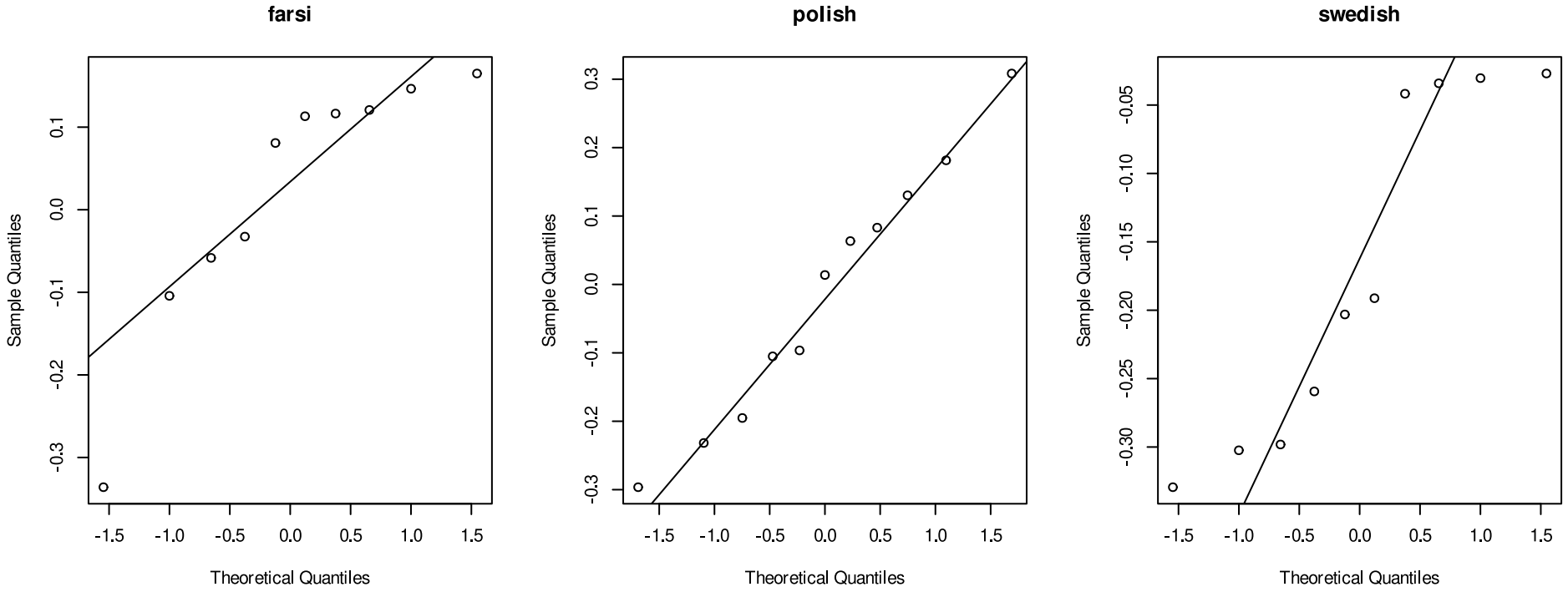
```
leveneTest(PronDist ~ Language, data = dat3)
```

```
# Levene's Test for Homogeneity of Variance (center = median)
#       Df F value Pr(>F)
# group  2    0.69   0.51
#       28
```

- Levene's test shows the variances are similar

Assessing normality (1)

```
par(mfrow = c(1, 3))
for (lang in levels(dat3$Language)) {
    qqnorm(dat3[dat3$Language == lang, ]$PronDist, main = lang)
    qqline(dat3[dat3$Language == lang, ]$PronDist)
}
```



Assessing normality (2)

```
aggregate(PronDist ~ Language, data = dat3, function(x) shapiro.test(x)$p.value)
```

```
#   Language PronDist
# 1    farsi 0.035895
# 2   polish 0.922943
# 3  swedish 0.040296
```

```
table(dat3$Language)  # unequal sample sizes
```

```
#
#   farsi  polish swedish
#      10      11      10
```

- Non-normal and unequal sample sizes, so Kruskal-Wallis test should be used instead

Kruskal-Wallis rank sum test

```
kruskal.test(PronDist ~ Language, data = dat3)
```

```
#
#   Kruskal-Wallis rank sum test
#
# data:  PronDist by Language
# Kruskal-Wallis chi-squared = 6.44, df = 2, p-value = 0.04
```

Kruskal-Wallis rank sum test: post-hoc tests

```
library(PMCMR)
posthoc.kruskal.dunn.test(PronDist ~ Language, data = dat3)
```

```
#
#   Pairwise comparisons using Dunn's-test for multiple
#                           comparisons of independent samples
#
# data:  PronDist by Language
#
#         farsi polish
# polish  0.634 -
# swedish 0.051 0.099
#
# P value adjustment method: holm
```

- Note that even though the omnibus test shows there to be a significant effect of Language on Pronunciation Difference from English, none of the levels appear to differ significantly (i.e. they represent different tests)

Kruskal-Wallis: effect size

- Effect size for each pair can be obtained using Mann-Whitney U procedure
- For example:

```
pairs2 <- dat3[dat3$Language %in% c("swedish", "farsi"), ]
model <- wilcox.test(PronDist ~ Language, data = pairs2)
wilcox.effsize(model$p.value, nrow(pairs2))
```

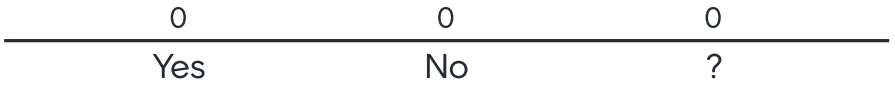
```
# [1] 0.5075
```

Question 8

Go to **www.menti.com/8a981a**

When doing a two-way ANOVA (instead of one-way), should the data be balanced?

Mentimeter



Press ENTER to show correct

Multi-way anova: first some remarks

- Multiple types if data is **unbalanced** (balanced data: all types equal)
  - Type I (used in aov): SS(A), SS(B | A), SS(A*B | B, A)
    - This approach is order-dependent and rarely tests a hypothesis of interest, as the effects (except for the final interaction) are obtained without controlling for the other effects in the model
  - Type II: SS(A | B), SS(B | A)
    - This approach is valid if no interaction is necessary
  - Type III: SS(A | B, A*B), SS(B | A, A*B)
    - (This is the default SPSS approach)
    - Note: main effects are rarely interpretable when the interaction is significant
    - If interactions are not significant, Type II is more powerful
    - Contrasts need to be orthogonal (default contrasts in R are not)

Present data not balanced

```
dat2 <- droplevels(dat[dat$Language %in% c("mandarin", "dutch"), ])  # new dataset
table(dat2$Language, dat2$Sex)
```

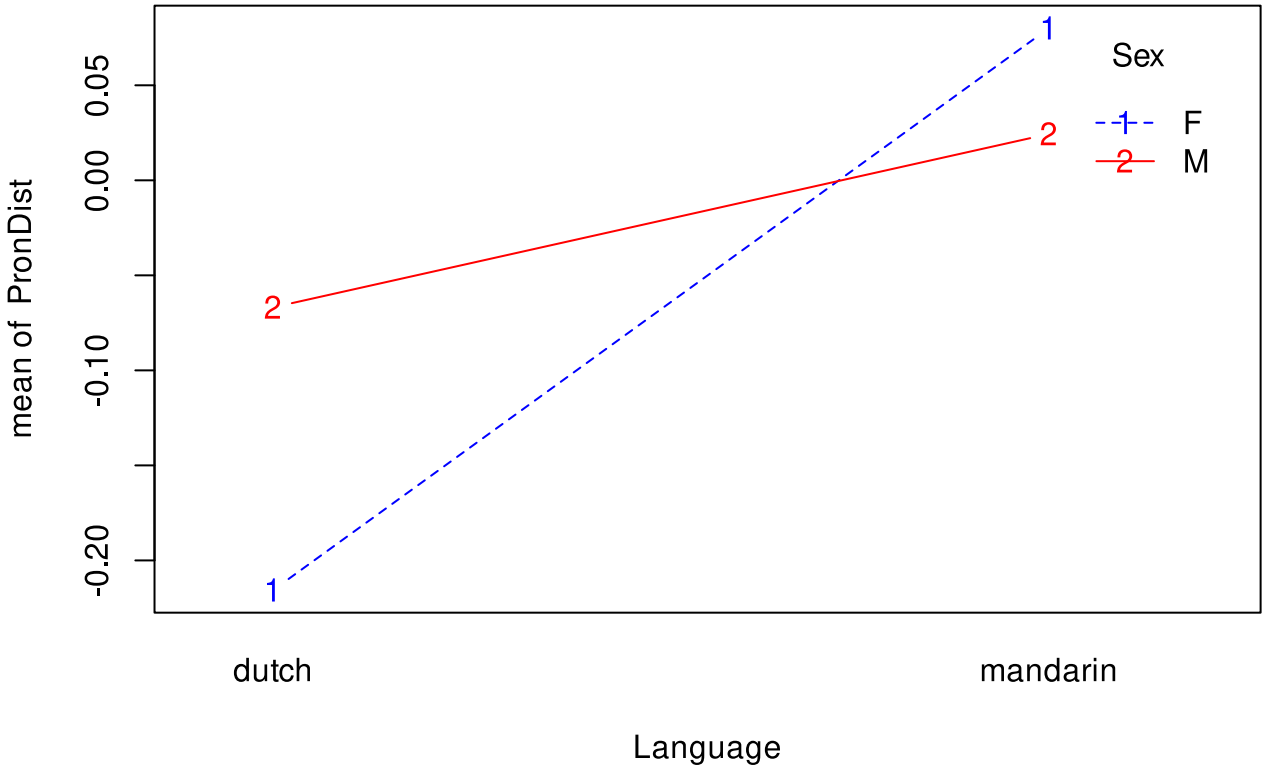
```
#
#             F  M
#   dutch     7  7
#   mandarin 14  9
```

```
# normality OK
aggregate(PronDist ~ Language, data = dat2, function(x) shapiro.test(x)$p.value)
```

```
#   Language PronDist
# 1    dutch  0.39841
# 2 mandarin  0.34953
```

Interaction plot

```
with(dat2, interaction.plot(Language, Sex, PronDist, col = c("blue", "red"), type = "b"))
```



Multi-way anova: Type I

```
summary(aov(PronDist ~ Language * Sex, data = dat2))
```

```
#              Df Sum Sq Mean Sq F value  Pr(>F)
# Language      1  0.347   0.347   20.06 8.5e-05 ***
# Sex           1  0.005   0.005    0.30    0.59
# Language:Sex  1  0.089   0.089    5.17    0.03 *
# Residuals    33  0.570   0.017
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(result <- aov(PronDist ~ Sex * Language, data = dat2))
```

```
#              Df Sum Sq Mean Sq F value  Pr(>F)
# Sex           1  0.000   0.000    0.00    0.95
# Language      1  0.352   0.352   20.36 7.7e-05 ***
# Sex:Language  1  0.089   0.089    5.17    0.03 *
# Residuals    33  0.570   0.017
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multi-way anova: Type II

```
Anova(result <- aov(PronDist ~ Language * Sex, data = dat2), type = 2)  # from library(car), case sensitive!
```

```
# Anova Table (Type II tests)
#
# Response: PronDist
#              Sum Sq Df F value  Pr(>F)
# Language      0.352  1   20.36 7.7e-05 ***
# Sex           0.005  1    0.30    0.59
# Language:Sex  0.089  1    5.17    0.03 *
# Residuals     0.570 33
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
etaSquared(result, type = 2)
```

```
#                eta.sq eta.sq.part
# Language     0.3478048   0.3815433
# Sex          0.0051338   0.0090241
# Language:Sex 0.0883463   0.1354766
```

Multi-way anova: Type III (appropriate)

```
op <- options(contrasts = c("contr.sum", "contr.poly"))  # set orthogonal contrasts for unordered and ordered factors
Anova(result <- aov(PronDist ~ Language * Sex, data = dat2), type = 3)
```

```
# Anova Table (Type III tests)
#
# Response: PronDist
#              Sum Sq Df F value  Pr(>F)
# (Intercept)   0.068  1    3.92 0.05611 .
# Language      0.320  1   18.52 0.00014 ***
# Sex           0.019  1    1.07 0.30784
# Language:Sex  0.089  1    5.17 0.02959 *
# Residuals     0.570 33
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
etaSquared(result, type = 3)
```

```
#                eta.sq eta.sq.part
# Language     0.316338    0.359431
# Sex          0.018328    0.031486
# Language:Sex 0.088346    0.135477
```

Multi-way anova: interpretation

```
model.tables(result, type = "means")
```

```
# Tables of means
# Grand mean
#
# -0.017218
#
#  Language
#       dutch mandarin
#     -0.1413  0.05828
# rep 14.0000 23.00000
#
#  Sex
#           F         M
#     -0.0275 -0.003726
# rep 21.0000 16.000000
#
#  Language:Sex
#           Sex
# Language   F      M
#   dutch    -0.216 -0.067
#   rep       7.000  7.000
#   mandarin  0.080  0.024
#   rep      14.000  9.000
```

Multi-way anova: post-hoc tests

```
dat2$LangSex <- interaction(dat2$Language, dat2$Sex)
newresult <- aov(PronDist ~ LangSex, data = dat2)
posthocPairwiseT(newresult)  # from library(lsr)
```

```
#
#   Pairwise comparisons using t tests with pooled SD
#
# data:  PronDist and LangSex
#
#            dutch.F mandarin.F dutch.M
# mandarin.F 2e-04   -          -
# dutch.M    0.125   0.086      -
# mandarin.M 0.005   0.355      0.355
#
# P value adjustment method: holm
```

Multi-way ANOVA: reporting results

- Using an α-level of 0.05, a two-way ANOVA was conducted on the influence of two independent variables (language: Dutch and Mandarin, and sex: male and female) on the pronunciation differerence from English. The main effect of language was significant, $F(1, 33) = 18.52$ ($p < 0.001$), with a higher pronunciation difference from English for Mandarin speakers (mean: 0.058, sd: 0.147) than for Dutch speakers (mean: -0.141, sd: 0.12). The main effect for sex was not significant ($F(1, 33) = 1.07$, $p = 0.31$). However the interaction effect was significant ($F(1, 33) = 5.17$, $p = 0.03$) and indicated that while the female Dutch speakers had lower pronunciation differences compared to English than males, the effect was inverse for the Mandarin speakers. The effect size of language, η_p^2, was equal to 0.36 (large). The effect size of the interaction between sex and language, η_p^2, was equal to 0.14 (medium).

Variants of ANOVA

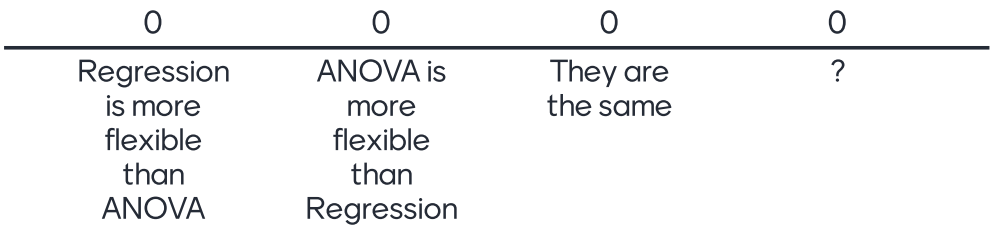
- There are several variants of ANOVA, e.g.:
  - ANCOVA: covariates can be added as control variables in the analysis
  - MANOVA: assessing the relationship between one or more predictors and **multiple** dependent variables
  - Repeated-measures ANOVA
- These are not covered further, as (mixed-effects) regression is more flexible

Question 9

Go to **www.menti.com/8a981a**

How do ANOVA and regression relate?

Mentimeter



Press ENTER to show correct

Recap

- In this lecture, we've covered:
  - The t-test (and non-parametric alternatives) for comparing means of 2 groups
  - The χ^2 test to assess the relationship between 2 categorical variables
  - ANOVA for comparing 3+ groups (and interactions between factorial predictors)
- Associated lab session:
  - https://www.let.rug.nl/wieling/Statistics/Basic-Tests/lab

Evaluation

Go to **www.menti.com/8a981a**

Please provide your opinion about this lecture in Mentimeter at most 3 words/phrases!

Questions?

Thank you for your attention!

university of groningen

http://www.martijnwieling.nl
m.b.wieling@rug.nl