# Computational notebooks in R

## Intall the following packages

```
install.packages("knitr")
install.packages("rmarkdown")
install.packages("biostatUZH", repos = "http://R-Forge.R-project.org")
install.packages("Hmisc")
```

## Libraries & Variables

```
library(knitr)
library(rmarkdown)
library(biostatUZH)
library(Hmisc)
library(ggplot2)
# PATH_HOME = path.expand("~")
# PATH = file.path(PATH_HOME, "Data", "workshop1")
```

## Load data

```
# data = read.csv(file = file.path(PATH, "data.csv"), header = TRUE, stringsAsFactors = TRUE)
data = read.csv(file = "data.csv", header = TRUE, stringsAsFactors = TRUE)
```

## Analysis 1a: Descriptives of the studies

Table with the number of studies, the median effect size, median sample size, range of publication years, and median number of citations per year per medicall specialty.

```
# number of studies per medical specialty
nrOfStudies = summary(data$specialty)
# median effect size per medical specialty
effect = tapply(data$effect.r, data$specialty, function (x) median(abs(x)))
effect_IRQ_lower = tapply(data$effect.r, data$specialty, function (x) quantile(abs(x), 0.25))
effect_IRQ_upper = tapply(data$effect.r, data$specialty, function (x) quantile(abs(x), 0.75))
# median sample size per medical specialty
sampleSize = tapply(data$effect.N, data$specialty, function (x) median(x))
# range of publication years
publicationYears = tapply(data$year, data$specialty,
                          function (x) sprintf("%d-%d",
                                               range(x, na.rm = TRUE)[1],
```

```
                                                range(x, na.rm = TRUE)[2]))
citations = tapply(data$cited.peryear, data$specialty, function (x) median(x, na.rm = TRUE))

# create a data frame
tab.descriptives = data.frame(nrOfStudies,
                              effect = sprintf("%.2f (%.2f-%.2f)", effect, effect_IRQ_lower, effect_IRQ
                              sprintf("%.0f", sampleSize), publicationYears,
                              citations = sprintf("%.1f", citations)
                              )

tab = tab.descriptives[order(tab.descriptives$nrOfStudies, decreasing = TRUE),]
colnames(tab) = c("studies", "effect (IRQ)", "N", "publication years", "citations")
kable(tab)
```

| | studies | effect (IRQ) | N | publication years | citations |
|---|---|---|---|---|---|
| Psychiatry & Mental Health | 8091 | 0.16 (0.07-0.31) | 76 | 1953-2019 | 4.3 |
| Pregnancy & Childbirth | 5031 | 0.19 (0.08-0.34) | 109 | 1922-2019 | 2.2 |
| Gynaecology & Urology | 4736 | 0.19 (0.07-0.35) | 90 | 1960-2019 | 3.0 |
| Anaesthesia & Pain | 4686 | 0.26 (0.12-0.42) | 61 | 1962-2019 | 3.6 |
| Heart & Hypertension | 4250 | 0.20 (0.08-0.35) | 100 | 1950-2019 | 3.9 |
| Lungs | 4191 | 0.17 (0.07-0.31) | 83 | 1923-2019 | 3.1 |
| Gastroenterology | 3466 | 0.20 (0.08-0.36) | 94 | 1957-2019 | 4.4 |
| Emergency & Trauma | 3281 | 0.19 (0.08-0.36) | 64 | 1956-2018 | 4.5 |
| Neurology | 2662 | 0.19 (0.08-0.34) | 60 | 1958-2019 | 5.5 |
| Oral Health, Eyes & ENT | 2566 | 0.21 (0.09-0.40) | 61 | 1951-2019 | 3.6 |
| Kidney & Transplant | 2554 | 0.21 (0.09-0.37) | 60 | 1968-2018 | 2.8 |
| Oncology | 2511 | 0.14 (0.06-0.30) | 120 | 1963-2018 | 4.2 |
| Spine & Muscles | 2457 | 0.19 (0.08-0.33) | 63 | 1955-2018 | 5.9 |
| Skin & Wounds | 2438 | 0.24 (0.09-0.42) | 73 | 1964-2018 | 2.7 |
| Infectious Diseases | 2212 | 0.19 (0.08-0.37) | 128 | 1931-2018 | 3.1 |
| Public Health & Work | 1986 | 0.15 (0.06-0.29) | 113 | 1946-2018 | 5.7 |
| Hepato-Biliary | 1976 | 0.22 (0.09-0.36) | 63 | 1966-2018 | 3.2 |
| Genetics & Endocrinology | 1751 | 0.19 (0.07-0.36) | 65 | 1961-2017 | 5.7 |
| Neonatal | 1600 | 0.18 (0.08-0.35) | 60 | 1954-2018 | 3.0 |

## Analysis 1b: Proportion of significant effects per specialty

```
numberSign = tapply(data$effect.p, data$specialty, function (x) sum(x < 0.05))
total = tapply(data$effect.p, data$specialty, length)

myCI =  binconf(x=numberSign, n=total, method = "wilson")

tab.effect = data.frame(
  specialty = names(numberSign),
  prop = numberSign/total,
  lower = myCI[,2],
  upper = myCI[,3],
  proportion = sprintf("%.2f (%.2f-%.2f)",numberSign/total, myCI[,2], myCI[,3]))
rownames(tab.effect) = NULL
kable(tab.effect[order(tab.effect$prop, decreasing = TRUE),c("specialty","proportion")], row.names = FA
```

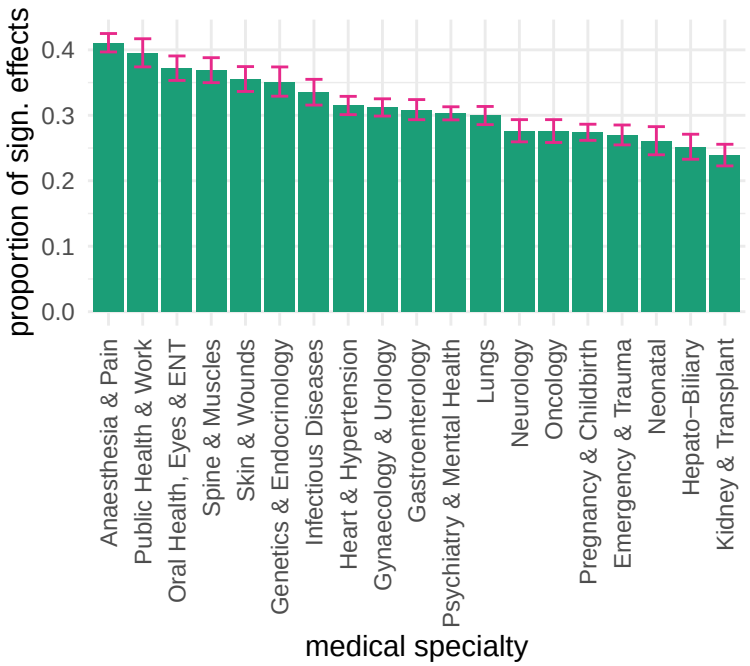
| specialty | proportion |
|---|---|
| Anaesthesia & Pain | 0.41 (0.40-0.42) |
| Public Health & Work | 0.40 (0.37-0.42) |
| Oral Health, Eyes & ENT | 0.37 (0.35-0.39) |
| Spine & Muscles | 0.37 (0.35-0.39) |
| Skin & Wounds | 0.36 (0.34-0.37) |
| Genetics & Endocrinology | 0.35 (0.33-0.37) |
| Infectious Diseases | 0.33 (0.32-0.35) |
| Heart & Hypertension | 0.31 (0.30-0.33) |
| Gynaecology & Urology | 0.31 (0.30-0.33) |
| Gastroenterology | 0.31 (0.29-0.32) |
| Psychiatry & Mental Health | 0.30 (0.29-0.31) |
| Lungs | 0.30 (0.29-0.31) |
| Neurology | 0.28 (0.26-0.29) |
| Oncology | 0.28 (0.26-0.29) |
| Pregnancy & Childbirth | 0.27 (0.26-0.29) |
| Emergency & Trauma | 0.27 (0.25-0.29) |
| Neonatal | 0.26 (0.24-0.28) |
| Hepato-Biliary | 0.25 (0.23-0.27) |
| Kidney & Transplant | 0.24 (0.22-0.26) |

```
stat = prop.test(numberSign, total, correct = FALSE)
paste("p", formatPval(stat$p.value))
```

```
## [1] "p < 0.0001"
```

```
# simple plot:
# barplot(tab.effect$prop)
# order levels
idx = order(tab.effect$prop, decreasing = TRUE)
tab.effect$specialty = factor(tab.effect$specialty, levels = tab.effect$specialty[idx])

ggplot(tab.effect, aes(x=specialty, y=prop)) +
  geom_bar(stat="identity", fill="#1b9e77") +
  geom_errorbar(aes(ymin=lower, ymax=upper), width=.5, color="#e7298a") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  ylab("proportion of sign. effects") +
  xlab("medical specialty")
```




## Analysis 1c: Logistic model of statistical significance of a study

```
data$specialty = relevel(data$specialty, ref = "Gastroenterology")

fit = glm((effect.p < 0.05) ~ specialty + SJR.Best.Quartile + log2(effect.N) +
              log2(cited.peryear) + year, data = data, family = "binomial")

tab = tableOR(fit, latex = FALSE, short = TRUE,
              refLevels = c("Gastroenterology", "Q1"), Wald = TRUE)
kable(tab)
```

| | OR | 95% CI | z value | p-value |
|---|---|---|---|---|
| specialty: Gastroenterology | 1 | – | – | – |
| Anaesthesia & Pain | 2.05 | from 1.79 to 2.34 | 10.36 | $< 0.0001$ |
| Emergency & Trauma | 0.97 | from 0.84 to 1.12 | -0.41 | 0.68 |
| Genetics & Endocrinology | 1.44 | from 1.21 to 1.72 | 4.08 | $< 0.0001$ |
| Gynaecology & Urology | 1.28 | from 1.12 to 1.47 | 3.55 | 0.0004 |
| Heart & Hypertension | 1.07 | from 0.93 to 1.23 | 0.92 | 0.36 |
| Hepato-Biliary | 0.68 | from 0.57 to 0.82 | -4.12 | $< 0.0001$ |
| Infectious Diseases | 1.12 | from 0.94 to 1.33 | 1.28 | 0.20 |
| Kidney & Transplant | 0.84 | from 0.71 to 0.98 | -2.16 | 0.031 |
| Lungs | 1.09 | from 0.95 to 1.26 | 1.23 | 0.22 |
| Neonatal | 0.98 | from 0.83 to 1.17 | -0.19 | 0.85 |
| Neurology | 0.90 | from 0.77 to 1.06 | -1.22 | 0.22 |
| Oncology | 0.98 | from 0.83 to 1.16 | -0.23 | 0.81 |
| Oral Health, Eyes & ENT | 1.52 | from 1.30 to 1.78 | 5.29 | $< 0.0001$ |
| Pregnancy & Childbirth | 0.95 | from 0.83 to 1.09 | -0.78 | 0.44 |
| Psychiatry & Mental Health | 1.14 | from 1.00 to 1.29 | 2.02 | 0.043 |
| Public Health & Work | 1.61 | from 1.37 to 1.89 | 5.80 | $< 0.0001$ |

| | OR | 95% CI | z value | p-value |
|---|---|---|---|---|
| Skin & Wounds | 1.57 | from 1.34 to 1.83 | 5.68 | $< 0.0001$ |
| Spine & Muscles | 1.67 | from 1.43 to 1.96 | 6.43 | $< 0.0001$ |
| SJR.Best.Quartile: Q1 | 1 | – | – | – |
| Q2 | 1.26 | from 1.18 to 1.35 | 6.54 | $< 0.0001$ |
| Q3 | 1.33 | from 1.20 to 1.47 | 5.58 | $< 0.0001$ |
| Q4 | 1.43 | from 1.22 to 1.68 | 4.46 | $< 0.0001$ |
| log2(effect.N) | 1.12 | from 1.11 to 1.14 | 16.24 | $< 0.0001$ |
| log2(cited.peryear) | 1.07 | from 1.06 to 1.09 | 9.88 | $< 0.0001$ |
| year | 0.99 | from 0.99 to 0.99 | -6.85 | $< 0.0001$ |