MRI-Together 2021-https://mritogether.github.io/

A White Hat's Guide to p-Hacking

Dr. Xeni Deligianni- University of Basel

xeni.deligianni@unibas.ch https://github.com/XDeligianni

Imagine, we have two data distributions, one the **controls** and one a distribution that we want to compare to the control. This could be the quantitative values of a volunteers' group (controls="ctrl"), let's call them **qmr** and the respective values of a different group e.g. patients ("pat") with a certain condition.

Let's assume for now that the distributions are **normal** Gaussian distributions. So let's assume we checked the distributions. And let' s assume, that BEFORE starting our analysis, we have a **hypothesis** that in disease presence qmr values are increased. And let's imagine we have estimated that we should measure and we measured under identical and ideal conditions **50 healthy controls** and **50 patients**.

Let' s agree on some definitions first

α: Significance level.A p-value below this will lead to the null hypothesis being rejected.

## Ready? Let's create some fictional data.

Let's create two perfectly **normal distributions**. One with **mean** 40 and **standard deviation** (sd) 5 and one with mean 10 and standard deviation 2. And let's visualize at their distributions, as well as the distribution of their difference.

```
## Loading required package: viridisLite
```
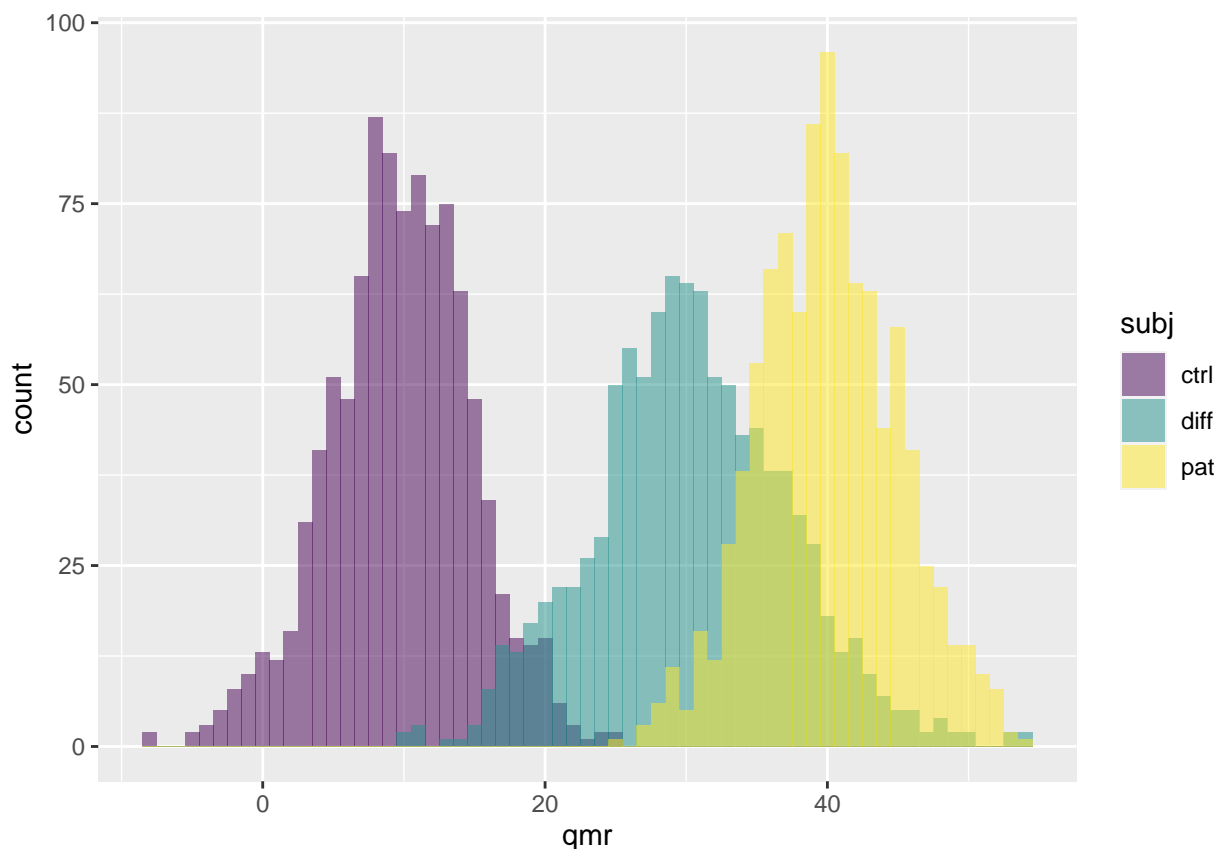
That looks clear. Doesn't it? Let's also look at the boxplot visualization.

```
library(ggplot2)# A great library for visualization
library(viridis)
load('New_Data.RData')
#qmr <-c(rnorm(1000, mean=10, sd=5),rnorm(1000, mean=40, sd=5))
#subj <- c(rep("ctrl",1000),rep("pat",1000),rep("diff",1000))
#d.exp <- data.frame(qmr,subj)
x<-d.exp[d.exp$subj=='pat',]$qmr
bw <- 2 * IQR(x) / length(x)^(1/3)#decide for the nr of bins-Friedman-Diaconis rule
ggplot(data = d.exp,mapping = aes(y = qmr,x = subj, fill=subj)) + geom_boxplot()+scale_fill_viridis_d()
```
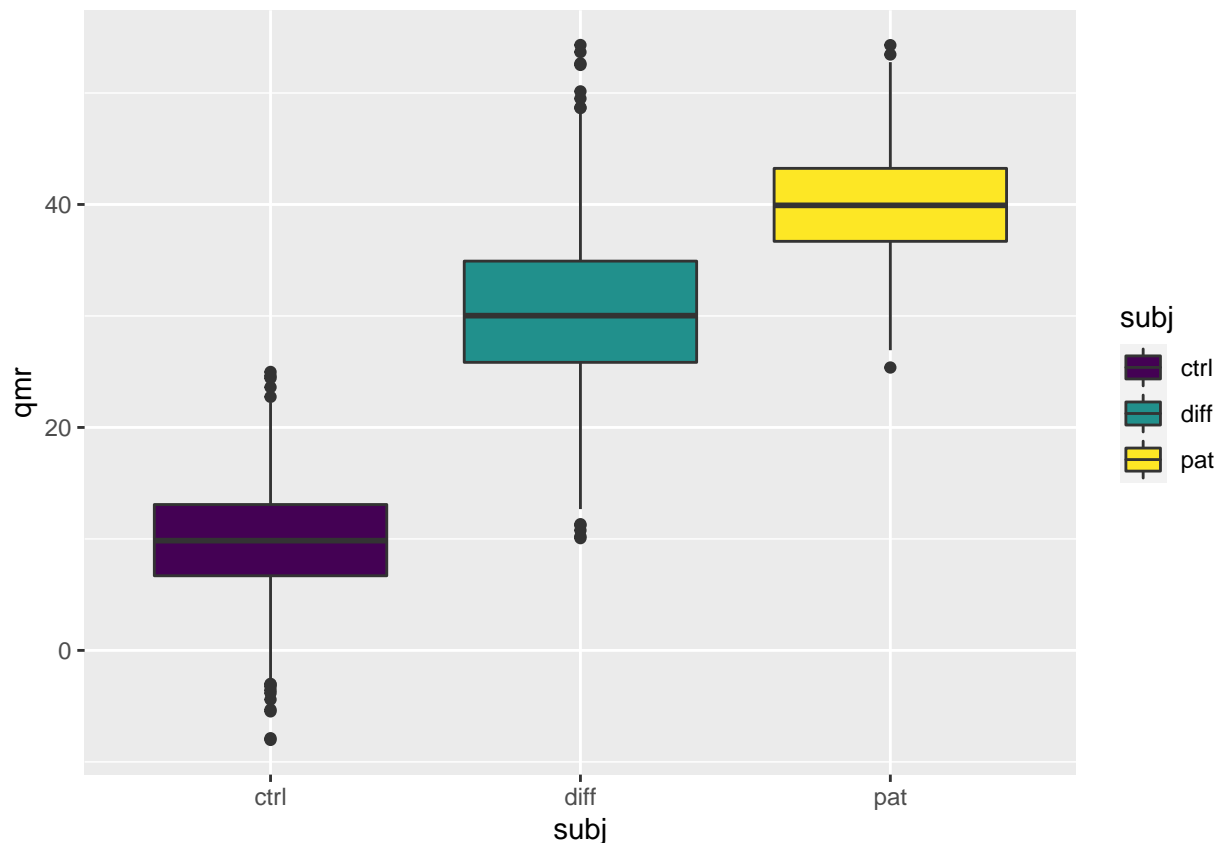
```
#colorblind accessible colors
```

So what's next? How likely is that this difference is 0.

What would we get by testing whether there is a difference in these 2 distributions. Having all our assumptions, most of us would go for a t-test.

```
load('New_Data.RData')

#qmr <-c(rnorm(1000, mean=10, sd=5),rnorm(1000, mean=40, sd=5))
#subj <- c(rep("ctrl",1000),rep("pat",1000),rep("diff",1000))
#d.exp <- data.frame(qmr,subj)
# Randomly choose number of samples (1000/50/25/15/5)

nr_s<-4

p_series<-vector()
for (i in 1:100) {
  sample_rows_ctrl<-sample(nrow(d.exp[d.exp$subj=="ctrl",]),nr_s)
  sample_rows_pat<-1000+sample(nrow(d.exp[d.exp$subj=="pat",]),nr_s)
  d.exp_sub<-rbind(d.exp[sample_rows_ctrl,],d.exp[sample_rows_pat,])
  p_Res<-t.test(d.exp_sub[d.exp_sub$subj=="pat",]$qmr, d.exp_sub[d.exp_sub$subj=="ctrl",]$qmr, paired =
alternative = "two.sided", conf.level = 0.95,var.equal=TRUE)
  p_series [i]<-p_Res$p.value
}
```

```
plot(p_series,ylim=c(0,0.06))
abline(h=0.05)
```
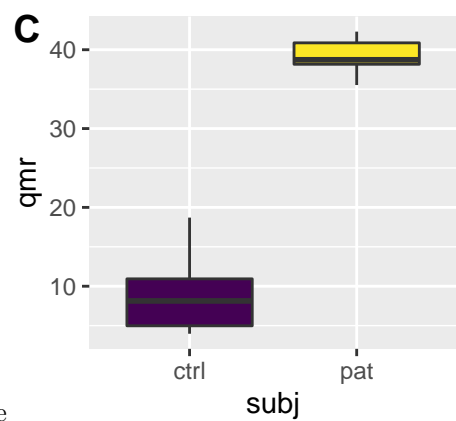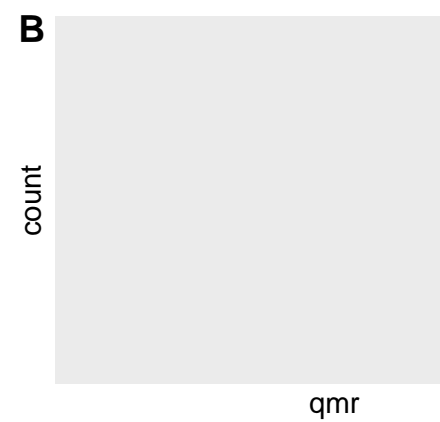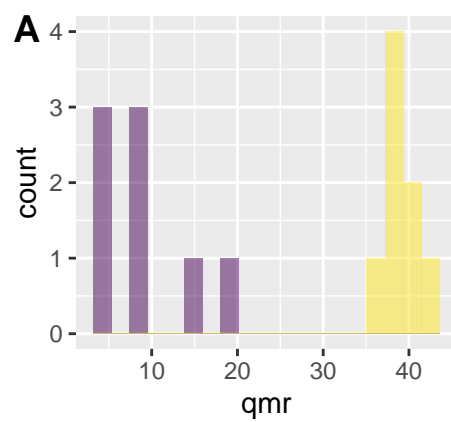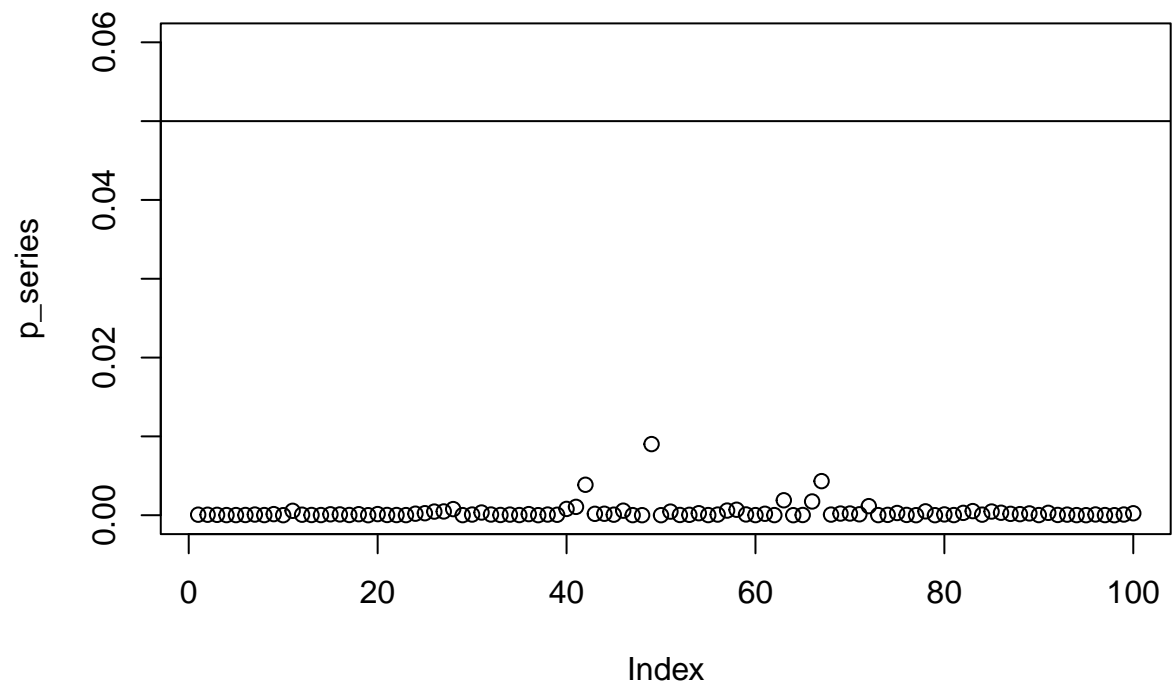


So let's sum up and continue

##

```
##  Paired t-test
##
## data:  d.exp_sub[d.exp_sub$subj == "pat", ]$qmr and d.exp_sub[d.exp_sub$subj == "ctrl", ]$qmr
## t = 14.295, df = 7, p-value = 1.949e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   25.00649 34.91887
## sample estimates:
## mean of the differences
##                 29.96268
```

But what if we don't have 1000 samples or 1000 representative samples?

Trap Nr. 1: Be real! 5 data points is not 5000

Topic 1: The number of samples is important What if we have 15 samples of some less different data. Let's get them & repeat the steps

p-values change but our difference is still constantly higher than 0. ## But what if things are not so clear? Let's modify bring the mean values closer and increase the standard deviation. Keeping the assumptions for t-test we keep standard deviation the same.
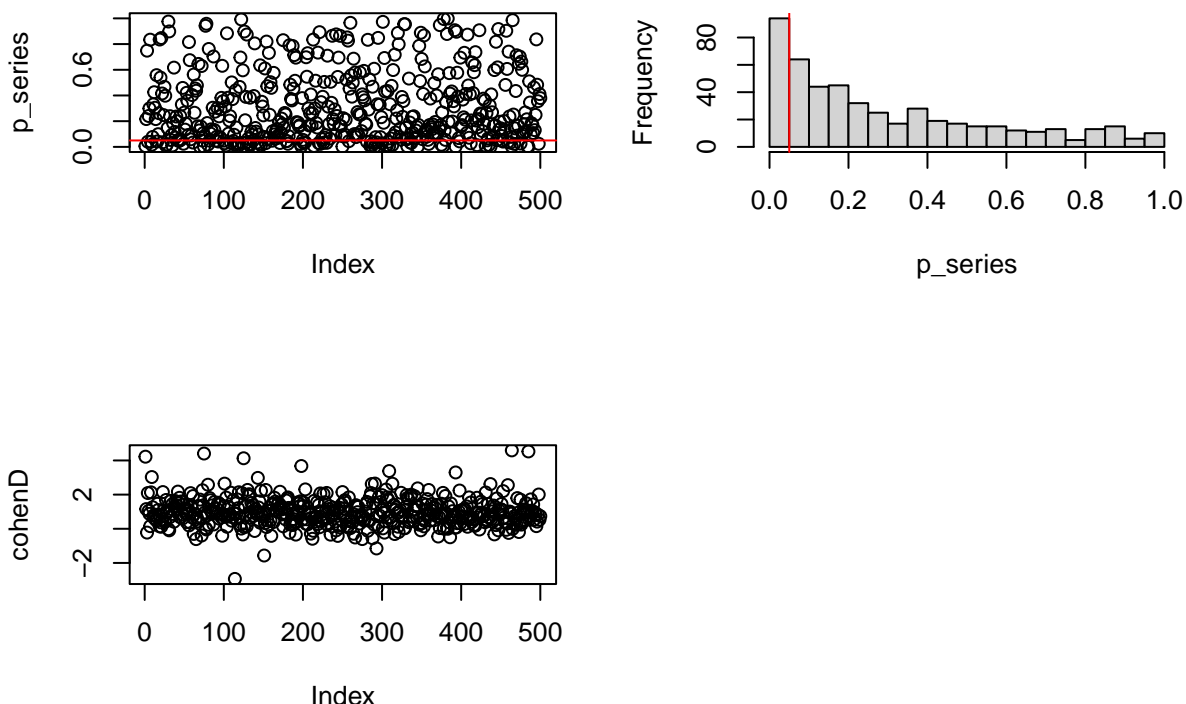


```
##
##  Paired t-test
##
## data:  d.exp[d.exp$subj == "pat", ]$qmr and d.exp[d.exp$subj == "ctrl", ]$qmr
```

5
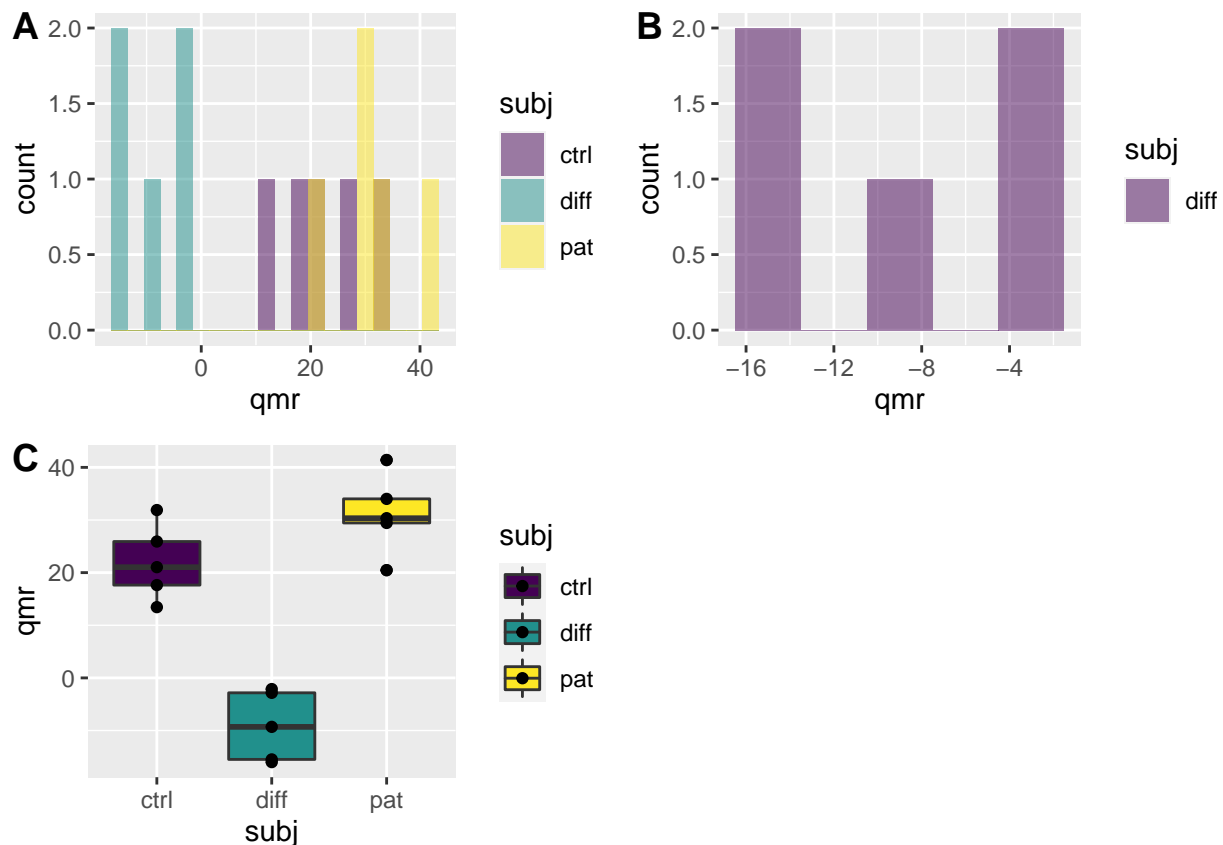
```
## t = 3.0807, df = 4, p-value = 0.03691
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##    0.9036059 17.3941725
## sample estimates:
## mean of the differences
##                9.148889


##
## Cohen's d
##
## d estimate: 1.237462 (large)
## 95 percent confidence interval:
##       lower      upper
## -0.3544583  2.8293833
```


Histogram of p_series

What can we do?

Decide in advance the sample size we need! If we don' t know maybe it is an exploratory analysis after all
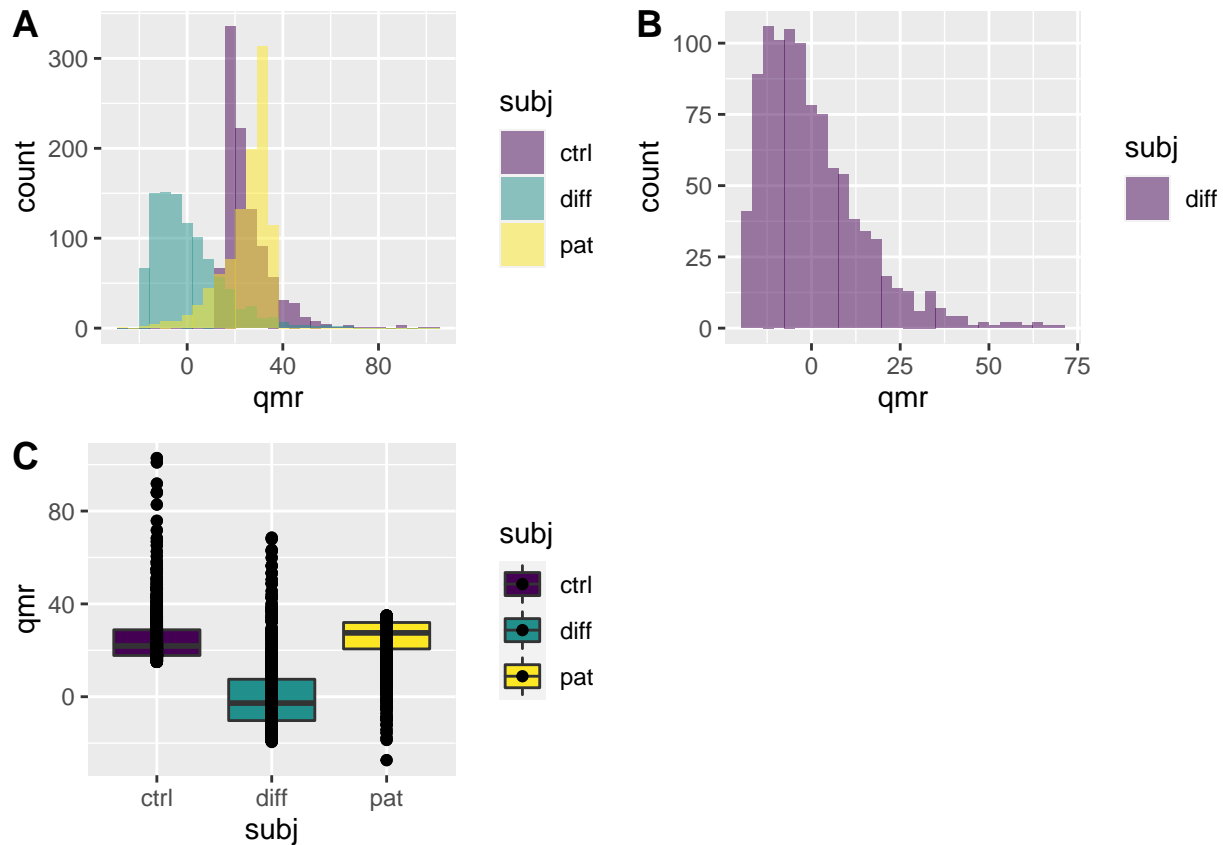
###Clinical vs statistical significance (to the slides and back)

Trap Nr. 2: Step 1: Look at your data

**Not all datasets are normal, t-test is not for everything!**

The same way as mean is not always the proper statistic.

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

```
## 
##  Paired t-test
## 
## data:  d.exp[d.exp$subj == "pat", ]$qmr and d.exp[d.exp$subj == "ctrl", ]$qmr
## t = -0.76922, df = 999, p-value = 0.4419
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.2379758  0.5407345
## sample estimates:
## mean of the differences
##               -0.3486206


## 
##  Wilcoxon signed rank test with continuity correction
## 
## data:  d.exp[d.exp$subj == "pat", ]$qmr and d.exp[d.exp$subj == "ctrl", ]$qmr
## V = 276158, p-value = 0.00457
## alternative hypothesis: true location shift is not equal to 0
```
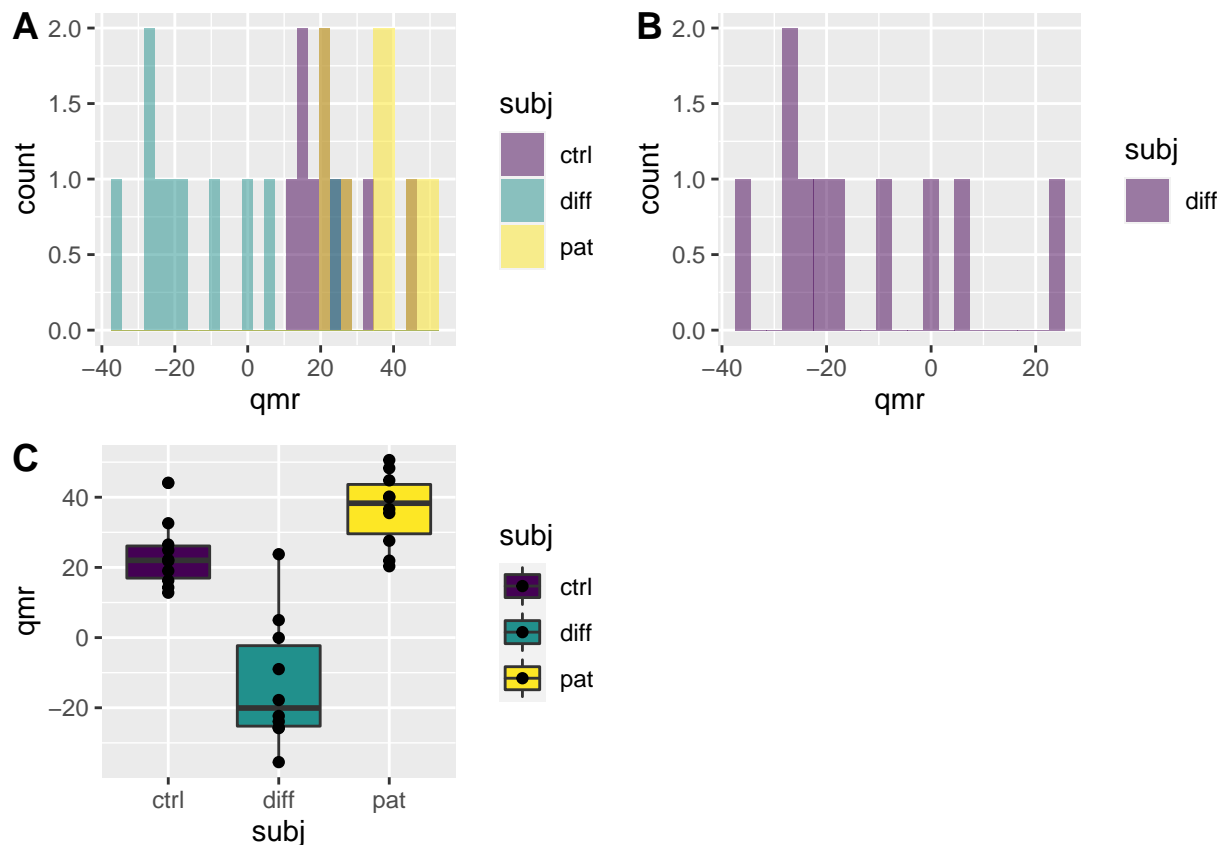
```
## 
##  Paired t-test
## 
## data:  d.exp[d.exp$subj == "pat", ]$qmr and d.exp[d.exp$subj == "ctrl", ]$qmr
## t = 2.306, df = 9, p-value = 0.04654
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.249618 26.007024
## sample estimates:
## mean of the differences
##                 13.12832


## 
## Cohen's d
## 
## d estimate: 1.320048 (large)
## 95 percent confidence interval:
##     lower     upper
## 0.2831986 2.3568978
```
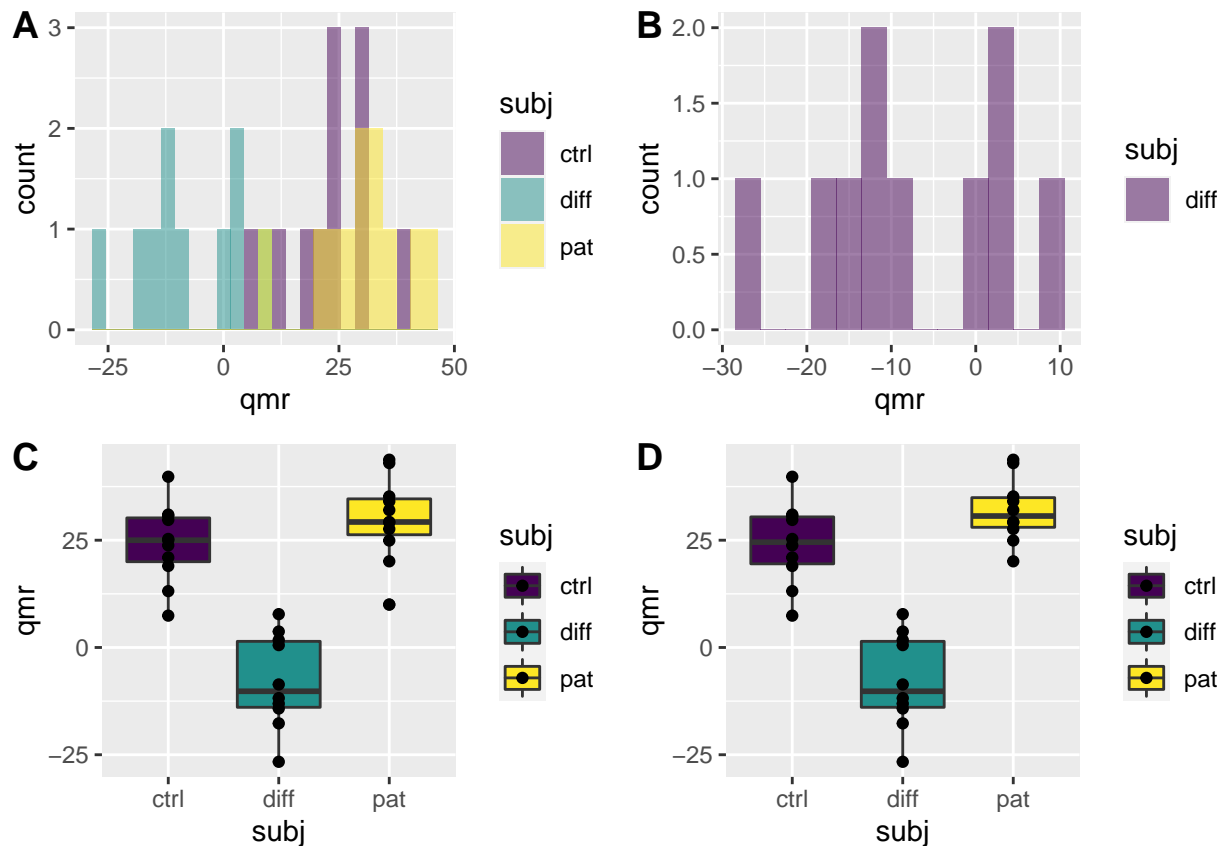
Histogram of p_series
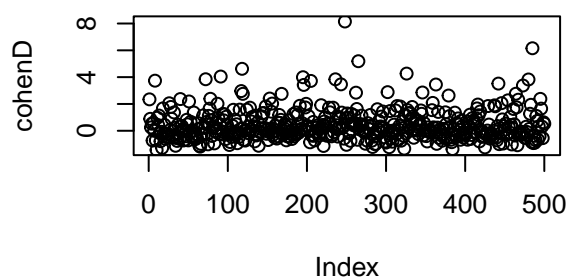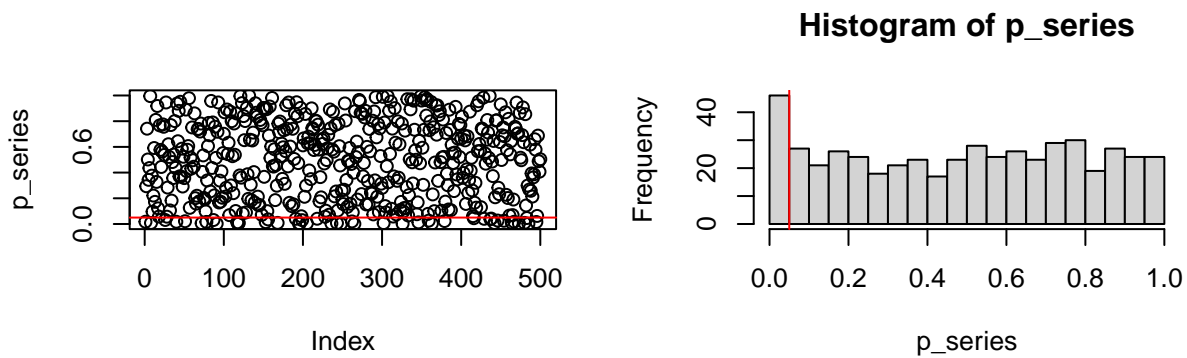
Trap 3: Just one more experiment then! It almost looks good. Let's repeat it. Trap of adding data on
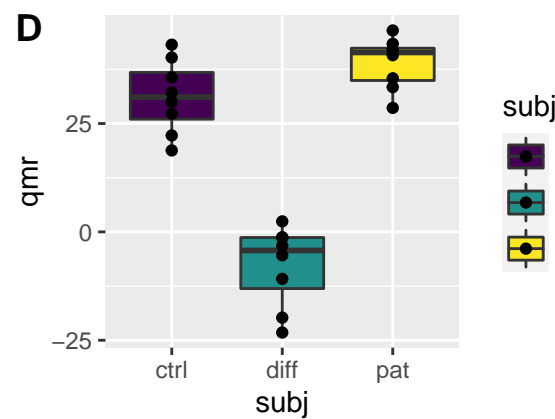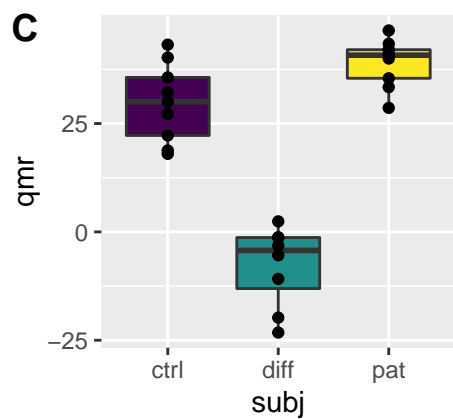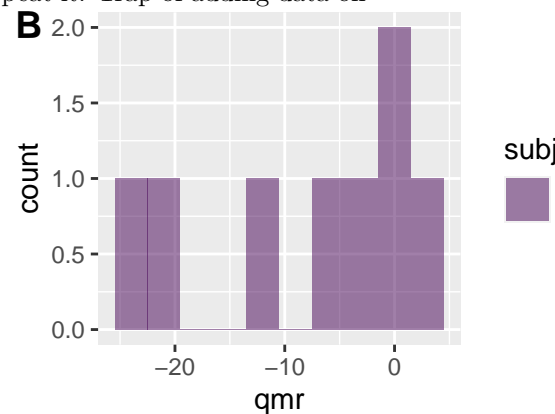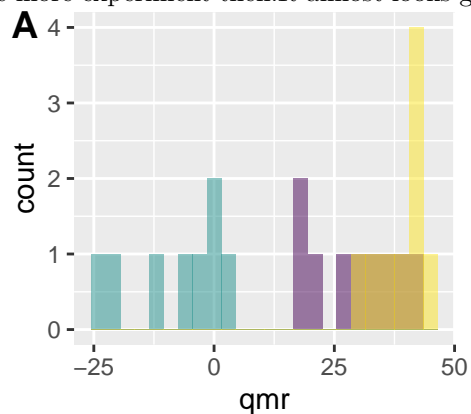
marginal distributions.

9

## Trap Nr. 3: This one looks wrong. Let's remove it

##TO DO:post-hoc data selection, Keep adjusting the data collection removing outliers, 1) use different threshold, 2) remove outliers and test till you get you result

Trap Nr. 4: Problem of repeated sequential testing

Bonferroni and other corrections

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##   Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##   1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##   Median :5.800   Median :3.000   Median :4.350   Median :1.300
##   Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##   3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##   Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##         Species
##   setosa    :50
##   versicolor:50
##   virginica :50
##
##
##
```
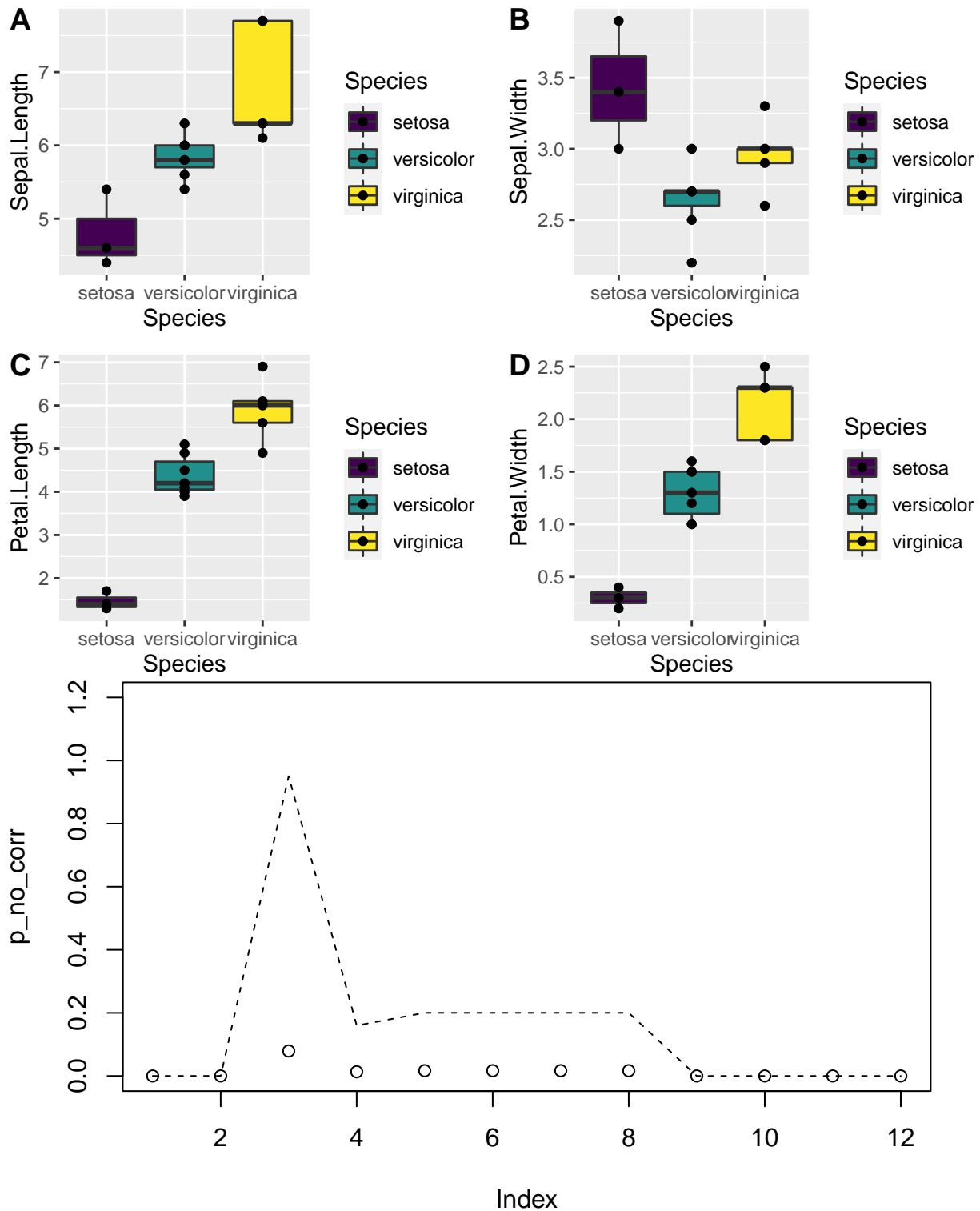
Trap Nr. 5: TODO POST HOC hypothesis

Be ware of one-sided tests,

###Check the assumptions: (different variation, reaching normality)

[1] 0.04930856 0.97534572 0.02465428

##References

Loading required namespace: bibtex

[1] _False-Positive Psychology: Undisclosed Flexibility in Data
Collection and Analysis Allows Presenting Anything as Significant -
Joseph P. Simmons, Leif D. Nelson, Uri Simonsohn, 2011_. Dez. 11, 2021.
<URL: https://journals.sagepub.com/doi/full/10.1177/0956797611417632>
(visited on 12/11/2021).
##
[2] _Understanding The New Statistics : Geoff Cumming :_. Dez. 11,
2021. (Visited on 12/11/2021).
##
[3] M. J. Campbell, D. Machin, and S. J. Walters. _Medical Statistics_.
Dez. 11, 2021. (Visited on 12/11/2021).
##
[4] Geoff Cumming. _Intro Statistics 9 Dance of the p Values_. Sep.
2013. <URL: https://www.youtube.com/watch?v=5OL1RqHrZQ8> (visited on
12/11/2021).
##
[5] N. L. Kerr. "HARKing: Hypothesizing After the Results are Known".
En. In: _Personality and Social Psychology Review_ 2.3 (Aug. 1998).
Publisher: SAGE Publications Inc, pp. 196-217. ISSN: 1088-8683. DOI:
10.1207/s15327957pspr0203_4. <URL:
https://doi.org/10.1207/s15327957pspr0203_4> (visited on 12/11/2021).
##
[6] S. Lee and D. Lee. "What is the proper way to apply the multiple
comparison test?" In: _Korean Journal of Anesthesiology_ 73 (Dez.
2020), pp. 572-572. DOI: 10.4097/kja.d.18.00242.e1.
##
[7] C. Pernet. _Hacking, HARKing and SHARKING your research: a
tutorial_. En. presentation. Publisher: figshare. Sep. 2017. DOI:
10.6084/m9.figshare.5451067.v1. <URL:
https://figshare.com/articles/presentation/Hacking_HARKing_and_SHARKING_your_research_a_tutorial/545
(visited on 12/11/2021).
##
[8] S. Schwab and L. Held. "Different worlds Confirmatory versus
exploratory research". En. In: _Significance_ 17.2 (2020). _ eprint:
https://onlinelibrary.wiley.com/doi/pdf/10.1111/1740-9713.01369, pp.
8-9. ISSN: 1740-9713. DOI: 10.1111/1740-9713.01369. <URL:
https://onlinelibrary.wiley.com/doi/abs/10.1111/1740-9713.01369>
(visited on 12/11/2021).
##
[9] StatQuest with Josh Starmer. _p-hacking: What it is and how to
avoid it!_ Mai. 2020. <URL:
https://www.youtube.com/watch?v=HDCOUXE3HMM> (visited on 12/11/2021).
##
[10] G. M. Sullivan and R. Feinn. "Using Effect Size-or Why the P Value
Is Not Enough". In: _Journal of Graduate Medical Education_ 4.3 (Sep.
2012), pp. 279-282. ISSN: 1949-8349. DOI: 10.4300/JGME-D-12-00156.1.
<URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3444174/> (visited

on 12/11/2021).
##
[11] T. L. Weissgerber, O. Garcia-Valencia, V. D. Garovic, et al. "Why
we need to report more than 'Data were Analyzed by t-tests or ANOVA'".
In: _eLife_ 7 (Dez. 2018). Ed. by M. D. Teare and P. A. Rodgers.
Publisher: eLife Sciences Publications, Ltd, p. e36163. ISSN:
2050-084X. DOI: 10.7554/eLife.36163. <URL:
https://doi.org/10.7554/eLife.36163> (visited on 12/11/2021).