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 controls. 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 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 **nr__s healthy controls** and **nr__s patients**. We consider these independent measurements.

Let' s agree on some definitions first

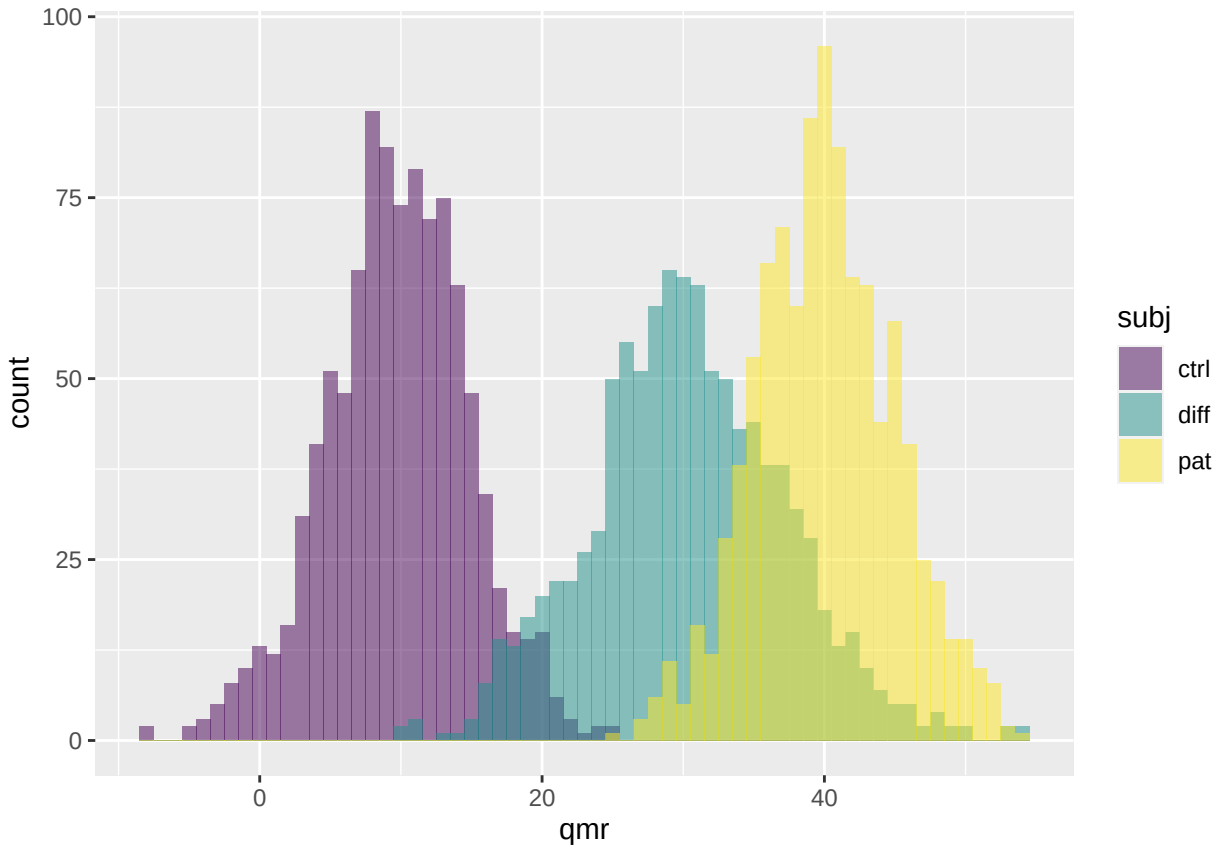
α: Significance level.

p-value:the probability of obtaining the observed difference, or one more extreme, if the null hypothesis is true. A p-value below α will lead to the null hypothesis being rejected.

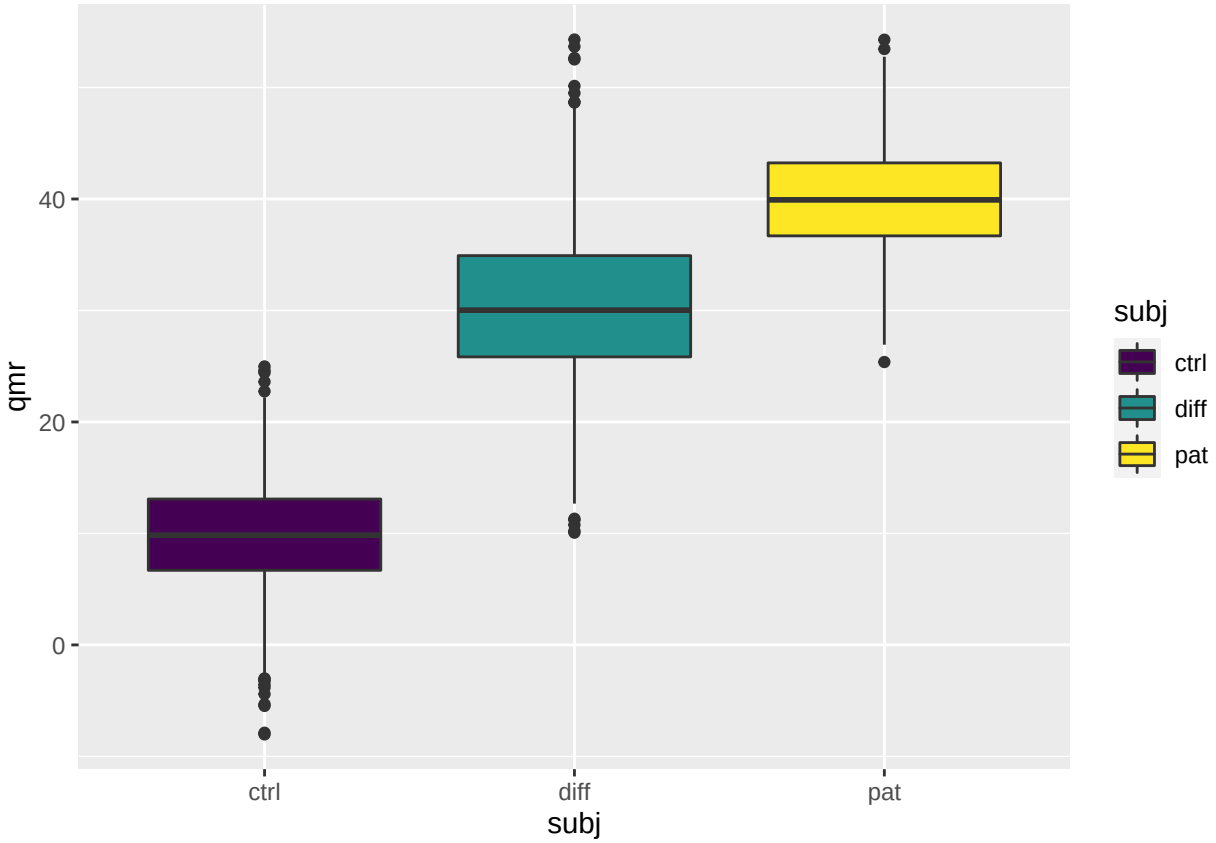
Ready? Let's create some fictional data.

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. We start with 1000 samples for each ditribution.

```
## Loading required package: viridisLite
```

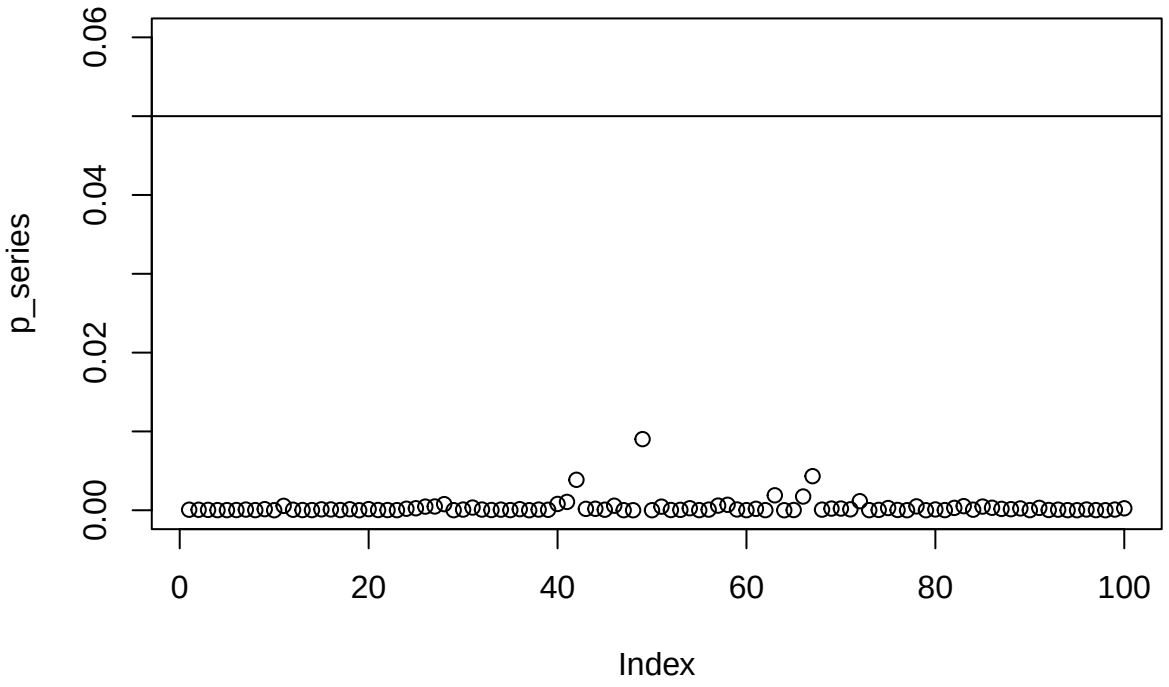

100
75
50
25
0
count
0
20
40
qmr
subj
ctrl
diff
pat

That looks clear. Doesn't it? Let's also look at the boxplot visualization.

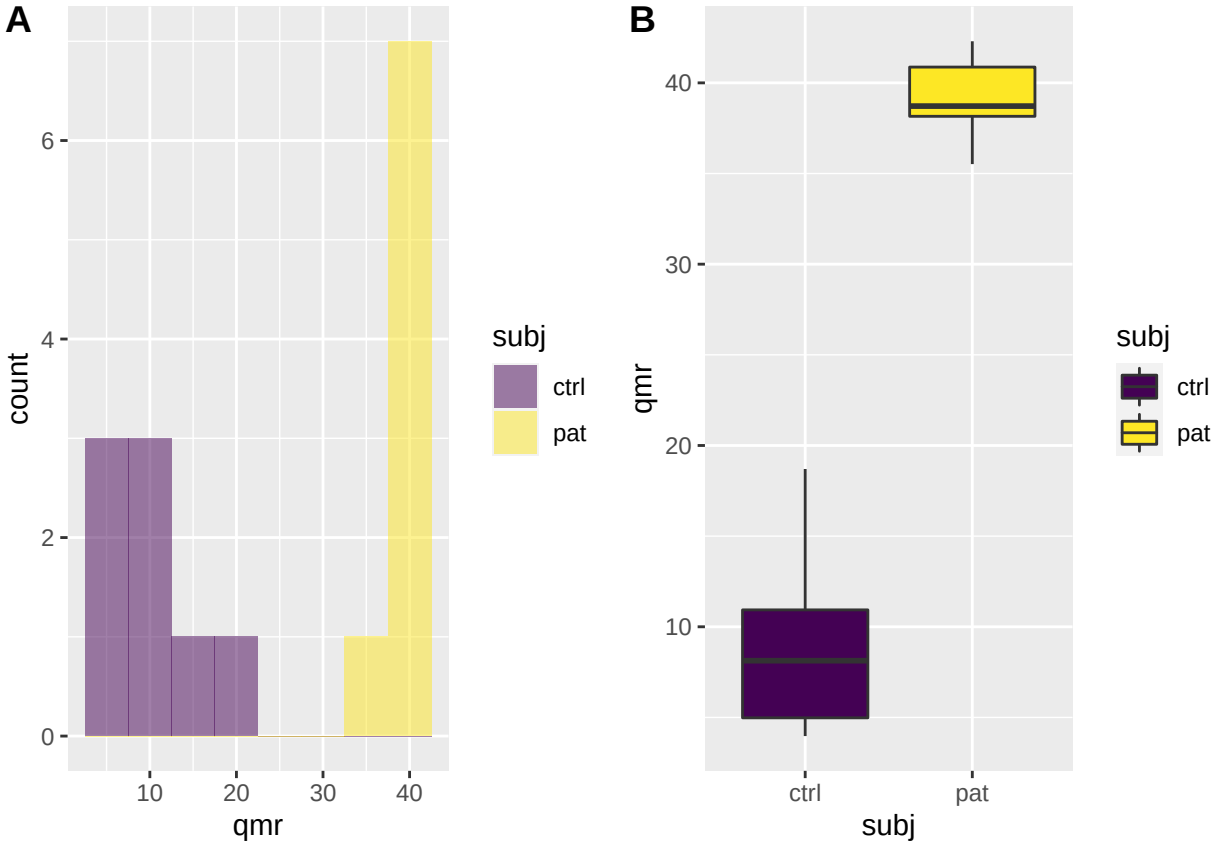


So what's next? What result would we get from a t-test?

What would we get by testing whether there is a difference in these 2 distributions and if the difference is different than zero. And let's see what would happen if we wouldn't have 1000 samples for each distribution, but less. Try to change the **nr_s** and plot.



###So let's sum up and continue Let's look at one of the subdatasets with 8 samples per distribution




```
##
##  Two Sample t-test
##
```

```
## data:  d.exp_sub[d.exp_sub$subj == "pat", ]$qmr and d.exp_sub[d.exp_sub$subj == "ctrl", ]$qmr
## t = 14.547, df = 14, p-value = 7.63e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  25.54505 34.38031
## sample estimates:
## mean of x mean of y
## 39.160019  9.197341
```

But what if the differences were less clear?

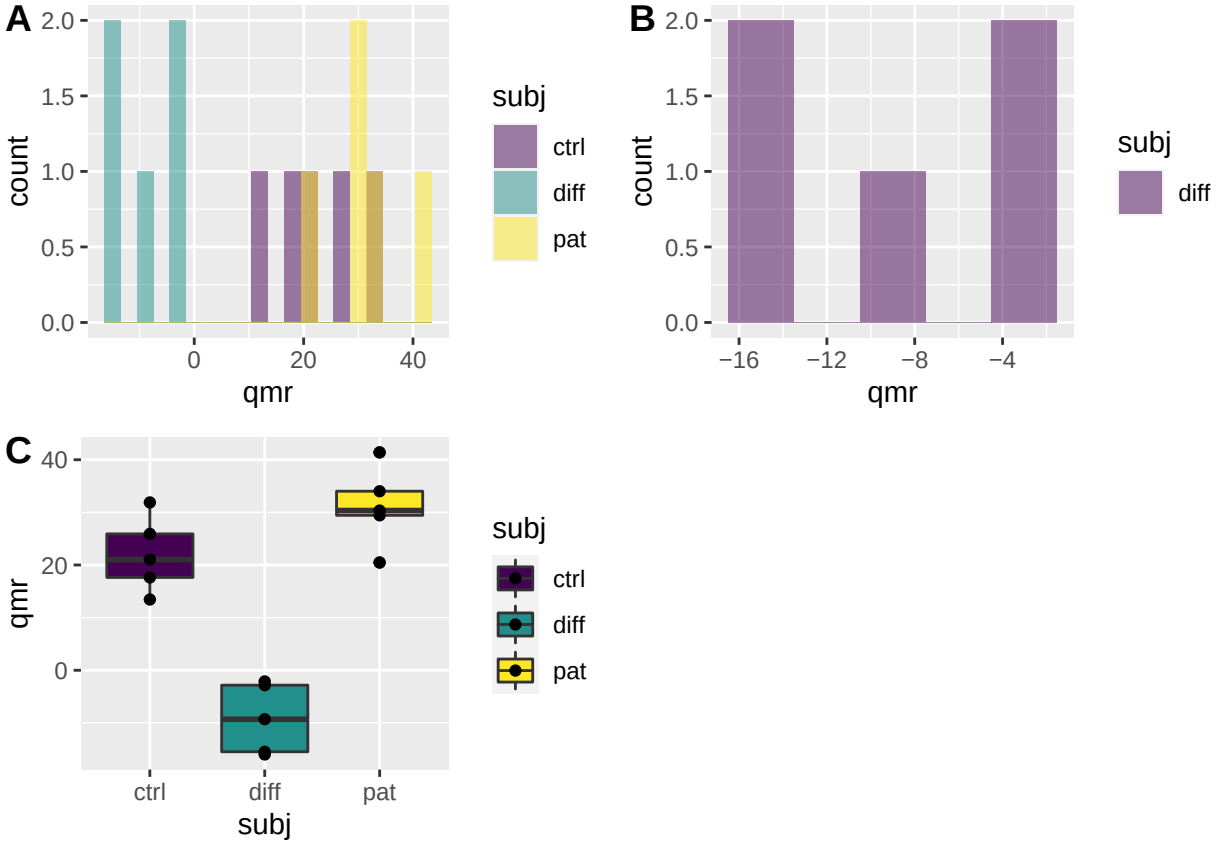
Trap Nr. 1: Be real! 5 data points is not 5000

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 change distributions, bring the mean values closer and increase the standard deviation. Keeping the assumptions for t-test, we keep standard deviation the same.

Here, you can try again to change the number of samples (nr_s) of each distribution and observe the effect: 1) on the distributions, 2) on the t-test performed once.

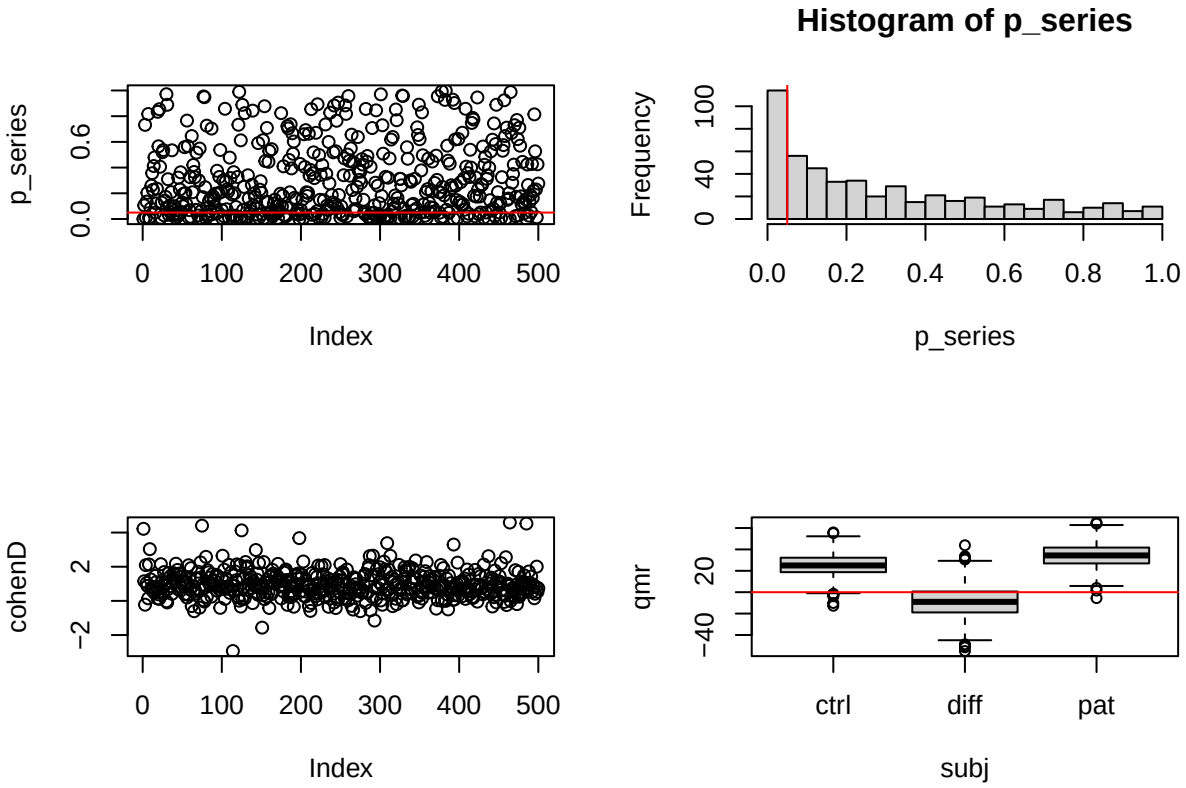



```
##
##  Two Sample t-test
##
```

```
## data:  d.exp[d.exp$subj == "pat", ]$qmr and d.exp[d.exp$subj == "ctrl", ]$qmr
## t = 1.9566, df = 8, p-value = 0.08611
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.633783 19.931561
## sample estimates:
## mean of x mean of y
##  31.13354  21.98465

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

So let's draw different number of samples **nr_s** (from the bigger normal distribution) and let's look at the p_value. Let's also look how the test result change, if we draw again nr_s new samples!



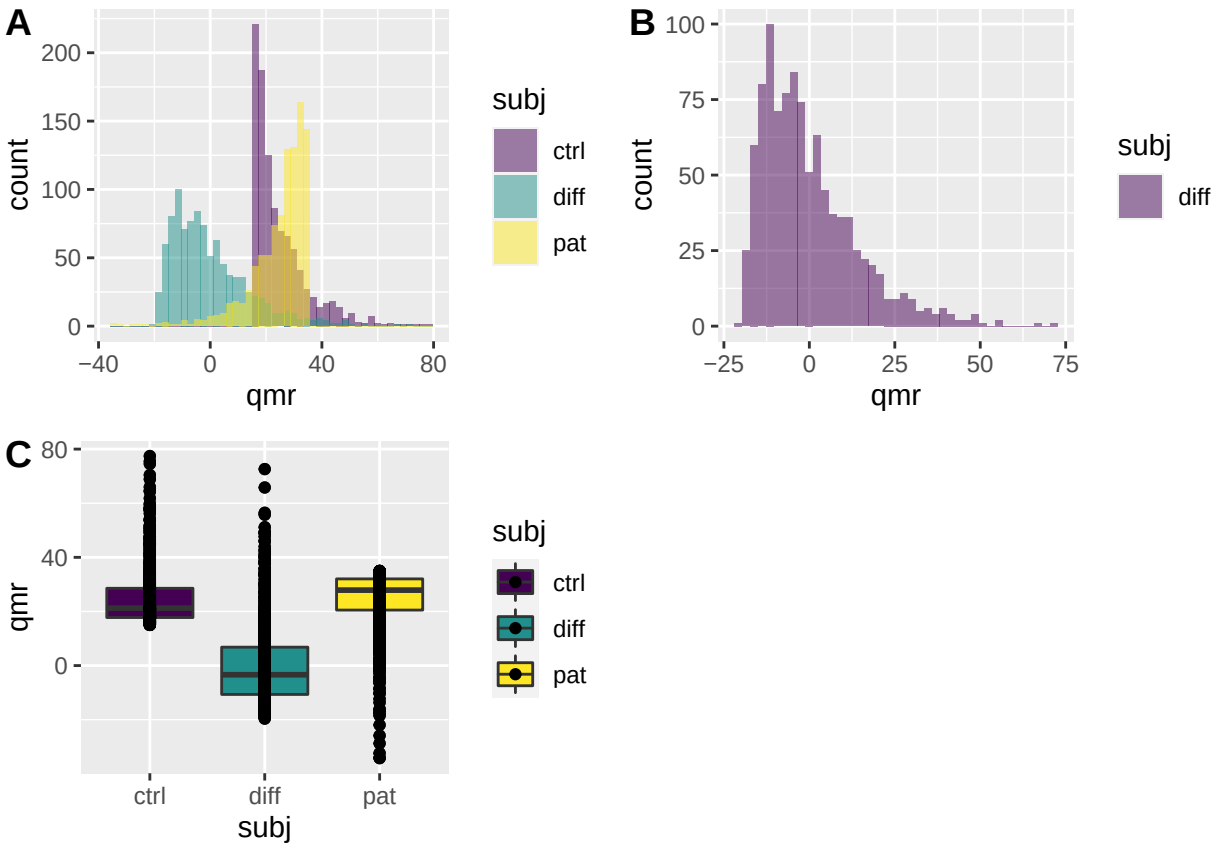
What can we do?

Decide in advance the sample size we need! If we don' t know maybe it is an exploratory analysis after all.

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

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

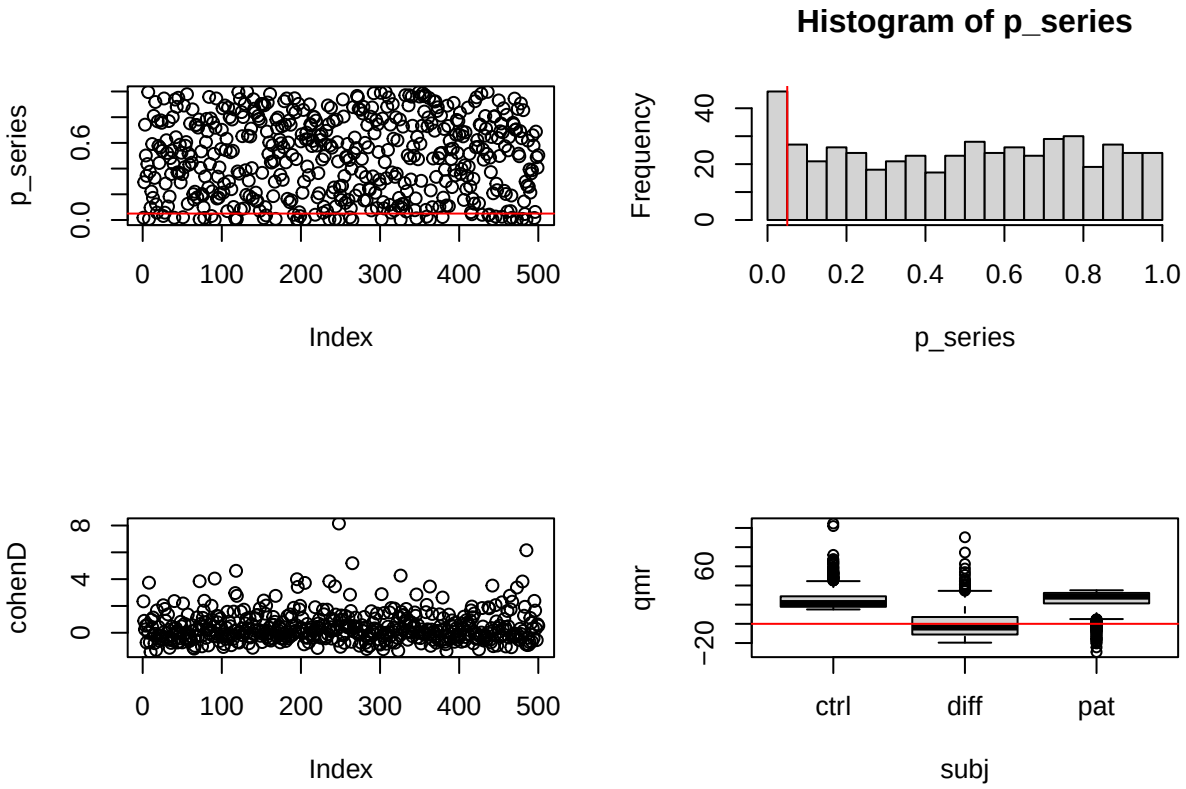
The same way as mean is not always the proper statistic. Let's now look at some different distributions. Let's start with 1000 samples and try to reduce it!



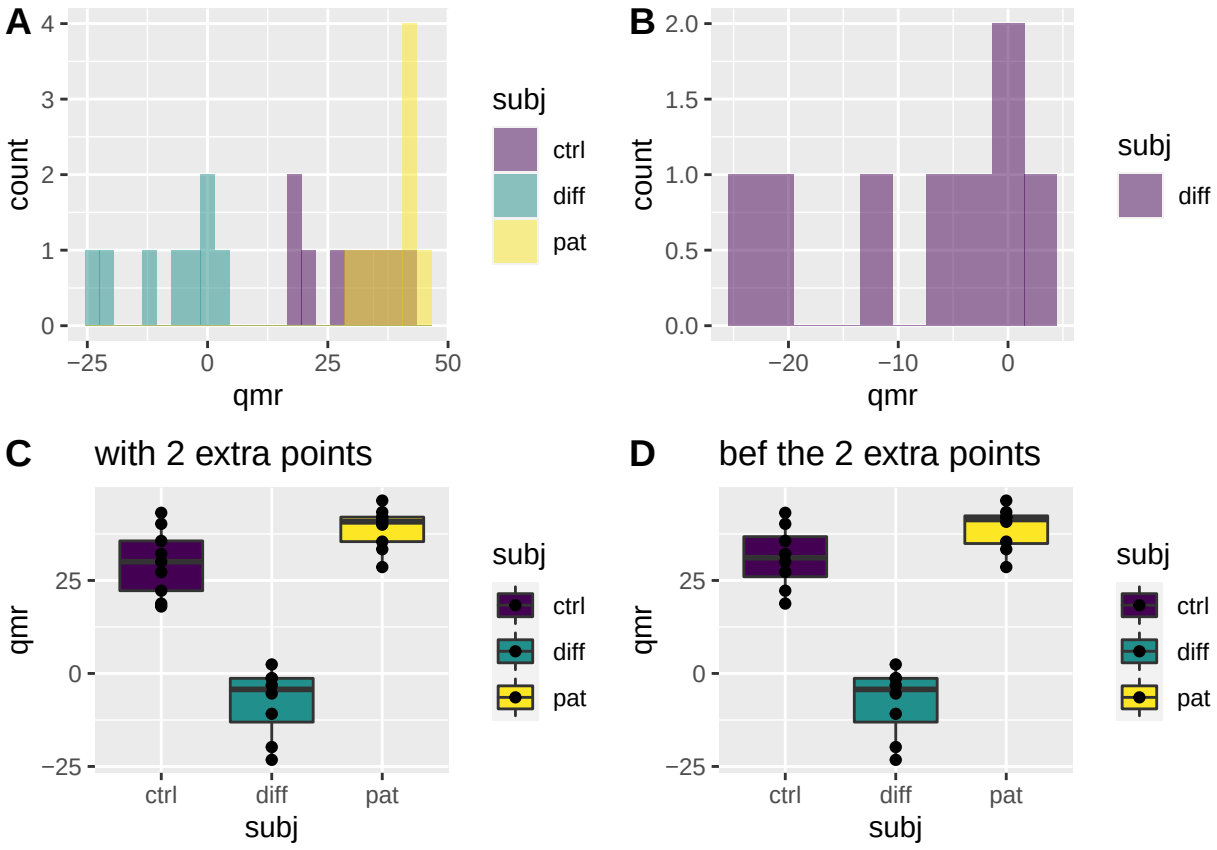

```
##
##  Welch Two Sample t-test
##
## data:  d.exp[d.exp$subj == "pat", ]$qmr and d.exp[d.exp$subj == "ctrl", ]$qmr
## t = 0.14116, df = 1996.9, p-value = 0.8878
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8210709  0.9484353
## sample estimates:
## mean of x mean of y
##  24.80365  24.73997
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  d.exp[d.exp$subj == "pat", ]$qmr and d.exp[d.exp$subj == "ctrl", ]$qmr
## W = 598415, p-value = 2.512e-14
## alternative hypothesis: true location shift is not equal to 0
```

Let's load an example of these distributions, and experiment with taking samples **nr__s** and observe the effect.



Trap 3: Just one more experiment then!It almost looks good. Let's repeat it. Trap of adding data on marginal distributions.

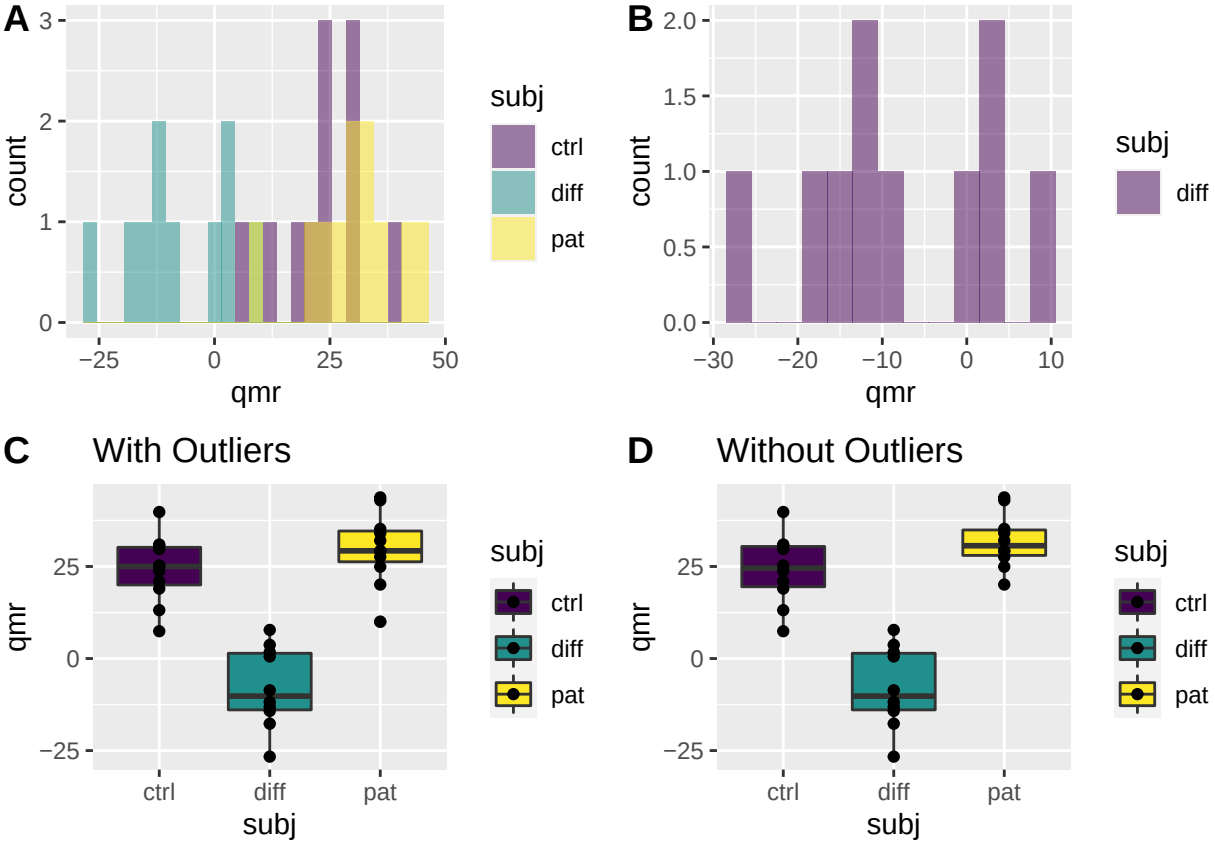


```
## [1] "p-value with 2 extra points" "0.0172692270370792"
```

```
## [1] "p-value without points" "0.0501597788149334"
```

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




```
## [1] "p-value wo Outliers:" "0.0544741516793329"   "0.163547659953177"
```

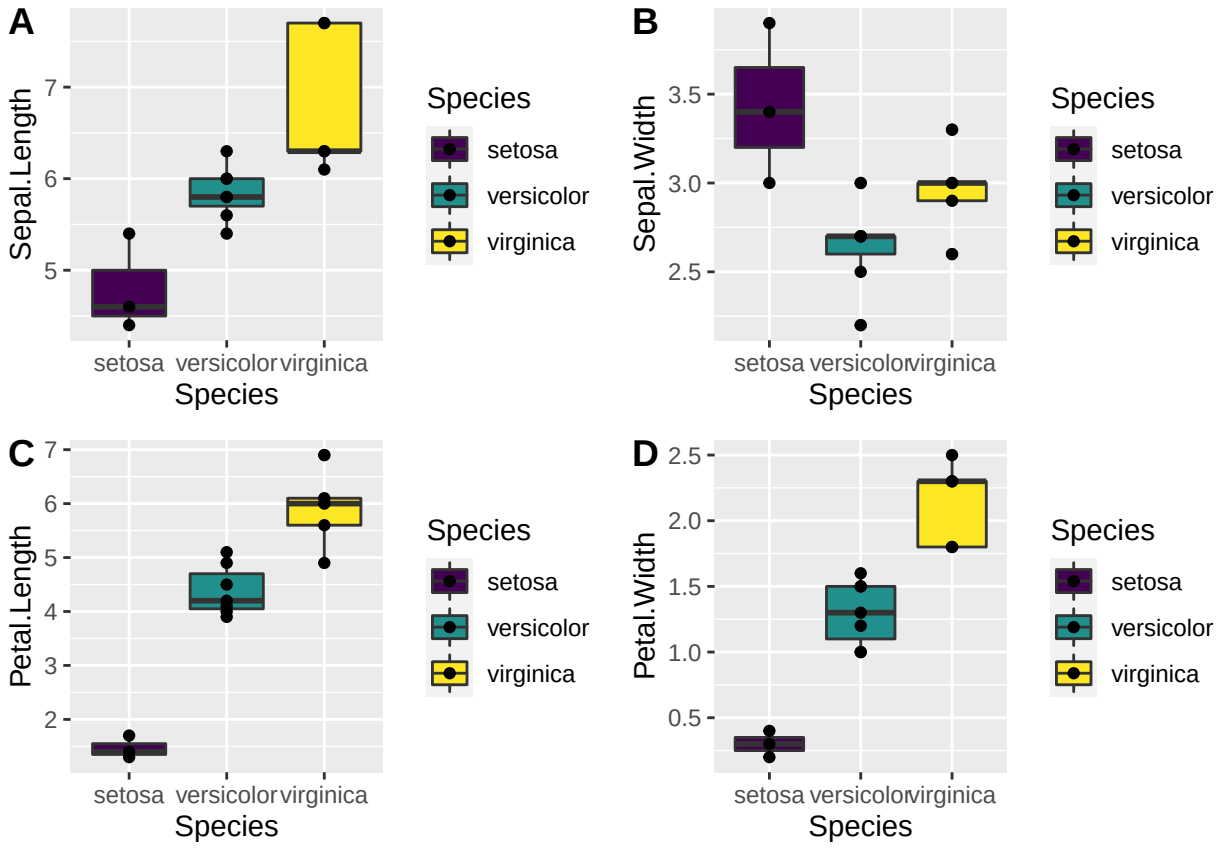
```
## [1] "p-value with Outliers:" "0.163547659953177"
```

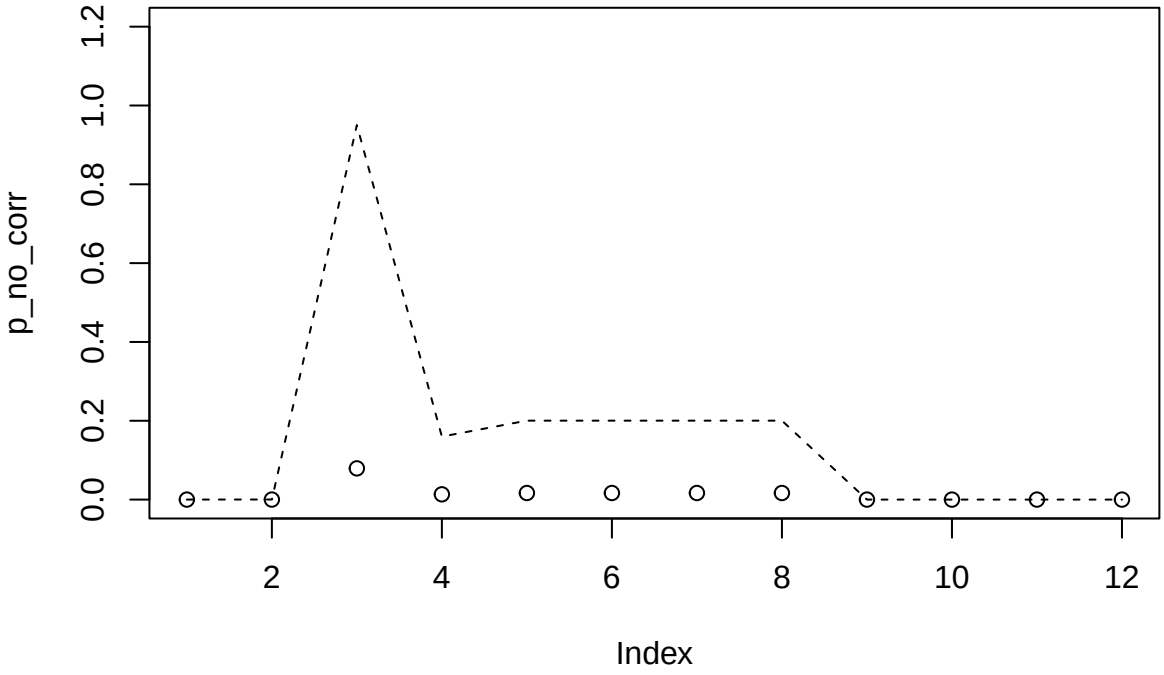
Trap Nr. 4: Problem of repeated sequential testing

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
##
##
##
```

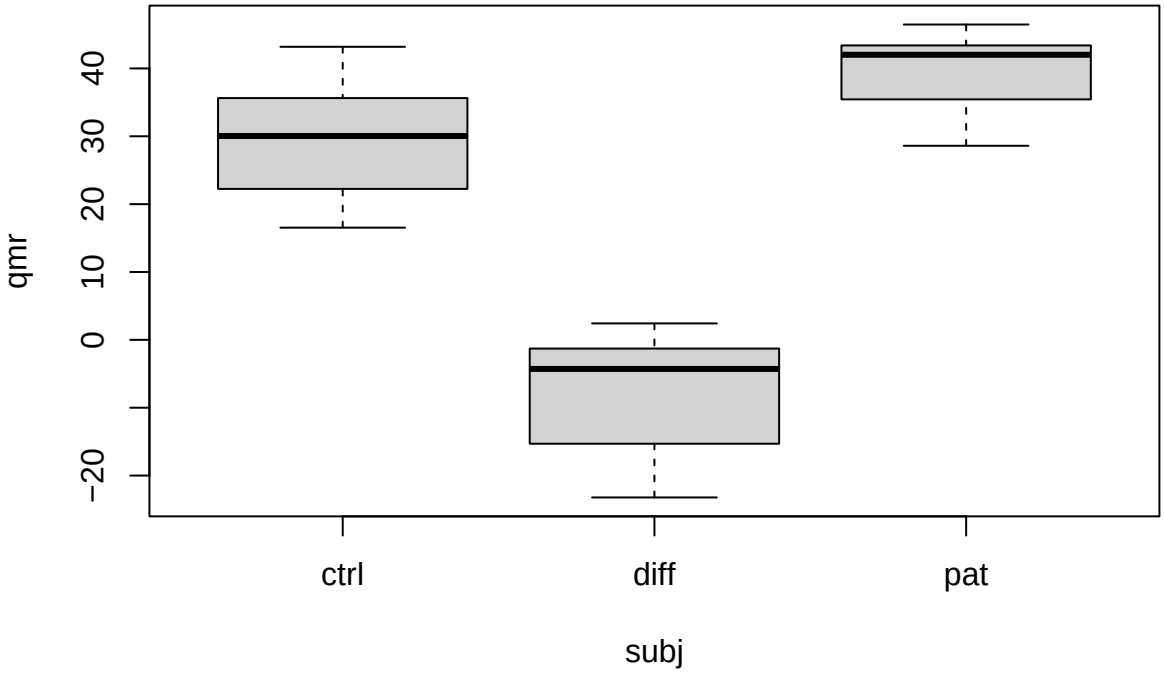




Trap Nr. 5: POST HOC hypothesis

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).