universität
innsbruck

Methods of Empirical Finance

Seminar (UE)

Christoph Huber
University of Innsbruck

Master in Banking and Finance
Winter term 2019/20 (this version: 2019-11-19)

Hypothesis Testing

Methods of Empirical Finance

Hypothesis Testing

Things to consider:

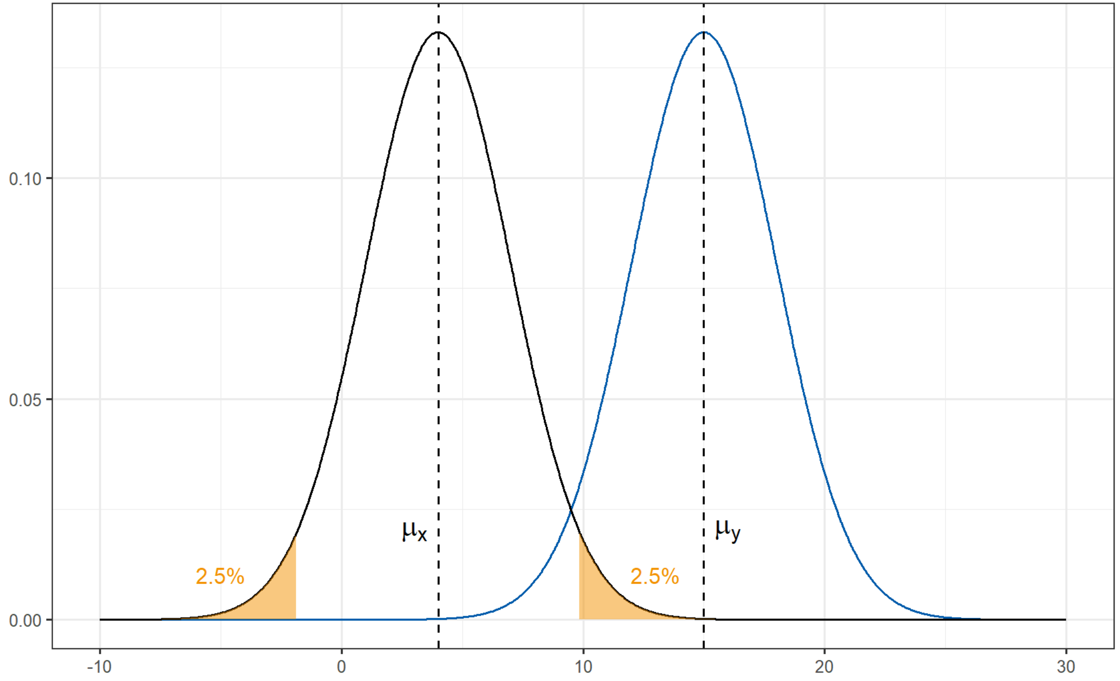
- Null Hypothesis (H_0) vs. Alternative Hypothesis (H_1, H_A)
- Exploratory vs. Directional Hypotheses
    - Exploratory: e.g. H_1: $\mu_x \neq \mu_y$
    - Directional: e.g. H_1: $\mu_x > \mu_y$
- Type I and Type II errors
- Statistical significance and power

Hypothesis Testing

Review

Exploratory Hypothesis: H_1: $\mu_x \neq \mu_y$

$\alpha = 0.05$

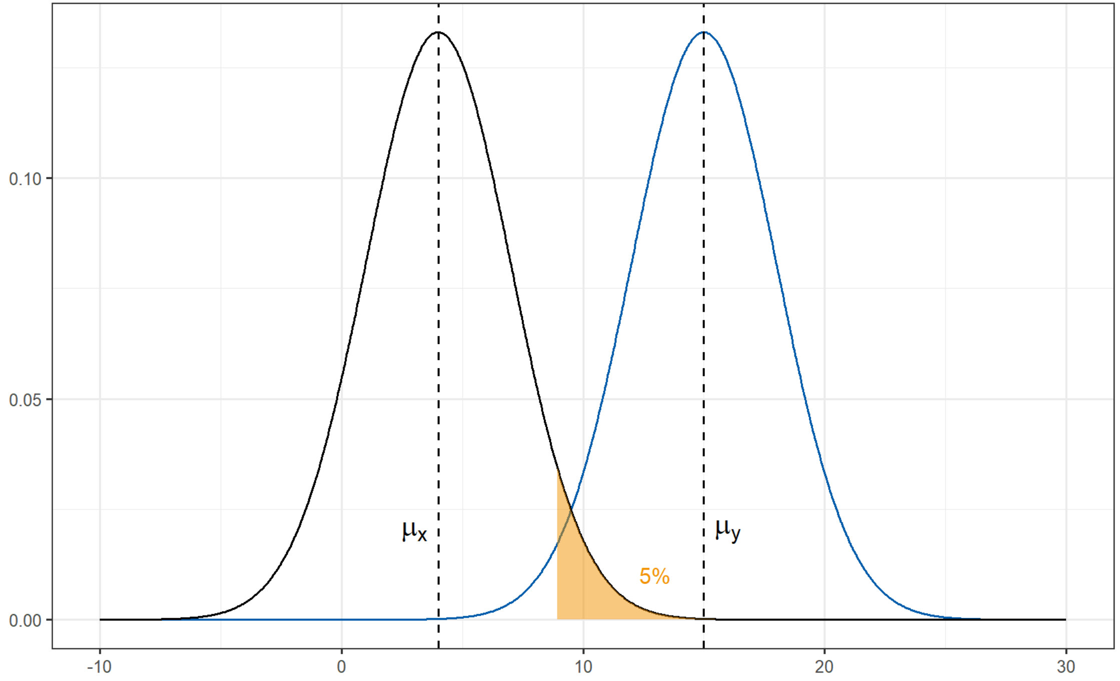


Hypothesis Testing

Review

Directional Hypothesis: H_1: $\mu_x < \mu_y$

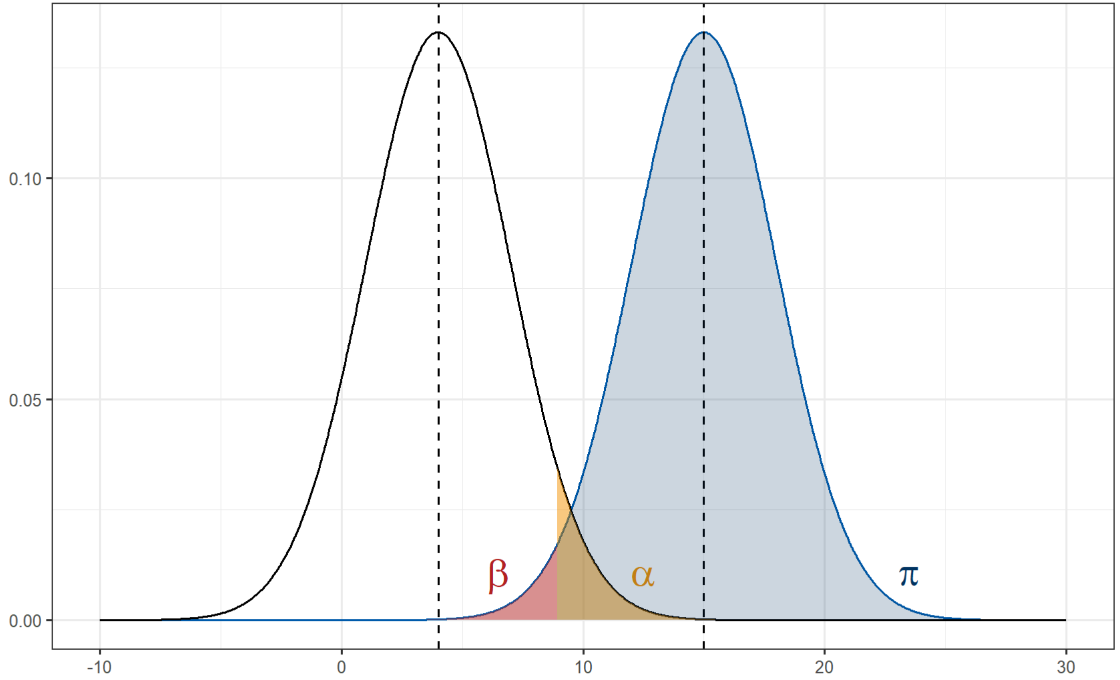
$\alpha = 0.05$



Hypothesis Testing

Review

α, β, and π



Hypothesis Testing

Review

Type I and Type II errors

		Null hypothesis H_0 is …	
		True	*False*
Decision about H_0 …	*Don't reject*	Correct inference	Type II error (false negative) β
	Reject	Type I error (false positive) α	Correct inference

Power Analysis

A test's statistical power depends on the following factors:

- the significance criterion ($\alpha \uparrow \rightarrow \pi \uparrow$)
- the sample size used to detect the effect ($n \uparrow \rightarrow \pi \uparrow$)
- the magnitude of the effect, i.e. the effect size ($d \uparrow \rightarrow \pi \uparrow$)

Power Analysis

Several scenarios:

- Calculate the required *sample size* for a given power:
    - given: power, effect size (Cohen's d), significance level (α), type of test
- Calculate the effect size you are able to detect for a given power and sample size:
    - given: power, number of observations (sample size n), significance level (α), type of test

Be careful: true power $\neq$ observed power
$\rightarrow$ do not use post-hoc power analyses!

Power Analysis

Software packages

R:

```r
require("pwr")
library(pwr)
```

`power.*` or `pwr.*` commands

Stata:

`power` command (see `help power`)

Power Analysis

Example 1

Kirchler, M., Palan, S. (2018) Immaterial and monetary gifts in economic transactions: evidence from the field. Experimental Economics 21. Link

From Table 1, we see:

$\bar{x}_{NORMAL} = 106.03, s_{NORMAL} = 18.78, n_{NORMAL} = 36 \; \bar{x}_{COMPLIMENT} = 117.20, s_{COMPLIMENT} = 20.55, n_{COMPLIMENT} = 36$

Do not use this data for post-hoc power analysis (see here), but, e.g. for *calculating the required sample size* or *calculating the targeted power* for a new analysis.

Power Analysis

Example 1

Calculate d: $d = \frac{\bar{x}_{NORMAL} - \bar{x}_{COMPLIMENT}}{\sqrt{s^2_{NORMAL} + s^2_{COMPLIMENT}/2}}$

```
d = (106.03 - 117.20)/(((18.78^2 + 20.55^2)/2)^(1/2))
print(d)
```

```
## [1] -0.5674399
```

Calculate power for $n = 80$:

```
pwr.t.test(n = 80, d = -0.5674399, sig.level = 0.05)
```

```
##
##      Two-sample t test power calculation
##
##               n = 80
##               d = 0.5674399
##       sig.level = 0.05
##           power = 0.9459666
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Power Analysis

Example 2

What sample size do you need to detect a medium effect size of $d = 0.5$ with 90% power if you use $\alpha = 0.05$?

(using a two-sided t-test)

```
pwr.t.test(d = 0.5, sig.level = 0.05, power = 0.9, type = c("two.sample"))
```

```
##
##      Two-sample t test power calculation
##
##               n = 85.03128
##               d = 0.5
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Power Analysis

What's an appropriate power?

- most fields use $\pi = 0.80$ as a the conventional/standard power
- this implies a probability of a Type II error (false negative) of $\beta = 1 - \pi = 0.2$ and therefore, with a conventional $\alpha = 0.05$ a 4-to-1 trade-off between β- and α-risk
- however, depending on the field and research question, this might be inappropriate $\rightarrow$ depending on field and research question, think about: *what is more important - avoiding false positives or false negatives?*

p-values

Suppose you have an hypothesis that U.S. public companies with small boards of directors outperform companies with large boards. You create two value-weighted portfolios and test for differences in mean returns. The key parameter of interest, the mean performance difference, is significant with $p = 0.01$.

Consider the following six statements (true/false):

- (i) You have disproved the null hypothesis (no difference in mean performance).
- (ii) You have found the probability of the null hypothesis being true.
- (iii) You have proved the hypothesis that firms with small boards outperform firms with large boards.
- (iv) You can deduce the probability of your hypothesis (small better than large) being true.
- (v) If you reject the null hypothesis (no difference), you know the probability that you are making a mistake.
- (vi) You have a reliable finding in the sense that if, hypothetically, the experiment were repeated a large number of times, you would obtain a significant result 99% of the time.

All of them are false!

p-values

Definition

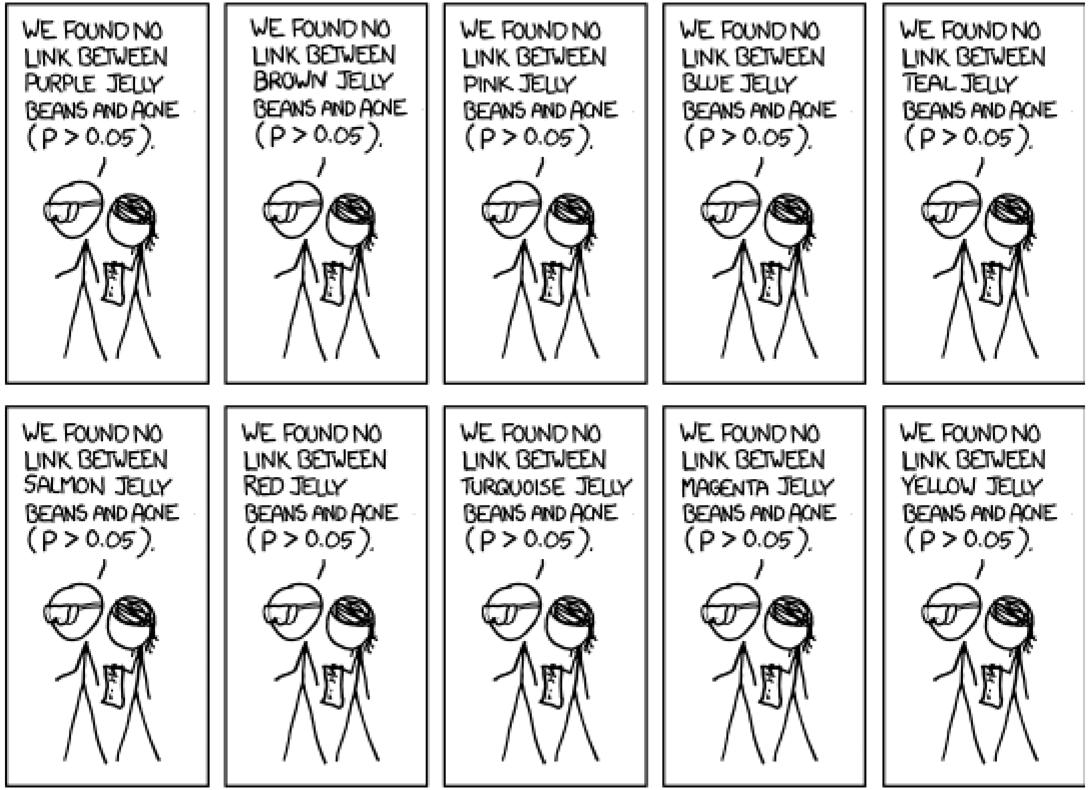
> A p-value is the probability of the observed data (or of more extreme data points), *given that the null hypothesis H_0 is true*.
>
> The p-value indicates the probability of observing an effect, D, (or greater) given the null hypothesis H_0 is true, that is,

$$p(D|H_0)$$

not $p(H_0|D)$!

Multiple Hypothesis Testing

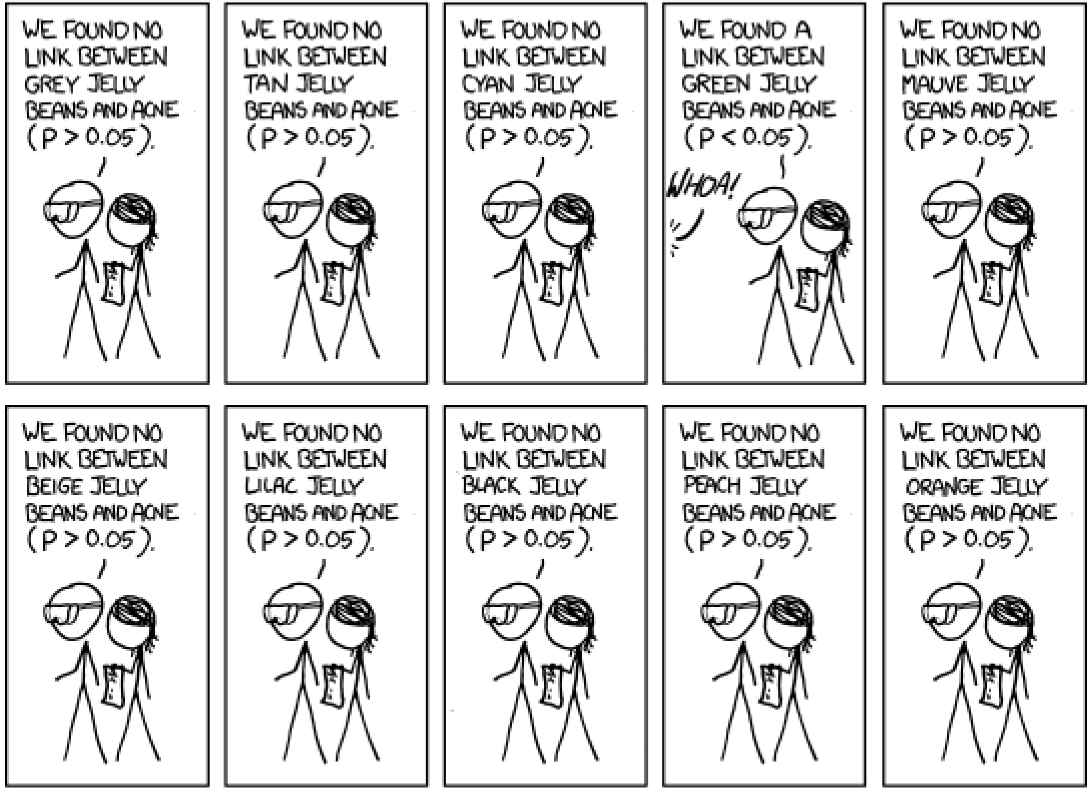
A comical illustration



From https://xkcd.com/882/

Multiple Hypothesis Testing

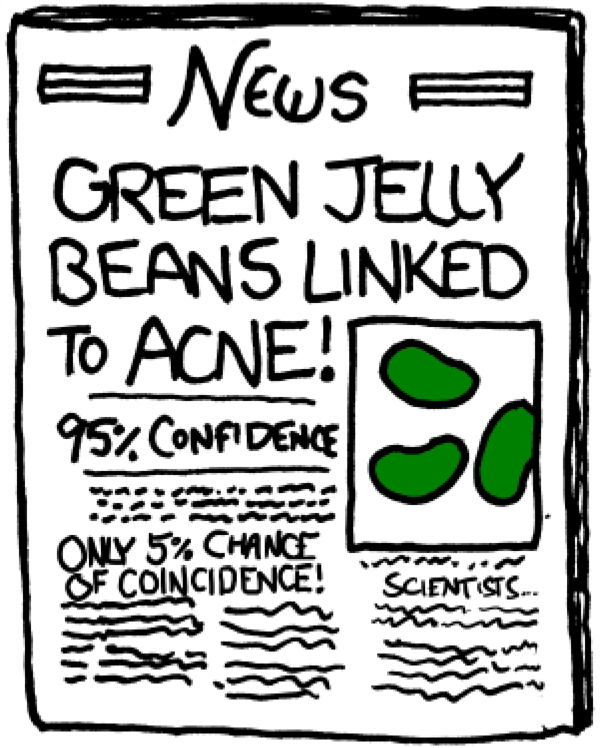
A comical illustration



From https://xkcd.com/882/

Multiple Hypothesis Testing

A comical illustration



From https://xkcd.com/882/

Multiple Hypothesis Testing

$\rightarrow$ testing multiple hypotheses simultaneously can be a problem

$Pr(\text{making a type I error}) = \alpha$

$Pr(\text{not making a type I error}) = 1 - \alpha$

$Pr(\text{not making a type I error } \mathbf{in\ m\ tests}) = (1 - \alpha)^{\mathbf{m}}$

Thus, with $\alpha = 0.05$ and 20 hypotheses, we get a *family-wise error rate (FWER)*

$FWER = 1 - (1 - 0.05)^{20} = 0.642$,

i.e. the probability of at least 1 type I error is 64.2%!

Multiple Hypothesis Testing

Definitions

Family-wise error rate (FWER)

> *probability* of false discoveries …

False-discovery rate (FDR)

> *proportion* of false discoveries …

Multiple Hypothesis Testing

Possible solutions: p-value corrections

- Bonferroni correction
  - controls FWER
  - set significance cut-off at α/m (m … number of tests)
  - applied to example above: $\alpha/m = \frac{0.05}{20} = 0.0025$, $FWER = 1 - (1 - 0.0025)^{20} = 0.049$
  - very conservative

```
pvalues <- runif(20)/10    # 20 random p-values
sort(pvalues)
```

```
##  [1] 0.0008945796 0.0073144469 0.0100053522 0.0260427771 0.0277374958
##  [6] 0.0286000621 0.0293739612 0.0415607119 0.0455102418 0.0583987980
## [11] 0.0585800305 0.0724405893 0.0754675027 0.0761702403 0.0813574215
## [16] 0.0906092151 0.0949040221 0.0954068775 0.0962204624 0.0971055656
```

```
p.adjust(sort(pvalues), method = "bonferroni")
```

```
##  [1] 0.01789159 0.14628894 0.20010704 0.52085554 0.55474992 0.57200124
##  [7] 0.58747922 0.83121424 0.91020484 1.00000000 1.00000000 1.00000000
## [13] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
## [19] 1.00000000 1.00000000
```

Multiple Hypothesis Testing

Possible solutions: p-value corrections

- Bonferroni correction
- Holm–Bonferroni correction
  - controls FWER
  - order the p-values from lowest to highest: $p_1 \leq p_2 \leq \ldots \leq p_k$
  - set significance cut-off at $\frac{\alpha}{m+1-k}$ (with k being the rank of the respective p-value)
  - applied to example above:

$\frac{\alpha}{m+1-k} = \frac{0.05}{20+1-1} = 0.0025$ for the smallest p ($k = 1$),
$\frac{\alpha}{m+1-k} = \frac{0.05}{20+1-2} = 0.00263$ for the second-smallest p ($k = 2$), and so on until
$\frac{\alpha}{m+1-k} = \frac{0.05}{20+1-20} = 0.05$ for the largest p ($k = 20$)

```r
p.adjust(sort(pvalues), method = "holm")
```

```
##  [1] 0.01789159 0.13897449 0.18009634 0.44272721 0.44379993 0.44379993
##  [7] 0.44379993 0.54028925 0.54612290 0.64238678 0.64238678 0.65196530
## [13] 0.65196530 0.65196530 0.65196530 0.65196530 0.65196530 0.65196530
## [19] 0.65196530 0.65196530
```

Multiple Hypothesis Testing

Possible solutions: p-value corrections

- Bonferroni correction
- Holm–Bonferroni correction
- Benjamini, Hochberg, and Yekutieli correction
  - controls FDR
  - order the p-values from lowest to highest: $p_1 \leq p_2 \leq \ldots \leq p_k$
  - starting from k, identify the first i such that $p_i < \frac{i}{k}\alpha$
  - declare all tests $1, \ldots, i$ significant, tests $i+1, \ldots, k$ not significant

```
p.adjust(sort(pvalues), method = "BH")
```

```
##  [1] 0.01789159 0.06670235 0.06670235 0.08392560 0.08392560 0.08392560
##  [7] 0.08392560 0.09710557 0.09710557 0.09710557 0.09710557 0.09710557
## [13] 0.09710557 0.09710557 0.09710557 0.09710557 0.09710557 0.09710557
## [19] 0.09710557 0.09710557
```

- Several other correction procedures

Related Literature

- Hoenig, J. M., & Heisey, D. M. (2001). The Abuse of Power. The American Statistician, 55(1), 19–24.
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. Nature Human Behaviour, 2(1), 6–10.
- Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the Cross-Section of Expected Returns. Review of Financial Studies, 29(1), 5–68.
- Harvey, C. R. (2017). The Scientific Outlook in Financial Economics. The Journal of Finance, 72(4), 1399–1440.

Your turn

- Load the data in `sp500_data.csv` again
- Think about interesting questions and formulate hypotheses you could test with these data
- Test your hypothesis(-es) using appropriate tests
- Can you reject the H_0?
- Comment on upcoming issues regarding the statistical power, multiple hyptheses, etc.