Hypothesis, Research Design, Correlation

Hypotheses

Are assumptions about structural properties of reality generally valid: they go beyond an individual situation or event must be falsifiable => they can be disproved

Good example: The monetary return to education is higher in natural sciences than in arts.

Bad example: COVID-19 infections can improve health in the long-term.

t-tests

Often, we would like to make statements on equality of groups, values etc. e.g. gpa's of students at the TUM and the LMU are different across study programmes.

So calculate differences of gpa's for each study programme between TU and LMU with Nullhypothesis (H0): mean of gpa's at TUM = mean of gpa's at LMU and Alternative hypothesis (H1): mean of gpa's at TUM differ from mean of gpa's at LMU

Calculate the test statistic: t = (Xbar-mu)/(s/sqrt(n)) where:

t = test statistics

Xbar = sample mean

mu = mean of the population under the Nullhypothesis

s = standard deviation of the SAMPLE

n = number of observations

Obtain critical value from t-table based on a level of significance (alpha).

Make a decision to reject H0 or not based on whether the test statistic exceeds the critical value.

We will never accept a hypothesis but only reject a hypothesis!

Research and survey design

Populations are usually (too) large. Therefore, we select a sample that we would like to study. There are different selection methods which require us to adjust our analysis accordingly.

Central limit theorem: for sample sizes above 30, the mean of the sample is approximately normally distributed.

Let us check whether gpa’s at TUM and LMU are different from each other:

We first import a take a look of the relevant data

```
library(readxl)
gpadat <- read_excel("D:/data/Empirical Research/gpadat.xlsx")

gpadat

## # A tibble: 37 × 3
##    `Study programme`      TUMgpa LMUgpa
##    <chr>                   <dbl>  <dbl>
##  1 Aerospace engineering     3.4    2.4
##  2 Agricultural sciences     2.8    1.6
##  3 Biology                   2.4    2.6
##  4 Architecture              1.4    2.2
##  5 Automotive engineering    2      2.8
##  6 Chemistry                 2.8    2.2
##  7 Mathematics               2.2    1.8
##  8 Physics                   2.6    2.2
##  9 Medicine                  1.4    1.8
## 10 Economics                 1.8    1.6
## # ℹ 27 more rows
```

Then we create a variable containing the differences between gpa’s

```
 gpadat$gpadif = gpadat$TUMgpa-gpadat$LMUgpa
 gpadat

## # A tibble: 37 × 4
##    `Study programme`      TUMgpa LMUgpa gpadif
##    <chr>                   <dbl>  <dbl>  <dbl>
##  1 Aerospace engineering     3.4    2.4  1
##  2 Agricultural sciences     2.8    1.6  1.2
##  3 Biology                   2.4    2.6 -0.200
##  4 Architecture              1.4    2.2 -0.8
##  5 Automotive engineering    2      2.8 -0.8
##  6 Chemistry                 2.8    2.2  0.600
##  7 Mathematics               2.2    1.8  0.4
##  8 Physics                   2.6    2.2  0.4
##  9 Medicine                  1.4    1.8 -0.4
## 10 Economics                 1.8    1.6  0.2
## # ℹ 27 more rows
```

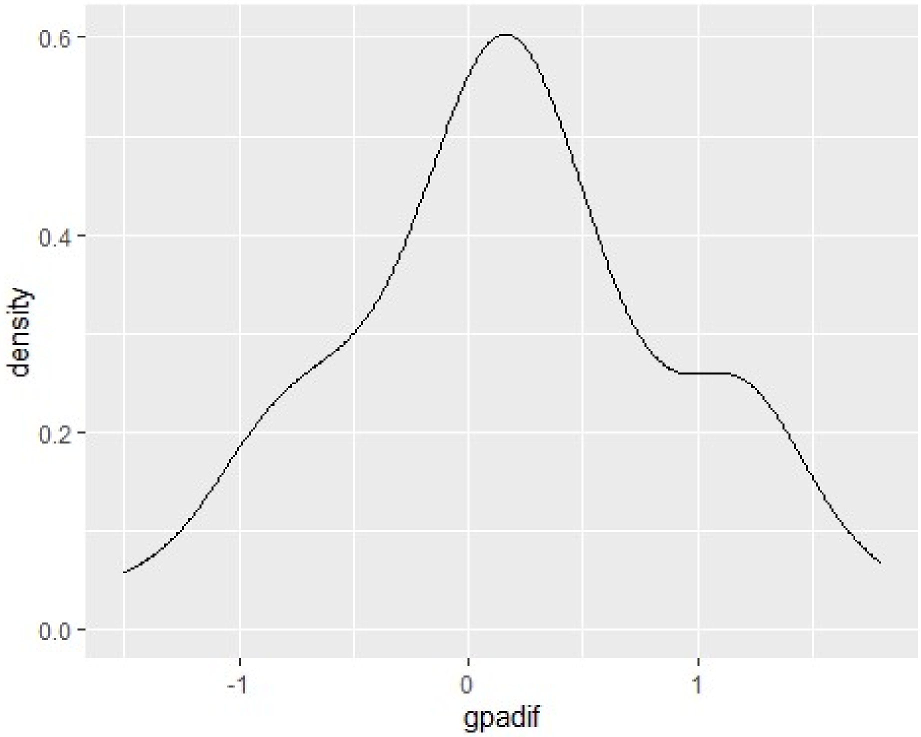
We can first investigate this the distribution of the difference in gpa’s grafically .

```
library(psych)
library(ggplot2)
describe(gpadat$gpadif)

##    vars  n mean   sd median trimmed  mad  min max range skew
kurtosis   se
```

```
## X1     1 37 0.18 0.74     0.2     0.18 0.59 -1.5 1.8    3.3     0     -
0.45 0.12

densip = ggplot() + geom_density(data=gpadat,aes(x=gpadif))
densip
```



The empirical distribution seems symmetric and centered around 0. So we can conclude quite safely that there is no significant difference. However we must verify this by performing a t-test. So,

we first calculate the t-statistic

```
library(gdata)
t = (mean(gpadat$gpadif)-
0)/(sd(gpadat$gpadif)/(nobs(gpadat$gpadif)^0.5))
```

and then we calculate the critical values for a series of significance levels (note that we have to half the level of significance since R assumes that we carry out a one-sided test by default!)

```
tcrit001 = qt(0.0005,nobs(gpadat$gpadif)-1, lower.tail=FALSE)
tcrit010 = qt(0.005,nobs(gpadat$gpadif)-1, lower.tail=FALSE)
tcrit050 = qt(0.025,nobs(gpadat$gpadif)-1, lower.tail=FALSE)
tcrit100 = qt(0.05,nobs(gpadat$gpadif)-1, lower.tail=FALSE)
```

H0: The mean difference in gpa's is equal to zero.

H1: The mean difference in gpa's is not equal to zero.

Decision Rule: Reject H0 at the level α% if the t-statistic exceeds the critical value of α% significance.

```
t

## [1] 1.458546

tcrit001

## [1] 3.58215

tcrit010

## [1] 2.719485

tcrit050

## [1] 2.028094

tcrit100

## [1] 1.688298
```

Hence, we cannot reject H0 and do not find evidence for a difference in mean gpa's on any reasonable level of significance.

Equivalently we could use an R command for the t-test, which verifies our manual result

```
 t.test(gpadat$gpadif, mu=0)

## 
##  One Sample t-test
## 
## data:  gpadat$gpadif
## t = 1.4585, df = 36, p-value = 0.1534
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.06965498  0.42641174
## sample estimates:
## mean of x 
## 0.1783784
```

Correlation

Oftentimes we are interested in analyzing the relationship of at least two variables.

Correlation analysis is a common tool for this. e.g. What is the correlation between the average daily temperature and the ice cream #sales in a supermarket?

There are different correlation coefficients (e.g. Spearman and Bravais-Pearson). The choice mostly depends on the type of data the you have. The most important/frequent are the Spearman rank correlation coefficient and the Bravais-pearson correlation coefficient.

The Bravais-Pearson correlation coefficient for two variables X and Y is calculated by dividing the covariance of X and Y by the product of the standard deviations of X and Y:

BPC = cov(X,Y)/sd(X)*sd(Y) where

cov = covariance and sd = standard deviation

It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation).

The Spearman rank correlation coefficient is calculated similarly but ranks are used.

For interval scaled data, use B-P but… in the presence of outliers … non-normally distributed data … highly non-linear relationships use the Spearman correlation coefficient.

Does renewable electricity production leads to increased price volatility on the electricity spot market?

or in other words

Is there a significant correlation between solar power production and the coefficient of variation of the electricty spot market price?

Let us first import the relevant data from the csv file "solardat.csv"

```
library(readr)
solardat <- read.csv("D:/data/Empirical Research/solardat.csv",sep=";")
head(solardat)

##         Date  CV_daily PV_daily_MWh
## 1 06.01.2015 0.2706780     32887.25
## 2 07.01.2015 0.4032656     17114.75
## 3 08.01.2015 0.3745514      8598.25
## 4 09.01.2015 0.4617914      6823.75
## 5 10.01.2015 0.3316949     20475.00
## 6 11.01.2015 4.5214577     19811.25
```

We start with the Bravais-Pearson correlation. We calculate the covariance of cv and solar power production

```
covcvs = cov(solardat$CV_daily,solardat$PV_daily_MWh)
covcvs

## [1] -2544.674
```

and the standard deviations of each variable

```
sdcv = sd(solardat$CV_daily)
sdcv
```

```
## [1] 0.9396159

sds = sd(solardat$PV_daily_MWh)
sds

## [1] 64764.98
```

So the Bravais-Pearson correlation coefficient becomes

```
bpman = covcvs/(sdcv*sds)
bpman

## [1] -0.04181592
```

But is this significant? Let's use a t-test to check it. To do so we walculate test statistic

```
library(gdata)
tbpcor = (bpman/((1-bpman^2)^0.5))*(nobs(solardat$CV_daily)-2)^0.5
tbpcor

## [1] -1.509595
```

as well as the critical values for various significance levels(we have to half the level of significance since R assumes that we carry out a one-sided test by default!)

```
tcritbp001 = qt(0.0005,nobs(solardat$CV_daily)-2, lower.tail=TRUE)
tcritbp010 = qt(0.005,nobs(solardat$CV_daily)-2, lower.tail=TRUE)
tcritbp050 = qt(0.025,nobs(solardat$CV_daily)-2, lower.tail=TRUE)
tcritbp100 = qt(0.05,nobs(solardat$CV_daily)-2, lower.tail=TRUE)

tcritbp001

## [1] -3.298021

tcritbp010

## [1] -2.579614

tcritbp050

## [1] -1.961789

tcritbp100

## [1] -1.646026
```

H0: The bpman equal to zero.

H1: The bpman is not equal to zero.

Reject H0 at the α% significance level if the t-statistic exceeds the the respective critical value.

As tbpcor does not exceed any of them even at the lowest 90% level of significance we fail to reject H0 and conclude that the correlation is zero at the 90% significance level.

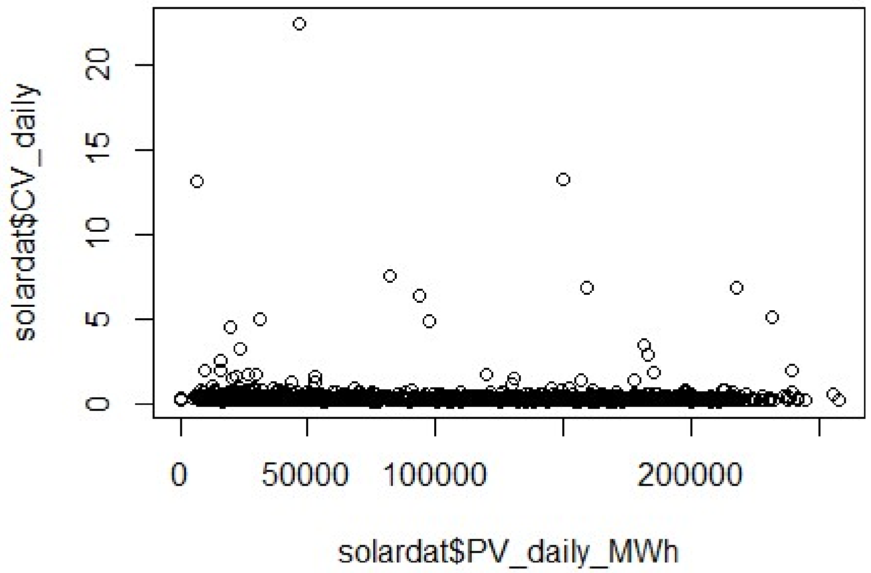
Again, R offers the opportunity to avoid doing the correlation analysis manually

```
library(Hmisc)
rcorr(as.matrix(solardat[,2:3]))
```

```
##              CV_daily PV_daily_MWh
## CV_daily         1.00        -0.04
## PV_daily_MWh    -0.04         1.00
##
## n= 1303
##
##
## P
##              CV_daily PV_daily_MWh
## CV_daily              0.1314
## PV_daily_MWh 0.1314
```

However,

```
plot(solardat$PV_daily_MWh,solardat$CV_daily)
```



Now, let us use the Spearman-rank correlation

```
rcorr(as.matrix(solardat[,2:3]), type = c("spearman"))
```

```
##              CV_daily PV_daily_MWh
## CV_daily         1.00        -0.26
## PV_daily_MWh    -0.26         1.00
##
## n= 1303
##
##
## P
##              CV_daily PV_daily_MWh
## CV_daily               0
## PV_daily_MWh  0
```