Numeric Data

Mandy Vogel

June 14, 2015

Table of Contents

Numeric summaries

To describe data we need a proper way to summarize them for easier understanding. Therefore we focus on three main areas:

- parameters of location (today)
- spread (today) and
- shape (later)

Exercises

1. load the data *ZA5240_v2-0-0.sav* from the data directory using the `spss.get()` function from the Hmisc package (data source: gesis.org, General Social Survey 2014), assign the data set to a variable with a appropriate name.
2. the file *variablenliste.txt* contains a list with the variable description (only available in German)
3. how many rows? (`nrow()`)
4. how many columns? (`ncol()`)

Location parameters

- a location parameter is a central or typical value for a distribution
- typical location parameters are:
  - mean (`mean()`)
  - trimmed means (`mean()`)
  - median (`median()`)
  - mode

Mean and median I

How to interpret the mean?

- graphically, it is the visual balance point of the given values (physics formula for the center of mass)
- this demonstrates a weakness of the mean when used to represent *center*
- the *trimmed mean* tries to make the mean more stable by trimming from both sides a certain percentage of the most extreme values
- if mean and the trimmed mean are substantially different the data are very likely to be skewed
- the trimmed mean pushed to its limits (by trimming 50% of the data from each end) leaves us with basically a single value: the median
- so the median is the value which divides the values into the 50% lowest and the 50% highest values

Exercises

1. the column V417 contains the net income, calculate the mean using the `mean()` function! What is the problem?
2. again calculate the mean but now the trimmed version (using the `trim=T` argument). What is the conclusion?

Other measures of position

- the concept of the median can be generalized: as the median splits the data in half (with half the data smaller and the other half larger), the pth quantile is basically the value in the data set for which $100 \cdot p$ is less than the value and $100 \cdot (1 - p)$ is more (so the median is the 0.5 quantile); special cases of quantiles are percentiles, quartiles and quintiles (`quantiles()`)
- hinges (not often mentioned but used in boxplots; `fivenum()`)
- and of course min and max (`min()`, `max()`)

Exercises

1. summarize the net income using `summary()`, `quantile()` and `fivenum`!
2. make a boxplot by using the following syntax! (we also use column V81 - gender and V86 - graduation)

```
require(ggplot2)
ggplot(x, aes(x=V86, y=V417)) +
    geom_boxplot()
```

3. add gender as coloring

```
ggplot(x, aes(x=V86, y=V417, fill=V81)) +
    geom_boxplot()
```

So what now?

- so we se - there is a difference in income between male and female people
- the next step would be to test if this difference is statistically significant, but therefore we need also the measure of spread (even if you do not use it explicitly in testing) so there are

Spread I

- these are parameters which measure the variability in the data
- there is e.g. the range (`range()`)
- the sample variance (`var()`) and
- the sample standard deviation (`sd()`) which is simply the square root of the variance
- the coefficient of variation which is the standard deviation normalized by the mean
- the IQR (interquartile range; `IQR` - there are nine ways to calculate it - so different statistics software can have different IQRs - depending on the method)
- the mad (median absolute deviation)

four possible situations

		Situation	
		H_0 is true	H_0 is false
Conclusion	H_0 is not rejected	Correct decision	Type II error
	H_0 is rejected	Type I error	Correct decision

Common symbols

symbol	meaning

Common symbols

symbol	meaning
n	number of observations (sample size)
K	number of samples (each having n elements)

Common symbols

symbol	meaning
n	number of observations (sample size)
K	number of samples (each having n elements)
α	level of significance
ν	degrees of freedom
σ	standard deviation (population)
s	standard deviation (sample)

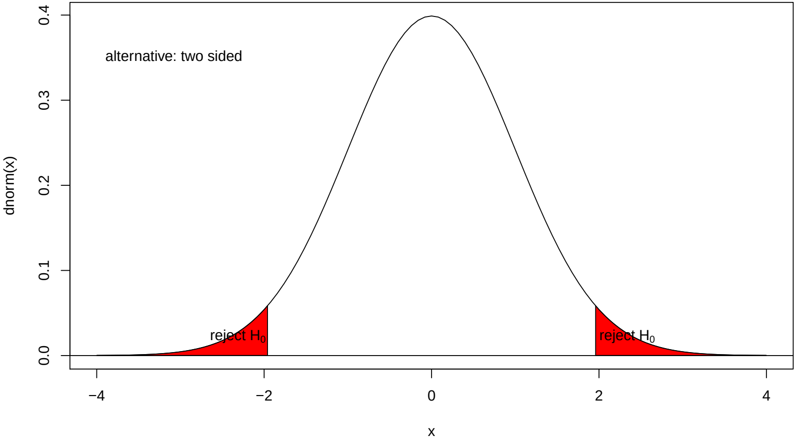
Common symbols

`symbol`	meaning
n	number of observations (sample size)
K	number of samples (each having n elements)
α	level of significance
ν	degrees of freedom
σ	standard deviation (population)
s	standard deviation (sample)
μ	population mean
$\bar{x}$	sample mean
ρ	population correlation coefficient
r	sample correlation coefficient

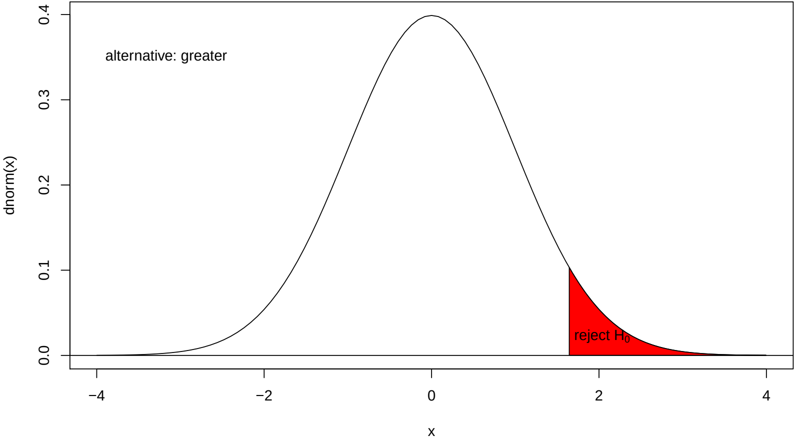
Common symbols

symbol	meaning
n	number of observations (sample size)
K	number of samples (each having n elements)
α	level of significance
ν	degrees of freedom
σ	standard deviation (population)
s	standard deviation (sample)
μ	population mean
$\bar{x}$	sample mean
ρ	population correlation coefficient
r	sample correlation coefficient
Z	standard normal deviate

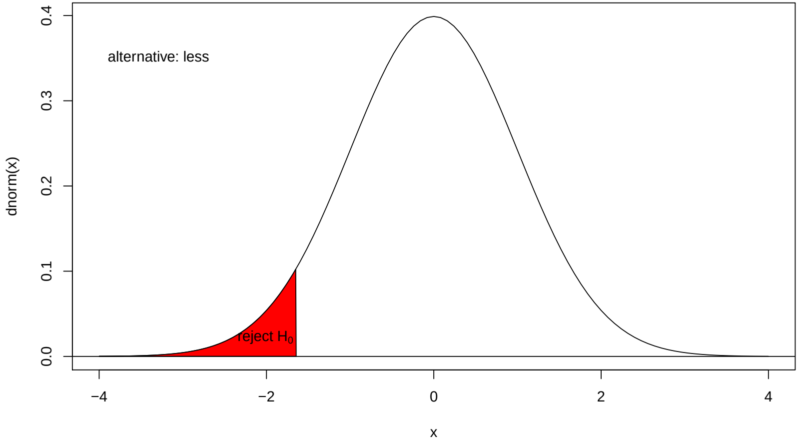
Alternatives



Alternatives



Alternatives



p-value

Note:

The p-value is the probability of the sample estimate (of the respective estimator) under the null. We do not know anything of probabilities of the null or of estimates under the alternative!

William Gosset

- the t-test is called t-test because its test statistic is t distributed (this distribution was discovered or invented or whatsoever by William Gosset)
- so we take our data
- calculate some mystic t-value
- compare this value with the distribution of t-values under the null
- and then we decide: is our t value so unlikely, that we can reject the null (wow! lower than 5% - is 5% really that low...??? - this cutoff level was suggested by Fisher in the 1920s, so it is not a part of the decalogue and - therefore - there is no guarantee linked to it)

t-test

There is not only one t-test, there is

- the one sample t-test
- the two sample t-test assuming equal variances
- the two sample t-test without the former assumption
- the paired t-test (in fact: this is a one sample t-test against 0)

But: there is only one command in R: `t.test()`

t-tests

- *one sample t-test*: test a sample mean against a population mean
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$
where $\bar{x}$ is the sample mean, s is the sample standard deviation and n is the sample size. The degrees of freedom used in this test is $n - 1$

One Sample t-test

```
> set.seed(1)
> x <- rnorm(12)
> t.test(x,mu=0) ## population mean 0

One Sample t-test

data:  x
t = 1.1478, df = 11, p-value = 0.2754
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.2464740  0.7837494
sample estimates:
mean of x
0.2686377
```

One Sample t-test

```
> t.test(x,mu=1) ## population mean 1

One Sample t-test

data:  x
t = -3.125, df = 11, p-value = 0.009664
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
 -0.2464740  0.7837494
sample estimates:
mean of x
0.2686377
```

Two Sample t-tests

There are two ways to perform a two sample t-test in R:

- given two vectors `x` and `y` containing the measurement values from the respective groups `t.test(x,y)`
- given one vector `x` containing all the measurement values and one vector `g` containing the group membership $t.test(x \sim g)$ (read: `x` dependend on `g`)

Two Sample t-tests: two vector syntax

```
> set.seed(1)
> x <- rnorm(12)
> y <- rnorm(12)
> g <- sample(c("A","B"),12,replace = T)
> t.test(x,y)

Welch Two Sample t-test

data:  x and y
t = 0.5939, df = 20.012, p-value = 0.5592
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5966988  1.0717822
sample estimates:
 mean of x  mean of y
0.26863768 0.03109602
```

Two Sample t-tests: formula syntax

```
> t.test(x ~ g)

Welch Two Sample t-test

data:  x by g
t = -0.6644, df = 6.352, p-value = 0.5298
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.6136329  0.9171702
sample estimates:
mean in group A mean in group B
      0.1235413       0.4717726
```

Welch/Satterthwaite vs. Student

- if not stated otherwise `t.test()` will not assume that the variances in the both groups are equal
- if one knows that both populations have the same variance set the `var.equal` argument to TRUE to perform a student's t-test

Student's t-test

```
> t.test(x, y, var.equal = T)

Two Sample t-test

data:  x and y
t = 0.5939, df = 22, p-value = 0.5586
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5918964  1.0669797
sample estimates:
 mean of x  mean of y
0.26863768 0.03109602
```

t-test

- the t-test, especially the Welch test is appropriate whenever the values are normally distributed
- it is also recommended for group sizes ≥ 30 (robust against deviation from normality)