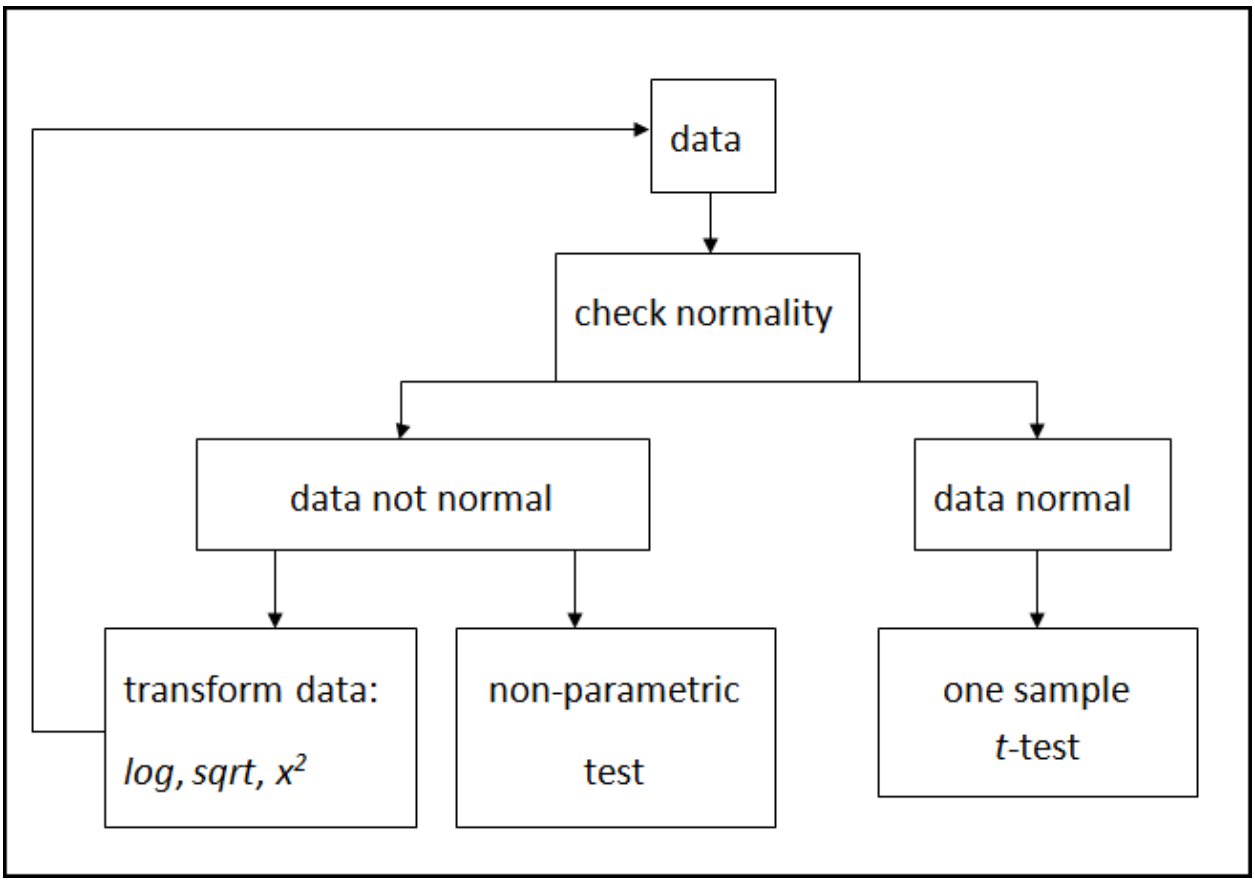
Chapter 8: Statistical inference for continuous data

Contents

1 One sample

The scheme of the analysis is as follows:

1. When sample size is large enough, we can use the CLT which assures that the average is normally distributed. In this situation, the one sample t-test can be used.
2. When sample size is small, then we have to use the scheme below. Check for normality (e.g. Shapiro-Wilk test). If normality is not rejected, we can use the one sample t-test. When normality is rejected, a non-parametric alternative or a transformation can be used.



We are using the data set **`normtemp`** from the package `UsingR`.

normtemp {UsingR} R Documentation

Body temperature and heart rate of 130 health individuals

Description

A data set used to investigate the claim that "normal" temperature is 98.6 degrees.

Usage

```
data(normtemp)
```

Format

A data frame with 130 observations on the following 3 variables.

temperature

normal body temperature

gender

Gender 1 = male, 2 = female

hr

Resting heart rate

```
head(normtemp)
```

```
##   temperature gender hr
## 1        96.3      1 70
## 2        96.7      1 71
## 3        96.9      1 74
## 4        97.0      1 80
## 5        97.1      1 73
## 6        97.1      1 75
```

We want to test the following hypothesis:

$H_0 : \mu_{temp} = 100$
$H_1 : \mu_{temp} \neq 100$

Descriptive statistics

```
library(rstatix)
summary_result <- normtemp %>%
  get_summary_stats(temperature, hr, show = c("n", "mean", "sd", "median"))
summary_result
```

```
## # A tibble: 2 x 5
##   variable        n  mean    sd median
```

```
##   <chr>       <dbl> <dbl> <dbl>  <dbl>
## 1 hr            130  73.8 7.06    74
## 2 temperature   130  98.2 0.733   98.3
```

1.1 One sample t-test

If the sample size is large enough, the CLT can be applied and one sample t-test can be used.

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

There are 3 options:

`conf.level`	Confidence level for CI (confidence interval)
`mu`	population mean under the null hypothesis
`alternative`	alternative hypothesis: can be `two-sided`, `greater`, `less`

The sample size of the data `normtemp` is large ($n = 130$), hence CLT can be applied and the one sample t-test can be used to test the earlier mentioned hypothesis.

```
temp <- normtemp$temperature
t.test(temp, mu = 100)
```

```
##
##  One Sample t-test
##
## data:  temp
## t = -27.226, df = 129, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 100
## 95 percent confidence interval:
##  98.12200 98.37646
## sample estimates:
## mean of x
##  98.24923
```

The output returns `p-value < 2.2e-16` hence the null hypothesis H_0 is rejected and the average temperature is significant different from 100.

1.2 Non-parametric test

In case CLT cannot be used or the normality is not fulfilled, a non-parametric test can be applied

```
wilcox.test(x, y = NULL,
            alternative = c("two.sided", "less", "greater"),
            mu = 0, paired = FALSE, exact = NULL, correct = TRUE,
            conf.int = FALSE, conf.level = 0.95,
            tol.root = 1e-4, digits.rank = Inf, ...)
```

By using the Wilcoxon signed rank test (for one sample) you can check if the distribution of x is symmetric around `mu`.

```
wilcox.test(temp, mu = 100)
```

```
##
```

```
##  Wilcoxon signed rank test with continuity correction
##
## data:  temp
## V = 8.5, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to 100
```

Remark:

1. You can also use the `t_test` in `rstatix`

```
# One sample t-test
normtemp %>% t_test(temperature ~ 1, mu = 100)
```

```
## # A tibble: 1 x 7
##   .y.         group1 group2         n statistic    df        p
## * <chr>       <chr>  <chr>      <int>     <dbl> <dbl>    <dbl>
## 1 temperature 1      null model   130     -27.2   129 2.54e-55
```

2. You can use the `wilcox_test` in `rstatix`

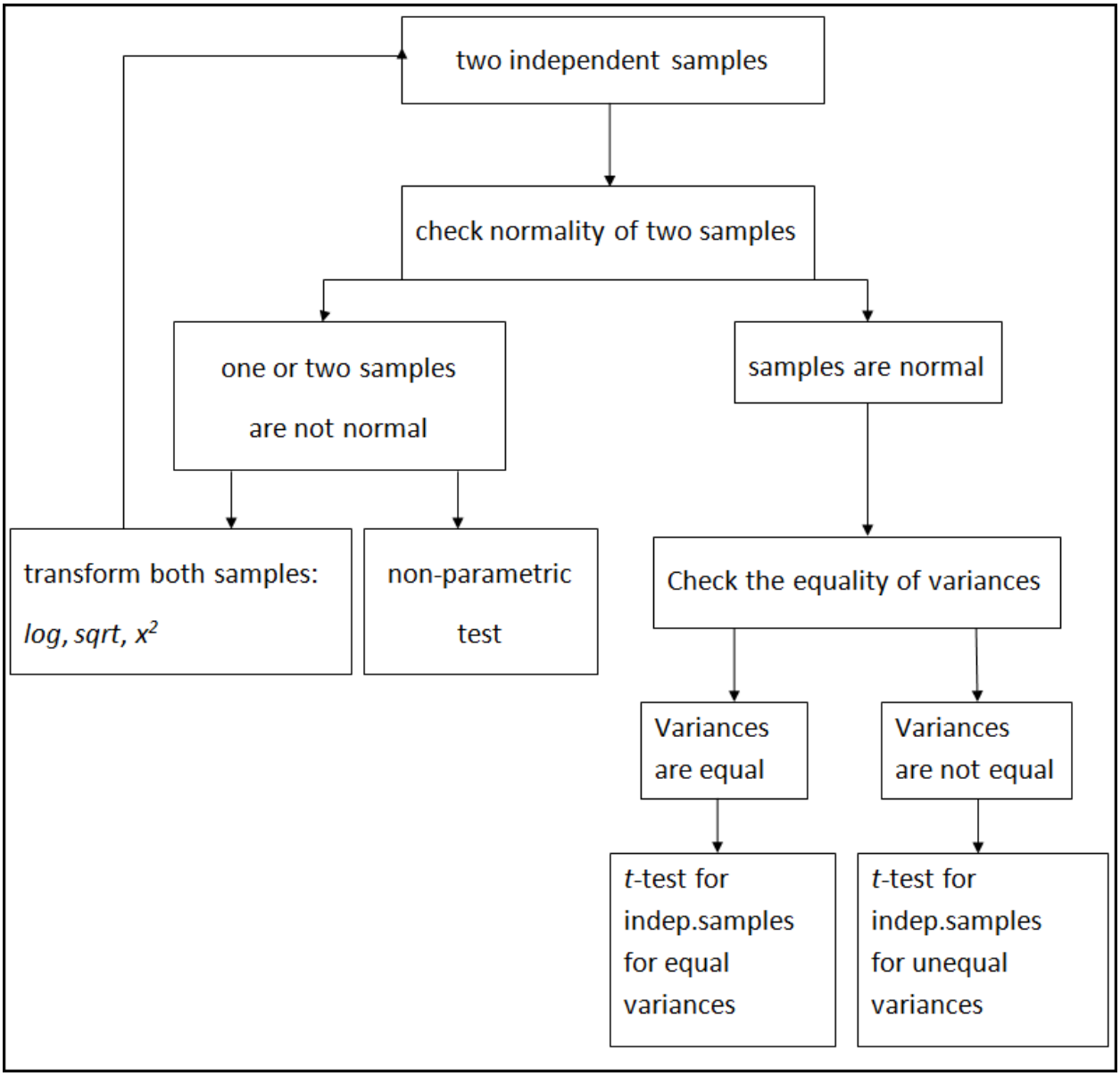
```
normtemp %>% wilcox_test(temperature ~ 1, mu = 100)
```

```
## # A tibble: 1 x 6
##   .y.         group1 group2         n statistic        p
## * <chr>       <chr>  <chr>      <int>     <dbl>    <dbl>
## 1 temperature 1      null model   130       8.5 7.74e-23
```

2 Two samples

2.1 Two sample t-test

1. When both sample sizes are large enough, we can use CLT and apply the two sample t-test.
2. When one (or both) have a small sample size, we have to test for normality and follow the scheme below.



We want to test whether the average body temperature for men is equal to the average body temperature for women. We first create two new vectors with the body temperature for men and women separately.

```
mentemp <- normtemp[normtemp$gender==1, 'temperature']
length(mentemp)
```

```
## [1] 65
```

```
womentemp <- normtemp[normtemp$gender==2, 'temperature']
length(womentemp)
```

```
## [1] 65
```

Here, both sample sizes are large enough to use the CLT. Below we show how to use the Shapiro Wilk test in case sample sizes are small.

Ask for some descriptive statistics:

```r
# Summary statistics by gender
gender_result <- normtemp %>%
  group_by(gender) %>%
  get_summary_stats(temperature, show = c("n", "mean", "sd", "median"))
gender_result
```

```
## # A tibble: 2 x 6
##   gender variable        n  mean    sd median
##    <int> <chr>       <dbl> <dbl> <dbl>  <dbl>
## 1      1 temperature    65  98.1 0.699   98.1
## 2      2 temperature    65  98.4 0.743   98.4
```

2.1.1 Test normality in both groups

Test normality in both groups is not necessary in this example, but it is given for illustrative purposes.

H_0 : **distribution of the data is normal**
H_1 : **distribution of the data is not normal**

2.1.1.1 Shapiro-Wilk test: `shapiro.test` test normality in both groups

```r
shapiro.test(mentemp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mentemp
## W = 0.98941, p-value = 0.8545
```

```r
shapiro.test(womentemp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  womentemp
## W = 0.96797, p-value = 0.09017
```

The p-value is large in both groups. There is no evidence that the data is not normally distributed.

2.1.2 Test equality of variances: `var.test`

$H_0 : \sigma^2_{men} = \sigma^2_{women}$
$H_1 : \sigma^2_{men} \neq \sigma^2_{women}$

Test if variances are equal

```r
var.test(mentemp, womentemp)
```

```
##
##  F test to compare two variances
##
## data:  mentemp and womentemp
## F = 0.88329, num df = 64, denom df = 64, p-value = 0.6211
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5387604 1.4481404
## sample estimates:
## ratio of variances
```

```
##          0.8832897
```

Another way to test if variances are equal

```
var.test(normtemp$temperature ~ normtemp$gender)
```

```
##
##  F test to compare two variances
##
## data:  normtemp$temperature by normtemp$gender
## F = 0.88329, num df = 64, denom df = 64, p-value = 0.6211
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5387604 1.4481404
## sample estimates:
## ratio of variances
##          0.8832897
```

2.1.3 Test equality of means

$H_0 : \mu_{men} = \mu_{women}$
$H_1 : \mu_{men} \neq \mu_{women}$

```
t.test(mentemp, womentemp, var.equal = TRUE)
```

2.1.3.1 T-test when equal variances can be assumed

```
##
##  Two Sample t-test
##
## data:  mentemp and womentemp
## t = -2.2854, df = 128, p-value = 0.02393
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.53963938 -0.03882216
## sample estimates:
## mean of x mean of y
##  98.10462  98.39385
```

Another possibility is:

```
t.test(normtemp$temperature ~ normtemp$gender, var.equal = T)
```

```
t.test(mentemp, womentemp, var.equal = FALSE)
```

2.1.3.2 In case the variances are not equal, apply the t-test for unequal variances

```
##
##  Welch Two Sample t-test
##
## data:  mentemp and womentemp
## t = -2.2854, df = 127.51, p-value = 0.02394
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.53964856 -0.03881298
## sample estimates:
```

```
## mean of x mean of y
##  98.10462  98.39385
```

```
wilcox.test(mentemp, womentemp)
```

2.1.3.3 A non-parametric alternative

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  mentemp and womentemp
## W = 1637, p-value = 0.02676
## alternative hypothesis: true location shift is not equal to 0
```

Remark:
We can also use functions of the `rstatix` package

2.1.3.4 Functions of `rstatix` package[1]

```
normtemp %>%
  group_by(gender) %>%
  shapiro_test(temperature)
```

2.1.3.4.1 To test normality

```
## # A tibble: 2 x 4
##   gender variable    statistic      p
##    <int> <chr>           <dbl>  <dbl>
## 1      1 temperature     0.989 0.855
## 2      2 temperature     0.968 0.0902
```

```
normtemp %>%
  levene_test(temperature ~ as.factor(gender))
```

2.1.3.4.2 Test homogeneity of variances

```
## # A tibble: 1 x 4
##     df1   df2 statistic     p
##   <int> <int>     <dbl> <dbl>
## 1     1   128    0.0635 0.801
```

2.1.3.4.3 Two sample t-test

Two sample t-test, assuming equal variances

```
normtemp %>%
  t_test(temperature ~ gender, var.equal = T)
```

```
## # A tibble: 1 x 8
##   .y.         group1 group2    n1    n2 statistic    df      p
## * <chr>       <chr>  <chr>  <int> <int>     <dbl> <dbl>  <dbl>
## 1 temperature 1      2         65    65     -2.29   128 0.0239
```

[1] https://cran.r-project.org/web/packages/rstatix/readme/README.html#:~:text=rstatix,Kruskal%2DWallis%20and%20correlation%20analyses

2.1.3.4.4 A non-parametric alternative Wilcoxon non-parametric test

```
normtemp %>% wilcox_test(temperature ~ gender)
```

```
## # A tibble: 1 x 7
##   .y.         group1 group2    n1    n2 statistic      p
## * <chr>       <chr>  <chr>  <int> <int>     <dbl>  <dbl>
## 1 temperature 1      2         65    65      1637 0.0268
```

2.2 Correlation analysis

To estimate the **linear** association between two continuous variables: function `cor_test()` or `cor_mat()`.

- `cor_test()`: correlation test between two or more variables using Pearson, Spearman or Kendall methods.
- `cor_mat()`: compute correlation matrix with p-values (in case you have more than 2 numeric variables). Returns a data frame containing the matrix of the correlation coefficients. The output has an attribute named `pvalue`, which contains the matrix of the correlation test p-values.

To test for significant association

$H_0 : \rho = 0$
versus
$H_1 : \rho \neq 0$

Correlation analysis

```
normtemp %>% cor_test(temperature, hr)
```

```
## # A tibble: 1 x 8
##   var1        var2    cor statistic       p conf.low conf.high method
##   <chr>       <chr> <dbl>     <dbl>   <dbl>    <dbl>     <dbl> <chr>
## 1 temperature hr     0.25      2.97 0.00359   0.0852     0.408 Pearson
```

In case we want to use the non-parametric Spearman correlation

```
normtemp %>% cor_test(temperature, hr, method = "spearman")
```

```
## # A tibble: 1 x 6
##   var1        var2    cor statistic       p method
##   <chr>       <chr> <dbl>     <dbl>   <dbl> <chr>
## 1 temperature hr     0.28   263288. 0.00121 Spearman
```

3 Exercises

1. Use the `chol` data set. (import the *chol.txt* file from Toledo)
   a) Create a new variable `group` based on the variable `SMOKE` of the `chol` data set. The variable `group` has two possible values `nonsmoke` or `smoke`. In the group `smoke` we have the sigare and pipe smokers.
   Hint: Use the function `ifelse()` to create a new variable with the value `nonsmoke`/`smoke`.
   b) Make a grouped boxplot of `chol` value by `group`.
2. Generate, on the same graphical window, tho histograms for the `CHOL` values of these groups.
3. Use the data set `chol` to detect if there is a significant difference in average cholesterol (`CHOL`) between the two groups `smoke` and `nonsmoke`. Give comment on the methods you are using.