

Testing of Hypothesis using R

Dr. Kapil Kumar

Department of Statistics
Chaudhary Charan Singh University, Meerut, India
kapilstats@gmail.com

One Week Online National Workshop on Statistical Data Analysis using SPSS & R
(SDASR), 5th Edition (11–15 March 2024)

Department of Statistics, Central University of Haryana, Mahendergarh

12 March 2024 · Lecture 1 of 2 (online)

Plan of the session

1. The logic of a statistical test
2. One-sample tests: blood pressure
3. Comparing two means: independent and paired samples
4. Tests for categorical data
5. Choosing a test, and practice

By the end you can state H_0 and H_1 for a research question, run z , t , paired t , χ^2 and Fisher tests in R, compute a p -value by hand, check the assumptions, and report the result in plain words with a confidence interval.

Packages and data used in this session

- **stats** (loaded with R): `t.test()`, `chisq.test()`, `fisher.test()`, `pt()`, `pnorm()`
- **BSDA**: `z.test()` when σ is known
- **Base graphics**: `hist()`, `boxplot()`, `curve()`
- Data files sit in a `data/` folder next to the script and are read with `read.csv("data/. . .")`, so the script runs on any computer without drive-letter paths

```
# install once
# install.packages("BSDA")
library(BSDA)           # z.test()

# two colours used in all plots
col_main <- "#1F5FA8"
col_alt  <- "#B45309"
```

PART 1

The logic of a statistical test

From a scientific question to a statistical hypothesis

Scientific hypotheses say how we think the world works:

- the coin is biased
- average body temperature is 98.6°F
- average systolic blood pressure is 120 mmHg
- wheat yield is higher with fertiliser
- an advertising campaign increased sales

Statistical hypotheses are statements about a model parameter:

Coin: $X_i = 1$ if toss i is a head,
 $E(X_i) = p$; unbiased means $p = 0.5$

BP: X_i = blood pressure of person i ,
 $E(X_i) = \mu$; is $\mu = 120$ mmHg?

Remember Hypotheses concern the *population*, never one individual or the sample in hand.

Null and alternative hypotheses

- The **null hypothesis** H_0 is the “no effect, no difference” statement. It contains the equality and is retained unless the data give strong evidence against it.
- The **alternative** H_1 is what we conclude when H_0 is rejected.
- A **two-sided** H_1 uses $\neq$; a **one-sided** H_1 uses $<$ or $>$ and matches words such as “less than” or “greater than”.

Coin: $H_0 : p = 0.5$ vs $H_1 : p \neq 0.5$

BP: $H_0 : \mu = 120$ vs

two-sided $H_1 : \mu \neq 120$

left-tailed $H_1 : \mu < 120$

right-tailed $H_1 : \mu > 120$

Rule Fix the direction of H_1 from the research question *before* looking at the data.

Two kinds of error

Decision	H_0 true	H_0 false
Reject H_0	Type I error (probability α)	correct (power $1 - \beta$)
Do not reject H_0	correct ($1 - \alpha$)	Type II error (probability β)

- A Type I error claims an effect that does not exist; a Type II error misses a real one.
- The **level of significance** α is fixed in advance: usually 0.05, sometimes 0.01 or 0.10.
- For a fixed sample size, lowering α raises β ; a larger sample lowers both.

Test statistic and p -value

- A **test statistic** ($Z, t, \chi^2, \dots$) is computed from the sample and has a known distribution when H_0 is true.
- The **p -value** is the probability, computed assuming H_0 is true, of a test statistic at least as extreme as the one observed.
- A small p -value means the data would be surprising if H_0 were true.

Decision rule Reject H_0 at level α if $p \leq \alpha$. Otherwise say “do not reject H_0 ”, not “accept H_0 ”.

Common misreadings p is *not* the probability that H_0 is true, and a small p does not mean a large or important effect. Report an estimate and a confidence interval as well.

Computing a p -value

Observed statistic t ; under H_0 , $T \sim t_\nu$.

Left-tailed: $p = P(T \leq t)$

Right-tailed: $p = P(T \geq t) = 1 - P(T \leq t)$

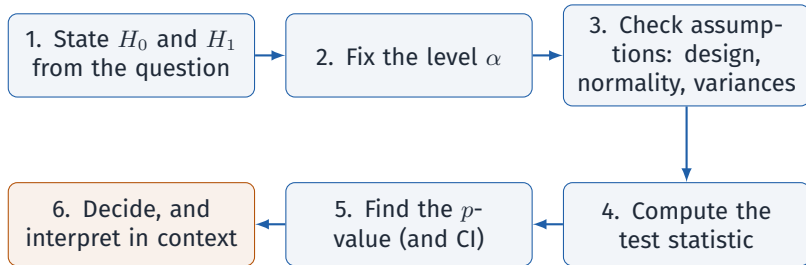
Two-tailed: $p = 2 P(T \leq -|t|)$

The two-tailed formula relies on symmetry, so it holds for the t and normal distributions.

Skewed distributions For F or χ^2 statistics, with $F(x) = P(X \leq x)$, use $p = 2 \min\{F(x), 1 - F(x)\}$, which never exceeds 1.

In R `pt(t, df)` gives $P(T \leq t)$; likewise `pnorm()`, `pchisq()`, `pf()`. `qt()`, `qnorm()` give critical values.

The steps of every test



In R steps 4 and 5 are one function call. Steps 1–3 and 6 are the statistician's job.

PART 2

One-sample tests: blood pressure

The blood pressure data

- Systolic blood pressure (mmHg) of 55 persons: 25 healthy (status = 0) and 30 with hypertension (status = 1)
- Illustrative data from the workshop materials
- Question: is the population mean 120 mmHg?

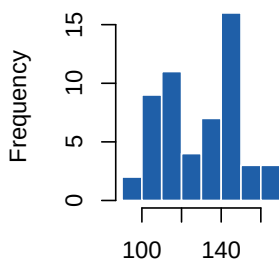
```
bp <- read.csv("data/bp_data.csv")
head(bp, 3)
tapply(bp$mmhg, bp$status, length)
tapply(bp$mmhg, bp$status, mean)
```

##	no	status	mmhg
## 1	1	0	120
## 2	2	0	115
## 3	3	0	94
##	0	1	
##	25	30	
##		0	1
##	112.9	144.2	

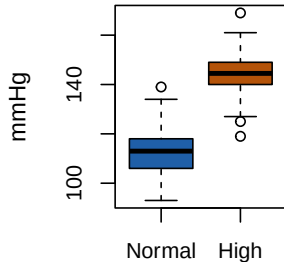
Look before you test

```
par(mfrow = c(1, 2), cex = 0.9,  
    mar = c(4, 4, 1, 1))  
hist(bp$mmhg, breaks = 10, main = "",  
     col = col_main, border = "white",  
     xlab = "Systolic BP (mmHg)")  
boxplot(mmhg ~ status, data = bp,  
        names = c("Normal", "High"),  
        col = c(col_main, col_alt),  
        xlab = "Group", ylab = "mmHg",  
        cex.axis = 0.9)
```

The overall mean (130 mmHg) hides two groups.



Systolic BP (mmHg)



Group

Group means 112.9 and 144.2 mmHg: compared in Part 3.

Z test when σ is known: by hand

$H_0 : \mu = 120$ vs $H_1 : \mu \neq 120$, with
 $\sigma = 20$ given and $\alpha = 0.05$.

$$Z = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} \sim N(0, 1)$$

under H_0

Reject H_0 if $|z| > z_{0.025} = 1.96$;

$$p = 2\Phi(-|z|)$$

```
xbar <- mean(bp$mmhg)
n <- length(bp$mmhg)
mu0 <- 120 # value under H0
sigma <- 20 # known sigma
z <- sqrt(n) * (xbar - mu0) / sigma
p_value <- 2 * pnorm(-abs(z))
z_tab <- qnorm(0.05 / 2, lower.tail = FALSE)
round(c(z = z, z_tab = z_tab,
        p_value = p_value), 4)

##      z    z_tab p_value
## 3.7081 1.9600 0.0002
```

The same test with BSDA: : z.test()

```
z.test(bp$mmhg, mu = 120, sigma.x = 20)

##
## One-sample z-Test
##
## data: bp$mmhg
## z = 3.7081, p-value = 0.0002088
## alternative hypothesis: true mean is not equal to 120
## 95 percent confidence interval:
## 124.7144 135.2856
## sample estimates:
## mean of x
## 130
```

Reading it $z = 3.71 > 1.96$
and $p = 0.00021$: reject H_0 at
the 5% level. The 95% interval
(124.7, 135.3) mmHg does not
contain 120.

In practice σ is rarely known.
When it is estimated from the
sample, use the t test.

One-sample t test: σ unknown

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1} \text{ under } H_0$$

Assumptions: a random sample from an approximately normal population (or n large enough for the central limit theorem).

```
tt <- t.test(bp$mmhg, mu = 120)
tt

##
## One Sample t-test
##
## data: bp$mmhg
## t = 3.8693, df = 54, p-value = 0.0002961
## alternative hypothesis: true mean is not equal to 120
## 95 percent confidence interval:
## 124.8185 135.1815
## sample estimates:
## mean of x
## 130
```

Reading the t -test output

- $t = 3.87$ on 54 degrees of freedom
- $p = 0.0003 < 0.05$: reject $H_0 : \mu = 120$
- 95% CI for μ : (124.8, 135.2) mmHg; it excludes 120, which is the same decision
- In words: the mean systolic BP in this population is about 10 mmHg above 120

Checking the p -value by hand with `pt()`:

```
t_obs <- unname(tt$statistic)
df <- unname(tt$parameter)
2 * pt(-abs(t_obs), df)

## [1] 0.0002961
```

Duality A two-sided test at level α rejects $H_0 : \mu = \mu_0$ exactly when μ_0 lies outside the $100(1 - \alpha)\%$ confidence interval.

One-sided alternatives

```
# left-tailed, H1:  $\mu < 120$ 
less <- t.test(bp$mmhg, mu = 120, alternative = "less")
# right-tailed, H1:  $\mu > 120$ 
more <- t.test(bp$mmhg, mu = 120, alternative = "greater")
c(less = less$p.value, greater = more$p.value)

##      less      greater
## 0.9998519 0.0001481

more$conf.int[1] # lower 95% bound for  $\mu$ 

## [1] 125.7
```

- Right-tailed p is half the two-sided p , because $\bar{x} = 130$ lies on the side named in H_1
- Left-tailed $p \approx 1$: the data point the other way, so there is no evidence that $\mu < 120$

Do not shop for a side

Running both one-sided tests and keeping the smaller p doubles the real Type I error rate.

PART 3

Comparing two means

Two independent samples: pooled and Welch t

Pooled (equal variances assumed)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$
$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

R: `t.test(x, y, var.equal = TRUE)`

Welch (variances may differ)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

approximately t_ν , ν from the Welch–Satterthwaite formula

R: `t.test(x, y)` (the default)

Choosing Both test $H_0 : \mu_1 = \mu_2$ for independent samples from normal populations. With similar spreads they agree; with unequal spreads and sizes the pooled test can mislead, so Welch is the safer default.

Healthy vs hypertensive: pooled t test

```
t.test(mmhg ~ status, data = bp, var.equal = TRUE) # formula: response ~ group

##
## Two Sample t-test
##
## data: mmhg by status
## t = -10.468, df = 53, p-value = 1.66e-14
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
## -37.31328 -25.31339
## sample estimates:
## mean in group 0 mean in group 1
## 112.9200 144.2333
```

mmhg ~ status splits the response by group. $p < 0.001$: the two group means differ.

Welch test and the equal-variance check

```
S0 <- bp$mmhg[bp$status == 0] # healthy
S1 <- bp$mmhg[bp$status == 1] # hypertensive
welch <- t.test(S0, S1)       # R's default
round(welch$conf.int[1:2], 2)

## [1] -37.33 -25.30

var.test(S0, S1)$p.value      # H0: equal variances

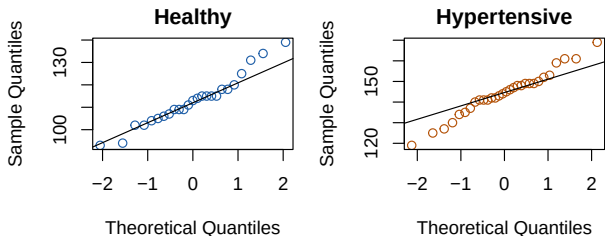
## [1] 0.918
```

Reading it Welch: $t = -10.45$ on 50.9 df, $p < 0.001$. Hypertensive persons average 31.3 mmHg more than healthy persons (95% CI 25.3 to 37.3 mmHg).

Which one? The F test finds no difference in spread ($p = 0.92$), so pooled and Welch agree. (F is itself sensitive to non-normality.)

Checking normality in each group

```
par(mfrow = c(1, 2), mar = c(4, 4, 2, 1))
qqnorm(S0, main = "Healthy", col = col_main)
qqline(S0)
qqnorm(S1, main = "Hypertensive", col = col_alt)
qqline(S1)
```



```
shapiro.test(S0)$p.value
```

```
## [1] 0.4249
```

```
shapiro.test(S1)$p.value
```

```
## [1] 0.8179
```

Verdict Points follow the lines and both Shapiro–Wilk p -values exceed 0.05: the normality assumption is reasonable in each group.

Small samples: scores of two batches

- Scores of 5 students in batch A: 20, 20.5, 13, 16, 11
- Scores of 5 other students in batch B: 40, 42, 38, 56, 57
- $H_0 : \mu_A = \mu_B$ VS $H_1 : \mu_A \neq \mu_B$

With $n = 5$ Normality cannot be checked well from five values; the test leans on the assumption.

```
A1 <- c(20, 20.5, 13, 16, 11)
B1 <- c(40, 42, 38, 56, 57)
t.test(A1, B1)$p.value      # Welch

## [1] 0.0006718

fit.t <- t.test(A1, B1, var.equal = TRUE)
fit.t$p.value               # pooled

## [1] 0.0001415
```

A test result is a list: extract what you report

```
names(fit.t)
fit.t$statistic
fit.t$parameter
round(fit.t$conf.int[1:2], 2)

## [1] "statistic" "parameter" "p.value"
## [4] "conf.int" "estimate" "null.value"
## [7] "stderr" "alternative" "method"
## [10] "data.name"
## t
## -6.774
## df
## 8
## [1] -40.88 -20.12
```

Reading it Batch B scores 30.5 marks more on average; $t = -6.77$ on 8 df, $p = 0.00014$. The 95% CI for $\mu_A - \mu_B$ is $(-40.9, -20.1)$.

Pulling numbers out of the object (as this slide does) avoids copying errors in reports.

The formula interface: `response ~ group`

- Data in “long” form: one column of scores, one column naming the group
- `gl(2, 5, labels = ...)` makes a factor with 2 levels of 5 each
- The same call works on a data frame: `t.test(mmhg ~ status, data = bp)`

```
score <- c(A1, B1)
group <- gl(2, 5, labels = c("batchA", "batchB"))
group
res <- t.test(score ~ group)      # Welch
res$estimate
res$p.value      # same as t.test(A1, B1)

## [1] batchA batchA batchA batchA batchA batchB batchB
## [8] batchB batchB batchB
## Levels: batchA batchB
## mean in group batchA mean in group batchB
##                      16.1                      46.6
## [1] 0.0006718
```

Paired design: does an anti-smoking film help?

- Mean daily cigarettes of 8 heavy smokers in the month before and the month after seeing the film
- The *same* people are measured twice, so we analyse the differences

$$d_i = \text{before}_i - \text{after}_i$$

$$H_0 : \mu_d = 0 \text{ vs } H_1 : \mu_d > 0 \text{ (a decline)}$$

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} \sim t_{n-1}$$

```
smoke <- read.csv("data/smoker_data.csv")
smoke$diff <- smoke$Before - smoke$After
smoke
```

##	Before	After	diff
## 1	28	26	2
## 2	29	27	2
## 3	31	32	-1
## 4	44	44	0
## 5	35	35	0
## 6	20	16	4
## 7	50	47	3
## 8	25	23	2

Paired t test in R

```
tp <- t.test(smoke$Before, smoke$After, paired = TRUE)
c(tp$statistic, tp$parameter, p = tp$p.value)
round(tp$conf.int[1:2], 2)
# H1: mean difference > 0, i.e. a decline
t.test(smoke$Before, smoke$After, paired = TRUE,
       alternative = "greater")$p.value

##      t      df      p
## 2.5100 7.0000 0.0404
## [1] 0.09 2.91
## [1] 0.0202
```

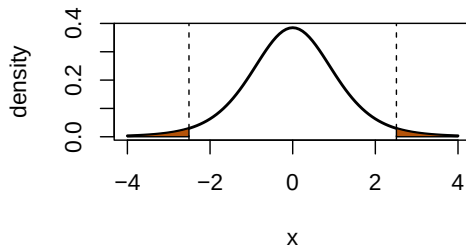
Reading it Mean fall 1.5 cigarettes a day (95% CI 0.09 to 2.91). The question is about a *decline*, so the one-sided $p = 0.02$ applies: evidence of a decline at the 5% level.

But Only 8 smokers; is a fall of 1.5 cigarettes a day useful in practice?

Where the p -value comes from

```
t_obs <- unname(tp$statistic)
df <- unname(tp$parameter)
2 * pt(-abs(t_obs), df)      # two-tailed
1 - pt(t_obs, df)           # right-tailed
par(mar = c(4, 4, 1, 1))
curve(dt(x, df), -4, 4, lwd = 2, ylab = "density")
xr <- seq(t_obs, 4, length.out = 50)
ys <- c(dt(xr, df), 0 * xr)
polygon(c(xr, rev(xr)), ys, col = col_alt)
polygon(-c(xr, rev(xr)), ys, col = col_alt)
abline(v = c(-t_obs, t_obs), lty = 2)

## [1] 0.0404
## [1] 0.0202
```



Picture Each tail beyond ± 2.51 has area 0.0202: one tail = one-sided p , both = two-sided p .

What if we ignore the pairing?

```
# wrong: treats the 16 values as
# two independent samples
t.test(smoke$Before, smoke$After)$p.value

## [1] 0.7735

# right: uses the 8 differences
t.test(smoke$Before, smoke$After,
       paired = TRUE)$p.value

## [1] 0.0404

cor(smoke$Before, smoke$After)

## [1] 0.9878
```

- Smokers differ a lot from each other (20 to 50 a day), but each smoker's two values are very close: the correlation is 0.99
- Pairing removes the between-smoker variation, so the same mean difference becomes detectable

Rule Analyse the design you used:
paired data need a paired test.

IQ before and after training: testing a non-zero difference

- IQ of 20 students before (IQ1) and after (IQ2) a training
- Claim: training raises IQ by 10 points on average

$$d = \text{IQ1} - \text{IQ2}$$

$$H_0 : \mu_d = -10$$

$$H_1 : \mu_d \neq -10$$

$\text{mu} = -10$ sets the value under H_0 .

```
iq <- read.csv("data/iq_data.csv")
t.test(iq$IQ1, iq$IQ2, mu = -10, paired = TRUE)

##
## Paired t-test
##
## data: iq$IQ1 and iq$IQ2
## t = -1.2854, df = 19, p-value = 0.2141
## alternative hypothesis: true mean difference is not equal to -10
## 95 percent confidence interval:
## -11.051338 -9.748662
## sample estimates:
## mean difference
## -10.4
```

Reading the IQ result

```
d <- iq$IQ1 - iq$IQ2
table(d)

## d
## -16 -12 -10
##    1    1   18
```

Eighteen of the 20 students gained exactly 10 points: the data are illustrative, not a real experiment.

Reading it Mean gain 10.4 points; 95% CI for the gain (9.75, 11.05). $p = 0.21$: no evidence that the average gain differs from 10.

Careful “Not rejected” is not “proved”. To *establish* a gain of about 10 we would need an equivalence test or rely on the narrow interval.

PART 4

Tests for categorical data

Diabetes, age and gender: a small hypothetical data set

- 20 hypothetical persons, three categorical variables
- `stringsAsFactors = TRUE` stores the text columns as factors
- `table()` counts each category

```
sugar <- read.csv("data/sugar_data.csv",  
                  stringsAsFactors = TRUE)  
head(sugar, 3)  
summary(sugar)      # counts for each factor  
  
##   diabetes   age gender  
## 1        No   Old Female  
## 2        No Young Female  
## 3        Yes   Old  Male  
##   diabetes   age      gender  
##  No :13   Old  : 8   Female:11  
##  Yes: 7   Young:12  Male   : 9
```

Contingency tables and proportions

```
tab <- with(sugar, table(diabetes, age))
tab
```

```
##           age
## diabetes Old Young
##      No    2    11
##      Yes    6     1
```

```
round(prop.table(tab, 2), 2) # by column
```

```
##           age
## diabetes Old Young
##      No  0.25  0.92
##      Yes 0.75  0.08
```

- `prop.table(tab, 2)` divides each column by its total; 1 would use rows
- 75% of old persons are diabetic, against 8% of young persons
- `fable(table(diabetes, age, gender))` prints a three-way table compactly

Question Is diabetes associated with age? With gender?

Chi-square test of independence

H_0 : diabetes and age are independent

$$\chi^2 = \sum_{i,j} (O_{ij} - E_{ij})^2 / E_{ij}, \quad E_{ij} = R_i C_j / n$$

df = $(r - 1)(c - 1)$; Yates' correction for 2×2

R warns “Chi-squared approximation may be incorrect”: 2 of 4 expected counts are below 5. Use an exact test rather than `suppressWarnings()`.

```
ct <- chisq.test(tab)
ct$p.value
ct$expected

## [1] 0.009773
##           age
## diabetes Old Young
##      No   5.2   7.8
##      Yes   2.8   4.2
```

$\chi^2 = 6.68$ on 1 df, $p = 0.01$, but the approximation is doubtful with expected counts this small.

Small counts: Fisher's exact test

```
fisher.test(tab)

##
## Fisher's Exact Test for Count Data
##
## data:  tab
## p-value = 0.004438
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.0006152406 0.5491794538
## sample estimates:
## odds ratio
## 0.03977064
```

```
tab2 <- with(sugar,
             table(diabetes, gender))
fisher.test(tab2)$p.value

## [1] 0.3742
```

Reading it Age:
 $p = 0.0044$, a clear
association; the young
have far lower odds of
diabetes (OR 0.04). Gender:
 $p = 0.37$, no evidence.

PART 5

Choosing a test, and practice

Which test? A quick guide

Question	Example today	R function
One mean, σ known	BP vs 120, $\sigma = 20$	<code>BSDA::z.test(x, mu, sigma.x)</code>
One mean, σ unknown	BP vs 120	<code>t.test(x, mu = 120)</code>
Two independent means	healthy vs hypertensive	<code>t.test(y ~ g)</code> (Welch)
Two paired means	smokers before / after	<code>t.test(x, y, paired = TRUE)</code>
Equality of two variances	healthy vs hypertensive	<code>var.test(x, y)</code>
Two categorical variables	diabetes by age	<code>chisq.test()</code> , <code>fisher.test()</code>
Small, clearly non-normal samples	–	<code>wilcox.test()</code>

Every one of these returns an object of class `htest` with `$statistic`, `$p.value` and (usually) `$conf.int`.

Exercises

1. Blood donations by blood group in 1995 and 2000. Draw a multiple bar diagram with `barplot(donations, beside = TRUE)` and test whether the distribution over blood groups changed (`chisq.test(donations)`).

```
donations <- matrix(c(576, 1125, 775, 1280, 1154, 1700, 115, 560), nrow = 2,  
                    dimnames = list(c("1995", "2000"), c("A", "B", "O", "AB")))
```

2. `marks.csv` holds marks of 20 students in two papers (m_1, m_2). Plot $m_1 - m_2$ first, then test whether the mean marks differ. What do students 1 and 2 do to the result?
3. Repeat the healthy vs hypertensive comparison with `t.test(mmhg ~ status, data = bp)` and with `wilcox.test()`. Do the conclusions agree?

Key take-aways

1. **The question sets the hypotheses:** write H_0 , H_1 and α before looking at the data.
2. **Plot first:** the BP histogram showed two groups before any test did.
3. **Match the test to the design:** independent groups, paired measurements and categorical counts each need their own test.
4. **Check assumptions:** normality, equal variances, expected counts ≥ 5 ; act on R's warnings, do not suppress them.
5. **Report more than p :** give the estimate, a confidence interval and a sentence in plain words.

Next session: correlation and regression, where the same tests are applied to relationships between two variables.

References

- Dalgaard, P. (2008). *Introductory Statistics with R*, 2nd Ed., Springer, New York.
- Jones, O., Harden, S., and Crawley, M. J. (2022). *The R Book*, 3rd Ed., Wiley, Hoboken.
- Verzani, J. (2014). *Using R for Introductory Statistics*, 2nd Ed., CRC Press, Boca Raton.
- Shapiro, S. S., and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4), 591–611.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. <https://www.R-project.org>

Thank you

Questions and discussion