

Applications of Statistical Software R

Dr. Kapil Kumar

Assistant Professor, Department of Statistics
Central University of Haryana, Mahendergarh, India
kapilstats@gmail.com

23 March 2021

One-Week Online National Workshop on Applications of Mathematical and Statistical Tools
(20–24 March 2021)

Departments of Mathematics and Statistics and School of Education, Central University of
Haryana, under PMMMNMTT

Expert talk (online)

Plan of the session

- 1 R for applied work: objects and data import
- 2 Case study 1: categorical data
- 3 Case study 2: comparing two groups
- 4 Case study 3: analysis of variance
- 5 Case study 4: regression
- 6 Choosing a method and reporting it

By the end you can import data from Excel or CSV, choose a suitable test for a research question, check its assumptions, run it in R and write the result as a sentence for a paper or a thesis.

Audience Teachers and researchers who know the statistics and want to do it in R.

- 1 R for applied work
- 2 Categorical data
- 3 Two groups
- 4 ANOVA
- 5 Regression
- 6 Choosing and reporting

Why R for teaching and research?

- Free, open source, and the same on Windows, macOS and Linux
- Every analysis is a **script**: it can be rerun, checked, shared and corrected
- Thousands of contributed packages on CRAN cover almost every applied field
- Publication-quality graphics from the same code
- RStudio adds an editor, console, plot window and data viewer in one place

Today's approach Four small case studies. For each: the question, a first look at the data, the assumptions, the test or model, and a plain-language reading of the output.

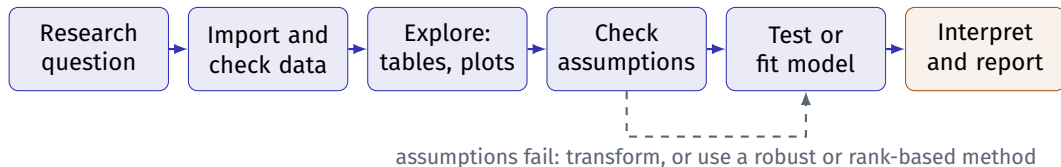
R:

<https://cran.r-project.org>

RStudio:

<https://posit.co/download/rstudio-desktop/>

One workflow for every analysis



The question comes first What is the outcome, what is the comparison, and are the observations independent or paired? These three answers choose the method.

Report An estimate, its confidence interval and a p -value, in words a reader outside statistics can follow.

Packages and colours used in this session

- Base R: `table()`, `fisher.test()`, `t.test()`, `wilcox.test()`, `aov()`, `TukeyHSD()`, `lm()`
- **readxl**: `read_excel()` reads `.xls` and `.xlsx` files
- **arm**: `display()`, a compact regression summary
- Built-in data: Titanic, sleep, swiss

```
# install once
# install.packages(c("readxl", "arm"))
library(readxl)
library(arm)
col_main <- "#3333B3" # blue
col_alt  <- "#B45309" # amber
```

The four data objects you need

Object	Holds
vector	values of one type
factor	a categorical variable
data frame	a data set
list	anything, e.g. results

Results of tests and models are
lists: \$ extracts a part, as in
res\$p.value.

```
marks <- c(56, 71, 48, 90)           # vector
sex <- factor(c("F", "M", "F", "F")) # factor
dat <- data.frame(marks, sex)        # data frame
str(dat)

## 'data.frame': 4 obs. of  2 variables:
## $ marks: num  56 71 48 90
## $ sex : Factor w/ 2 levels "F","M": 1 2 1 1

res <- t.test(marks)                 # a list
round(res$conf.int[1:2], 1)

## [1] 36.8 95.7
```

Importing from Excel and from CSV

```
book <- read_excel("data/Book1.xlsx",  
                  sheet = 1)  
book          # a tibble, 5 x 2
```

```
## # A tibble: 5 x 2
```

```
##       A       B
```

```
##   <dbl> <dbl>
```

```
## 1     1     2
```

```
## 2     2     5
```

```
## 3     3     8
```

```
## 4     5     9
```

```
## 5     6     6
```

```
names(read_csv("data/Book1.csv"))
```

```
## [1] "X...A" "B"
```

```
book_csv <- read_csv("data/Book1.csv",  
                    fileEncoding = "UTF-8-BOM")  
names(book_csv)
```

```
## [1] "A" "B"
```

```
all.equal(as.data.frame(book), book_csv)
```

```
## [1] TRUE
```

A common trap Excel's "CSV UTF-8" adds an invisible byte-order mark that garbles the first column name. `fileEncoding = "UTF-8-BOM"` removes it.

- 1 R for applied work
- 2 Categorical data**
- 3 Two groups
- 4 ANOVA
- 5 Regression
- 6 Choosing and reporting

Case study 1: is diabetes associated with age?

```
data4 <- read.table("data/data4.txt", header = TRUE,
                    stringsAsFactors = TRUE)
data4$Sugar <- factor(data4$Sugar,
                      levels = c("yes", "no"))
summary(data4)
```

```
##  Sugar      Age      Gender
##  yes: 7   old : 8   female:11
##  no :13  young:12  male  : 9
```

- Hypothetical data on 20 patients: Sugar (diabetic yes/no), Age group, Gender
- stringsAsFactors = TRUE makes text columns factors
- Putting "yes" first makes later odds ratios refer to having sugar

Question Are older patients more likely to be diabetic?

A first look: counts, percentages and a mosaic plot

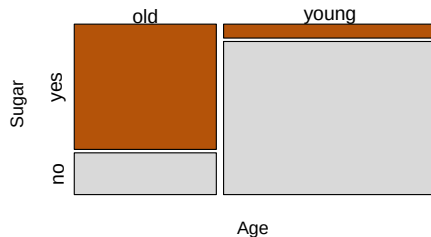
```
tab <- xtabs(~ Age + Sugar, data = data4)
addmargins(tab)
```

```
##           Sugar
## Age      yes no Sum
##  old       6  2   8
##  young     1 11  12
##  Sum       7 13  20
```

```
round(100 * prop.table(tab, 1), 1) # row %
```

```
##           Sugar
## Age      yes  no
##  old    75.0 25.0
##  young   8.3 91.7
```

```
par(mar = c(3, 3, 1, 1))
mosaicplot(tab, col = c(col_alt, "grey85"),
            main = "", cex.axis = 1.1)
```



Widths: age groups; heights: share with sugar, 75% of old against 8% of young patients.

Chi-square or Fisher? Check the expected counts

```
round(chisq.test(tab)$expected, 1)
```

```
##           Sugar
## Age      yes  no
##  old     2.8 5.2
##  young   4.2 7.8
```

Rule The χ^2 test needs expected counts of at least 5. Two cells fail, so R warns that the approximation may be incorrect: use Fisher's exact test.

Table $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$: rows old, young; columns yes, no

$$\text{OR} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

H_0 : OR = 1, i.e. no association

Here Sample OR
= $(6 \times 11)/(2 \times 1) = 33$: the odds of sugar are far higher in the old group.

Fisher's exact test

```
fisher.test(tab)

##
##  Fisher's Exact Test for Count Data
##
## data:  tab
## p-value = 0.004
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##      1.821 1625.380
## sample estimates:
## odds ratio
##      25.14
```

`fisher.test()` reports the conditional maximum-likelihood estimate of the odds ratio, which with so few patients is smaller than the simple cross-product ratio 33.

Reading case study 1

```
ft <- fisher.test(tab)
ft_gender <- fisher.test(xtabs(~ Gender + Sugar, data = data4))
round(c(p_age = ft$p.value, p_gender = ft_gender$p.value), 4)

##      p_age p_gender
## 0.0044    0.3742
```

- **Age:** $p = 0.0044$. The odds of having sugar are estimated to be about 25 times higher for old patients
- The 95% CI, 1.8 to 1,625, is very wide: 20 patients say little about *how* strong the association is
- **Gender:** $p = 0.37$, no evidence of an association

For a report Diabetes was more common among older patients (6/8 vs 1/12; Fisher's exact test, $p = 0.004$).

Case study 1b: class and survival on the Titanic

```
tit <- margin.table(Titanic, c(1, 4))
tit
```

```
##           Survived
## Class    No Yes
## 1st    122 203
## 2nd    167 118
## 3rd    528 178
## Crew   673 212
```

```
round(100 * prop.table(tit, 1), 1)
```

```
##           Survived
## Class    No  Yes
## 1st    37.5 62.5
## 2nd    58.6 41.4
## 3rd    74.8 25.2
## Crew   76.0 24.0
```

```
ct <- chisq.test(tit)
ct

##
## Pearson's Chi-squared test
##
## data:  tit
## X-squared = 190, df = 3, p-value <2e-16
```

Large sample Titanic is a built-in four-way table (class, sex, age, survival) for 2201 people. `margin.table()` keeps class and survival. All expected counts are far above 5, so the χ^2 test is fine.

Where does the association come from?

```
round(ct$stdres, 1) # standardised residuals
```

##		Survived	
##	Class	No	Yes
##	1st	-12.6	12.6
##	2nd	-3.5	3.5
##	3rd	4.9	-4.9
##	Crew	6.9	-6.9

$$r_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}(1 - p_{i\cdot})(1 - p_{\cdot j})}}$$

approximately $N(0, 1)$ under H_0

- $X^2 = 190.4$ on 3 df: survival depended strongly on class
- 62.5% of first-class passengers survived, against 25.2% in third class and 24% of the crew
- Residuals beyond ± 2 mark the cells that drive the result: many more first-class survivors than independence predicts (+12.6)

But Class is mixed up with sex and age (“women and children first”). A fair comparison would stratify, e.g. `mantelhaen.test()`.

- 1 R for applied work
- 2 Categorical data
- 3 Two groups**
- 4 ANOVA
- 5 Regression
- 6 Choosing and reporting

Which test for two groups?

Design	Data roughly normal	Skewed, outliers or ordinal
Independent groups	Welch t -test <code>t.test(y ~ g)</code>	Wilcoxon rank-sum test <code>wilcox.test(y ~ g)</code>
Same units twice	paired t -test <code>t.test(x, y, paired = TRUE)</code>	Wilcoxon signed-rank test <code>wilcox.test(x, y, paired = TRUE)</code>

Design first Paired data arise from the design: before/after on the same person, matched pairs, two treatments given to each subject.

Normality and sample size t -tests are robust to mild non-normality when groups are moderate in size. With five per group, normality tests have little power: plot the data.

Case study 2a: two batches of students

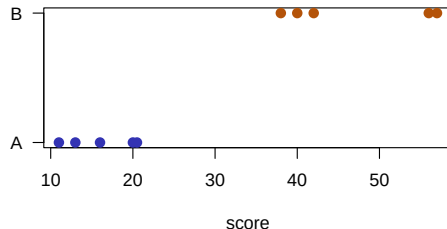
```
batches <- data.frame(  
  score = c(20, 20.5, 13, 16, 11,  
            40, 42, 38, 56, 57),  
  group = gl(2, 5, labels = c("A", "B"))  
)  
tapply(batches$score, batches$group, mean)
```

```
##      A      B  
## 16.1 46.6
```

```
tapply(batches$score, batches$group, sd)
```

```
##      A      B  
## 4.189 9.154
```

```
par(mar = c(4, 3, 1, 1))  
stripchart(score ~ group, data = batches,  
  pch = 16, cex = 1.3, las = 1,  
  col = c(col_main, col_alt))
```



Scores of five students in each of two batches of a course. Batch B is higher and more spread out.

Checking the assumptions

```
scoreA <- batches$score[batches$group == "A"]
scoreB <- batches$score[batches$group == "B"]
shapiro.test(scoreA)$p.value      # normality, A

## [1] 0.4828

shapiro.test(scoreB)$p.value      # normality, B

## [1] 0.09969

var.test(score ~ group, data = batches)$p.value

## [1] 0.1591
```

Reading it No test rejects normality ($p = 0.48$ and 0.10) or equal variances ($p = 0.16$).

But “Not rejected” is not “confirmed”: five values per group give these tests little power. Welch’s test does not need equal variances, so use it by default.

Welch's t -test, with a rank-based check

```
t.test(score ~ group, data = batches)

##
##  Welch Two Sample t-test
##
## data:  score by group
## t = -6.8, df = 5.6, p-value = 0.0007
## alternative hypothesis: true difference in means between group A and group B is not equal to 0
## 95 percent confidence interval:
##  -41.71 -19.29
## sample estimates:
## mean in group A mean in group B
##           16.1           46.6

wilcox.test(score ~ group, data = batches)$p.value # Wilcoxon rank-sum (exact)

## [1] 0.007937
```

- Mean difference A – B: –30.5 points
- 95% CI –41.7 to –19.3: excludes 0
- Welch $t = -6.77$, $df = 5.6$, $p = 0.0007$
- Rank-sum $p = 0.0079$: every B score exceeds every A score, the smallest p possible with 5 + 5 observations

The formula `score ~ group` expects the usual layout of a data file: one column of values, one column of group labels.

For a report Batch B scored on average 30.5 points higher than batch A (95% CI 19.3 to 41.7; Welch's t -test, $p < 0.001$).

Case study 2b: two sleeping drugs on the same patients

```
drug1 <- sleep$extra[sleep$group == 1]
drug2 <- sleep$extra[sleep$group == 2]
d <- drug2 - drug1          # per patient
data.frame(drug1, drug2, d)[1:5, ]
```

```
##   drug1 drug2  d
## 1   0.7   1.9 1.2
## 2  -1.6   0.8 2.4
## 3  -0.2   1.1 1.3
## 4  -1.2   0.1 1.3
## 5  -0.1  -0.1 0.0
```

- Built-in sleep data, the classic example used by “Student” (1908): extra hours of sleep for 10 patients
- Each patient took *both* drugs: rows with the same ID belong together
- The natural quantity is the difference d_i for each patient

Question Does drug 2 give more extra sleep than drug 1?

Same data, two analyses

```
t.test(drug2, drug1)$p.value # wrong: unpaired
```

```
## [1] 0.07939
```

```
t.test(drug2, drug1, paired = TRUE)$p.value
```

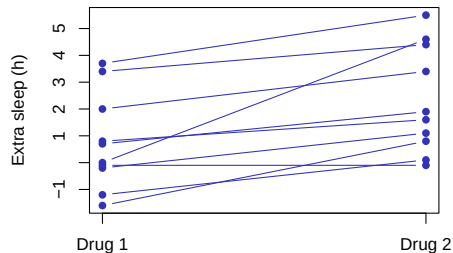
```
## [1] 0.002833
```

```
c(sd_drug1 = sd(drug1), sd_diff = sd(d))
```

```
## sd_drug1 sd_diff
```

```
## 1.789 1.230
```

```
par(mar = c(3, 4, 1, 1))  
matplot(rbind(drug1, drug2), type = "b",  
        pch = 16, lty = 1, col = col_main,  
        xaxt = "n", ylab = "Extra sleep (h)")  
axis(1, at = 1:2, labels = c("Drug 1",  
                              "Drug 2"))
```



Reading case study 2b

```
shapiro.test(d)$p.value      # differences  
  
## [1] 0.03334  
  
wilcox.test(drug2, drug1, paired = TRUE,  
            exact = FALSE)$p.value  
  
## [1] 0.009091  
  
round(t.test(d)$conf.int[1:2], 2)  
  
## [1] 0.70 2.46
```

- Patients differ a lot (SD 1.79 h); the drug effect within a patient is consistent (SD of d : 1.23 h). Pairing removes the patient-to-patient noise: p falls from 0.079 to 0.0028
- The differences look non-normal ($p = 0.033$, one patient gained 4.6 h), but the signed-rank test agrees ($p = 0.009$)

For a report Drug 2 gave 1.58 h more sleep on average (95% CI 0.7 to 2.46).

- 1 R for applied work
- 2 Categorical data
- 3 Two groups
- 4 ANOVA**
- 5 Regression
- 6 Choosing and reporting

Case study 3: three batches

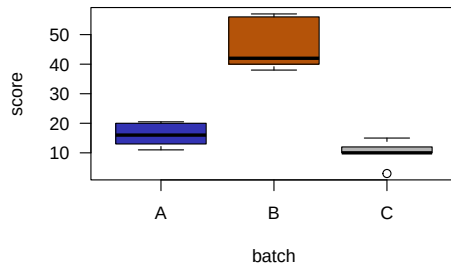
```
scores3 <- data.frame(  
  score = c(20, 20.5, 13, 16, 11,  
            40, 42, 38, 56, 57,  
            10, 15, 12, 10, 3),  
  batch = gl(3, 5, labels = c("A", "B", "C")))  
round(tapply(scores3$score, scores3$batch,  
             mean), 1)
```

```
##    A    B    C  
## 16.1 46.6 10.0
```

```
round(tapply(scores3$score, scores3$batch,  
             sd), 1)
```

```
##    A    B    C  
## 4.2 9.2 4.4
```

```
par(mar = c(4, 4, 1, 1))  
boxplot(score ~ batch, data = scores3,  
        col = c(col_main, col_alt, "grey70"),  
        las = 1)
```



Question: do the three batches have the same mean score?

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2)$$

independent

$$SS_{\text{total}} = SS_{\text{between}} + SS_{\text{within}}$$

$$F = \frac{SS_{\text{between}} / (k - 1)}{SS_{\text{within}} / (n - k)}$$

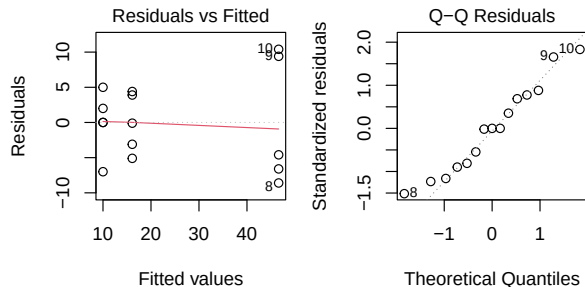
H_0 : all $k = 3$ batch means are equal.
A large F means the batches differ more than students within a batch do.

```
fit3 <- aov(score ~ batch, data = scores3)
print(summary(fit3), signif.stars = FALSE)
```

##	Df	Sum Sq	Mean Sq	F value	Pr(>F)
## batch	2	3845	1923	47.7	0.0000019
## Residuals	12	483	40		

Checking the ANOVA assumptions

```
par(mfrow = c(1, 2), mar = c(4, 4, 2, 1))  
plot(fit3, which = 1:2)
```



```
bartlett.test(score ~ batch,  
              data = scores3)$p.value
```

```
## [1] 0.2269
```

```
shapiro.test(residuals(fit3))$p.value
```

```
## [1] 0.6125
```

Reading it Residual spread looks similar across batches and the Q-Q plot is close to the line; neither test rejects (equal variances, normal residuals).

Which batches differ? Tukey's HSD

```
TukeyHSD(fit3)

##    Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = score ~ batch, data = scores3)
##
## $batch
##      diff      lwr      upr  p adj
## B-A   30.5   19.79  41.209 0.0000
## C-A   -6.1  -16.81   4.609 0.3167
## C-B  -36.6  -47.31 -25.891 0.0000
```

Why Tukey Each row gives a difference with a 95% *family-wise* interval and adjusted *p*-value; three separate 5% *t*-tests would push the chance of a false positive above 5%.

Reading case study 3

```
ss <- summary(fit3)[[1]][, "Sum Sq"]
ss[1] / sum(ss)          # eta squared

## [1] 0.8883

kruskal.test(score ~ batch,
              data = scores3)$p.value

## [1] 0.003698

oneway.test(score ~ batch,          # Welch
             data = scores3)$p.value

## [1] 0.000264
```

- $\eta^2 = 0.89$: batch membership accounts for 89% of the variation in scores
- Rank-based Kruskal–Wallis and Welch’s ANOVA (no equal-variance assumption) agree

For a report Mean scores differed between batches ($F_{2,12} = 47.7, p < 0.001$). Batch B was higher than A and C (Tukey); A and C did not differ (adjusted $p = 0.32$).

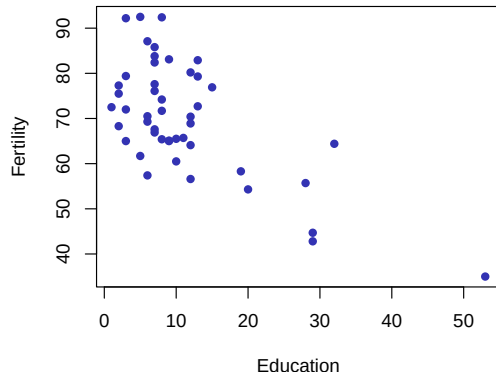
- 1 R for applied work
- 2 Categorical data
- 3 Two groups
- 4 ANOVA
- 5 Regression**
- 6 Choosing and reporting

Case study 4: fertility and education in Switzerland, 1888

- Built-in swiss data: 47 French-speaking provinces
- Fertility: standardised fertility index (0–100)
- Education: % of army draftees with education beyond primary school
- Also Agriculture (% of males in agriculture), Catholic (%), Infant.Mortality (% of live births dying within a year)

Question How is fertility related to education?

```
par(mar = c(4, 4, 1, 1))  
plot(Fertility ~ Education, data = swiss,  
     pch = 16, col = col_main)
```



$$\text{Fertility}_i = \beta_0 + \beta_1 \text{Education}_i + \varepsilon_i$$

$\varepsilon_i \sim N(0, \sigma^2)$, independent

`display()` shows each coefficient with its standard error; a coefficient about two standard errors from 0 is significant at roughly 5%. `confint()` gives the 95% intervals.

```
m1 <- lm(Fertility ~ Education, data = swiss)
display(m1)

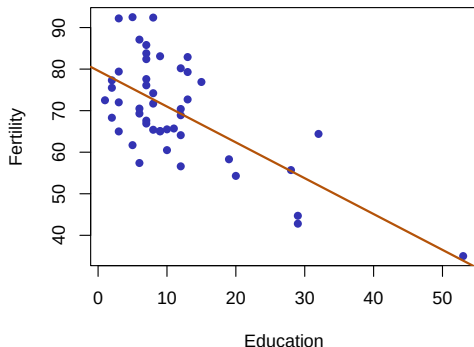
## lm(formula = Fertility ~ Education, data = swiss)
##               coef.est coef.se
## (Intercept)  79.61      2.10
## Education   -0.86      0.14
## ---
## n = 47, k = 2
## residual sd = 9.45, R-Squared = 0.44

round(confint(m1), 2)

##               2.5 % 97.5 %
## (Intercept)  75.37  83.85
## Education   -1.15 -0.57
```

Reading the simple regression

```
par(mar = c(4, 4, 1, 1))
plot(Fertility ~ Education, data = swiss,
     pch = 16, col = col_main)
abline(m1, col = col_alt, lwd = 2)
```



- Each extra percentage point of education goes with a fertility index lower by 0.86 (95% CI 0.57 to 1.15)
- $R^2 = 0.44$: education alone accounts for 44% of the variation between provinces
- The intercept (79.6) is the predicted index at 0% education

Careful These are province averages from observational data: an association, not a causal effect, and not a statement about individual women.

Multiple regression: holding other factors fixed

```
m2 <- lm(Fertility ~ Education + Agriculture + Catholic + Infant.Mortality, data = swiss)
display(m2)

## lm(formula = Fertility ~ Education + Agriculture + Catholic +
##      Infant.Mortality, data = swiss)
##              coef.est coef.se
## (Intercept)    62.10    9.60
## Education      -0.98    0.15
## Agriculture     -0.15    0.07
## Catholic         0.12    0.03
## Infant.Mortality 1.08    0.38
## ---
## n = 47, k = 5
## residual sd = 7.17, R-Squared = 0.70

round(summary(m2)$adj.r.squared, 3)

## [1] 0.671
```

Reading the multiple regression

- **Education:** -0.98 per percentage point, for provinces with the same agriculture, religion and infant mortality
- **Infant mortality:** $+1.08$ per point; **Catholic:** $+0.12$ per point
- **Agriculture:** -0.15 per point, once the others are in the model
- R^2 rises from 0.44 to 0.70

Partial effects A coefficient in a multiple regression compares provinces that differ in that variable alone. The education effect stayed close to its simple-regression value.

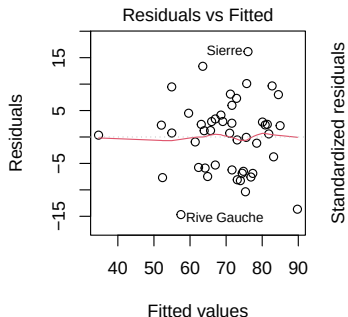
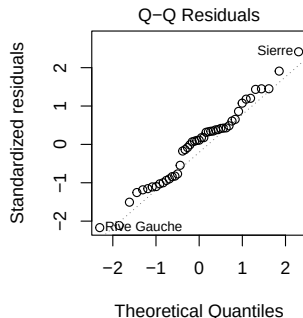
Choose variables with care
Pick predictors from subject knowledge, not by trying every combination.

Checking the model

```
par(mfrow = c(1, 2), mar = c(4, 4, 2, 1))  
plot(m2, which = 1:2, id.n = 2)
```

```
shapiro.test(residuals(m2))$p.value
```

```
## [1] 0.459
```



- Residuals vs fitted: no curve and no funnel, so linearity and constant variance look reasonable
- Q-Q plot close to the line; Shapiro-Wilk $p = 0.46$
- `id.n = 2` labels the two largest residuals: check those provinces

Prediction: confidence or prediction interval?

```
new <- data.frame(Education = c(5, 20))  
predict(m1, new, interval = "confidence")
```

```
##      fit   lwr   upr  
## 1 75.30 72.02 78.58  
## 2 62.36 58.54 66.19
```

```
predict(m1, new, interval = "prediction")
```

```
##      fit   lwr   upr  
## 1 75.30 55.99 94.60  
## 2 62.36 42.96 81.77
```

- **Confidence interval:** where the *mean* fertility of provinces with 5% education lies: 72 to 78.6
- **Prediction interval:** where *one new* province with 5% education is likely to fall: 56 to 94.6

Stay in range Do not predict far outside the observed education values (1 to 53%).

- 1 R for applied work
- 2 Categorical data
- 3 Two groups
- 4 ANOVA
- 5 Regression
- 6 Choosing and reporting**

Which method for which question?

Outcome and comparison	Method	R
categorical by categorical	χ^2 or Fisher's exact test	<code>chisq.test()</code> , <code>fisher.test()</code>
numeric, two groups	Welch t / rank-sum	<code>t.test()</code> , <code>wilcox.test()</code>
numeric, paired measurements	paired t / signed-rank	<code>t.test(paired = TRUE)</code>
numeric, three or more groups	one-way ANOVA + Tukey	<code>aov()</code> , <code>TukeyHSD()</code>
same, doubtful assumptions	Kruskal-Wallis / Welch ANOVA	<code>kruskal.test()</code> , <code>oneway.test()</code>
numeric by numeric	linear regression	<code>lm()</code> , <code>display()</code>

Same pattern every time Look at the data, check the assumptions, run the method, then report an estimate with its interval as well as the p -value.

Writing up: sentences from today's output

- **Categorical:** Diabetes was more common among older patients (6/8 vs 1/12; Fisher's exact test, $p = 0.004$).
- **Two groups:** Batch B outscored batch A by 30.5 points (95% CI 19.3 to 41.7; Welch $t = 6.77$, $df = 5.6$).
- **Paired:** Drug 2 gave 1.58 h more sleep than drug 1 (95% CI 0.7 to 2.46; paired $t_9 = 4.06$, $p = 0.003$).
- **ANOVA:** Batch means differed ($F_{2,12} = 47.7$, $\eta^2 = 0.89$).
- **Regression:** Each percentage point of education was associated with a 0.86-point lower fertility index (95% CI 0.57 to 1.15; $R^2 = 0.44$).

Good habits Name the test, give the statistic with its df, the exact p -value (not just " $p < 0.05$ "), and the estimate with a confidence interval in the units of the data.

Key take-aways

- 1 **The design chooses the test:** independent or paired, two groups or more, categorical or numeric outcome.
- 2 **Check assumptions, but read them sensibly:** small samples give weak tests; plots matter more.
- 3 **Have a robust fallback:** Fisher, Welch and rank-based tests are one line away.
- 4 **Ignoring pairing throws information away** (p went from 0.079 to 0.0028).
- 5 **Report estimates with intervals**, and say what the numbers mean in words.
- 6 **Association is not causation**, above all in observational data.

- Dalgaard, P. (2008). *Introductory Statistics with R*, 2nd Ed., Springer, New York.
- Faraway, J. J. (2014). *Linear Models with R*, 2nd Ed., CRC Press, Boca Raton.
- Fox, J., and Weisberg, S. (2019). *An R Companion to Applied Regression*, 3rd Ed., Sage, Thousand Oaks.
- Gelman, A., and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, Cambridge.
- Verzani, J. (2014). *Using R for Introductory Statistics*, 2nd Ed., CRC Press, Boca Raton.
- 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