Course 1 · Week 1 — Toolchain & data hygiene

Cheatsheet — biostats_courses

R. Heller

The research workflow

`Question → Measurements → Design → Acquisition → Description → Analysis → Interpretation → Validation → Knowledge`

Biological vs statistical significance: "$p < 0.05$" ≠ "effect matters".

R / RStudio / Quarto / renv

Task	Command
Install packages	`install.packages("pkg")`
Pin project	`renv::init()` / `renv::snapshot()` / `renv::restore()`
Render one file	`quarto render path/to/file.qmd`
Preview with live reload	`quarto preview`
Render only slides	`quarto render file.qmd --to revealjs`

Data types and scales

Scale	Example	Statistic
Nominal	sex, site	counts, proportions
Ordinal	pain 0–10	median, IQR, rank tests
Interval	°C	mean, SD
Ratio	kg, time	mean, SD, ratios

Accuracy ≠ precision. Bias shifts the target; variance spreads the shots.

Tidy data

1. One variable per column.
2. One observation per row.
3. One observational unit per table.

`dplyr` verbs

```
df |>
  filter(age >= 18) |>
  select(id, age, sbp) |>
  mutate(sbp_z = (sbp - mean(sbp)) / sd(sbp)) |>
  group_by(arm) |>
  summarise(mean_sbp = mean(sbp, na.rm = TRUE), n = n())
```

Joins: `left_join`, `inner_join`, `anti_join` by a key.

Missingness: `is.na()`, `drop_na()`, `tidyr::complete()`.

`ggplot2` grammar

```
ggplot(data, aes(x, y, colour = group)) +
  geom_point() +
  facet_wrap(~ factor) +
  labs(x = "x label", y = "y label") +
  theme_minimal()
```

Layer order: `data → aes → geom → facet → scales → labels → theme`.

Plot families

- One variable continuous ⟶ **histogram**, **density**.
- Two continuous ⟶ **scatter**, sometimes binned / smoothed.
- One categorical + one continuous ⟶ **boxplot**, **violin**, **strip**.
- Two categorical ⟶ **bar with fill**, **mosaic**.

Decision rule for Week 1

- Cannot read the data into a tibble ⟶ fix data entry first.
- Tibble reads but has missing or mixed types ⟶ clean before modelling.
- Plot is ugly or cluttered ⟶ trust the plot, not the summary statistic.

Common pitfalls

- Running statistics on untyped / partially-missing data.
- Adjusting colour scales instead of simplifying the chart.
- Reporting mean + SD for a skewed variable.
- Forgetting `set.seed()` in any simulation chunk.

Further reading

- Wickham, *R for Data Science* (2e).
- Posit ggplot2 and dplyr cheatsheets.