Course 1 · Week 2 — Describing data, probability, distributions

Cheatsheet — biostats_courses

R. Heller

Descriptive statistics

Situation	Summary
Roughly symmetric, no outliers	mean ± SD
Skewed or has outliers	median, IQR, range
Categorical	counts and proportions
Grouped	Table 1 via `gtsummary::tbl_summary()`

Contingency tables

```
table(df$x, df$y)
prop.table(table(df$x, df$y), margin = 1)
```

Bayes' theorem

$P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$

In odds form: **posterior odds = prior odds × likelihood ratio.**

Diagnostic testing

Metric	Formula
Sensitivity	TP / (TP + FN)
Specificity	TN / (TN + FP)
PPV	TP / (TP + FP)
NPV	TN / (TN + FN)
LR+	Se / (1 − Sp)
LR−	(1 − Se) / Sp

PPV collapses at low prevalence — even a 95%-accurate test is mostly false positives.

Discrete distributions

Distribution	Use	R
Bernoulli(p)	single trial	`rbinom(n, 1, p)`
Binomial(n, p)	# successes	`dbinom / pbinom / rbinom`
Poisson(λ)	rare events	`dpois / ppois / rpois`
Negative binomial	overdispersed counts	`MASS::rnegbin`, `dnbinom`

Binomial ⟶ Poisson as n grows and p shrinks with $np = \lambda$.

Continuous distributions

Distribution	Use
Normal(μ, σ)	almost everything via the CLT
Student $t(\nu)$	inference about means, small n
$\chi^2(\nu)$	variance, counts, GoF
F(ν_1, ν_2)	variance ratios, ANOVA
Exponential(λ)	time between events, memoryless

R uses the `d / p / q / r` prefix: density, CDF, quantile, random.

Q-Q plot

```
ggplot(df, aes(sample = x)) + stat_qq() + stat_qq_line()
```

S-shape ⟶ heavy tails. U-shape ⟶ skew. Plot before the test.

Decision rule for Week 2

- Reporting a mean? First check a histogram.
- Testing a proportion? First sketch a 2×2 table.
- Choosing a distribution? Simulate before believing.

Common pitfalls

- Quoting PPV without disclosing prevalence.
- Assuming normality because $n > 30$.
- Using mean ± SD on skewed data.

Further reading

- Altman, *Practical Statistics for Medical Research*.
- Gelman et al., *Bayesian Data Analysis*, ch. 1.