Course 1 · Week 3 — Sampling, estimation, one-sample inference

Cheatsheet — biostats_courses

R. Heller

Populations vs samples

- **Parameter** = truth about the population (μ, σ, π).
- **Estimator** = statistic computed on a sample ($\bar{x}$, s, $\hat{p}$).
- **Bias** = $E[\hat{\theta}] - \theta$. **MSE** = bias² + variance.

Central limit theorem

For iid samples with finite variance, as $n \to \infty$: $\bar{X} \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$

Holds roughly by $n \approx 30$ for most non-pathological distributions.

Standard error

$\mathrm{SE}(\bar{X}) = \sigma/\sqrt{n}$, estimated by $s/\sqrt{n}$.

Bootstrap

```r
B <- 2000
boot_mean <- replicate(B, mean(sample(x, replace = TRUE)))
quantile(boot_mean, c(0.025, 0.975))   # 95% percentile CI
sd(boot_mean)                           # bootstrap SE
```

Permutation test

```r
obs  <- mean(a) - mean(b)
pool <- c(a, b)
null <- replicate(5000, {
  idx <- sample(seq_along(pool), length(a))
  mean(pool[idx]) - mean(pool[-idx])
})
mean(abs(null) >= abs(obs))   # two-sided p
```

Maximum likelihood

Given data x and model $f(x;\theta)$:

- $\ell(\theta) = \sum \log f(x_i;\theta)$.
- $\hat{\theta}_{\mathrm{MLE}} = \arg\max_\theta \ell(\theta)$.
- Fisher information $I(\theta)$; asymptotic SE = $1/\sqrt{I(\hat{\theta})}$.

One-sample tests (five-step template)

Hypothesis → Visualise → Assumptions → Conduct → Conclude.

Test	R	Assumptions
One-sample t	`t.test(x, mu = 0)`	roughly normal or large n
One-proportion	`prop.test(k, n, p = 0.5)` or `binom.test`	$np, n(1-p) \geq 5$ for `prop.test`

Hypothesis-testing vocabulary

- **Type I**: reject true H_0 (probability α).
- **Type II**: accept false H_0 (probability β).
- **Power** = $1 - \beta$.
- A *p*-value is **not** the probability that H_0 is true.

Decision rule for Week 3

- Want a CI? Bootstrap first, then ask whether a formula applies.
- Want a test? State H_0 and α in writing before running it.
- Want to know if *n* is big enough? Simulate the CLT for your outcome.

Common pitfalls

- Reporting SE when you meant SD (or vice versa).
- Using the 95% CI to "prove" there is no effect.
- Running many tests and quoting the smallest *p*-value.

Further reading

- Harrell, *Biostatistics for Biomedical Research*, ch. 3–5.
- Efron & Tibshirani, *An Introduction to the Bootstrap*.