1. Fundamentals

GHV Chapters 4-5

DATA 335 – Univerrsity of Calgary – Winter 2025

Statistical models and statistical inference

- A *statistical model* is a probability distribution.
- A statistical model is characterized by unknown and often unknowable numbers called *parameters*. They are our quantities of interest.
- Statistical models facilitate *statistical inference* – procedures for turning data into parameters estimates, avatars for their uncertainty.
  - Frequentist inference: point estimation, standard errors, confidence intervals, hypothesis tests
  - Bayesian inference: posterior distribution

Estimators for mean and variance

- Let $x_0, \ldots, x_{n-1}$ be a *random sample*[1] from the a model (distribution) F with mean μ and variance σ^2.
- The *sample mean*

$$\bar{x} = \frac{x_0 + \cdots + x_{n-1}}{n}$$

  estimates μ.
- The *sample variance*

$$s^2 = \frac{1}{n-1} \sum_{i<n} (x_i - \bar{x})^2$$

  estimates σ^2.

[1] independent and identically distributed

Estimators have distributions

- Since the x_i are random variables, the estimators $\bar{x}$ and s^2 are computed computed from them are, too.
- In particular, they have distributions.
- Distributions of random variables computed from random samples from other distributions are called *sampling distributions*.
- **(Demo)** Visualizing sampling distributions

Standard error

- The *standard error* of a random variable x, denoted $\text{se}(x)$, is the standard deviation of its distribution.
- $\text{se}(x)$ is the fundamental numerical distillation of the uncertainty in x.

The sampling distribution of the mean

- If $x_0, \ldots, x_{n-1}$ is a random sample drawn from a distribution with mean μ and standard deviation σ, then

$$\mathrm{se}(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$

- By the *Central Limit Theorem*, the distribution of $\bar{x}$ is approximately[2] normal, with mean μ and standard deviation $\mathrm{se}(\bar{x})$.
- Said differently,

$$z = \frac{\bar{x} - \mu}{\mathrm{se}(\bar{x})} \longrightarrow N(0, 1)$$

as $n \to \infty$.

[2] the larger the sample size, the better the approximation

Normal approximation to the binomial proportion

- When $y \sim \text{Bin}(n, p)$, we estimate the binomial proportion p by
$$\hat{p} = \frac{y}{n}.$$
- $\text{Bin}(n, p)$-RVs are *sums* of $\text{Ber}(p)$-RVs, making $\hat{p}$ the *average* of $\text{Ber}(p)$-RVs. Thus, $\hat{p}$ is a *sample mean*.
- Since $\text{Ber}(p)$ has standard deviation $\sqrt{p(1-p)}$,
$$\text{se}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$
and, by the central limit theorem,
$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right) \qquad \text{(approx.)}.$$

Normal approximation to the binomial distribution

- Since $y = n\hat{p}$, we have

$$\mathrm{Bin}(n, p) = \text{distribution of } y \approx N\left(np, np(1-p)\right).$$

- **(Demo)** Normal approximation to the binomial distribution

Cumulative distribution and percent point functions

- The *(cumulative) distribution function* of a random variable x is defined by
$$\mathrm{cdf}_x(u) = \mathbb{P}[x \leq u].$$
- Its inverse function is called the *percent point function*.
$$\mathrm{ppf}_x(v) = u \Longleftrightarrow \mathbb{P}[x \leq u] = v$$
Also known as the *quantile function* or *inverse (cumulative) distribution function*.

Confidence intervals for sample means

- Define
$$z_{\alpha/2} = \mathrm{ppf}_{N(0,1)}(1 - \alpha/2).$$
- If n is sufficiently large, then
$$\frac{\bar{x} - \mu}{\mathrm{se}(\bar{x})} \sim N(0,1) \qquad \text{(approx.)}$$
by the central limit theorem, implying
$$\begin{aligned} 1 - \alpha &\approx \mathbb{P}\left[\left|\frac{\bar{x} - \mu}{\mathrm{se}(\bar{x})}\right| < z_{\alpha/2}\right] \\ &= \mathbb{P}[\bar{x} - z_{\alpha/2}\,\mathrm{se}(\bar{x}) < \mu < \bar{x} + z_{\alpha/2}\,\mathrm{se}(\bar{x})]. \end{aligned}$$
- The interval with endpoints $\bar{x} \pm z_{\alpha/2}\,\mathrm{se}(\bar{x})$ is called the $100(1-\alpha)\%$-*confidence interval* for μ associated to $\bar{x}$.
- **(DEMO)** Confidence intervals for sample means

Estimating the standard error

- Since σ is typically unknown, so is

$$\mathrm{se}(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$

- To estimate it, plug in the sample standard deviation for σ:

$$\widehat{\mathrm{se}}(\bar{x}) = \frac{s}{\sqrt{n}}.$$

- If n is sufficiently large, you can use this estimate to constructing confidence intervals as above:

$$100(1-\alpha)\%\text{-CI} = [\bar{x} - z_{\alpha/2}\widehat{\mathrm{se}}(\bar{x}),\ \bar{x} + z_{\alpha/2}\widehat{\mathrm{se}}(\bar{x})]$$

  If n is small, you can't!

- **(DEMO)** Bad coverage for small n

Confidence intervals from the t-distribution

- **If the common distribution of the x_i is normal**, then

$$\frac{\bar{x} - \mu}{\widehat{\mathrm{se}}(\bar{x})} \sim t_{n-1}$$

  (t-distribution with $n - 1$ degrees of freedom).
- **(DEMO)** Simulating t_{n-1}
- We can use the percent-point function of t_{n-1} to construct confidence intervals for $\bar{x}$. Set

$$t_{n-1,\alpha/2} = \mathrm{ppf}_{t_{n-1}}(1 - \alpha/2).$$

  and define

$$100(1-\alpha)\%\text{-CI} = [\bar{x} - t_{n-1,\alpha/2}\widehat{\mathrm{se}}(\bar{x}),\ \bar{x} + t_{n-1,\alpha/2}\widehat{\mathrm{se}}(\bar{x})].$$

- This is valid for small n! For large n, we have $t_{n-1} \approx N(0,1)$ and $z_{\alpha/2} \approx t_{n-1,\alpha/2}$.