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*

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

  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 partifcular, 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 $\mathrm{se}(x)$, is the standard deviation of its distribution.
- $\mathrm{se}(x)$ is the fundamental numerical distillation of the uncertainty in x.

The sampling distribution of the mean

- If x_0, ..., 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,

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

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

Normal approximation to the binomial proportion

- When $y \sim \mathrm{Bin}(n, p)$, we estimate the binomial proportion p by

$$\hat{p} = \frac{y}{n}.$$

- $\mathrm{Bin}(n, p)$-RVs are *sums* of $\mathrm{Ber}(p)$-RVs, making $\hat{p}$ the *average* of $\mathrm{Ber}(p)$-RVs. Thus, $\hat{p}$ is a *sample mean*.
- Since $\mathrm{Ber}(p)$ has standard deviation $\sqrt{p(1-p)}$,

$$\mathrm{se}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

and, by the central limit theorem,

$$\text{distribution of } \hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right).$$

Normal approximation to the binomial distribution

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

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

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

Cumulative distribution 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*.
$$\mathrm{ppf}_x(v) = u \iff \mathbb{P}[x \leq u] = v$$
Also known as the *quantile function* or *inverse (cumulative) distribution function*.

Confidence intervals for sample means

- Define
$$z_{\alpha/2} = \text{ppf}_{N(0,1)}(1 - \alpha/2).$$
- Let $\bar{x}$ be a sample mean with expected value μ. Then it's approximately normally distributed by the CLT and
$$\mathbb{P}[\bar{x} - z_{\alpha/2}\,\text{se}(\bar{x}) < \mu < \bar{x} + z_{\alpha/2}\,\text{se}(\bar{x})] = 1 - \alpha$$
- The interval with endpoints $\bar{x} - z_{\alpha/2}\,\text{se}(\bar{x})$ is called the $100(1-\alpha)\%$-confidence interval for μ associated to $\bar{x}$.
- **(DEMO)** Confidence intervals for sample means