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.
- Standard error of the mean:

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

Standard error of the binomial proportion

- When $y \sim \mathrm{Bin}(n, p)$, we estimate the binomial proportion p by
$$\hat{p} = \frac{y}{n}.$$
- $\hat{p}$ is a $\mathrm{Ber}(p)$-sample mean: $\mathrm{Bin}(n, p)$-RVs are *sums* of $\mathrm{Ber}(p)$-RVs, making $\hat{p}$ the *average* of such.
- $\mathrm{Ber}(p)$ has standard deviation $\sigma = \sqrt{p(1-p)}$, so
$$\mathrm{se}(\hat{p}) = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}.$$
- As p is unknown, we estimate $\mathrm{se}(\hat{p})$ by plugging in $\hat{p}$ for p:
$$\mathrm{se}(\hat{p}) \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Example: Inference for binomial proportions

- In a survey of university students, 57 out of 146 of male respondents say they regularly tweeze their eyebrows.
- Model the number y of male student tweezers by $\text{Bin}(n, p)$ with $n = 146$.
- Estimate p:
$$\hat{p} = \frac{y}{n} = \frac{57}{146} = 0.39$$
- Estimate $\text{se}(\hat{p})$:
$$\text{se}(\hat{p}) \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.39(1-0.39)}{146}} = 0.04$$

Approximate normality of sample means

- By the *Central Limit Theorem* (CLT), the distribution of $\bar{x}$ is approximately $N(\mu, \text{se}(\bar{x})^2)$-distributed if n is sufficiently large.
- Samples means being approximately normal, we can use *normal theory* to perform related inference tasks.
- **(DEMO)** Illustrate CLT for sample means

Confidence intervals for binomial proportions

- Use standard z-table values to construct confidence intervals for $\hat{p}$. With $z_{\alpha/2} = \text{ppf}_{N(0,1)}(1-\alpha/2)$,

$$\begin{aligned} 100(1-\alpha)\%\text{-CI} &= [\hat{p} \pm z_{\alpha/2}\,\text{se}(\hat{p})] \\ &\approx \left[\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]. \end{aligned}$$

- Example: The 95%-CI for the proportion of male student eyebrow tweezers is

$$[0.39 \pm 1.96 \cdot 0.04] = [0.39 \pm 0.08].$$

Combining means and proportions

- Stantard errors of independent random variables combine according to the Pythagoream theorem:

$$\mathrm{se}(x \pm y) = \sqrt{\mathrm{se}(x)^2 + \mathrm{se}(y)^2}.$$

- More generally, the standard error of a *weighted sum* of independent random variables is:

$$\mathrm{se}\left(\sum_i w_i x_i\right) = \sqrt{\sum_i w_i^2 \,\mathrm{se}(x_i)^2}$$

Example: Gender gap (GHV §4.2, pp. 52-53)

- In a survey of voting intentions, 57% of 400 men 45% of 600 women say they plan to vote for the Republican candidate in an upcoming election.
- Model the men's and women's counts by $\text{Bin}(400, p)$ and $\text{Bin}(600, q)$, respectively.
- Estimate p, q, $\text{se}(p)$, and $\text{se}(q)$:

$$\hat{p} = 0.57, \quad \text{se}(\hat{p}) = 0.025, \quad \hat{q} = 0.45, \quad \text{se}(\hat{q}) = 0.020$$

- We get corresponding estimates for the *gender gap* and its standard error:

$$\hat{p} - \hat{q} = 0.12, \quad \text{se}(\hat{p} - \hat{q}) = \sqrt{0.025^2 + 0.020^2} = 0.032$$

- The 95%-CI for this gender gap is

$$[(\hat{p} - \hat{q}) \pm 1.96\,\text{se}(\hat{p} - \hat{q})] = [0.12 \pm 0.06]\,.$$

Example: A goodness of fit test (cf. GHV §4.6)

- The 1000 votes in an election with two candidates, A and B, are tallied batches of 100. The counters report the following batch tallies for candidate A:

  61, 64, 54, 61, 59, 58, 65, 62, 61, 59

  Candidate B protests, suggesting that these results exhibit implausible uniformity. Does he have a case?

- Let y_i be the i-th tally and let $\bar{y}$ be their average.
- Implausible uniformity would manifest as an implausibly small value of the *test statistic*

$$t = \sum_i (y_i - \bar{y})^2.$$

- The observed vote tallies give $t = 88$.
- We assess the implausibility observing $t = 88$ by studying the distribution of t under the assumption of a fair election in which candidate A has 60% support.
- **(DEMO)** Goodness of fit