Concepts of Statistics

STAT 135

Instructor: Rebecca Barter

Kelvin Lee

UC Berkeley

Contents

1 Introduction to Inference

1.1 Parameters, populations, and estimates

Definition 1.1.1 (Population)**.** A *population* is the complete set of individuals or entities that we are interested in. We usually only have data on a subset of them.

Definition 1.1.2 (Parameter)**.** A *parameter* is any quantifiable feature of a population.

1.1.1 Common parameters of interest in statistics

The most common population parameters we are interested in are:

1. *Mean*
2. *Proportions* (averages of binary data)

1.1.2 Inference

Definition 1.1.3 (Inference)**.** *Inference* involves using data to compute an estimate of a population parameter of interest.

Remark. The population should always be defined in the context of where the results will be applied. Accurate inference is only possible when the data is representative of the population (i.e., the data is **unbiased**).

1.2 Bias in data

Example 1.2.1 (Survivorship bias)**.** In the figure below, each dot corresponds to a place that a returning plane has been hit. Where should you reinforce the plane's armor?

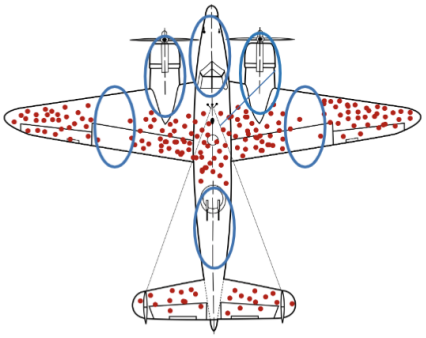

Figure 1.1: If the bullets hit the top circled area, the plane goes down and does not return. The data is a biased representation of where the planes are getting hit.

Definition 1.2.2 (Biased)**.** Data is *biased* if it does not reflect the population it was designed to represent. **Biased data leads to biased results**.

Example 1.2.3. If AI-driven skin cancer detection is built only using patients with light skin tones but is used to detect skin cancer in racially diverse patients, the algorithm might be biased.

Random Variables We use **random variables** to represent all possible values that an unknown quantity could take when we observe it.

1.3 Evaluating Estimators

1.3.1 Parameter bias

Definition 1.3.1 (Bias)**.** The *bias* of an estimate, $\hat{\theta}$, of population parameter, θ, is

$$Bias(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta.$$

A parameter estimate is **unbiased** if the bias is 0.

Example 1.3.2 (Sample mean is unbiased)**.** The sample mean, $\hat{\mu} = \frac{1}{n}\sum_{i=1} X_i$ is an unbiased estimate of μ.

Proof.

$$\begin{aligned}
\mathbb{E}[\hat{\mu}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] \\
&= \frac{1}{n} n\mu \\
&= \mu.
\end{aligned}$$

□

Question. A parameter estimate from a sample is biased if it is not equal to the underlying population quantity it is supposed to represent?

Answer. False. Even if the parameter estimate is unbiased, there is no guarantee that the parameter computer from a specific sample of data points will be exactly equal to the underlying population parameter.

Remark. Unbiasedness is referring to the expected value of the estimate, not the sample estimate itself.

1.3.2 Parameter variance

Definition 1.3.3 (Variance)**.** The *variance* of a parameter estimate tells us how much it generally changes across alternative equivalent versions of the data. The *variance* of an estimate, $\hat{\theta}$, of population parameter, θ, is

$$\text{Var}(\hat{\theta}) = \mathbb{E}[\hat{\theta}^2] - \mathbb{E}[\hat{\theta}]^2.$$

Theorem 1.3.4 (Variance of sample mean)**.**
The variance of sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is $\frac{\sigma^2}{n}$.

Proof.

$$\begin{aligned}\text{Var}(\hat{\mu}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left(X_i\right)\\ &= \frac{1}{n^2}n\sigma^2\\ &= \frac{\sigma^2}{n}.\end{aligned}$$

□

1.3.3 Mean Square Error

Definition 1.3.5. The Mean Squared Error (MSE) is a measure of how "good" an estimate $\hat{\theta}$ is. The MSE is

$$MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2].$$

Theorem 1.3.6 (Bias-Variance Decomposition of MSE)**.**
The MSE can be decomposed into the sum of squared bias and the variance of $\hat{\theta}$:

$$MSE(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2.$$

Proof.

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\ &= \mathbb{E}[\hat{\theta}^2] - 2\theta\mathbb{E}[\hat{\theta}] + \theta^2 \\ &= \mathrm{Var}(\hat{\theta}) + \mathbb{E}[\hat{\theta}]^2 \\ &= \mathrm{Var}(\hat{\theta}) + \mathbb{E}[\hat{\theta}]^2 - 2\theta\mathbb{E}[\hat{\theta}] + \theta^2 \qquad = \mathrm{Var}(\hat{\theta}) + (\mathbb{E}[\hat{\theta}] - \theta)^2 \\ &= \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta})^2. \end{aligned}$$

□

1.4 Techniques for estimating bias, variance, and MSE from a single data sample

1.4.1 Non-parametric bootstrap

- Treat the original sample as the population.
- Treat the bootstrapped sample as the sample.
- Draw samples from our sample **with replacement** (to ensure same size as the original sample).
- Use these to estimate the bias and variance

$$Bias(\hat{\mu}) \approx \frac{1}{N}\sum_{k=1}^{N} \hat{\mu}_k^* - \hat{\mu},$$

$$Var(\hat{\mu}) \approx \frac{1}{N}\sum_{k=1}^{N} (\hat{\mu}_k^* - \overline{\hat{\mu}^*})^2$$

where N is the number of bootstrapped samples and $\hat{\mu}_k^*$ is the mean of kth bootstrapped sample.

1.4.2 Parametric bootstrap

- Data distribution is known.
- Approximate distribution using $\hat{\mu}$ and $\hat{\sigma}$.
- Use the distribution with the estimated parameters to draw **parametric bootstrap** samples.
- The formulae for bias and variance estimates for parametric bootstrap are the same as the non-parametric version.

1.5

Theorem 1.5.1 (Law of Large Numbers).
If $X_1, X_2, \ldots, X_n$ is an i.i.d sample, then

$$\overline{X} \to \mathbb{E}[X_1]$$