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?

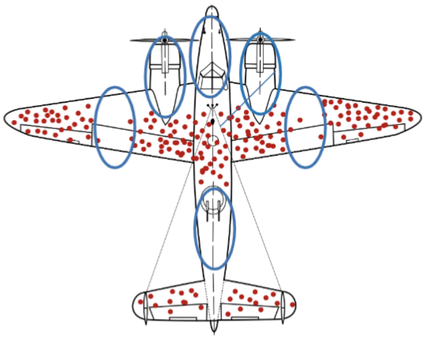

Figure 1.1: If the bullets hit the top circled area, the plane goes down and does not return. The data is a biased representation of where the planes are getting hit.

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

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

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.

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

□

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 $\hat{\theta}$:

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

Proof.

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

□

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.

1.4.3 Law of Large Numbers

Theorem 1.4.1 (Law of Large Numbers)**.**
If $X_1, X_2, \ldots, X_n$ is an IID sample, then

$$\overline{X}_n \xrightarrow{P} \mathbb{E}[X_1] \qquad \text{as } n \to \infty.$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} \mathbb{E}[X_1^k] \qquad \text{as } n \to \infty.$$

1.4.4 Central Limit Theorem

Theorem 1.4.2 (Central Limit Theorem)**.**
If $X_1, \ldots, X_n$ is an IID sample form a population with mean μ and standard deviation σ, then

$$\overline{X}_n \xrightarrow{D} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{as } n \to \infty.$$

2 Maximum Likelihood Estimation

2.1 Likelihood Functions

Definition 2.1.1 (Maximum Likelihood Estimation)**.** *Maximum likelihood estimation* is a generating technique for identifying reasonable estimates of the parameters form any distribution. The idea is choose the value parameter based that is most likely to have led to our observed data.

Definition 2.1.2. The *likelihood function* (θ) corresponds to the probability of observing the particular data in our sample for various values of θ.

$$\begin{aligned} lik(\theta) &= \mathbb{P}(X_1 = x_1, \dots, X_n = x_n) \\ &= \prod_{i=1}^{n} f_\theta(X_i). \end{aligned}$$

Definition 2.1.3 (Maximum Likelihood Estimate)**.** The *maximum likelihood estimate* $\hat{\theta}_{MLE}$ of a parameter θ is the value that maximizes the likelihood function based on the observed data.

2.2 Steps for performing MLE

1. $lik(\theta) = \prod_i f_\theta(X_i)$.
2. $\ell(\theta) = \log\left(\prod_i f_\theta(X_i)\right) = \sum_i \log(f_\theta(X_i))$.
3. Differentiate the log-likelihood function with respect to θ, set to zero, and solve for θ.

Example 2.2.1. Normal(μ) If $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\begin{aligned} lik(\mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{(2\pi\sigma)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right\}. \end{aligned}$$

$$\begin{aligned} \ell(\mu) &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \\ \implies \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}. \end{aligned}$$

We see that sample mean is actually a MLE estimator.

2.3 Properties of MLE Estimators

- **Consistency**: as the sample size gets larger the MLE approaches the true parameter value.
- **Normality**: as the sample size gets larger the distribution of the MLE (as in if you were able to compute various versions of the MLE from many different random samples) becomse Normal.

2.3.1 Consistency

Definition 2.3.1 (Consistent). An estimate $\hat{\theta}_n$ of θ is *consistent* if

$$\hat{\theta}_n \xrightarrow{P} \theta \qquad \text{as } n \to \infty.$$

where $\hat{\theta} \xrightarrow{P} \theta$ means that for all $\epsilon > 0$,

$$\mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon) \to 0 \qquad \text{as } n \to \infty.$$

Theorem 2.3.2 (Consistency of the MLE).
The MLE $\hat{\theta}_{MLE,n}$ is a **consistent estimator** of the parameter, θ, that it is estimating, which means that

$$\hat{\theta}_{MLE,n} \xrightarrow{P} \theta \qquad \text{as } n \to \infty.$$

Sketch. □

The consistency of the MLE implies that the MLE is **asymtotically unbiased**:

$$\mathbb{E}[\hat{\theta}_{MLE,n}] \to \theta \qquad \text{as } n \to \infty.$$

Remark. Hence, we see that consistency is a stronger statement.

Theorem 2.3.3 (Continuous mapping theorem).
For any continuous function g, if $\hat{\theta} \xrightarrow{P} \theta$ as $n \to \infty$, then

$$g(\hat{\theta}) \xrightarrow{P} g(\theta) \qquad \text{as } n \to \infty.$$

Example 2.3.4. $\overline{X} \xrightarrow{P}\to \mu$ as $n \to \infty$, implies that $\overline{X}^2 \xrightarrow{P} \mu^2$ as $n \to \infty$.

2.3.2 Asymptotic normality of the MLE

Theorem 2.3.5 (The MLE is asymptotically Normal)**.**
The MLE is asymptotically *normal.* If $\hat{\theta}_{ML,n}$ is the ML estimate of a parameter θ whose true value is θ_0, then as $n \to \infty$, we have that

$$\hat{\theta}_{ML,n} \xrightarrow{D} \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right) \qquad \text{as } n \to \infty,$$

where $I(\theta_0)$ is the *Fisher Information.*

The mean follows from the consistency of the MLE.

Definition 2.3.6 (Fisher information)**.** The *Fisher information* is defined by

$$I(\theta_0) = \mathbb{E}\left[\left(\frac{d}{d\theta}\log(f_\theta(x))\bigg|_{\theta_0}\right)^2\right]$$

or

$$I(\theta_0) = -\mathbb{E}\left[\frac{d^2}{d^2\theta}\log(f_\theta(x))\bigg|_{\theta_0}\right].$$

It measures how "peaked" some function $\ell(\theta)$ is around θ_0. If $I(\theta_0)$ is large, then it is easier to detect θ_0, which implies lower variance.

2.3.3 Delta Method

Theorem 2.3.7 (Delta method)**.**
By CLT, we know that $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$ as $n \to \infty$.
For any function g such that $g'(\mu)$ exists and is non-zero, then

$$\sqrt{n}(g(\overline{X}_n)) - g(\mu)) \xrightarrow{D} \mathcal{N}(0, \sigma^2 g'(\mu)^2)$$

3 Method of Moments

3.1 Moments

Definition 3.1.1 (Moment)**.** The k-th moment of X is

$$\mu_k = \mathbb{E}[X^k].$$

Another way to formulate a parameter estimate is by relating **sample moments** to the **theoretical moments**. For example,

$$\text{Theoretical moment: } \mathbb{E}[X] \qquad \text{Sample moment: } \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

3.2 MOM vs MLE

Theorem 3.2.1 (MOM estimators are consistent)**.**

$$\hat{\theta}_{MOM} \xrightarrow{P} \theta_0 \qquad \text{as } n \to \infty.$$

Proof. This follows from the LLN for moments: if $X_1, \dots, X_n$ is an IID sample, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i^K \xrightarrow{P} \mathbb{E}[X_1^k] \qquad \text{as } n \to \infty.$$

□

Remark. MOM estimators don't have limiting distribution results like the MLE. That's why MLE are used more often.

3.3 Cramer-Rao lower bound

While the MLE and MOM often yield the same estimatros, they will sometimes differ.

Question. How should we compare two possible estimators for the same parameter?

Answer. Compare their bias/ variance.

Theorem 3.3.1 (Cramer-Rao lower bound).
If X_i are IID from a distribution with density f_θ, under smoothness conditions on f_θ, we have that: if $\hat{\theta}$ is an unbiased estimator for θ, then

$$\operatorname{Var}(\hat{\theta}) \geq \underbrace{\frac{1}{nI(\theta)}}_{\text{variance of MLE}}.$$

Interpretation: This result essentially states that the price to pay for having an unbiased estimator is a certain amount of variance.

Remark. This means that the MLE has the lowest possible variance among unbiased estimators!

3.4 Efficiency

Definition 3.4.1 (Efficiency). Given two estimators $\hat{\theta}$ and $\tilde{\theta}$ of a parameter θ, the *efficiency* of $\hat{\theta}$ relative to $\tilde{\theta}$ is

$$\operatorname{eff}(\hat{\theta}, \tilde{\theta}) = \frac{\operatorname{Var}(\tilde{\theta})}{\operatorname{Var}(\hat{\theta})}.$$

If $\operatorname{eff}(\hat{\theta}, \tilde{\theta}) \leq 1$, then $\operatorname{Var}(\hat{\theta}) \geq \operatorname{Var}(\tilde{\theta})$, which implies that $\hat{\theta}$ is *less efficient* that $\tilde{\theta}$.

3.4.1 Efficient estimators

Definition 3.4.2 (Efficient estimator). An *unbiased* estimator that achieves the Cramer-Rao lower bound is called *efficient*. The Cramer-Rao lower bound is

$$\operatorname{Var}(\hat{\theta}) = \frac{1}{nI(\theta)}.$$

Remark. Unbiased estimators cannot do better in terms of variance than the Cramer-Rao lower bound. If an estimator actually does better than the lower bound, then it is biased.

Remark. The MLE is **asymptotically efficient**. (not necessarily efficient for finite samples)

3.5 Sufficiency

$X_1, \dots, X_n$ can be high-dimensional and might be expensive to store. It'd be neat if there was a function T of the data (statistic) that contains all of the information about a parameter of interest.

Definition 3.5.1 (Sufficient). A statistic T is *sufficient* for θ if $\mathbb{P}((X_i)_{i=1}^n \mid T(X_1, \dots, X_n) = t)$ does not depend on θ for any t.

Example 3.5.2 (Examples of statistics). $\overline{X}_n, \mathrm{Var}(X_n), \max\{X_1, \dots, X_n\}$.

Suppose that $X_1, \dots, X_n \sim F(\theta)$. Then $T(X_1, \dots, X_n)$ is a *sufficient statistic* for θ if the statistician who knows the value of T can do just a good job of estimating the unknown parameter θ as the statistician who knows the entire random sample.

Theorem 3.5.3 (The Factorization Theorem).
A necessary and sufficient condition for T to be sufficient for θ is

$$f_\theta(x_1, \dots, x_n) = g_\theta(T)h(x_1, \dots, x_n).$$

The density can be factors into a product such that one factor h, which does not depend on θ, and another factor, g, which does depend on θ, and depends on $(x_1, \dots, x_n)$ only through T.

Ways to show that a statistic T is sufficient for θ

1. Calculate $\mathbb{P}(X_1 = x_1, \dots, X_n = x_n \mid T(X_1, \dots, X_n))$ and show it is independent of θ.

2. Use factorization theorem and show that the density can be factorized as

$$f_\theta(x_1, \dots, x_n) = g_\theta(T)h(x_1, \dots, x_n).$$

Remark. If we don't already have a sufficient statistic in mind, the factorization approach can be used to find sufficient statistics.

Example 3.5.4 (Finding sufficient statistic for Poisson). Consider X_i IID Poisson(λ) and that the parameter of interest is $= e^{-\lambda}$. The PMF is

$$\mathbb{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = -\frac{\theta \log(\theta)^x}{x!}.$$

$$f_\theta(x_1, \dots, x_n) = \prod_i \left(-\frac{\theta \log(\theta)^{x_i}}{x_i!}\right) = \theta^n(-\log\theta)^{\sum_i x_i} \cdot \frac{1}{\prod_i x_i!}$$

$$\implies g_\theta(T) = \theta^n(-\log\theta)^{\sum_i x_i}, \quad h(x) = \frac{1}{\prod_i x_i!}$$

So $T = \sum_i X_i$ is a sufficient statistic for θ.

Corollay 3.5.5. If T is sufficient for θ, then $\hat{\theta}_{MLE}$ is a function of T.

Proof. If $f_\theta(x_1, \dots, x_n) = g_\theta(T)h(x_1, \dots, x_n)$, then

$$\log(L(\theta)) = \log(g_\theta(T)) + \log(h(x_1, \dots, x_n)).$$

So $\log(h(x_1, \dots, x_n))$ plays no role in the maximization since it does not involve θ. □

3.6 Rao-Blackwell theorem and the bias-variance tradeoff

Theorem 3.6.1 (Rao-Blackwell Theorem)**.**
Suppose that $\hat{\theta}$ is an estimator for θ (with $\mathbb{E}[\hat{\theta}^2] < \infty$) and that T is a sufficient statistic for θ. If we define a new estimator to be

$$\tilde{\theta} = \mathbb{E}[\hat{\theta} \mid T].$$

Then $MSE(\tilde{\theta}) \leq MSE(\hat{\theta})$.

Interpretation: if we know a sufficient statistic T, and we have an estimator $\hat{\theta}$, then we can define an even better estimator $\tilde{\theta}$ for θ which has smaller MSE.

4 Confidence Intervals

Corollay 4.0.1. If $X_1, \dots, X_n$ is an IID sample from a population with mean μ and standard deviation σ, then

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1) \qquad \text{as } n \to \infty.$$