

---

# Concepts of Statistics

## STAT 135

---

Instructor: Rebecca Barter

KELVIN LEE

UC BERKELEY

# Contents

<b>1</b>	<b>Introduction to Inference</b>	<b>4</b>
1.1	Parameters, populations, and estimates . . . . .	4
1.1.1	Common parameters of interest in statistics . . . . .	4
1.1.2	Inference . . . . .	4
1.2	Bias in data . . . . .	4
1.3	Evaluating Estimators . . . . .	5
1.3.1	Parameter bias . . . . .	5
1.3.2	Parameter variance . . . . .	6
1.3.3	Mean Square Error . . . . .	6
1.4	Techniques for estimating bias, variance, and MSE from a single data sample . .	7
1.4.1	Non-parametric bootstrap . . . . .	7
1.4.2	Parametric bootstrap . . . . .	7
1.4.3	Law of Large Numbers . . . . .	8
1.4.4	Central Limit Theorem . . . . .	8
<b>2</b>	<b>Maximum Likelihood Estimation</b>	<b>9</b>
2.1	Likelihood Functions . . . . .	9
2.2	Steps for performing MLE . . . . .	9
2.3	Properties of MLE Estimators . . . . .	10
2.3.1	Consistency . . . . .	10
2.3.2	Asymptotic normality of the MLE . . . . .	11
2.3.3	Delta Method . . . . .	11
<b>3</b>	<b>Method of Moments</b>	<b>12</b>
3.1	Moments . . . . .	12
3.2	MOM vs MLE . . . . .	12
3.3	Cramer-Rao lower bound . . . . .	12
3.4	Efficiency . . . . .	13
3.4.1	Efficient estimators . . . . .	13
3.5	Sufficiency . . . . .	13
3.6	Rao-Blackwell theorem and the bias-variance tradeoff . . . . .	15
<b>4</b>	<b>Confidence Intervals</b>	<b>16</b>
4.1	Definition of confidence intervals . . . . .	16
4.1.1	Quantile . . . . .	16
4.2	Generating confidence intervals for general parameter estimates . . . . .	16
4.3	Confidence intervals for the MLE . . . . .	17
4.4	Confidence intervals for the mean: unknown population variance . . . . .	18
4.5	Coverage . . . . .	18
<b>5</b>	<b>Hypothesis Testing</b>	<b>19</b>
5.1	The null and alternative hypotheses . . . . .	19
5.2	Terminology . . . . .	19

5.3	The test statistic . . . . .	19
5.4	The p-value . . . . .	20
5.5	Critical value and statistical significance . . . . .	20
5.6	Rejection and acceptance regions . . . . .	20
5.7	Alternative hypothesis formats . . . . .	21

# 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)

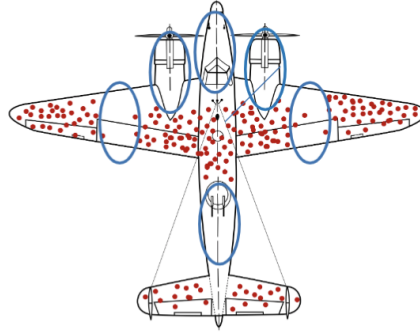
### 1.1.2 Inference

**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,  $\theta$ , is

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

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

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

*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 computed 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,  $\theta$ , 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.*

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

□

### 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.*

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

□

## 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

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

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

where  $N$  is the number of bootstrapped samples and  $\hat{\mu}_k^*$  is the mean of  $k$ th 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  $X_1, X_2, \dots, X_n$  is an IID sample, then

$$\bar{X}_n \xrightarrow{P} \mathbb{E}[X_1] \quad \text{as } n \rightarrow \infty.$$

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

### 1.4.4 Central Limit Theorem

**Theorem 1.4.2** (Central Limit Theorem).

If  $X_1, \dots, X_n$  is an IID sample from a population with mean  $\mu$  and standard deviation  $\sigma$ , then

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



## 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 from 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* ( $\theta$ ) corresponds to the probability of observing the particular data in our sample for various values of  $\theta$ .

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

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

### 2.2 Steps for performing MLE

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

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

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

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

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) become Normal.

### 2.3.1 Consistency

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

$$\hat{\theta}_n \xrightarrow{P} \theta \quad \text{as } n \rightarrow \infty.$$

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

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

#### **Theorem 2.3.2** (Consistency of the MLE).

The MLE  $\hat{\theta}_{MLE,n}$  is a **consistent estimator** of the parameter,  $\theta$ , that it is estimating, which means that

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

*Sketch.*

□

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

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

**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  $n \rightarrow \infty$ , then

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

**Example 2.3.4.**  $\bar{X} \xrightarrow{P} \mu$  as  $n \rightarrow \infty$ , implies that  $\bar{X}^2 \xrightarrow{P} \mu^2$  as  $n \rightarrow \infty$ .

### 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  $\theta$  whose true value is  $\theta_0$ , then as  $n \rightarrow \infty$ , we have that

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

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

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

or

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

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

### 2.3.3 Delta Method

**Theorem 2.3.7** (Delta method).

By CLT, we know that  $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$  as  $n \rightarrow \infty$ .

For any function  $g$  such that  $g'(\mu)$  exists and is non-zero, then

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

## 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,

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

### 3.2 MOM vs MLE

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

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

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

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

□

**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 estimators, 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  $f_\theta$ , under smoothness conditions on  $f_\theta$ , we have that: if  $\hat{\theta}$  is an unbiased estimator for  $\theta$ , then

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

**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  $\theta$ , the *efficiency* of  $\hat{\theta}$  relative to  $\tilde{\theta}$  is

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

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

#### 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

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

**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  $\theta$  if  $\mathbb{P}((X_i)_{i=1}^n \mid T(X_1, \dots, X_n) = t)$  does not depend on  $\theta$  for any  $t$ .

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

Suppose that  $X_1, \dots, X_n \sim F(\theta)$ . Then  $T(X_1, \dots, X_n)$  is a *sufficient statistic* for  $\theta$  if the statistician who knows the value of  $T$  can do just as good a job of estimating the unknown parameter  $\theta$  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  $\theta$  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  $\theta$ , and another factor,  $g$ , which does depend on  $\theta$ , and depends on  $(x_1, \dots, x_n)$  only through  $T$ .

**Ways to show that a statistic  $T$  is sufficient for  $\theta$**

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  $\theta$ .
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( $\lambda$ ) and that the parameter of interest is  $\theta = e^{-\lambda}$ . The PMF is

$$\begin{aligned} \mathbb{P}(X = x) &= \frac{e^{-\lambda} \lambda^x}{x!} = -\frac{\theta \log(\theta)^x}{x!}. \\ f_\theta(x_1, \dots, x_n) &= \prod_i \left( -\frac{\theta \log(\theta)^{x_i}}{x_i!} \right) = \theta^n (-\log \theta)^{\sum_i x_i} \cdot \frac{1}{\prod_i x_i!} \\ \implies g_\theta(T) &= \theta^n (-\log \theta)^{\sum_i x_i}, \quad h(x) = \frac{1}{\prod_i x_i!} \end{aligned}$$

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

**Corollary 3.5.5.** If  $T$  is sufficient for  $\theta$ , 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  $\theta$ . □

### 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  $\theta$  (with  $\mathbb{E}[\hat{\theta}^2] < \infty$ ) and that  $T$  is a sufficient statistic for  $\theta$ . 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  $\theta$  which has smaller MSE.

## 4 Confidence Intervals

**Corollary 4.0.1.** If  $X_1, \dots, X_n$  is an IID sample from a population with mean  $\mu$  and standard deviation  $\sigma$ , then

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

In addition,

$$\mathbb{P}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) \rightarrow \Phi(z) \quad \text{as } n \rightarrow \infty.$$

### 4.1 Definition of confidence intervals

#### 4.1.1 Quantile

Let  $Z \sim \mathcal{N}(0, 1)$ . Define  $z_\alpha$  to be the  $(1 - \alpha)$ -quantile of the  $\mathcal{N}(0, 1)$  distribution, then

$$\mathbb{P}(Z < z_\alpha) = 1 - \alpha.$$

**Definition 4.1.1.** A *confidence interval* is an interval that is calculated in such a way that it contains the true population value of  $\theta$  with some specified probability  $(1 - \alpha)$ , where  $(1 - \alpha)$  is the **coverage probability** or **confidence level**.

A common choice is  $\alpha = 0.05$ , which corresponds to a 95% confidence interval.

**Definition 4.1.2.** A  $(1 - \alpha)\%$  confidence interval  $[L, U]$  for a parameter  $\theta$ , is an interval calculated from a sample that contains  $\theta$  with probability

$$\mathbb{P}(L \leq \theta \leq U) \geq 1 - \alpha.$$

### 4.2 Generating confidence intervals for general parameter estimates

If our estimator approximately satisfies (by CLT, or MLE)

$$\frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim \mathcal{N}(0, 1).$$



Then we have approximately

$$\mathbb{P}\left(-1.96 < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < 1.96\right) = 0.95,$$

or more generally, that

$$\mathbb{P}\left(-z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < z_{\alpha/2}\right) = 1 - \alpha.$$

Rearranging gives

$$\mathbb{P}\left(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} < \theta < \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}\right) = 1 - \alpha.$$

Thus, a  $(1 - \alpha)\%$  CI for  $\theta$  (when  $\frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$  is approximately  $\mathcal{N}(0, 1)$ ), can be computed as

$$[\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}, \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}].$$

This interval contains the true  $\theta$  with probability  $1 - \alpha$ .

**Remark.** The interval is centered at the sample estimate,  $\hat{\theta}$ .

**Example 4.2.1.**  $X_1, \dots, X_n$  IID with mean  $\mu$  and standard deviation  $\sigma$ . Then the  $(1 - \alpha)\%$  CI for  $\mu$  can be computed as follows:

By CLT, we have

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{\text{approx}}{\sim} \mathcal{N}(0, 1).$$

Then

$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) \approx 1 - \alpha.$$

Rearranging gives

$$\mathbb{P}\left(\bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right) \approx 1 - \alpha.$$

Thus, the  $(1 - \alpha)\%$  CI for  $\mu$  is

$$\left[\bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right].$$

### 4.3 Confidence intervals for the MLE

$X_1, \dots, X_n$  (where  $n$  is fairly large) are IID from any distribution with parameter  $\theta$  and  $\hat{\theta}_{MLE}$  is the MLE estimate of  $\theta$ , a  $(1 - \alpha)\%$  CI for  $\theta$  is:

By the asymptotic normality of the MLE, we know that

$$\frac{\hat{\theta}_{MLE} - \theta}{\sigma_{\hat{\theta}_{MLE}}} = \frac{\hat{\theta}_{MLE} - \theta}{\sqrt{1/nI(\theta)}} \sim N(0, 1)$$

So a  $(1 - \alpha)\%$  confidence interval for  $\hat{\theta}_{MLE}$  is

$$\left[ \hat{\theta}_{MLE} - z_{\alpha/2} \sigma_{\hat{\theta}_{MLE}}, \hat{\theta}_{MLE} + z_{\alpha/2} \sigma_{\hat{\theta}_{MLE}} \right] = \left[ \hat{\theta}_{MLE} - \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta}_{MLE})}}, \hat{\theta}_{MLE} + \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta}_{MLE})}} \right].$$

However, we don't know the SD of the parameter estimate. For the mean  $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ . But we don't know  $\sigma$ . We can estimate it from the data using

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}.$$

**Remark.** For a general estimator,  $\hat{\theta}$ , we can estimate  $\sigma_{\hat{\theta}}$  using bootstrap.

## 4.4 Confidence intervals for the mean: unknown population variance

If the  $X_i$ 's are IID with unknown population variance  $\sigma^2$ , then an unbiased estimate is

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

It turns out that

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$$

where  $t_{n-1}$  is the  $t$ -distribution with  $n-1$  degrees of freedom. So the  $(1 - \alpha)\%$  CI is

$$\left[ \bar{X} - \frac{t_{n-1, \alpha/2}}{\sqrt{n}} \hat{\sigma}, \bar{X} + \frac{t_{n-1, \alpha/2}}{\sqrt{n}} \hat{\sigma} \right].$$

where  $t_{n-1, \alpha}$  is the value such that

$$\mathbb{P}(T \leq t_{n-1, \alpha}) = 1 - \alpha.$$

## 4.5 Coverage

**Definition 4.5.1** (Coverage). The *coverage* of  $(1 - \alpha)\%$  confidence interval is the (expected) proportion of the intervals that actually cover the true parameter.

# 5 Hypothesis Testing

## 5.1 The null and alternative hypotheses

**Definition 5.1.1** (Hypothesis testing). *Hypothesis testing* is a method of using inference to test a hypothesis.

**Example 5.1.2.** Suppose the DMV claims that the average waiting time is 20 minutes. We want to test whether the average waiting time at the DMV is actually more than 20 minutes.

We want to test the *null hypothesis*:

$$H_0 : \mu = 20$$

against the alternative hypothesis

$$H_1 : \mu > 20.$$

We will use data from a random sample of waiting times and determine whether we have enough evidence to show that the average waiting time for the population is more than 20 minutes.

## 5.2 Terminology

When conducting a hypothesis test, we either

1. have enough evidence to reject the null hypothesis ( $H_0 : \mu = 20$ ) in favor of the alternative hypothesis.
2. Don't have enough evidence to reject the null hypothesis.

**Remark.** We are never **proving** either hypothesis is true.

## 5.3 The test statistic

Suppose that our data  $X_1, \dots, X_n$  are IID from any distribution with variance  $\sigma^2$ . We want to test the null hypothesis:  $H_0 : \mu = \mu_0$  against the alternative hypothesis:  $H_1 : \mu > \mu_0$ . What statistic  $T(X_1, \dots, X_n)$  could give us evidence to suggest whether we have evidence in favor of the alternative hypothesis? The sample mean  $\bar{X}_n$ . But let's scale it: we call it the **Z-test statistic**:

$$Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}.$$

Under  $H_0$ , we have  $Z \sim \mathcal{N}(0, 1)$  by CLT. If our test statistic looks unlikely to have come from a  $\mathcal{N}(0, 1)$  distribution (e.g., because it's magnitude is very large), then this is evidence against  $H_0$ .

## 5.4 The p-value

**Question.** How do we determine what values of the test statistic  $z$  are big enough such that we can reasonably conclude that we have enough evidence against our null hypothesis?

**Definition 5.4.1** (p-value). The *p-value* is the probability of observing a test statistic that is “as or more extreme” than  $z$ , assuming the null hypothesis is true. (the definition of extreme is based on the alternative hypothesis).

$$\text{p-value} = \mathbb{P}(Z \geq z \mid H_0) = \mathbb{P}\left(Z \geq \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} \mid H_0\right).$$

**Remark.** The p-value is NOT the probability that the null is false nor is it the probability that the alternative is true.

## 5.5 Critical value and statistical significance

**Definition 5.5.1** (Critical value). The *critical value* or *significance level*  $\alpha$ , is the value beyond which we reject the null hypothesis, i.e. we reject the null hypothesis when the p-value is less than  $\alpha$ .

**Remark.** Convention says to reject the null hypothesis when the p-value is less than 0.05. We choose the significance level ourselves! In other words, the conventional significance level is  $\alpha = 0.05$ .

**Definition 5.5.2** (Statistical significance). When the p-value is less than the significance level, (e.g.,  $\text{p-value} < 0.05$ ), the result is said to be *statistically significant*.

## 5.6 Rejection and acceptance regions

**Definition 5.6.1** (Rejection/acceptance region). The set of values of for which  $H_0$  is rejected/not rejected is called the *rejection/acceptance region*.

**Remark.** Recall that we do not technically “accept” the null. We just gather evidence against it, and see if we have enough evidence to reject it.

## 5.7 Alternative hypothesis formats

There are several common forms of alternative hypotheses:

### 1. Composite hypotheses:

- a) *One-sided tests:*  $H_1 : \mu > \mu_0$  or  $H_1 : \mu < \mu_0$
- b) *Two-sided tests:*  $H_1 : \mu \neq \mu_0$ .

### 2. Simple hypothesis

- a)  $H_1 : \mu = \mu_1$ .

#### **Theorem 5.7.1.**

If  $X_1, \dots, X_n$  is an IID sample from a population with mean  $\mu$  and standard deviation  $\sigma$ , then

$$\mathbb{P}(|\bar{X} - \mu| \leq \delta) \approx 2\Phi\left(\frac{\sqrt{n}\delta}{\sigma}\right) - 1$$

regardless of the original distribution of the  $X_i$ .

*Proof.*

$$\mathbb{P}(|\bar{X} - \mu| \leq \delta) = \mathbb{P}\left(\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq \frac{\delta\sqrt{n}}{\sigma}\right) \approx \mathbb{P}\left(|Z| \leq \frac{\delta\sqrt{n}}{\sigma}\right) = 2\Phi\left(\frac{\sqrt{n}\delta}{\sigma}\right) - 1.$$

□