McGill

Statistics

MATH 324

McGill University, Montréal, Canada

Fall 2018

McGill

Introduction: Parameters of a distribution

- Consider a random variable X with a CDF F,

$$F(x) = P(X \leq x) \quad , \quad x \in \mathbb{R}.$$

- We are typically interested in quantities related to F which are called parameters, and are unknown.

- A parameter is either quantities such as expected value, variance, quantiles, etc; or if a parametric form is assumed for F, i.e.

$$F(x) = F(x; \theta)$$

then $\theta \in \mathbb{R}^d$ is called a parameter, and it has dimension $d \geq 1$.

McGill

Introduction: Parameters of a distribution

- Consider a random variable X with a CDF F,

$$F(x) = P(X \leq x) \quad , \quad x \in \mathbb{R}.$$

- We are typically interested in quantities related to F which are called parameters, and are unknown.

- A parameter is either quantities such as expected value, variance, quantiles, etc; or if a parametric form is assumed for F, i.e.

$$F(x) = F(x; \theta)$$

then $\theta \in \mathbb{R}^d$ is called a parameter, and it has dimension $d \geq 1$.

Parameters of a distribution

- In a statistical analysis, the data $(X_1, X_2, \ldots, X_n)$ is used to learn about the parameters.
- Here, learning means estimation.

The drone delivery time example: (revisited)

- A parameter of interest is the average delivery time:

$$\theta = E(X)$$

- If we assume that the delivery time (X) follows a normal distribution $N(\mu, \sigma^2)$, then:

$$\theta = (\mu, \sigma^2)^\top.$$

- Or if we assume that the delivery time (X) follows a gamma distribution Gamma(α, β), then:

$$\theta = (\alpha, \beta)^\top.$$

Point estimator of a parameter

- Consider the data $X_1, X_2, \ldots, X_n$.
- Point Estimator of θ:

  It is a statistic $T(X_1, X_2, \ldots, X_n)$. We denote this estimator by $\hat{\theta}_n$.
- For the observed data $x_1, x_2, \ldots, x_n$, the value of $\hat{\theta}_n$, i.e. $T(x_1, x_2, \ldots, x_n)$, is called an estimate of θ.

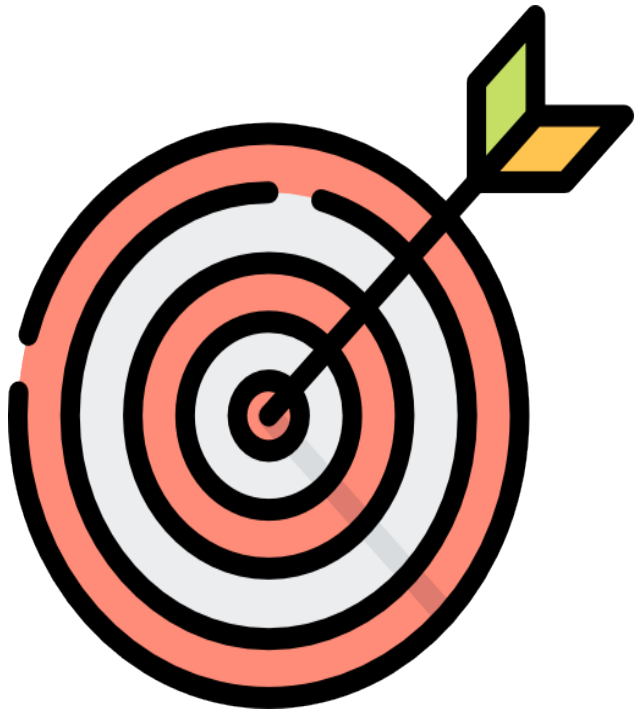
Examples

In many problems, we are interested in estimating the expected value (μ), variance (σ^2) or proportion (p).

- What are the classical estimators for these parameters?
- Answer: the sample version of these quantities!

How do we asses the accuracy/precision of an estimator?

- There may be several estimators for a parameter of interest. Which one do we choose?



Behaviour of an estimator

- In principle, the sampling distribution of an estimator $\hat{\theta}_n$ manages the behaviour of an estimator.
- However, in most cases this distribution may not be accessible: may not be possible to compute it directly.
- Instead, we use properties of the sampling distribution to assess the goodness of a an estimator.

CRITERION 1: BIAS, (Definition 8.2–8.3)

- Consider an estimator $\hat{\theta}_n$ of a parameter θ.
- Definition: the bias of the estimator $\hat{\theta}_n$ is given by

$$\text{Bias}(\hat{\theta}_n) \equiv B(\hat{\theta}_n) = E(\hat{\theta}_n) - \theta.$$

- $\hat{\theta}_n$ is *unbiased* if:

  $B(\hat{\theta}_n) = 0$.

- $\hat{\theta}_n$ is *asymptotically unbiased* if:

  $\lim_{n\to\infty} B(\hat{\theta}_n) = 0$.

Examples

- Derive the bias of sample mean, variance and proportion.

Another example: Exponential distribution

- Let $X_1, X_2, \ldots, X_n$ be a random sample from an exponential distribution with density function

$$f(x;\beta) = \frac{1}{\beta} e^{-x/\beta} \quad , \quad x > 0$$

  where $\beta > 0$ is the unknown parameter.

- Consider the following two estimators of β:

$$\hat{\beta}_n = \overline{X}_n \quad , \quad \tilde{\beta}_n = n \times \min(X_1, X_2, \ldots, X_n).$$

- Compare the biases of the two estimators.

CRITERION 2: VARIANCE

- From the above example we can see that there could be multiple *unbiased* estimators for a parameter of interest θ.
- Unbiasedness is desirable, but it does not tell the whole story.
- Among a family (or class) of unbiased estimators, we would choose the one that has the smallest variance !
- Let $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_M$ be M unbiased estimators of of θ. The best is:

$$\hat{\theta}_{\text{best}} = \text{argmin}_{1 \leq i \leq M} Var(\hat{\theta}_i)$$

Relative efficiency: (Definition 9.1)

- It is a simple way to compare two unbiased estimators.
- Consider two unbiased estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ of θ, with variances $Var(\hat{\theta}_n)$ & $Var(\tilde{\theta}_n)$, respectively. The efficiency of $\hat{\theta}_n$ relative to $\tilde{\theta}_n$ is

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

- **Decision making**:

  if $\text{eff}(\hat{\theta}_n, \tilde{\theta}_n) > 1$, choose $\hat{\theta}_n$; otherwise choose $\tilde{\theta}_n$.

Relative efficiency: (Definition 9.1)

- It is a simple way to compare two unbiased estimators.
- Consider two unbiased estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ of θ, with variances $Var(\hat{\theta}_n)$ & $Var(\tilde{\theta}_n)$, respectively. The efficiency of $\hat{\theta}_n$ relative to $\tilde{\theta}_n$ is

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

- **Decision making**:

  if $\text{eff}(\hat{\theta}_n, \tilde{\theta}_n) > 1$, choose $\hat{\theta}_n$; otherwise choose $\tilde{\theta}_n$.

Relative efficiency: (Definition 9.1)

- It is a simple way to compare two unbiased estimators.
- Consider two unbiased estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ of θ, with variances $Var(\hat{\theta}_n)$ & $Var(\tilde{\theta}_n)$, respectively. The efficiency of $\hat{\theta}_n$ relative to $\tilde{\theta}_n$ is

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

- **Decision making**:

  if $\text{eff}(\hat{\theta}_n, \tilde{\theta}_n) > 1$, choose $\hat{\theta}_n$; otherwise choose $\tilde{\theta}_n$.

Generalization of CRITERION 2: MEAN SQUARED ERROR

- This criterion combines the variance and bias of an estimator.
- The mean squared error of a point estimator $\hat{\theta}_n$ is: (Definition 8.4)

$$\text{MSE}(\hat{\theta}_n) = E\left[(\hat{\theta}_n - \theta)^2\right]$$

- Between two estimators, we would like to choose the one with smaller MSE.

Alternative representation of MSE

- For a point estimator $\hat{\theta}_n$, it can be shown that:

$$\text{MSE}(\hat{\theta}_n) = Var(\hat{\theta}_n) + \{B(\hat{\theta}_n)\}^2$$

- This will be verified in class.

Exponential example: revisted

- We considered two unbiased estimators of β:

$$\hat{\beta}_n = \overline{X}_n \quad , \quad \tilde{\beta}_n = n \times \min(X_1, X_2, \ldots, X_n).$$

- Find the MSE of the two estimators and select the best estimator (if possible). Will be done in class.

CRITERION III: CONSISTENCY

- The behaviour of $\hat{\theta}_n$ as we get more and more data (i.e. $n \to \infty$).
- Definition 9.2:

  The estimator $\hat{\theta}_n$ is said to be a *consistent estimator* of θ if

  $$\lim_{n\to\infty} \Pr(|\hat{\theta}_n - \theta| > \varepsilon) = 0.$$

  for any $\varepsilon > 0$.
- We write: $\hat{\theta}_n \xrightarrow{p} \theta$, as $n \to \infty$.

CRITERION III: CONSISTENCY

- The behaviour of $\hat{\theta}_n$ as we get more and more data (i.e. $n \to \infty$).
- Definition 9.2:

  The estimator $\hat{\theta}_n$ is said to be a *consistent estimator* of θ if

  $$\lim_{n\to\infty} \Pr(|\hat{\theta}_n - \theta| > \varepsilon) = 0.$$

  for any $\varepsilon > 0$.
- We write: $\hat{\theta}_n \xrightarrow{p} \theta$, as $n \to \infty$.

Examples of consistent estimators

- By the Weak Law of Large Numbers, the SAMPLE MEAN and SAMPLE PROPORTION are consistent estimators of their population counterparts. (based on a random sample of size n).

- Note:

  if $U_n \xrightarrow{p} U$ & $W_n \xrightarrow{p} W$ then, $U_n + W_n \xrightarrow{p} U + W$, as $n \to \infty$.

- This implies that the following estimators are also consistent:

  the difference of two sample means, the difference of two sample proportions, and the sample mean.

Examples of consistent estimators

- By the Weak Law of Large Numbers, the SAMPLE MEAN and SAMPLE PROPORTION are consistent estimators of their population counterparts. (based on a random sample of size n).

- Note:

  if $U_n \xrightarrow{p} U$ & $W_n \xrightarrow{p} W$ then, $U_n + W_n \xrightarrow{p} U + W$, as $n \to \infty$.

- This implies that the following estimators are also consistent:

  the difference of two sample means, the difference of two sample proportions, and the sample mean.

Examples of consistent estimators

- By the Weak Law of Large Numbers, the SAMPLE MEAN and SAMPLE PROPORTION are consistent estimators of their population counterparts. (based on a random sample of size n).

- Note:

  if $U_n \xrightarrow{p} U$ & $W_n \xrightarrow{p} W$ then, $U_n + W_n \xrightarrow{p} U + W$, as $n \to \infty$.

- This implies that the following estimators are also consistent:

  the difference of two sample means, the difference of two sample proportions, and the sample mean.

Example of an inconsistent estimator

- Consider the exponential example discussed on page 11 of the notes. An unbiased estimator of β is

$$\tilde{\beta}_n = n \times \min(X_1, X_2, \ldots, X_n)$$

and we showed that $\tilde{\beta}_n/n = \min(X_1, X_2, \ldots, X_n) \sim \text{Exp}(\beta/n)$.

- It can be shown that:

$$\Pr(|\tilde{\beta}_n - \beta| > \varepsilon) = e^{-(\beta-\varepsilon)/\beta} - e^{-(\beta+\varepsilon)/\beta}$$

Clearly, the above expression does not converge to 0 as $n \to \infty$, which implies that $\tilde{\beta}_n$ is NOT consistent.

Example of an inconsistent estimator

- Consider the exponential example discussed on page 11 of the notes. An unbiased estimator of β is

$$\tilde{\beta}_n = n \times \min(X_1, X_2, \ldots, X_n)$$

and we showed that $\tilde{\beta}_n/n = \min(X_1, X_2, \ldots, X_n) \sim \text{Exp}(\beta/n)$.

- It can be shown that:

$$\Pr(|\tilde{\beta}_n - \beta| > \varepsilon) = e^{-(\beta-\varepsilon)/\beta} - e^{-(\beta+\varepsilon)/\beta}$$

Clearly, the above expression does not converge to 0 as $n \to \infty$, which implies that $\tilde{\beta}_n$ is NOT consistent.

Example of an inconsistent estimator

- Consider the exponential example discussed on page 11 of the notes. An unbiased estimator of β is

$$\tilde{\beta}_n = n \times \min(X_1, X_2, \ldots, X_n)$$

and we showed that $\tilde{\beta}_n/n = \min(X_1, X_2, \ldots, X_n) \sim \text{Exp}(\beta/n)$.

- It can be shown that:

$$\Pr(|\tilde{\beta}_n - \beta| > \varepsilon) = e^{-(\beta-\varepsilon)/\beta} - e^{-(\beta+\varepsilon)/\beta}$$

Clearly, the above expression does not converge to 0 as $n \to \infty$, which implies that $\tilde{\beta}_n$ is NOT consistent.

McGill

One way of proving consistency

Theorem 9.1: An unbiased estimator $\hat{\theta}_n$ of θ is consistent if

$$\lim_{n\to\infty} Var(\hat{\theta}_n) = 0.$$

PROOF. Will be discussed in class.

One way of proving consistency

Theorem 9.1: An unbiased estimator $\hat{\theta}_n$ of θ is consistent if

$$\lim_{n\to\infty} Var(\hat{\theta}_n) = 0.$$

PROOF. Will be discussed in class.

Example

- Let $X_1, X_2, \ldots, X_n$ be i.i.d from Unif$(0, \theta)$. Show that

$$\hat{\theta}_n = \frac{n+1}{n} \max(X_1, \ldots, X_n)$$

  is an unbiased estimator of θ. Is it a consistent estimator of θ? Compare $\hat{\theta}_n$ with the estimator $\tilde{\theta}_n = 2\overline{X}_n$.

- Will be discussed in class.

Example

- Let $X_1, X_2, \ldots, X_n$ be i.i.d from Unif$(0, \theta)$. Show that

$$\hat{\theta}_n = \frac{n+1}{n} \max(X_1, \ldots, X_n)$$

  is an unbiased estimator of θ. Is it a consistent estimator of θ? Compare $\hat{\theta}_n$ with the estimator $\tilde{\theta}_n = 2\overline{X}_n$.

- Will be discussed in class.

Be careful with the bias when using Theorem 9.1

- Example:

  Let $X_1, X_2, \ldots, X_n$ be i.i.d from $N(0, \sigma^2)$ with unknown variance σ^2. We consider the following estimator of σ:

$$\hat{\sigma}_n = \frac{1}{n} \sum_{i=1}^{n} |X_i|.$$

- Is $\hat{\sigma}_n$ a consistent estimator of σ? Verify your answer.

- This will be discussed in class.

Consistent estimator of a function of a parameter

- Suppose that $\hat{\theta}_n$ is a consistent estimator of θ. If $g(\cdot)$ is a real-valued continuos function of θ, then we have

$$g(\hat{\theta}_n) \xrightarrow{p} g(\theta)$$

as $n \to \infty$.

- That is, $g(\hat{\theta}_n)$ is a consistent estimator of $g(\theta)$.

Example

- Let $X_1, \ldots, X_n$ be i.i.d from a distribution F with $\mu = E(X_i)$ and $\sigma^2 = Var(X_i) < \infty$. Find a consistent estimator of the standard deviation σ.

- Will be discussed in class.

Summary

- So far, we have been using our intuition to come up with possible estimator(s) for a parameter of interest.
- For example, based on a random sample $X_1, X_2, \ldots, X_n$, we use the sample mean $\overline{X}_n$ and variance S_n^2 to estimate the population mean μ and variance σ^2, respectively.
- We assess the "goodness" of an estimator using criteria such as MEAN SQUARED ERROR & CONSISTENCY

Sufficient Statistics

- Question:

  do we use all the information contained in $X_1, X_2, \ldots, X_n$ about a parameter θ when estimating θ by an estimator $\hat{\theta}_n$?

- The answer is "may" or "may not". It depends on the statistic $\hat{\theta}_n$.

- In this section, we propose methods for finding statistics (estimators) that in a sense summarize all the information contained in a sample about a target parameter θ.

- Such statistics are called *Sufficient Statistics*.

Sufficient Statistics

- Question:

  do we use all the information contained in $X_1, X_2, \ldots, X_n$ about a parameter θ when estimating θ by an estimator $\hat{\theta}_n$?

- The answer is "may" or "may not". It depends on the statistic $\hat{\theta}_n$.

- In this section, we propose methods for finding statistics (estimators) that in a sense summarize all the information contained in a sample about a target parameter θ.

- Such statistics are called *Sufficient Statistics*.

Sufficient Statistics

- Question:

  do we use all the information contained in $X_1, X_2, \ldots, X_n$ about a parameter θ when estimating θ by an estimator $\hat{\theta}_n$?

- The answer is "may" or "may not". It depends on the statistic $\hat{\theta}_n$.

- In this section, we propose methods for finding statistics (estimators) that in a sense summarize all the information contained in a sample about a target parameter θ.

- Such statistics are called *Sufficient Statistics*.

Sufficient Statistics

- Question:

  do we use all the information contained in $X_1, X_2, \ldots, X_n$ about a parameter θ when estimating θ by an estimator $\hat{\theta}_n$?

- The answer is "may" or "may not". It depends on the statistic $\hat{\theta}_n$.

- In this section, we propose methods for finding statistics (estimators) that in a sense summarize all the information contained in a sample about a target parameter θ.

- Such statistics are called *Sufficient Statistics*.

Example: a sufficient statistic

"For the Women's Health Study (J. Amer. Med. Assoc., vol. 295, pp. 306–313, 2006), heart attacks were reported for 198 of 19,934 subjects taking aspirin..."

If the goal is to estimate the probability of getting a heart attack when taking aspirin, based on a random sample $X_1, X_2, \ldots, X_n$:

- Is it sufficient to record the total number of heart attacks, $\sum_{i=1}^{n} X_i$?
- Do we need the actual data (yes/no, X_i) from each individual?

a sufficient statistic ...

- Let $X_1, X_2, \ldots, X_n$ be iid from Ber(θ), where θ is the probability of success. Suppose that we are given $T = \sum_{i=1}^{n} X_i$, i.e. the total number of success in the sample.
- Do we gain any further information about θ by looking at other functions of the sample?
- One way to answer this question is to look at the conditional distribution of the random sample given T.

a sufficient statistic ...

- It can be shown that

$$\Pr\left(X_1 = x_1, \ldots, X_n = x_n \middle| \sum_{i=1}^{n} X_i = t\right) = \begin{cases} \frac{1}{\binom{n}{t}} & \text{,if } \sum_{i=1}^{n} x_i = t; \\ 0 & \text{,otherwise.} \end{cases}$$

- This conditional distribution no longer depends on unknown θ !
- That is, once $T = \sum_{i=1}^{n} X_i$ is given, the sample $X_1, X_2, \ldots, X_n$ does not contain any additional information about θ.
- The statistic $T = \sum_{i=1}^{n} X_i$ is called a *sufficient statistic*.

a sufficient statistic ...

- It can be shown that

$$\Pr\left(X_1 = x_1, \ldots, X_n = x_n \middle| \sum_{i=1}^{n} X_i = t\right) = \begin{cases} \frac{1}{\binom{n}{t}} & \text{,if } \sum_{i=1}^{n} x_i = t; \\ 0 & \text{,otherwise.} \end{cases}$$

- This conditional distribution no longer depends on unknown θ !
- That is, once $T = \sum_{i=1}^{n} X_i$ is given, the sample $X_1, X_2, \ldots, X_n$ does not contain any additional information about θ.
- The statistic $T = \sum_{i=1}^{n} X_i$ is called a *sufficient statistic*.

a sufficient statistic ...

- It can be shown that

$$\Pr\left(X_1 = x_1, \ldots, X_n = x_n \middle| \sum_{i=1}^{n} X_i = t\right) = \begin{cases} \frac{1}{\binom{n}{t}} & \text{,if } \sum_{i=1}^{n} x_i = t; \\ 0 & \text{,otherwise.} \end{cases}$$

- This conditional distribution no longer depends on unknown θ !
- That is, once $T = \sum_{i=1}^{n} X_i$ is given, the sample $X_1, X_2, \ldots, X_n$ does not contain any additional information about θ.
- The statistic $T = \sum_{i=1}^{n} X_i$ is called a *sufficient statistic*.

a sufficient statistic ...

- It can be shown that

$$\Pr\left(X_1 = x_1, \ldots, X_n = x_n \middle| \sum_{i=1}^{n} X_i = t\right) = \begin{cases} \frac{1}{\binom{n}{t}} & \text{,if } \sum_{i=1}^{n} x_i = t; \\ 0 & \text{,otherwise.} \end{cases}$$

- This conditional distribution no longer depends on unknown θ !
- That is, once $T = \sum_{i=1}^{n} X_i$ is given, the sample $X_1, X_2, \ldots, X_n$ does not contain any additional information about θ.
- The statistic $T = \sum_{i=1}^{n} X_i$ is called a *sufficient statistic*.

Formal definition of a sufficient statistic

Definition 9.3:

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution $F(\cdot; \theta)$ with unknown parameter θ. A statistic $T(X_1, \ldots, X_n)$ is called sufficient if the conditional distribution of $X_1, X_2, \ldots, X_n$ given $T(X_1, \ldots, X_n) = t$ does NOT depend on θ.

Note:

The above definition only tells us how to check whether a statistic is sufficient, but it does not tell us to find one!

Formal definition of a sufficient statistic

Definition 9.3:

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution $F(\cdot; \theta)$ with unknown parameter θ. A statistic $T(X_1, \ldots, X_n)$ is called sufficient if the conditional distribution of $X_1, X_2, \ldots, X_n$ given $T(X_1, \ldots, X_n) = t$ does NOT depend on θ.

Note:

The above definition only tells us how to check whether a statistic is sufficient, but it does not tell us to find one!

The likelihood function

- We now state a theorem that gives us a tool for finding sufficient statistics. First, we need to define a likelihood function.

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution $F(\cdot;\theta)$ with a pmf (or pdf) $f(\cdot;\theta)$. The joint pmf (or pdf) of the X_i's is called a likelihood function and is given by

$$L(x_1, x_2, \ldots, x_n; \theta) = f(x_1;\theta) \times f(x_2;\theta) \times \ldots \times f(x_n;\theta)$$

- In the discrete case, the likelihood function is in fact the probability of observing the event $X_1 = x_1, X_2 = x_n, \ldots, X_n = x_n$.

The likelihood function

- We now state a theorem that gives us a tool for finding sufficient statistics. First, we need to define a likelihood function.

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution $F(\cdot;\theta)$ with a pmf (or pdf) $f(\cdot;\theta)$. The joint pmf (or pdf) of the X_i's is called a likelihood function and is given by

$$L(x_1, x_2, \ldots, x_n; \theta) = f(x_1;\theta) \times f(x_2;\theta) \times \ldots \times f(x_n;\theta)$$

- In the discrete case, the likelihood function is in fact the probability of observing the event $X_1 = x_1, X_2 = x_n, \ldots, X_n = x_n$.

The likelihood function

- We now state a theorem that gives us a tool for finding sufficient statistics. First, we need to define a likelihood function.

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution $F(\cdot;\theta)$ with a pmf (or pdf) $f(\cdot;\theta)$. The joint pmf (or pdf) of the X_i's is called a likelihood function and is given by

$$L(x_1, x_2, \ldots, x_n; \theta) = f(x_1;\theta) \times f(x_2;\theta) \times \ldots \times f(x_n;\theta)$$

- In the discrete case, the likelihood function is in fact the probability of observing the event $X_1 = x_1, X_2 = x_n, \ldots, X_n = x_n$.

The Fisher–Neyman Factorization Theorem

Theorem 9.4:

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution $F(\cdot;\theta)$ with a pmf (or pdf) $f(\cdot;\theta)$. Then, the statistic $T(X_1, X_2, \ldots, X_n)$ is sufficient for θ if and only if

$$L(x_1, x_2, \ldots, x_n; \theta) = g(t;\theta) \times h(x_1, x_2, \ldots, x_n)$$

where $t = T(x_1, x_2, \ldots, x_n)$, and

(1) $g(t;\theta)$ depends on $x_1, x_2, \ldots, x_n$ only through $t = T(x_1, x_2, \ldots, x_n)$,

(2) $h(x_1, x_2, \ldots, x_n)$ does NOT depend on θ.

McGill

Bernoulli example: revisited

- Note that:

$$L(x_1, x_2, \ldots, x_n; \theta) = \theta^{x_1}(1-\theta)^{1-x_1} \times \ldots \times \theta^{x_n}(1-\theta)^{1-x_n}$$

$$= \theta^{\sum_{i=1}^n x_i} \times (1-\theta)^{n-\sum_{i=1}^n x_i}.$$

- By the factorization theorem,

$$T = \sum_{i=1}^n X_i$$

is a sufficient statistic for θ, or for the parametric family $\mathrm{Ber}(\theta)$.

Bernoulli example: revisited

- Note that:

$$L(x_1, x_2, \ldots, x_n; \theta) = \theta^{x_1}(1-\theta)^{1-x_1} \times \ldots \times \theta^{x_n}(1-\theta)^{1-x_n}$$

$$= \theta^{\sum_{i=1}^n x_i} \times (1-\theta)^{n-\sum_{i=1}^n x_i}.$$

- By the factorization theorem,

$$T = \sum_{i=1}^{n} X_i$$

is a sufficient statistic for θ, or for the parametric family Ber(θ).

McGill

Example 2

- Let $X_1, X_2, \ldots, X_n$ be iid from Unif$(0, \theta)$. Find a sufficient statistic for this parametric family.
- Will be discussed in class.

Bernoulli example: revisited

- The likelihood function can be written as

$$L(x_1, x_2, \ldots, x_n; \theta) = \theta^{\left[\sum_{i=1}^{m} x_i + \sum_{i=m+1}^{n} x_i\right]} \times (1-\theta)^{n-\left[\sum_{i=1}^{m} x_i + \sum_{i=m+1}^{n} x_i\right]}$$

  for any fixed integer m belonging to $\{1, 2, \ldots, n\}$.

- Thus, by the factorization theorem, we conclude that $T_m^* = (\sum_{i=1}^{m} X_i, \sum_{i=m+1}^{n} X_i)$ is a sufficient statistic for θ.

  We also showed that $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic for θ.

Bernoulli example: revisited

- The likelihood function can be written as

$$L(x_1, x_2, \ldots, x_n; \theta) = \theta^{\left[\sum_{i=1}^{m} x_i + \sum_{i=m+1}^{n} x_i\right]} \times (1-\theta)^{n-\left[\sum_{i=1}^{m} x_i + \sum_{i=m+1}^{n} x_i\right]}$$

  for any fixed integer m belonging to $\{1, 2, \ldots, n\}$.

- Thus, by the factorization theorem, we conclude that $T_m^* = (\sum_{i=1}^{m} X_i, \sum_{i=m+1}^{n} X_i)$ is a sufficient statistic for θ.

  We also showed that $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic for θ.

Note

- We see that there could be many sufficient statistics for a parametric family. Some provide more reduction than others:

  such as $T = \sum_{i=1}^{n} X_i$ compared to $T_m^* = (\sum_{i=1}^{m} X_i, \sum_{i=m+1}^{n} X_i)$ in the Binomial example.

- How far can we go in this reduction? In other words, is there a sufficient statistic that provides "maximal reduction" of the data?

- The answer is: minimal sufficient statistic.

Note

- We see that there could be many sufficient statistics for a parametric family. Some provide more reduction than others:

  such as $T = \sum_{i=1}^{n} X_i$ compared to $T_m^* = (\sum_{i=1}^{m} X_i, \sum_{i=m+1}^{n} X_i)$ in the Binomial example.

- How far can we go in this reduction? In other words, is there a sufficient statistic that provides "maximal reduction" of the data?

- The answer is: minimal sufficient statistic.

Note

- We see that there could be many sufficient statistics for a parametric family. Some provide more reduction than others:

  such as $T = \sum_{i=1}^{n} X_i$ compared to $T_m^* = (\sum_{i=1}^{m} X_i, \sum_{i=m+1}^{n} X_i)$ in the Binomial example.

- How far can we go in this reduction? In other words, is there a sufficient statistic that provides "maximal reduction" of the data?

- The answer is: minimal sufficient statistic.

Minimal sufficient statistic

- Definition:

  Let $T = T(X_1, X_2, \ldots, X_n)$ be a sufficient statistic for a parametric family. Then T is called a minimal sufficient statistic if and only if, for any other sufficient statistic $U = U(X_1, X_2, \ldots, X_n)$, there exists a function $g(\cdot)$ such that $T = g(U)$.

- If T and U are both minimal sufficient statistics, then the function g in the above definition is going to be one-to-one. That is, the minimal sufficient statistic is unique.

Minimal sufficient statistic

- Definition:

  Let $T = T(X_1, X_2, \ldots, X_n)$ be a sufficient statistic for a parametric family. Then T is called a minimal sufficient statistic if and only if, for any other sufficient statistic $U = U(X_1, X_2, \ldots, X_n)$, there exists a function $g(\cdot)$ such that $T = g(U)$.

- If T and U are both minimal sufficient statistics, then the function g in the above definition is going to be one-to-one. That is, the minimal sufficient statistic is unique.

How to find the minimal sufficient statistic?

- Lehmann–Scheffé Criterion:

  Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pmf or pdf $f(x;\theta)$. A statistic $T = T(X_1, X_2, \ldots, X_n)$ is minimal sufficient if for any two sample points $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$,

$$\frac{L(x_1, x_2, \ldots, x_n; \theta)}{L(y_1, y_2, \ldots, y_n; \theta)}$$

  does NOT depend on θ if and only if

$$T(x_1, x_2, \ldots, x_n) = T(y_1, y_2, \ldots, y_n).$$

Examples

- We will derive minimal sufficient statistic for the Binomial and Uniform families in class.

Fisher–Neyman Factorization Theorem for muti-parameter cases

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution $F(\cdot;\theta)$ with a pmf (or pdf) $f(\cdot;\theta)$, with parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$.
Then, the statistics
$(T_1(X_1, X_2, \ldots, X_n), T_2(X_1, X_2, \ldots, X_n), \ldots, T_d(X_1, X_2, \ldots, X_n))$ is sufficient for θ if and only if

$$L(x_1, x_2, \ldots, x_n; \theta) = g(t_1, t_2, \ldots, t_d; \theta) \times h(x_1, x_2, \ldots, x_n)$$

where $t_j = T_j(x_1, x_2, \ldots, x_n)$, for $j = 1, 2, \ldots, d$, and

(1) $g(t_1, t_2, \ldots, t_d; \theta)$ depends on $x_1, x_2, \ldots, x_n$ only through $t_j = T_j(x_1, x_2, \ldots, x_n), j = 1, \ldots, d$,

(2) $h(x_1, x_2, \ldots, x_n)$ does NOT depend on θ.

Example

- Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, where the unknown parameter is $\theta = (\mu, \sigma^2)$. Find a sufficient statistic for this family. Is your statistic minimal sufficient?

- This will be discussed in class.

Sufficiency and point estimation

- Recall the examples that we have discussed so far.
- In all the examples, the point estimators of the parameters were functions of sufficient statistics !
- Question:

  Can we take advantage of sufficient statistics and construct “good” estimator(s) for the parameter(s) of interest in a parametric family?

The Rao–Blackwell Theorem

Theorem 9.5:

Let $\hat{\theta}_n$ be an unbiased estimator for a parameter θ such that $Var(\hat{\theta}_n) < \infty$. If T is a sufficient statistic for θ, the statistic

$$\tilde{\theta}_n = E(\hat{\theta}_n | T)$$

is an unbiased estimator of θ and $Var(\tilde{\theta}_n) \leq Var(\hat{\theta}_n)$.

- The proof will be discussed in class.

The Rao–Blackwell Theorem

Theorem 9.5:

Let $\hat{\theta}_n$ be an unbiased estimator for a parameter θ such that $Var(\hat{\theta}_n) < \infty$. If T is a sufficient statistic for θ, the statistic

$$\tilde{\theta}_n = E(\hat{\theta}_n | T)$$

is an unbiased estimator of θ and $Var(\tilde{\theta}_n) \leq Var(\hat{\theta}_n)$.

- The proof will be discussed in class.

- The Rao–Blackwell Theorem shows us how to improve on a given unbiased estimator of a parameter.
- If the unbiased estimator $\hat{\theta}$ is NOT a function of T, then applying the Rao-Blackwell theorem will provide a new estimator with smaller variance.
- If the unbiased estimator $\hat{\theta}$ is already a function of T, then applying the Rao-Blackwell theorem do nothing! Because,

$$\tilde{\theta}_n = E(\hat{\theta}_n|T) = \hat{\theta}_n.$$

- Thus, applying the Rao–Blackwell Theorem more than once has no effect !

- The Rao–Blackwell Theorem shows us how to improve on a given unbiased estimator of a parameter.
- If the unbiased estimator $\hat{\theta}$ is NOT a function of T, then applying the Rao-Blackwell theorem will provide a new estimator with smaller variance.
- If the unbiased estimator $\hat{\theta}$ is already a function of T, then applying the Rao-Blackwell theorem do nothing! Because,

$$\tilde{\theta}_n = E(\hat{\theta}_n|T) = \hat{\theta}_n.$$

- Thus, applying the Rao–Blackwell Theorem more than once has no effect !

- The Rao–Blackwell Theorem shows us how to improve on a given unbiased estimator of a parameter.
- If the unbiased estimator $\hat{\theta}$ is NOT a function of T, then applying the Rao-Blackwell theorem will provide a new estimator with smaller variance.
- If the unbiased estimator $\hat{\theta}$ is already a function of T, then applying the Rao-Blackwell theorem do nothing! Because,

$$\tilde{\theta}_n = E(\hat{\theta}_n | T) = \hat{\theta}_n.$$

- Thus, applying the Rao–Blackwell Theorem more than once has no effect !

Minimum variance unbiased estimators (MVUE)

For many well-known parametric families, an estimator $\hat{\theta}_n$ is MVUE if:

(i) $\hat{\theta}_n$ is unbiased,

(ii) $\hat{\theta}_n = g(T)$, where T is a (minimal) sufficient statistic.

- A precise formulation of this result, due to Lehmann and Scheffé, uses the notion of completeness, which goes beyond the scope of this course.

Minimum variance unbiased estimators (MVUE)

For many well-known parametric families, an estimator $\hat{\theta}_n$ is MVUE if:

(i) $\hat{\theta}_n$ is unbiased,

(ii) $\hat{\theta}_n = g(T)$, where T is a (minimal) sufficient statistic.

- A precise formulation of this result, due to Lehmann and Scheffé, uses the notion of completeness, which goes beyond the scope of this course.

Minimum variance unbiased estimators (MVUE)

For many well-known parametric families, an estimator $\hat{\theta}_n$ is MVUE if:

(i) $\hat{\theta}_n$ is unbiased,

(ii) $\hat{\theta}_n = g(T)$, where T is a (minimal) sufficient statistic.

- A precise formulation of this result, due to Lehmann and Scheffé, uses the notion of completeness, which goes beyond the scope of this course.

Minimum variance unbiased estimators (MVUE)

For many well-known parametric families, an estimator $\hat{\theta}_n$ is MVUE if:

(i) $\hat{\theta}_n$ is unbiased,

(ii) $\hat{\theta}_n = g(T)$, where T is a (minimal) sufficient statistic.

- A precise formulation of this result, due to Lehmann and Scheffé, uses the notion of completeness, which goes beyond the scope of this course.

Examples

- We will discuss examples on constructing MVUE in class.