Lecture 5
Point Estimator and Sampling Distribution

Fall 2013
Prof. Yao Xie, yao.xie@isye.gatech.edu
H. Milton Stewart School of Industrial Systems & Engineering
Georgia Tech
Road map

Point Estimation → Confidence Interval Estimation → Hypothesis Testing
Population, sample, statistics

Population: all possible observations

Sample: observed measurements taken from population

\{x_1, x_2, x_3, ..., x_n\}

Statistics: data summaries

S(x_1, x_2, x_3, ..., x_n)
Why need point estimator

Assume the data are actually random variables generated from a distribution

(e.g., Normal, Exponential, Binomial),

but the distribution’s PARAMETERS are unknown

(e.g., mean, variance, rate, proportion)
Point estimator

Goal of estimation: Create a best “approximation” of the unknown parameter using a statistic (summary or function of data) called point estimator

\[ S(x_1, x_2, \ldots, x_n) \]

Examples: normal, binomial
Sample proportion

An automobile manufacturer has developed a new type of bumper, which is supposed to absorb impacts with less damage than previous bumpers.

The manufacturer has used this bumper in a sequence of 25 controlled crashes against wall, each at 10 mph, using one of its compact models. Let $X$ be the number of crashes that result in no visible damage to the automobile. He observes that only 15 out of 25 cars have no damage after crash.

The manufacturer is interested in the proportion of all such crashes that result in no damage. We define this feature through the parameter $p$, which is the probability of no damage in a single crash.
Common point estimators

Estimating population mean

sample mean \( \hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \)

Estimating population variance

sample variance \( \hat{\sigma}^2 = S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \)

Estimating population proportion

sample proportion \( \hat{p} = \frac{X}{n} \) if \( X \sim \text{BIN}(n, p) \)
Sampling distribution

If we treat data $x_1,\ldots,x_n$ as realizations from random variables $X_1,\ldots,X_n$, then point estimators which are functions of RVs are also random variables, and so they also have distributions called **SAMPLING DISTRIBUTIONS**.

Example: Sampling distribution of $\hat{p}$ when $X \sim \text{Bin}(n=5, p=1/2)$
Sample mean: sampling distribution

For $X_1, \ldots, X_n \sim \text{mean } \mu \text{ and variance } \sigma^2$, estimate mean $\mu$ using the sample mean $\hat{\mu} = \frac{X_1 + \ldots + X_n}{n}$.

1. For $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ and $\sigma^2$ known, then the sampling distribution is $\hat{\mu} \sim N(\mu, \sigma^2 / n)$.

2. For large $n$, an approximate sampling distribution is $\hat{\mu} \sim N(\mu, \sigma^2 / n)$.

Distribution of sample difference?

Example: testing drug effectiveness.
The amount of time that a customer spends waiting in the airport security is a random variable with mean 8.2 minutes and standard deviation 1.5 minutes. Suppose that a random sample of \( n = 49 \) customers is observed.

Find the probability that the average waiting time for these customers is:

(a) Less than 10 minutes
(b) Between 5 and 10 minutes
Sample Mean: Sampling Distribution

Example: Standard error computation for the sample mean

1. For $\sigma^2$ known, the standard error is

$$\sqrt{V(\hat{\mu})} = \frac{\sigma}{\sqrt{n}}.$$ 

2. For $\sigma^2$ unknown, the standard error estimate is

$$\sqrt{V(\hat{\mu})} \approx \frac{\sigma}{\sqrt{n}} = \frac{S}{\sqrt{n}} = \frac{\sqrt{\sum(x_i - \bar{x})^2/(n-1)}}{\sqrt{n}}.$$
What is a good estimator?
What is a good estimator?

Suppose we are estimating the unknown parameter \( \theta \) with a point estimator \( \hat{\theta}(X_1, X_2, \ldots, X_n) \).

There are two properties for a good estimator:

**unbiased**

\[ E[\hat{\theta}(X_1, X_2, \ldots, X_n)] = \theta \]

| \( E(\hat{\theta}) - \theta \) | is called BIAS

**small variance**

\[ \text{VAR}(\hat{\theta}) \]

**Consistency**: unbiased and variance goes to 0 when \( n \) goes to infinity.
Evaluation of estimators: Example

Take two instruments measuring blood pressure. One of them has been calibrated, where the second one gives values larger than the true ones. For one individual, we measure the blood pressure for 10 times: $x_1, \ldots, x_{10}$ with the first instrument and $y_1, \ldots, y_{10}$ with the second instrument. Say $\theta$ is the blood pressure of the same individual. If we estimate $\theta$ by

$$\bar{x} = \frac{x_1 + \ldots + x_{10}}{10} \quad \text{and} \quad \bar{y} = \frac{y_1 + \ldots + y_{10}}{10}$$

based on the first and the second set of measurements. Which of the two estimates will be more biased?
Sample proportion

Our data is sampled from $\text{BIN}(n,p)$

**Point Estimation** of $p$: $\hat{p} = \frac{x}{n}$

Sampling Distribution of $\hat{p}$

$E(\hat{p}) = p, \quad \text{VAR}(\hat{p}) = \frac{p(1-p)}{n}$

1. As $X \sim \text{B}(n,p)$, the exact sampling distribution is:

   $$X = n\hat{p} \sim \text{Bin}(n,p)$$

2. For large $n$ (CLT), an approximate distribution is:

   $$\hat{p} \sim \text{N}(p, p(1-p)/n)$$
Car crash example

**Example:** Observe $x = 15$ cars that have no damage after crash, when $n = 25$ cars are tested. What is an approximate standard error for the estimated sample proportion?

If we want the standard error to be less than .1, how many tests are needed?
Sample mean: unbiasedness and consistency

For $X_1,\ldots,X_n \sim$ mean $\mu$ and variance $\sigma^2$, estimate mean $\mu$ using the sample mean and

$$E(\hat{\mu}) = \frac{E(X_1) + \ldots + E(X_n)}{n} = \frac{n\mu}{n} = \mu$$

So $\hat{\mu}$ is unbiased

As the bias is zero, we need to check whether the variance goes to zero:

$$\text{VAR}(\hat{\mu}) = \frac{\sigma^2}{n}$$

So $\hat{\mu}$ is consistent
Chi-square distribution

Let $Z_1, \ldots, Z_n \sim \mathcal{N}(0, 1)$ and independent

\[
Y = Z_1^2 + Z_2^2 + \ldots + Z_n^2 \sim \chi^2_n
\]

$f(x) > 0$ only for $x > 0$,

Chi-Square has one parameter, called it’s “degrees of freedom”

\[
X \sim \chi^2_k : \\
\begin{align*}
f(x \mid k) & = \frac{1}{2^{k/2} \Gamma(k/2)} x^{k/2 - 1} e^{-x/2}, \Gamma(k) = (k - 1)! \\
E(X) & = k, \quad \text{VAR}(X) = 2k
\end{align*}
\]
Chi-square distribution

If $X_1,\ldots,X_n \sim N(\mu,\sigma^2)$ are independent, what is the distribution of

$$Y = \sum_{i=1}^{n} \left( \frac{X_i - \mu}{\sigma} \right)^2$$

Let $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, then

$$W = \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 \sim \chi_{n-1}^2$$

$Y$ and $W$ differ only by one degree-of-freedom ($n$ vs. $n-1$)
Sample Variance: Sampling Distribution

If $X_1,\ldots,X_n \sim N(\mu, \sigma^2)$, then

$$\frac{S^2(n-1)}{\sigma^2} \sim \chi^2_{n-1}$$
Sample Variance: Unbiasedness and Consistency

\[ E(S^2) = \frac{\sigma^2}{n-1} \mathbb{E} \left\{ \frac{S^2(n-1)}{\sigma^2} \right\} = \frac{(n-1)\sigma^2}{n-1} = \sigma^2 \]

\[ \text{VAR}(S^2) = \frac{\sigma^4}{(n-1)^2} \text{V} \left\{ \frac{S^2(n-1)}{\sigma^2} \right\} \]

\[ = \frac{\sigma^4}{(n-1)^2} 2(n-1) = \frac{2\sigma^4}{n-1} \]

Therefore, \( S^2 \) is unbiased and consistent.
## Summary of Sampling Distributions

<table>
<thead>
<tr>
<th>Sample Variance</th>
<th>One population</th>
<th>Two populations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known Variance</td>
<td>$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$</td>
<td>$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim NID(0,1)$</td>
</tr>
<tr>
<td>Unknown Variance</td>
<td>$\frac{\overline{X} - \mu}{S / \sqrt{n}} \sim t(n-1)$</td>
<td>$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$</td>
</tr>
<tr>
<td>Sample Variance</td>
<td>$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$</td>
<td>$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$</td>
</tr>
</tbody>
</table>
Mean Square Error (MSE)

Quality of estimator usually measured by another quantity mean square error

\[ MSE(\hat{\theta}) = E|\hat{\theta} - \theta|^2 \]

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

\[ MSE(\Theta) = \left[ \text{Bias}(\Theta) \right]^2 + \text{var}(\Theta) \]
General Concepts of Point Estimation: Minimum Variance Unbiased Estimator (MVUE)

If we consider all unbiased estimators of \( \theta \), the one with the smallest variance is called the **minimum variance unbiased estimator (MVUE)**.

\( \hat{\theta} \) is an MVUE if:

a) It is unbiased estimator of \( \theta \),

b) It satisfies the following equality,

\[
\text{var}(\hat{\theta}) = \frac{1}{nE \left[ \left( \frac{\partial \ln f(x)}{\partial \theta} \right)^2 \right]} \quad \text{Cramér-Rao Bound}
\]

If \( X_1, X_2, \ldots, X_n \) is a random sample of size \( n \) from a normal distribution with mean \( \mu \) and variance \( \sigma^2 \), the sample mean \( \bar{X} \) is the MVUE for \( \mu \).