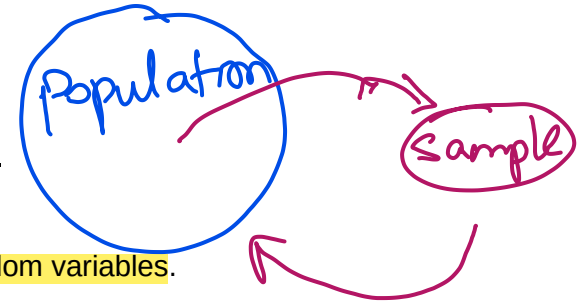
Statistical Inference



Ch 8.1: Introduction

So far, we were concerned with concepts of probability and random variables.

We used them to build models of nondeterministic (i.e., random) phenomena. We focused on distributions of random variables, including function of random variables such as the sample mean.

These distributions were sometimes used to determinethe likelihood of a particular event or the expected value/variance of the random variable.

In each of these instances, we assumed that we knew the correct probability model. Most often, however, we are interested in a numeric characteristic of a random phenomenon that cannot be computed/observed directly.

For example, we might be interested in knowing

Binomial (n, p) ?

- the percent of people in Oregon who buy organic food
- the effect of a teacher professional development program on student achievement
- the average change in blood sugar with a new diabetes drug
- the relationship between time spent playing video games and aggressive behavior
- the relationship between self-selected seating in a classroom and class performance

We typically do not know the true population parameter(s), such as the mean life lengths of patients who receive heart transplants and those who do not.

Instead, we observe one or more random variables whose distributions depend on the parameter of interest.

Then, we use **statistical inference** methods to analyze the observed values to gain knowledge about the unknown characteristic and/or make decisions based on the unknown characteristic. The rest of the term we will be focusing on developing these statistical inference methods.

Statistical Inference: (Definition)

- **Statistical Inference** refers generally to the methods by which one makes inferences or generalizations about a population and its parameters based on statistics calculated from sample data. These methods can be classified broadly into three forms:

  - **Point estimation** is the process by which an estimator of an unknown parameter, or function of parameters, is chosen and evaluated.

    The average age of PSU students is 27 years.

  - **Interval estimation** is the process by which an observed value of a statistic and the statistic's probability distribution are used to create a range of plausible values of an unknown parameter, or function of parameters.

    The average age of PSU students is between 25 and 27 years.
    $(25, 29)$

  - **Hypothesis testing** is the process by which an observed value of a statistic and the statistic's probability distribution are used to evaluate the plausibility of a statement made about the unknown value of a parameter.

    Test if the average age of PSU students is higher than 26 years.

Point Estimation

The basic situation is point estimation is as follows:

- We want to estimate an unknown population parameter θ (example, μ or λ; possibly vector-valued) or some function of parameters.
- We find an **estimator**, that is, a function of the random variables $Y_1, Y_2, \ldots, Y_n$, (example: $\bar{Y}$ or S^2). An estimator is denoted by $\hat{\theta}$.
- We observe random variables $Y_1, Y_2, \ldots, Y_n$ (not necessarily **iid**), where the distribution of $Y_1, Y_2, \ldots, Y_n$ depends on the unknown parameter θ.
- We calculate an **estimate**, that is, a function of the realized (observed) values $y_1, y_2, \ldots, y_n$ (example, $\bar{y}$ or s^2).

$\begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$

Big $\bar{Y}$ and S^2 are random variables

The small $\bar{y}$ and s^2 are single observed values.

During our discussion of point estimation, we are going to focus on two main topics:

1. Methods of evaluating the "goodness" of statistical estimators and
2. Methods of **finding estimators**, in particular Method of Moments Estimators (**MME**s) and Maximum Likelihood Estimators (**MLE**s).

Ch 8.2: The bias and Mean square Error of Point Estimators

Definition 8.2:

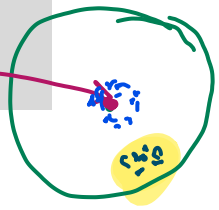
Let $\hat{\theta}$ be a point estimator for a parameter θ.

Then $\hat{\theta}$ is **an unbiased estimator** if $E(\hat{\theta}) = \theta$.

If $E(\hat{\theta}) \neq \theta$, then $\hat{\theta}$ is said to be **biased**.

$\Rightarrow$ In the long-run the average of $\hat{\theta}$ will reach θ.

target (θ)



We did prove $E(\bar{Y}) = \mu$ and therefore, $\bar{Y}$ is an unbiased estimator for μ.

When $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$, we can prove $E(S^2) = \sigma^2$ and therefore, S^2 is an unbiased estimator for σ^2.

Definition 8.3:

The **bias** of a point estimator $\hat{\theta}$ is $B(\hat{\theta}) = E(\hat{\theta}) - \theta$.

parameter

Let $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Then $E(\hat{\theta}) = \frac{(n-1)}{n}\sigma^2 \neq \sigma^2$

Therefore, $\hat{\theta}$ is a biased estimator for σ^2.

Bias is $B(\hat{\theta}) = E(\hat{\theta}) - \theta = \frac{(n-1)}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$

Since this bias is negative, $\hat{\theta}$ underestimates σ^2.

If there are **several unbiased estimators** for the same parameter, which one should we prefer?

The one which is most consistent, least variable and most precise.

The unbiased estimator with the smallest variance is called "most efficient" unbiased estimator.

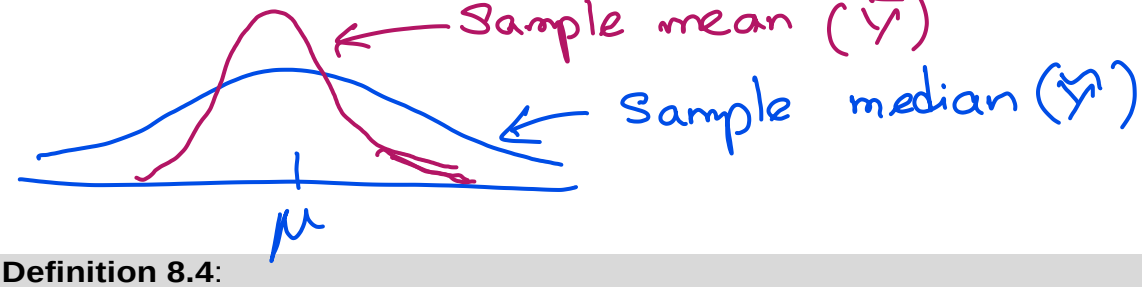
Sample median is middle value after sorting the values in increasing or decreasing order.

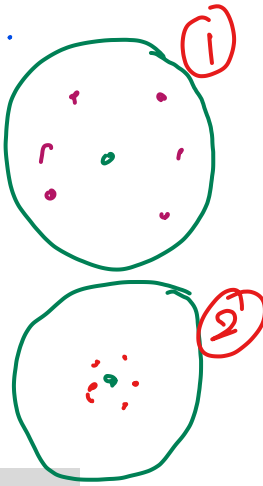
Let $\tilde{Y}$ is the sample median.

Sample median is also an unbiased estimator for population mean. $E(\tilde{Y}) = \mu$.

However, sample mean is the most efficient unbiased estimator of the population mean.

because $Var(\bar{Y}) < Var(\tilde{Y})$





Definition 8.4:

The **mean square error** of a point estimator $\hat{\theta}$ is $MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$.

Further, $MSE(\hat{\theta}) = Var(\hat{\theta}) + [B(\hat{\theta})]^2$

This gives the typical squared difference between estimator ($\hat{\theta}$) and the parameter (θ).

Ch 8.3: Some Common Unbiased Point Estimators

Let $Y_1, Y_2, \ldots, Y_n$ be an **iid** random sample from a population with $E(Y_i) = \mu$ and $Var(Y_i) = \sigma^2$.
(Note that a specific population distribution has not been provided!).

Target Parameter (θ)	Sample size(s)	Point Estimator ($\hat{\theta}$)	$E(\hat{\theta})$	Standard Error $\sigma_{\hat{\theta}}$
$\boldsymbol{\mu}$	n	$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$	μ	$\frac{\sigma}{\sqrt{n}}$
$\boldsymbol{\sigma^2}$	n	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$	σ^2	$\sqrt{\frac{2}{n-1}}\,\sigma$
$\boldsymbol{p}$	n	$\hat{p} = \frac{Y}{n}$	p	$\sqrt{\frac{p(1-p)}{n}}$
$\boldsymbol{\mu_1 - \mu_2}$	n_1 and n_2	$\overline{Y}_1 - \overline{Y}_2$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
$p_1 - p_2$	n_1 and n_2	$\hat{p}_1 - \hat{p}_2$	$\boldsymbol{p_1 - p_2}$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$