2. Random Variables and Distributions
Expected Value and Variance
Properties and Types of Estimators

Readings: Rosner: Ch. 4 and 5; Chihara and Hesterberg: Ch. 6
OpenIntro Statistics: 2.2, 2.5, 2.6

Homework: Homework 1 due by midnight on September 4

Overview

A) Definitions and Notation
B) Discrete: Probability Mass Function
C) Discrete: Expected Value (Mean), Variance, SD
D) Continuous: Probability Density Function
E) Continuous: Expected Value (Mean), Variance, SD
F) Joint, Marginal, Conditional Distributions
G) Properties and types of estimators

A) Definitions and Notation

Random variables and probability distributions are the theoretical or mathematical representations of data values and frequency distributions.

Random variable (r.v.): Quantity that takes on different values or sets of values with various probabilities. A numerical function that assigns a number to each possible outcome (i.e. point in the sample space) of a random trial.
Convention: capital letters: X, Y, etc. for r.v.; small letters for values of r.v., *x*, *y*, etc.

Discrete random variable: can only take on a finite or countable number of values
e.g. number of children in family, number of cases of disease

Continuous random variable: can take on any value in an interval
e.g. height, weight, food intake

Probability Distribution: describes the probabilities for each outcome for

- a discrete random variable (probability mass function, *pmf*)
- the probabilities of values in a range for a continuous random variable (probability density function, *pdf*)

Probability mass functions, probability density functions and cumulative distribution functions are useful as tools for describing the frequency distributions of random variables for the entire population of interest and as tools for providing probability statements about events involving random variables.

B) Discrete r.v.: Probability Mass Function

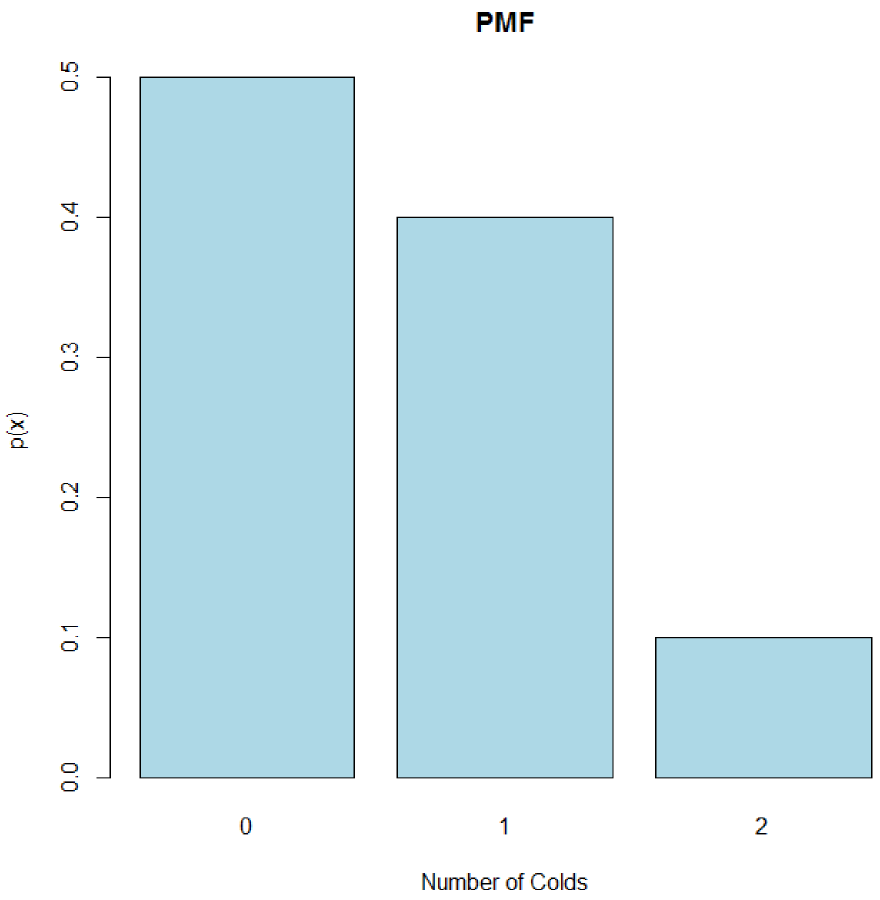
Let X be a discrete random variable and let *x* represent the values that X can take on.

Probability distribution of X is p(*x*) = P(X = *x*)

e.g. X = number of colds in a year caught by healthy adult = 0, 1, 2

Number of Colds:	0	1	2
P(X=*x*)	0.5	0.4	0.1

Note: All probabilities are nonnegative and the sum over all mutually exclusive and exhaustive values of the r.v. X is 1.



Cumulative Distribution Function (CDF): cumulative probability distribution of X

$F_x(x) = P(X \leq x) = \sum_{x=0}^{k} P(X = x)$ (accumulate probabilities from lowest to highest)

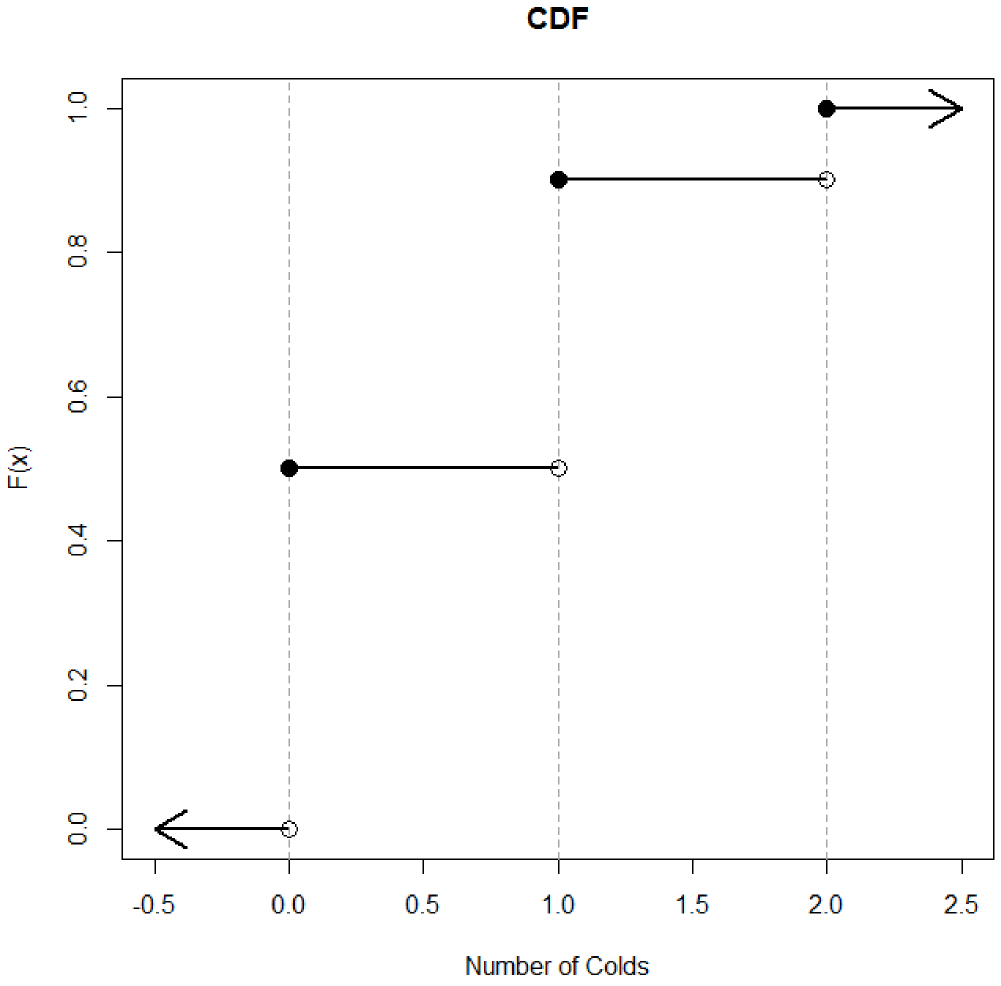
The CDF is monotone ↑, F(-∞) = 0, F(∞) = 1

# of Colds:	0	1	2
P(X=*x*)	0.5	0.4	0.1

F(0) = 0.5 (number of colds <= 0)

F(1) = 0.5 + 0.4 (number of colds <=1)

F(2) = 0.5 + 0.4 + 0.1 = 1.0

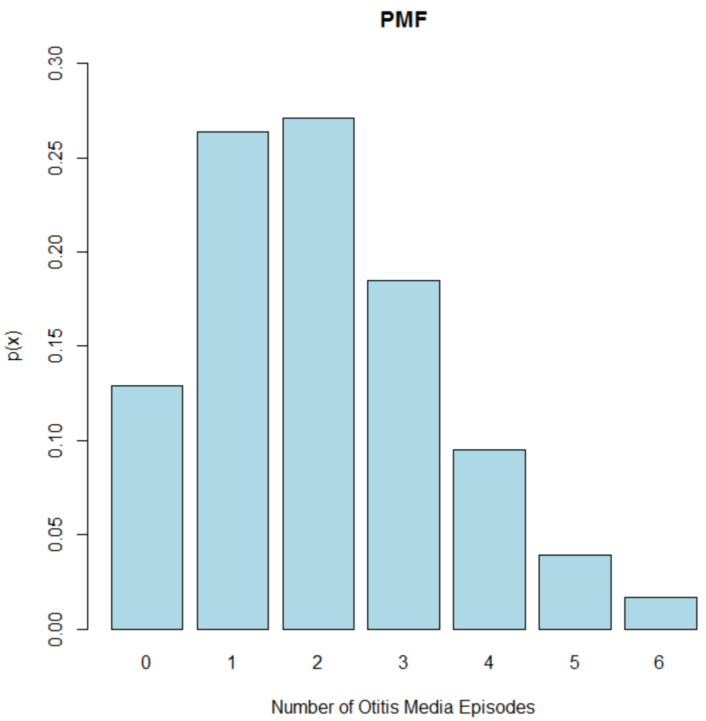


e.g. X = number of episodes of otitis media (disease of middle ear) in first 2 years of life:

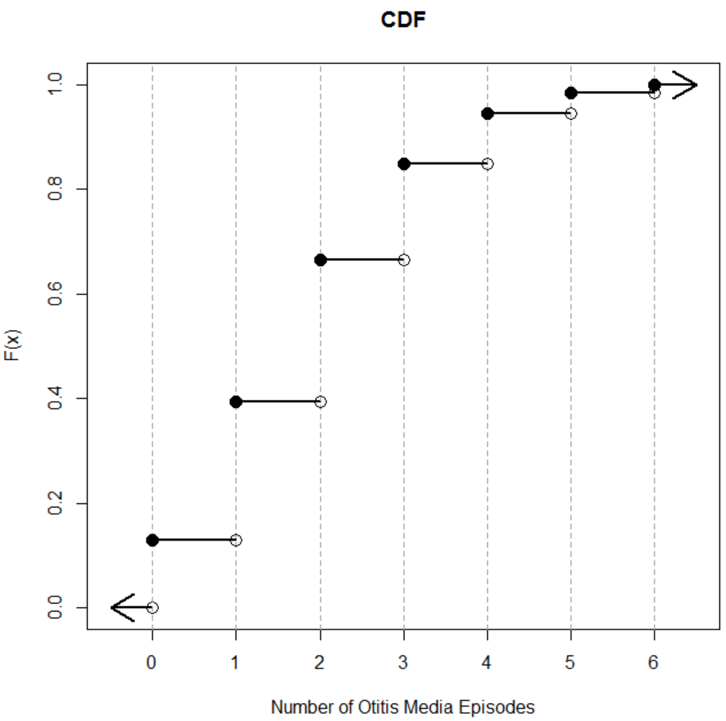
x	0	1	2	3	4	5	6
P(X=*x*)	0.129	0.264	0.271	0.185	0.095	0.039	0.017

First we want to check: $\sum p_i = 1$

PMF: P(X = *x*)



CDF: P(X ≤ *x*)



C) Discrete: Expected Value (Mean), Variance, Standard Deviation (SD)

There are summary values for random variables. For a discrete r.v. X:

Expected Value (also known as the Expectation, Mean)

$$\begin{aligned} E(X) &= \sum_{possible\ x} x\, P(X = x) &&= \mu \\ &= \sum_{x} x\, p(x) &&= \mu \end{aligned}$$

x represents possible values

$P(X = x)$, $p(x)$ represent the weight (probability) for a given x

$E(X)$ is the balance point (center of gravity in physics) on the graph

e.g. X = number of colds in a year caught by healthy adult = 0, 1, 2

Number of Colds:	0	1	2
P(X=*x*)	0.5	0.4	0.1

$\mu = E(X) =$

Variance and Standard Deviation

Var(X) = V(X) = $\sigma^2 = E\left[\left(X - E(X)\right)^2\right] = E[(X-\mu)^2]$

For discrete variables:

$$\begin{aligned} Var(X) &= \textstyle\sum_{possible\ x} (x-\mu)^2\, P(X=x) &&= \sigma^2 \\ &= \textstyle\sum_x \quad (x-\mu)^2\, p(x) &&= \sigma^2 \end{aligned}$$

$$SD(X) = \sqrt{\textstyle\sum_{possible\ x} (x-\mu)^2\, P(X=x)} = \sqrt{Var(X)} = \sigma$$

Number of Colds:	0	1	2		Recall:
P(X=*x*)	0.5	0.4	0.1		E(X)=0.6

Var(X) = σ^2 =

SD(X) = σ = $\sqrt{\quad\quad}$ = colds per year

Computational formula: Var(X) = E[X^2] – (E[X])^2

E(X^2) =

Var(X) =

D) Continuous r.v.: the Probability Density Function (pdf)

For a random variable measured on a continuous scale, there are an infinite number of possible values between any 2 adjacent points on the scale.

Thus the probability of observing any specific value on the scale is 0, or P(X = *x*) = 0.

To work with probabilities for continuous r.v., we must instead consider an interval of values on the scale, e.g. (a, b).

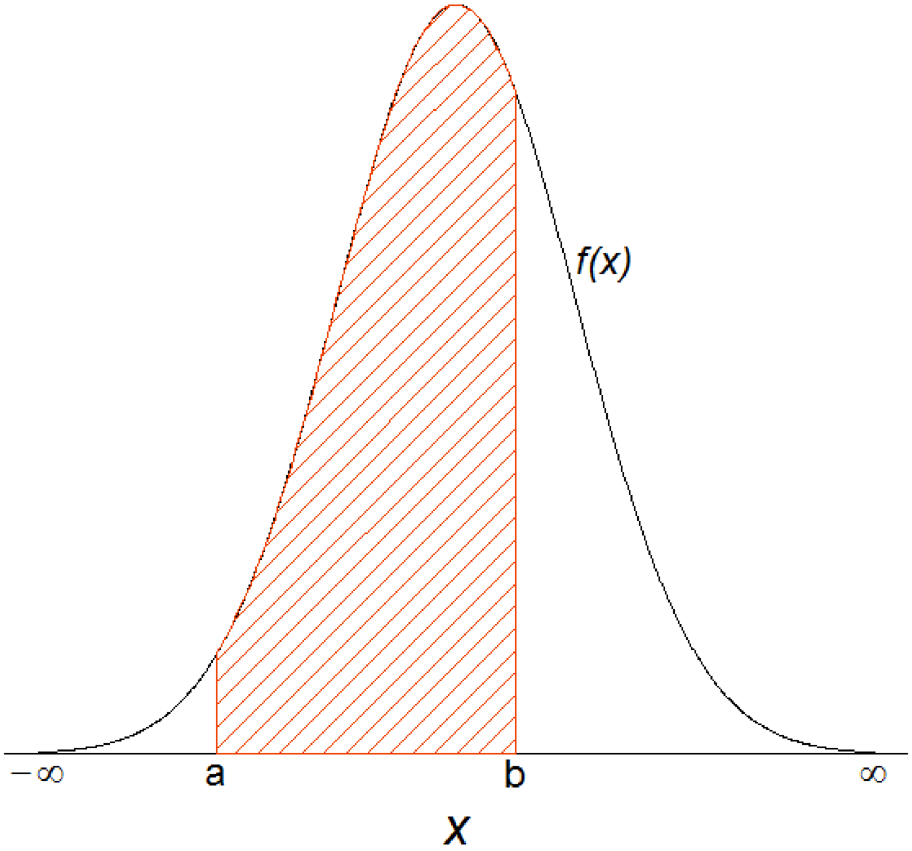
The idea of a probability mass function does not apply to continuous scale r.v. Instead we denote the function that assigns probability to values of the r.v. as *f(x)dx*. This is called the *probability density function* (pdf).

P(a ≤ X ≤ b) = area under the pdf, *f(x)dx*, from a to b = $\int_a^b f(x)dx$

Note: f(*x*) ≥ 0 for all *x,* and $\int_{-\infty}^{\infty} f(x)dx = 1$

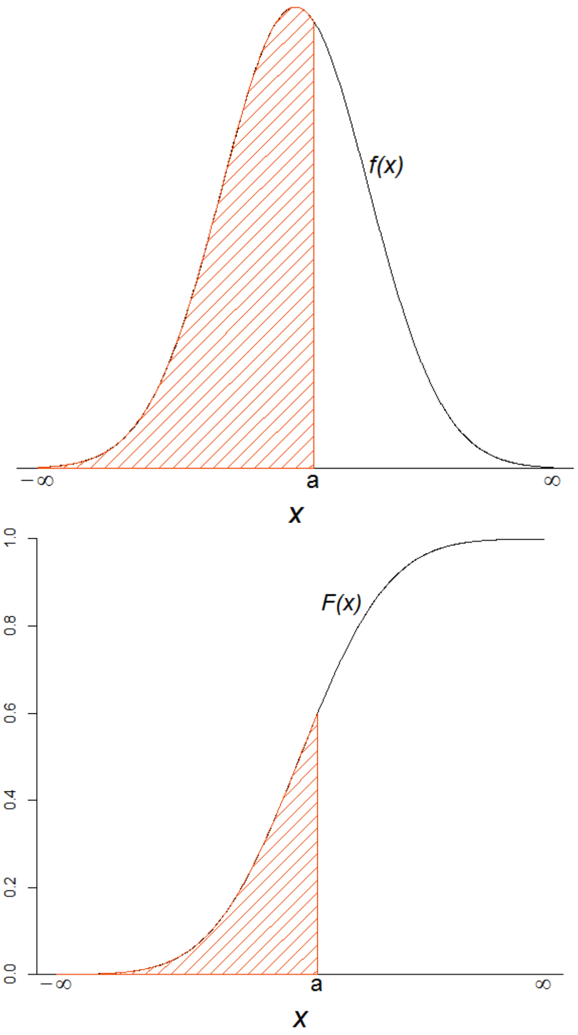
The pdf does not give probability values but, rather, tells what *set* of values is most likely:

$$P(a \leq X \leq b) = P(X \leq b) - P(X < a) = F(b) - F(a)$$



Cumulative Distribution Function: CDF

For a continuous r.v.: $\int_{-\infty}^{a} f(x)dx$ is the area under the curve from -∞ to *a*. We call this the *cumulative distribution function* (cdf) of X evaluated at the value *a*, F(*a*).



The cdf is a monotone increasing function.

F(-∞) = 0

F(+∞) = 1

CDF and PDF relationship: $f(x) = \frac{d}{dx}F(x)$

Every time we execute a statistical test or determine a p-value (level of significance), we will be using the cdf of a relevant continuous or discrete random variable.

E) Continuous: Expected Value (Mean), Variance, SD

Expected Value: Observations weighted by density function over the range of X:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{all\ x} xf(x)dx = \mu$$

Variance:

$$Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{all\ x} (x-\mu)^2 f(x)dx = \sigma^2$$

Standard Deviation:

$$s.d.(X) = \sqrt{Var(X)} = \sigma$$

F) Joint, Marginal and Conditional Distributions of r.v.

A respiratory disease (Y) and smoking (X) example:

	Non-Smoker (X = 0)	Smoker (X = 1)	Total
No respiration problem (Y = 0)	0.50	0.30	0.80
Respiration problem (Y = 1)	0.05	0.15	0.20
Total	0.55	0.45	1.00

This is a distribution defined by two r.v. – i.e. it's a *bivariate* distribution:
X = 1 if smoker, 0 if non-smoker
Y = 1 if respiratory problems, 0 if not

Joint distribution: P (X = x and Y = y) = P (X = $x \cap$ Y = y)

P (X = 0 $\cap$ Y = 1) = P (non-smoker and respiration problem) = 0.05

Marginal distributions: $P(X = x)$; $P(Y = y)$

	Non-Smoker (X = 0)	Smoker (X = 1)	Total
No respiration problem (Y = 0)	0.50	0.30	0.80
Respiration problem (Y = 1)	0.05	0.15	0.20
Total	0.55	0.45	1.00

$P(X = 0) = P(\text{non-smoker}) = 0.55$

$P(Y = 0) = P(\text{no respiration problem}) = 0.80$

Conditional distributions:

$P(X = x \mid Y = y) = \frac{P(X=x \cap Y=y)}{P(Y=y)}$

$P(Y = 1 \mid X = 0) = P(\text{resp problem given non-smoker}) = \frac{P(X=0 \cap Y=1)}{P(X=0)} =$

$P(Y = 1 \mid X = 1) = P(\text{resp problem given smoker}) = \frac{P(X=1 \cap Y=1)}{P(X=1)} =$

Independence of two r.v.: X and Y are independent *iff* (if and only if)

$$P(X = x \cap Y = y) = P(X = x) \times P(Y = y), \text{ or}$$
$$P(Y = y \mid X = x) = P(Y = y)$$

$P(X = 1 \cap Y = 1) = 0.15$

$P(X = 1) \times P(Y = 1) = (0.45) \times (0.20) = 0.09$

Therefore, X and Y are *not* independent.

Roughly, two random variables are independent if knowing the value of one does not change the probability distribution of the other. Which sequences $\{X_1, X_2, \ldots, X_n\}$ are independent?

1. X_i = high temperature in Denver on day *i*

2. X_i = color of car *i* in a row of parked cars

3. X_i = 1 if has flu, 0 if not for people working in an office building

4. X_i = religion of person *i*, where people are selected randomly from a phone book?

5. X_i = religion of person *i*, where people are selected in alphabetical order from a phone book?

G) Properties and Types of Estimators

Bias – the difference between the estimator's expected value and the true value of the parameter we are estimating

Unbiasedness – not sample size dependent, a bias of 0

e.g. E(X) = population mean μ and E($\bar{X}$) = μ, regardless of the sample used to obtain $\bar{X}$

Example:

Let X_1, X_2, X_3 be a random sample of size 3 from any distribution with mean parameter μ.

X – heights of sons whose fathers are over 5'10"

Estimator 1 = X_1, Estimator 2 = $(X_1 + X_2)/2$, Estimator 3 = $(X_1 + 2X_2)/3$.

Are these estimators unbiased? (fill-in during class …)

Consistency – sample size dependent unbiasedness, an estimator converges *in probability* to the true population parameter – asymptotic result, i.e. approaches true population parameter as sample size gets large

Median consistent estimator of the mean (for symmetric distributions) – Homework 2

Mean Square Error = Bias^2 + Variance – good for comparing biased estimators to each other, tradeoff can be important

Efficiency – variability (and power; more on power later)

Example: Let X_1, X_2, X_3 be a random sample of size 3 from any distribution with variance parameter σ^2. X – heights of sons whose fathers are over 5’10”

Estimator 1 = X_1, Estimator 2 = $(X_1 + X_2)/2$, Estimator 3 = $(X_1 + 2X_2)/3$.

Which estimator is most efficient? (i.e. has the smallest variance) (fill-in during class …)

Types of Estimators

Population parameters – mean, variance, proportions, etc.

Method of Moments (MoM) Estimators – based on powers of X_i

e.g. Sample mean is a function of X_i^1:

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

Sample variance is a function of X_i^2:

$$s^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n}$$

Sample skewness is a function of X_i^3:

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})^3}{n}$$

etc.

Maximum Likelihood Estimators (MLEs)

If X_1, X_2, …, X_n follow the same distribution $f_X(x; \theta)$, then the likelihood function *L* for a sample of n independent and identically distributed observations, x_1, x_2, …, x_n:

$$L \propto \prod_{i=1}^{n} f_X(x_i; \theta),$$

where θ is a population parameter(s) that define the distribution (e.g., μ and σ^2 for a normal distribution, λ for a Poisson, etc.).

By maximizing the function L with respect to θ we obtain $\hat{\theta}$ ("theta-hat"), the value of the population parameter that makes the data most likely to have been observed.

Steps:

1. Take first derivative with respect to θ and set equal to 0
2. Solve using numerical methods, sometimes closed form is possible
3. Check solution is a maximum by taking second derivative with respect to θ

$\hat{\theta}$ is known as the maximum likelihood estimator (MLE) of θ.

MLE Example:

Sample of size n, X – height of U.S. women over age 20 with mean μ, variance σ^2

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n};\ \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \hat{\mu})^2}{n}$$

MLE Notes:

- MLEs are sometimes not unbiased, but they are usually consistent and they converge to the true population parameter faster than MoM estimators

- MLEs often have the smallest variance compared with other estimators, like MoM estimators

Estimators for Regression Models

Ordinary Least Squares (OLS) for linear models – identical to MLE (will see later – Lecture 22)

For regression models that are not linear, MLE and weighted variants of OLS tend to have best properties