Econ 325: Introduction to Empirical Economics

Lecture 6

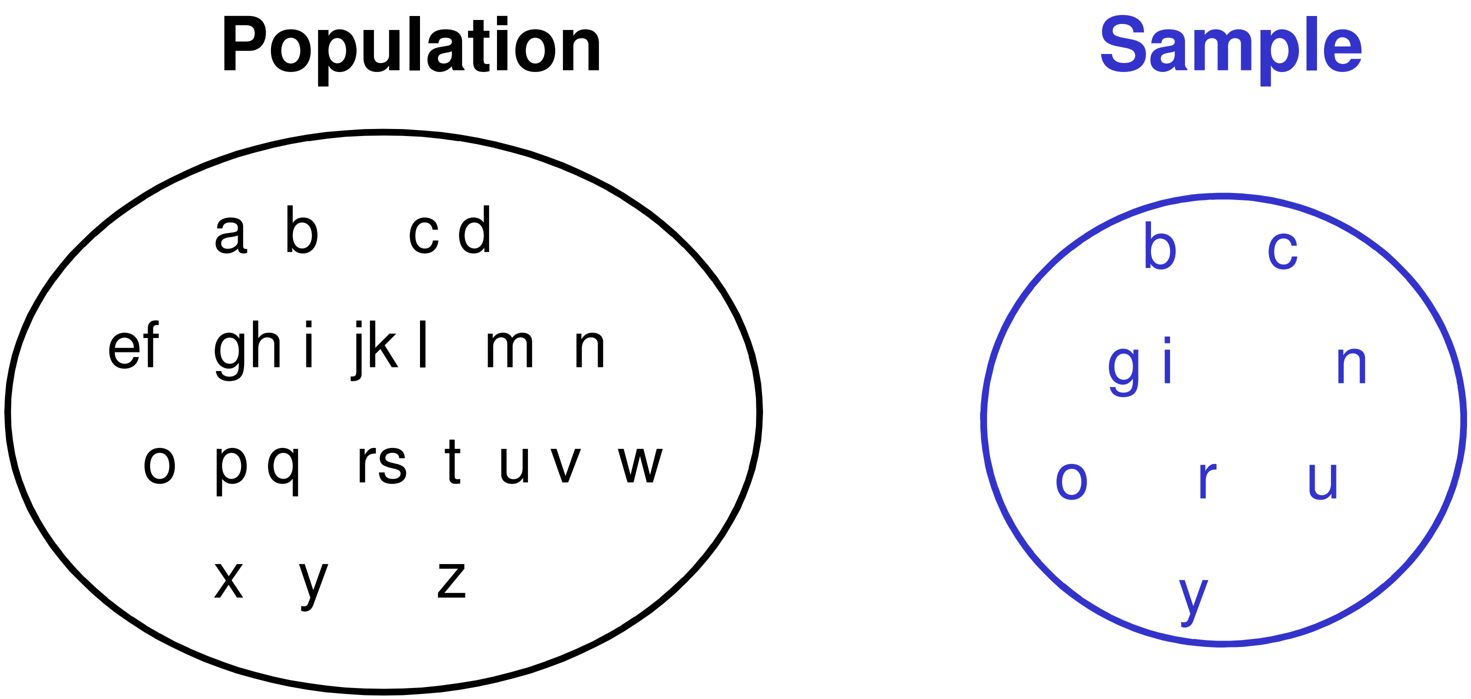
Sampling and
Sampling Distributions

Populations and Samples

- A **Population** is the set of all items or individuals of interest
  - Examples: All likely voters in the next election
    All parts produced today
    All sales receipts for November

- A **Sample** is a subset of the population
  - Examples: 1000 voters selected at random for interview
    A few parts selected for destructive testing
    Random receipts selected for audit

Population vs. Sample

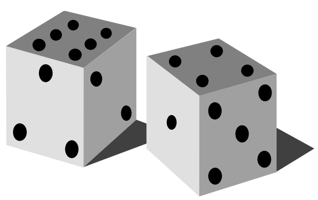


Why Sample?

- Less time consuming than a census
- Less costly to administer than a census

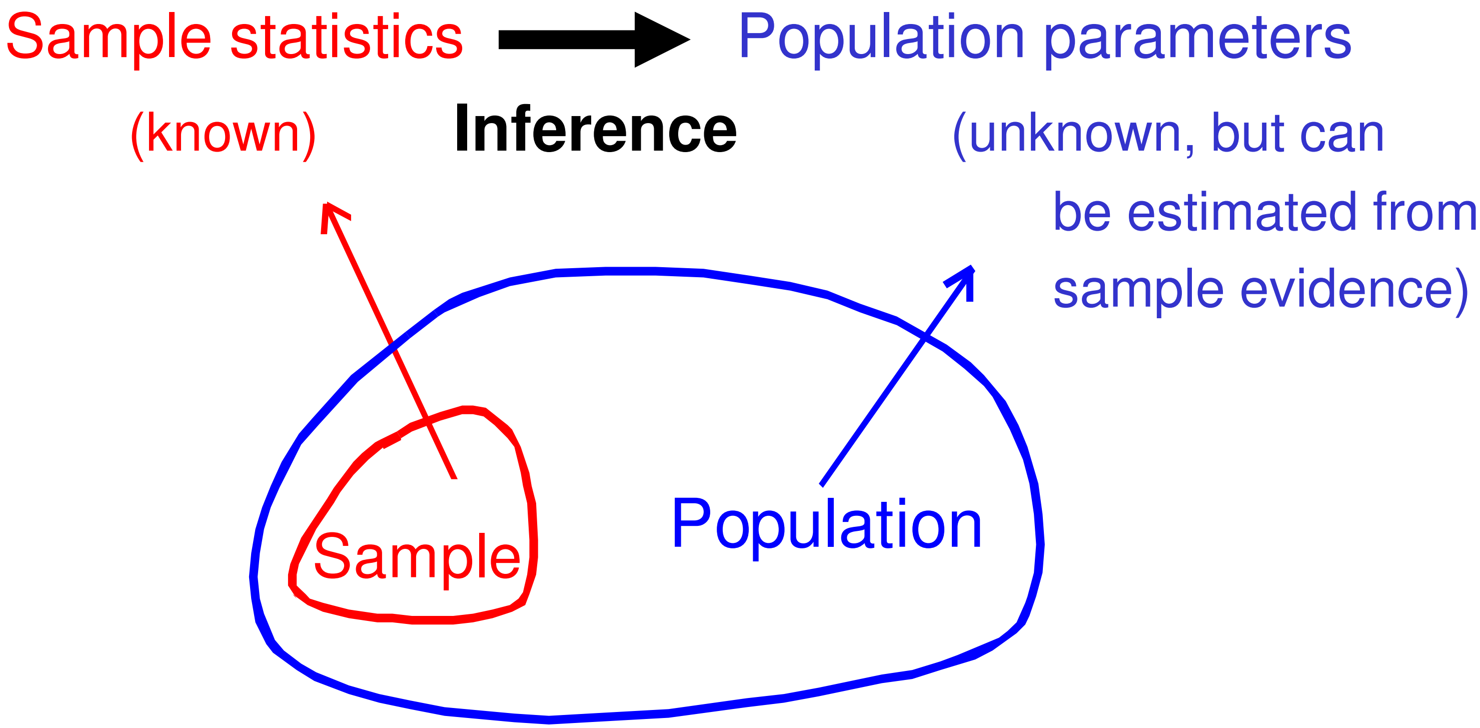
Random Sampling

- Every object in the population has an equal chance of being selected
- Objects are selected independently



Inferential Statistics

- Making statements about a population by examining sample results



Inferential Statistics

Drawing conclusions and/or making decisions concerning a population based on sample results.

- **Estimation**
  - e.g., Estimate the population mean weight using the sample mean weight
- **Hypothesis Testing**
  - e.g., Use sample evidence to test the claim that the population mean weight is 120 pounds

Sampling Distribution

- **Assume there is a population …**
- Four types of people
- Random variable, X, is age of individuals
- Possible Values of X:

  18, 20, 22, 24 (years)

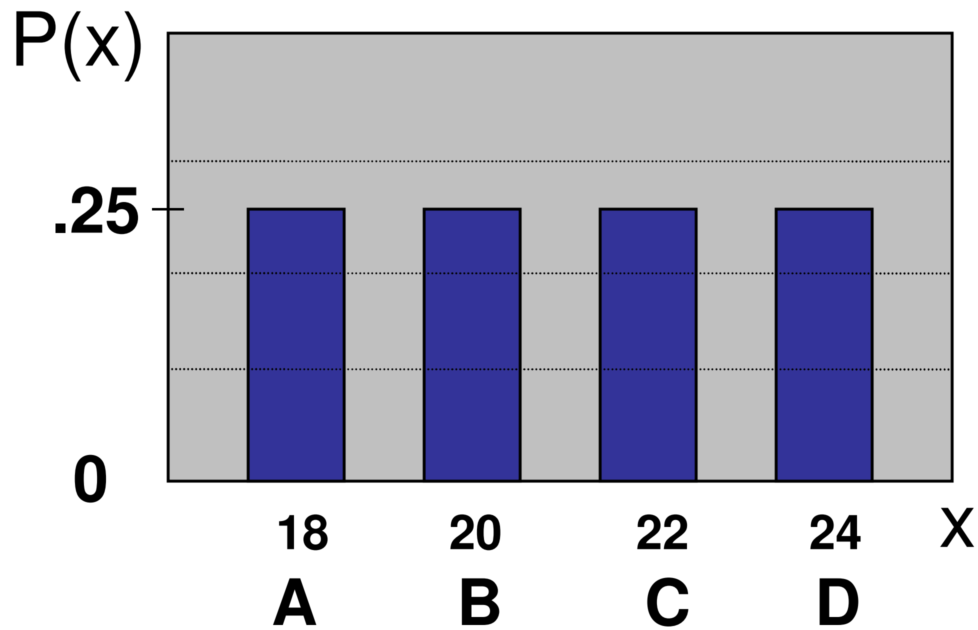
Sampling Distribution

(continued)

Summary Measures for the Population Distribution:

$$\mu = \frac{18 + 20 + 22 + 24}{4} = 21$$

$$\sigma^2 = \sqrt{\frac{\sum (X_i - \mu)^2}{4}} = 2.236$$



Uniform Distribution

Sampling Distribution

(continued)

Now consider all possible samples of size n = 2

1st Obs	2nd Observation			
	18	20	22	24
18	18,18	18,20	18,22	18,24
20	20,18	20,20	20,22	20,24
22	22,18	22,20	22,22	22,24
24	24,18	24,20	24,22	24,24

16 possible samples (sampling with replacement)

16 Sample Means

1st Obs	2nd Observation			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

Sampling Distribution

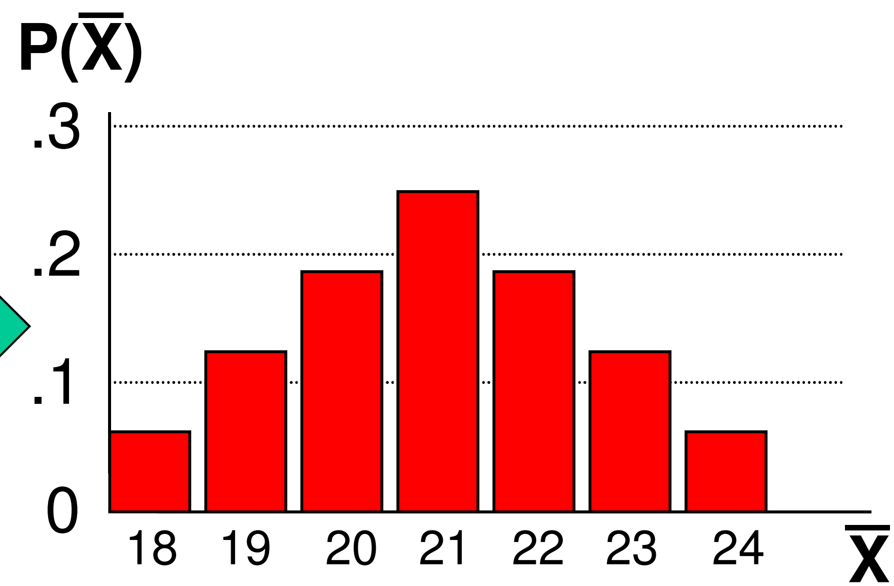
(continued)

Sampling Distribution of All Sample Means

16 Sample Means

1st Obs	2nd Observation 18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

Sample Means Distribution



(no longer uniform)

Developing a Sampling Distribution

(continued)

Summary Measures of this Sampling Distribution:

$$\mu_{\overline{X}} = \frac{18 + 2\times 19 + 3\times 20 + \cdots + 2\times 23 + 24}{16} = 21$$

$$\sigma_{\overline{X}} = \sqrt{\frac{(18\text{-}21)^2 + 2\times(19\text{-}21)^2 + \cdots + (24\text{-}21)^2}{16}} = 1.58$$

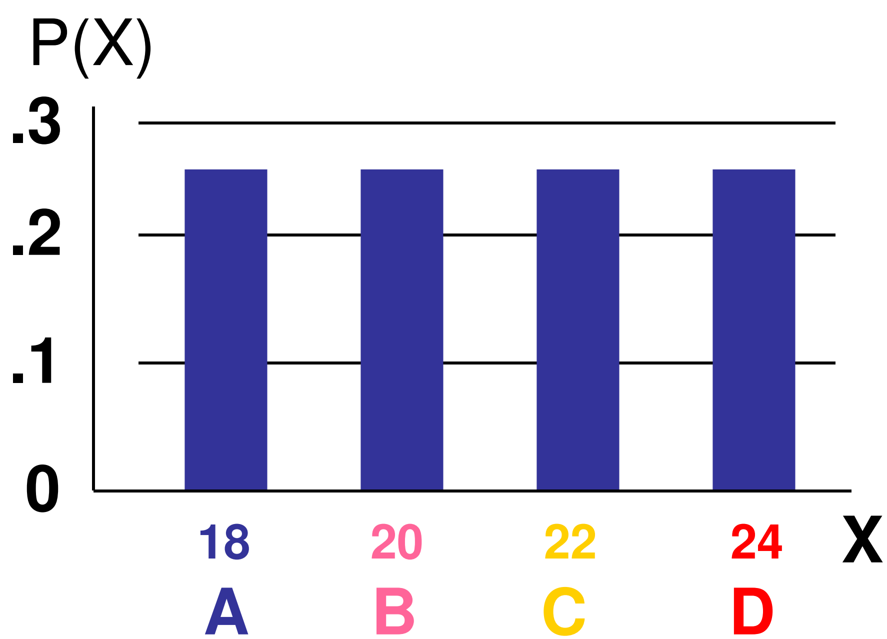
Comparing the Population with its Sampling Distribution

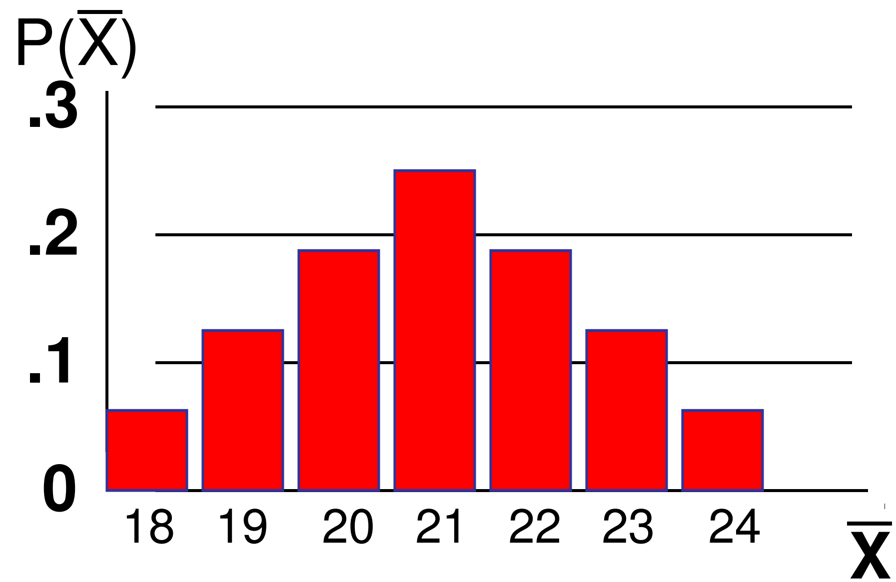
Population

$\mu = 21 \quad \sigma = 2.236$

Sample Means Distribution
n = 2

$\mu_{\bar{X}} = 21 \quad \sigma_{\bar{X}} = 1.58$





Expected Value of Sample Mean

- Let $X_1, X_2, \ldots X_n$ represent a random sample from a population

- The sample mean value of these observations is defined as

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

Standard Error of the Mean

- A measure of the variability in the mean from sample to sample is given by the Standard Error of the Mean:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

- If n=2, then

$$\sigma / \sigma_{\overline{X}} = \sqrt{2} = 1.4142$$

Clicker Question 6-1

- Suppose a random sample of size n = 36 is drawn from the population distribution with mean μ = 8 and standard deviation σ = 3. What is the standard deviation of the sample mean $\bar{X} = \frac{1}{36}\sum_{i=1}^{36} X_i$?

  A). 3

  B). 1/13

  C). 1/2

Clicker Question 6-2

- What will happen to the variance of the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ as the sample size n goes to infinity?

A). Var($\bar{X}$) goes to 1 as $n \to \infty$

B). Var($\bar{X}$) goes to 0 as $n \to \infty$

C). Var($\bar{X}$) goes to infinity as $n \to \infty$

Law of Large Numbers

Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $\{X_1, X_2, \ldots, X_n\}$ is a random sample with finite mean and finite variance.

Then, for any $\epsilon > 0$,

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| < \epsilon) = 1$$

Law of Large Numbers

We say that $\bar{X}_n$ **converges in probability to** $\boldsymbol{\mu}$, which is denoted as

$$\bar{X}_n \xrightarrow{p} \mu$$

What is the distribution of $\bar{X}_n$?

- As $n \to \infty$, $\bar{X}_n$ converges in probability to a constant value μ and, therefore, $\bar{X}_n$ is not random in the limit.

- However, when n is finite, $\bar{X}_n$ is a random variable.

- What is the distribution of $\bar{X}_n$ when n is finite?

If the Population is Normal

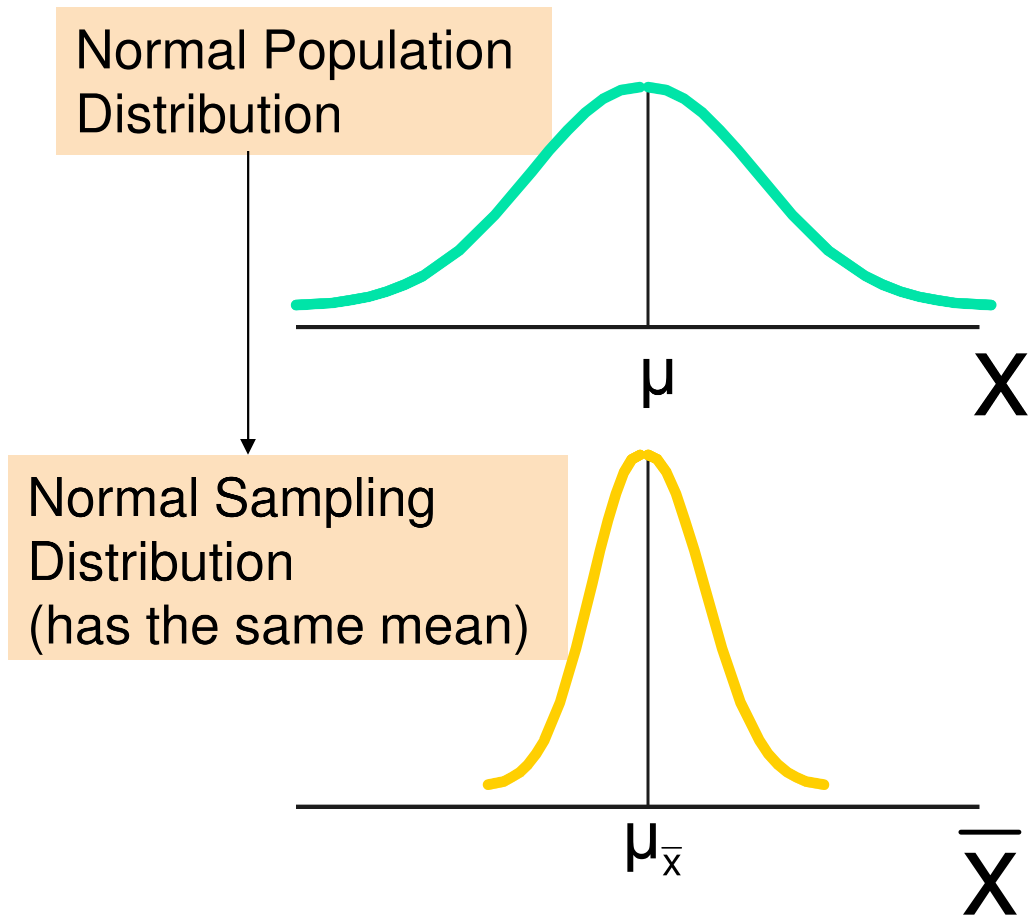
- If a population is normal, $\bar{X}_n$ is also normally distributed, i.e.,

$$\bar{X}_n \sim N(\mu\,, \sigma^2/n)$$

Sampling Distribution Properties

$$\mu_{\bar{X}} = \mu$$

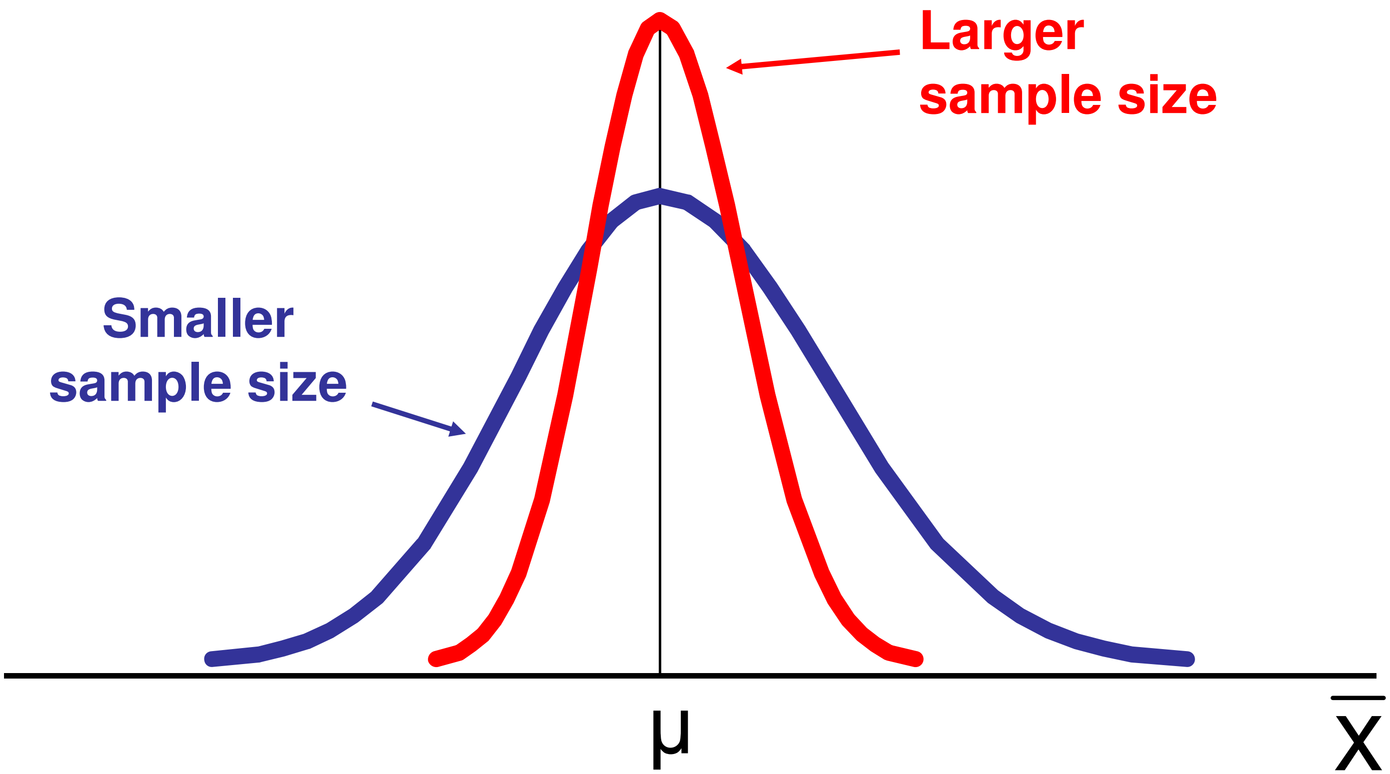
(i.e. $\bar{X}$ is unbiased)



Sampling Distribution Properties

(continued)

- As n increases, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ decreases!

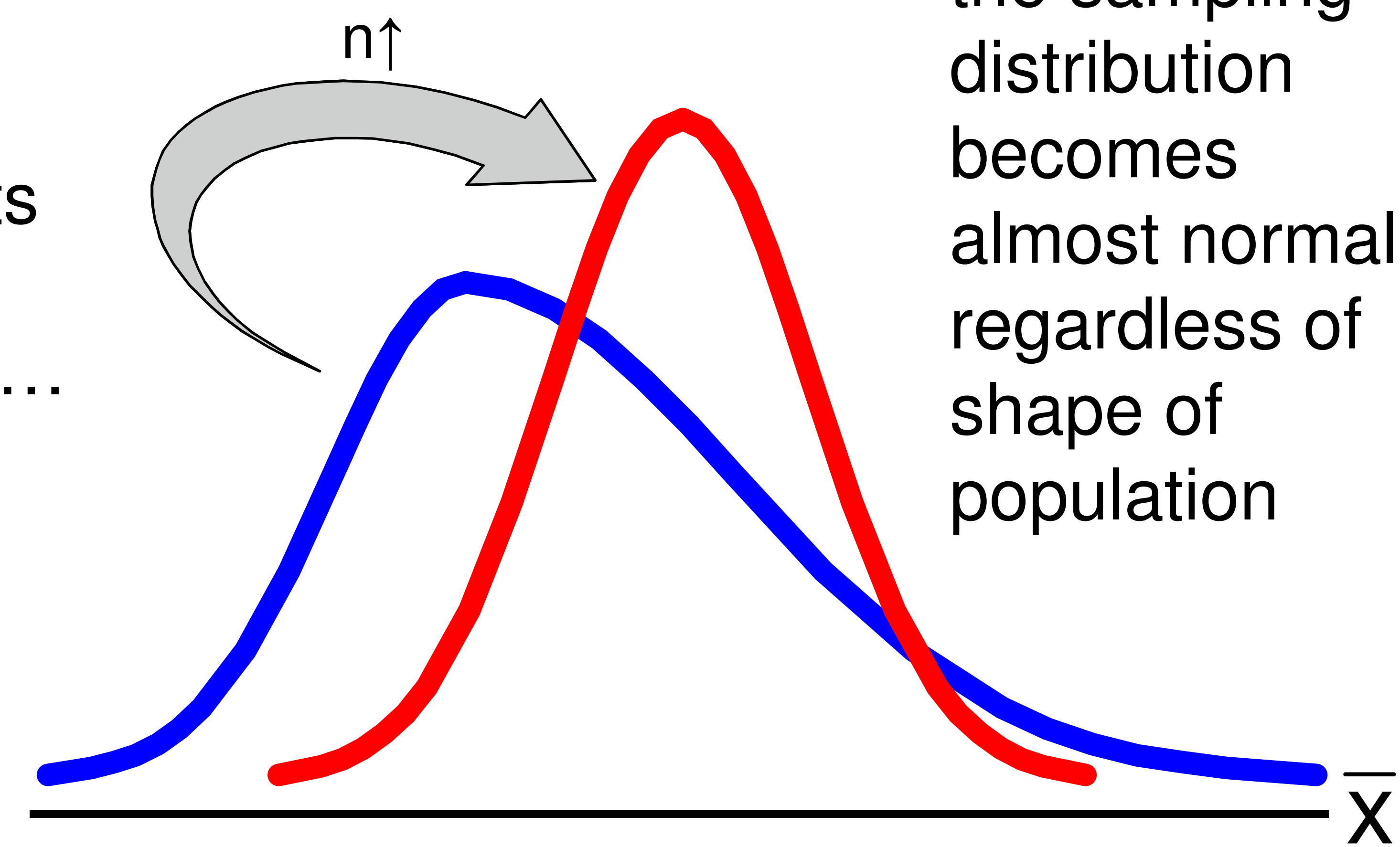


If the Population is **not** Normal

- We can apply the Central Limit Theorem:
  - Even if the population is not normal,
  - …sample means from the population will be approximately normal as long as the sample size is large enough.

Central Limit Theorem

As the sample size gets large enough…



the sampling distribution becomes almost normal regardless of shape of population

If the Population is **not** Normal

(continued)

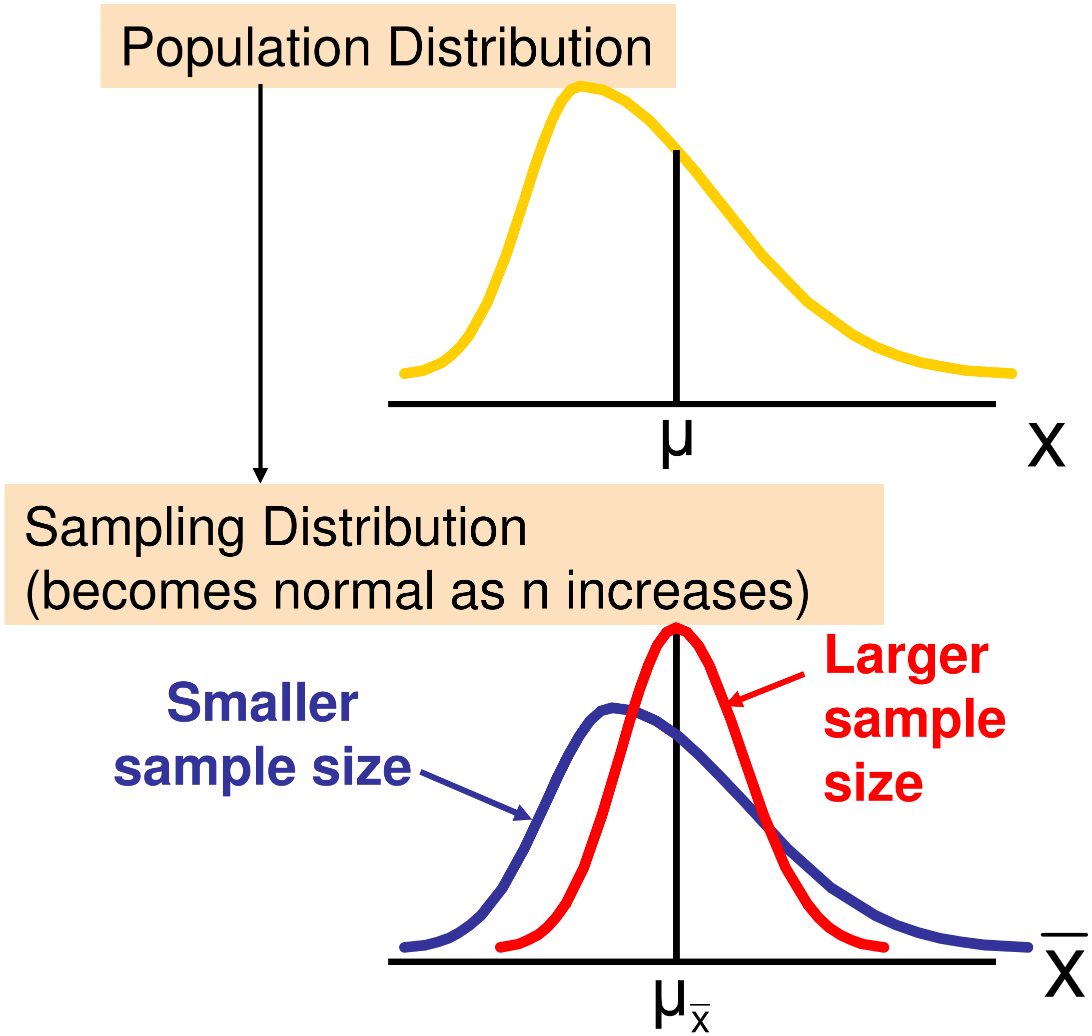
Sampling distribution properties:

Central Tendency

$$\mu_{\bar{x}} = \mu$$

Variation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$



Z-value for Sampling Distribution of the Mean

- Z-value for the sampling distribution of $\bar{X}$:

$$Z = \frac{(\bar{X} - \mu)}{\sigma / \sqrt{n}}$$

where $\bar{X}$ = sample mean

μ = population mean

σ = standard deviation of X_i

Central Limit Theorem

Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $\{X_1, X_2, \ldots, X_n\}$ is a random sample with finite mean and finite variance.

Define $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ and then,

$$\lim_{n \to \infty} P(Z_n < x) = \Phi(x)$$

where $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$

Central Limit Theorem

We say that $\sqrt{n}(\bar{X}_n - \mu)$ **converges in distribution** to a normal with mean 0 and variance σ^2, which is denoted as

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

Example

- Suppose a random sample of size $n = 36$ is drawn from the population distribution with mean $\mu = 8$ and standard deviation $\sigma = 3$.

- What is the approximated probability that the sample mean is between 7.8 and 8.2?

Example

(continued)

Solution:

- Even though the population is not normally distributed, we use the central limit theorem to get an approximated solution
- … the sampling distribution of $\bar{X}$ is approximately normal
- … with mean $\mu_{\bar{x}} = 8$
- …and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$

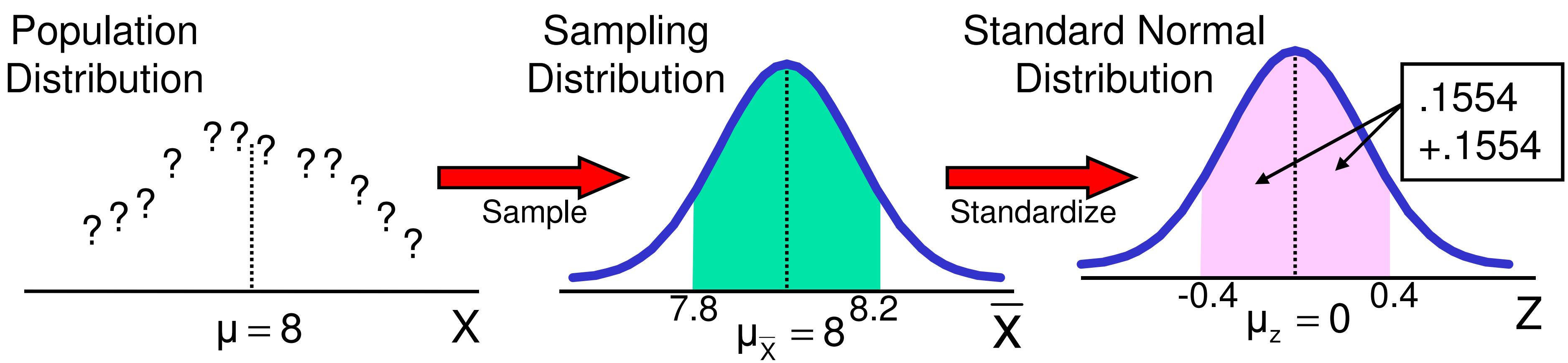
Example

(continued)

Solution (continued):

$$P(7.8 < \bar{X} < 8.2) = P\left(\frac{7.8-8}{3/\sqrt{36}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{8.2-8}{3/\sqrt{36}}\right)$$

$$= P(-0.4 < Z < 0.4) = 0.3108$$



Clicker Question 6-3

- Suppose a random sample of size $n = 36$ is drawn from the population distribution with mean $\mu = 8$ and standard deviation $\sigma = 3$. What is the probability that the sample mean is larger than 8.98?

A). 0.01　　B). 0.025　　C). 0.05　　D). 0.10

Clicker Question 6-4

- Suppose a random sample of size $n = 36$ is drawn from the population distribution with mean $\mu = 8$ and standard deviation $\sigma = 3$. Which of the following is true?

A). $P\left(8 - 1.96\left(\frac{1}{2}\right) < \bar{X} < 8 + 1.96\left(\frac{1}{2}\right)\right) = 0.90$

B). $P\left(8 - 1.96\left(\frac{1}{2}\right) < \bar{X} < 8 + 1.96\left(\frac{1}{2}\right)\right) = 0.95$

C). $P\left(8 - 1.96\left(\frac{1}{2}\right) < \bar{X} < 8 + 1.96\left(\frac{1}{2}\right)\right) = 0.99$

Acceptance Intervals

- Let $z_{\alpha/2}$ be the z-value that leaves area $\alpha/2$ in the upper tail of the normal distribution.

- Then,

$$P(\mu - z_{\alpha/2}\sigma_{\overline{X}} < \overline{X} < \mu + z_{\alpha/2}\sigma_{\overline{X}}) = 1 - \alpha$$

Sampling Distributions of Sample Proportions

p = the proportion of the population having some characteristic

- Sample proportion ($\hat{p}$) provides an estimate of p:

$$\hat{p} = \frac{\text{number of items in the sample having the characteristic of interest}}{\text{sample size}}$$

- Let $X_1, X_2, \dots, X_n$ be independent Bernouilli random variables with $E[X_i] = p$. Then,

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

Normal Approximation for the average of Bernoulli random var.

(continued)

- The shape of the average of independent Bernoulli is approximately normal if n is large

$$\hat{p} - p = \frac{1}{n}\sum_{i=1}^{n}(X_i - p) \to N\left(0, \frac{p(1-p)}{n}\right)$$

- Standardize to Z from the average of Bernoulli random variable:

$$Z = \frac{\hat{p} - \mathrm{p}}{\sqrt{\mathrm{Var}(\hat{p})}} = \frac{\hat{p} - \mathrm{p}}{\sqrt{\mathrm{p}(1-\mathrm{p})/\mathrm{n}}}$$

5.4

Normal Approximation for the average of Bernoulli random var.

- Bernoulli random variable X_i :
- By Central Limit Theorem,
  - X_i =1 with probability p
  - X_i =0 with probability 1-p

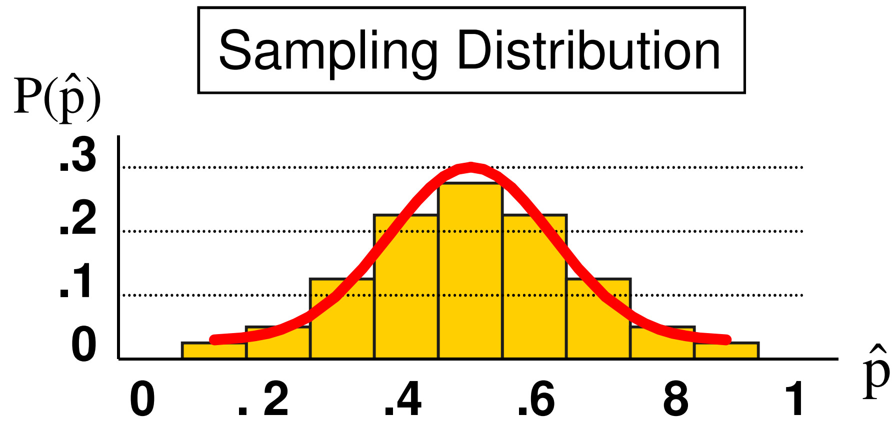
$$E(X_i) = p \qquad Var(X_i) = p(1\text{-}p)$$

- By Central Limit Theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - E(X_i)) \to N\left(0, Var(X_i)\right)$$

Sampling Distribution of $\hat{p}$

- Normal approximation:



Properties:

$$E(\hat{p}) = p$$

and

$$\sigma_{\hat{p}}^2 = \mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

(where p = population proportion)

Z-Value for Proportions

Standardize $\hat{p}$ to a Z value with the formula:

$$Z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Example

- 40% of all voters support ballot proposition A. What is the probability that between 0.40 and 0.45 fraction of voters indicate support in a sample of n = 100 ?

$$E(\hat{p}) = p = 0.40$$

$$\sqrt{\mathrm{Var}(\hat{p})} = \sqrt{\mathrm{p(1-p)/n}} = \sqrt{(0.40)(1-0.40)/100} = 0.049$$

$$\mathrm{P}(0.40 < \hat{p} < 0.45) = \mathrm{P}\left(\frac{0.40 - 0.40}{\sqrt{(0.4)(1-0.4)/100}} \le \mathrm{Z} \le \frac{0.45 - 0.40}{\sqrt{(0.4)(1-0.4)/100}}\right)$$

$$= \mathrm{P}(0 < \mathrm{Z} < 1.02)$$

$$= \Phi(1.02) - \Phi(0)$$

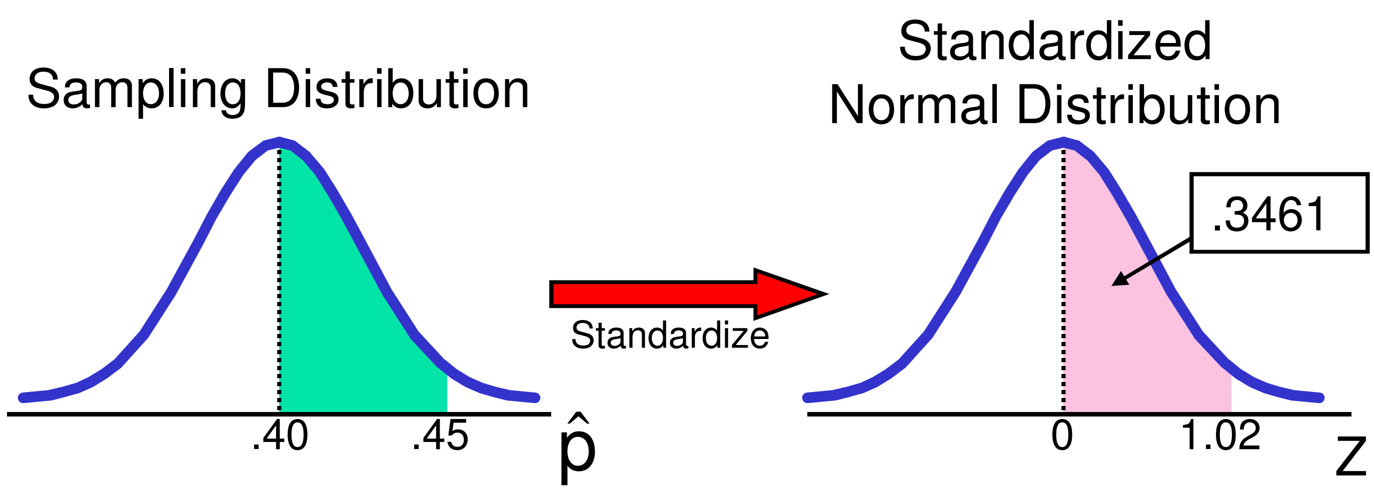
$$= 0.8461 - 0.5000 = 0.3461$$

Example

(continued)

- **if P = .4 and n = 100, what is**
  $P(.40 \le \hat{p} \le .45)$?

Use standard normal table: $P(0 \le Z \le 1.02) = .3461$



Clicker Question 6-5

- 40% of all voters support ballot proposition A. Let $\hat{p}$ be the sample fraction of voters who support proposition A in a sample of n=100. Which of the following is true?

A). $P\big(0.4 - 1.96(0.049) < \hat{p} < 0.4 + 1.96(0.049)\big) = 0.90$

B). $P\big(0.4 - 1.96(0.049) < \hat{p} < 0.4 + 1.96(0.049)\big) = 0.95$

C). $P\big(0.4 - 1.96(0.049) < \hat{p} < 0.4 + 1.96(0.049)\big) = 0.99$

Question

- 30% of all voters support ballot proposition A. Let $\hat{p}$ be the sample fraction of voters who support proposition A in a sample of n=100. What is the value of a?

$$P(0.3 - a < \hat{p} < 0.3 + a) = 0.95$$

Sample Variance

- Let $x_1, x_2, \ldots, x_n$ be a random sample from a population. The sample variance is

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

- the square root of the sample variance is called the sample standard deviation

- the sample variance is different for different random samples from the same population

Sampling Distribution of Sample Variances

- The sampling distribution of s^2 has mean σ^2

$$E(s^2) = \sigma^2$$

- **If the population distribution is normal,** then

$$\frac{(n-1)s^2}{\sigma^2}$$

has a χ^2 distribution with n – 1 degrees of freedom.

Chi-square distribution

- Consider Z_i for $i = 1, \dots, v$ independently drawn from the standard normal distribution, $N(0,1)$. Then,

$$\chi_k^2 = (Z_1)^2 + (Z_2)^2 + \cdots + (Z_v)^2$$

- When k = 1,

$$\chi_1^2 = (Z)^2$$

where $Z \sim N(0,1)$.

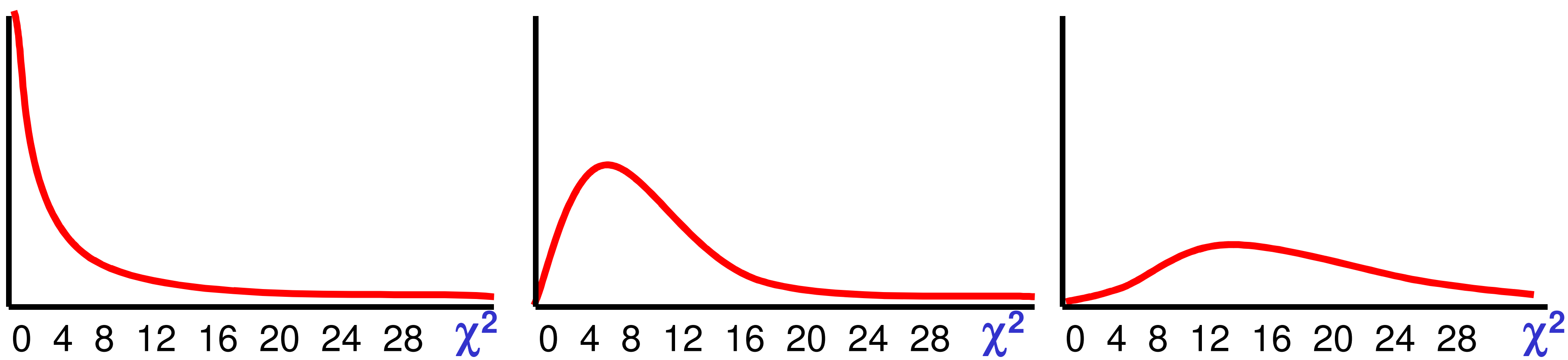
Question

- What is the value of b such that

$$P(\chi_1^2 > b) = 0.05?$$

The Chi-square Distribution

- The chi-square distribution is a family of distributions, depending on degrees of freedom:
- d.f. = n – 1



d.f. = 1 | d.f. = 5 | d.f. = 15

- Text Table 7 contains chi-square probabilities

Degrees of Freedom (df)

Idea: Number of observations that are free to vary after sample mean has been calculated

Example: Suppose the mean of 3 numbers is 8.0

Let $X_1 = 7$
Let $X_2 = 8$
What is X_3?

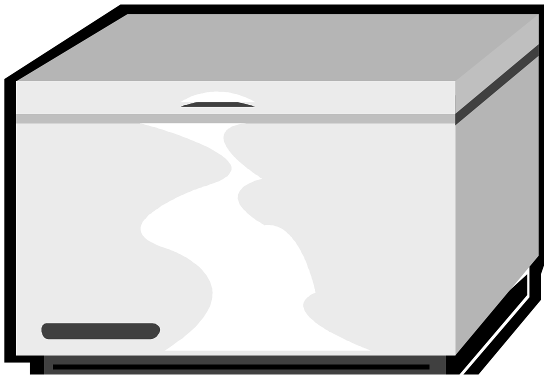
If the mean of these three values is 8.0,
then X_3 must be 9
(i.e., X_3 is not free to vary)

Here, n = 3, so degrees of freedom $= n - 1 = 3 - 1 = 2$

(2 values can be any numbers, but the third is not free to vary for a given mean)

Chi-square Example

- A commercial freezer must hold a selected temperature with little variation. Specifications call for a standard deviation of no more than 4 degrees (a variance of 16 degrees^2).
- A sample of 14 freezers is to be tested
- Suppose that the population variance is 16.
- What is the upper limit (K) for the sample variance such that the probability of exceeding this limit is less than 0.05?



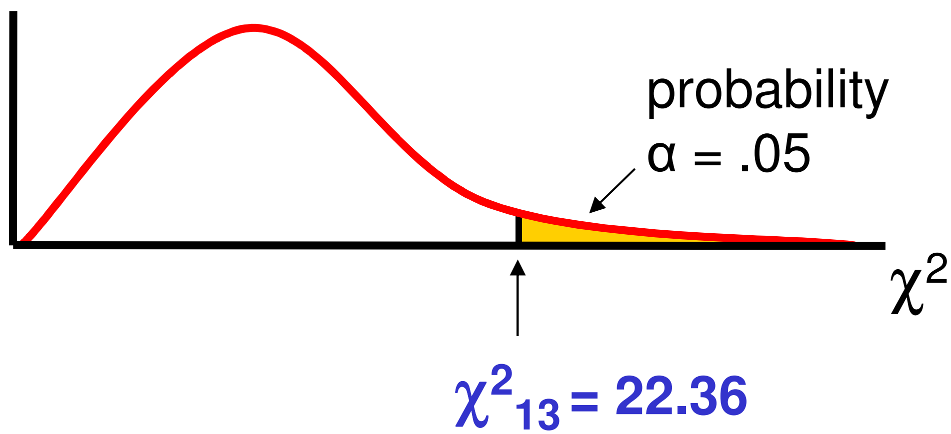
Finding the Chi-square Value

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

is chi-square distributed with (n – 1) = 13 degrees of freedom

- Use the the chi-square distribution with area 0.05 in the upper tail:

$$\chi^2_{13} = 22.36 \quad (\alpha = .05 \text{ and } 14 - 1 = 13 \text{ d.f.})$$



Chi-square Example

(continued)

$\chi^2_{13} = 22.36$ ($\alpha = .05$ and $14 - 1 = 13$ d.f.)

So:

$$P(s^2 > K) = P\left(\frac{(n-1)s^2}{16} > \chi^2_{13}\right) = 0.05$$

or $$\frac{(n-1)K}{16} = 22.36$$ (where n = 14)

so $$K = \frac{(22.36)(16)}{(14-1)} = 27.52$$

If s^2 from the sample of size n = 14 is greater than 27.52, there is strong evidence to suggest the population variance exceeds 16.

Question

- A commercial freezer must hold a selected temperature with little variation. Specifications call for a standard deviation of no more than 2 degrees.
- A sample of 10 freezers is to be tested
- Suppose that the population variance is 4.
- What is the upper limit (K) for the sample variance such that the probability of exceeding this limit is less than 0.05?