Ch. 6 - Sampling and Sampling Distributions

Ercan Karadas

New York University

ercan@nyu.edu

Summer, 2016

Introduction

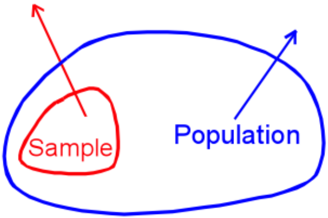
- **Descriptive statistics**
  - Collecting, presenting, and describing data
- **Inferential statistics**
  - Drawing conclusions and/or making decisions concerning a population based only on sample data

Inferential Statistics

- Making statements about a population by examining sample results

Sample statistics ➡ Population parameters

(known) **Inference** (unknown, but can be estimated from sample evidence)



Inferential Statistics

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

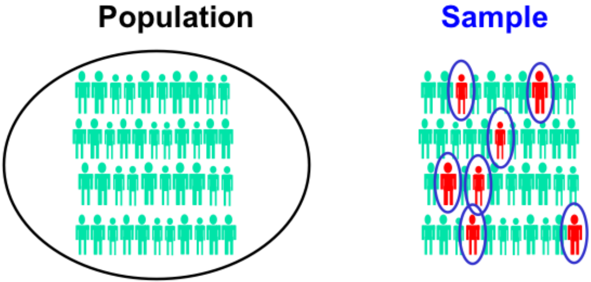
- **Estimation**
  - e.g., Estimate the population mean weight using the sample mean weight
- **Hypothesis Testing**
  - e.g., Use sample evidence to test the claim that the population mean weight is 120 pounds

Sampling from a Population

Sampling from a Population

- 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
- It is possible to obtain statistical results of a sufficiently high precision based on samples.

Simple Random Sample

- Every object in the population has the same probability of being selected
- Objects are selected independently
- Samples can be obtained from a table of random numbers or computer random number generators

- A simple random sample is the ideal against which other sampling methods are compared

Sampling Distributions

- A sampling distribution is a probability distribution of all of the possible values of a statistic for a given size sample selected from a population

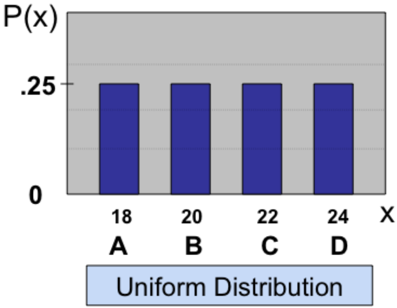
Developing a Sampling Distribution

- **Assume there is a population …**
- Population size N=4
- Random variable, X, is age of individuals
- Values of X: 18, 20, 22, 24 (years)

Developing a Sampling Distribution

(continued)

In this example the Population Distribution is uniform:



Developing a 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

Developing a Sampling Distribution

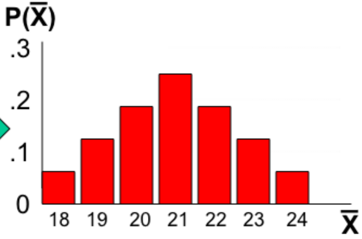
(continued)

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

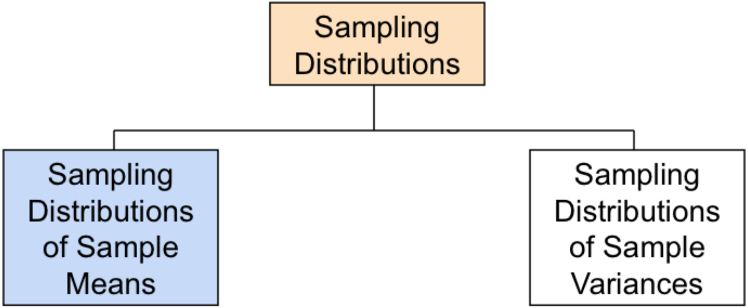
Distribution of Sample Means



(no longer uniform)

Introduction
Sampling from a Population
Sampling Distributions of Sample Means
Sampling Distribution of Sample Variances
Normal Approximation for Binomial

Sampling Distributions of Sample Means



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

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

Standard Error of the Mean

- Different samples of the same size from the same population will yield different sample means
- A measure of the variability in the mean from sample to sample is given by the Standard Error of the Mean:

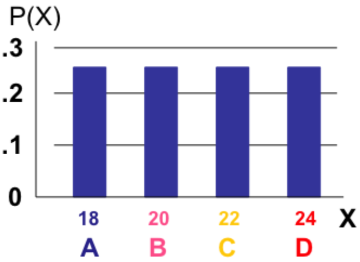
$$\boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$$

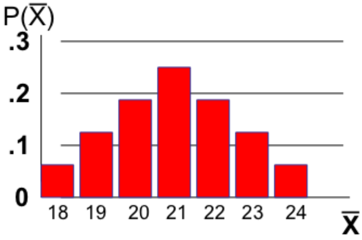
- Note that the standard error of the mean decreases as the sample size increases

Comparing the Population with its Sampling Distribution

Population
$N = 4$
$\mu = 21 \quad \sigma = 2.236$

Sample Means Distribution
$n = 2$
$\mu_{\bar{X}} = 21 \quad \sigma_{\bar{X}} = 1.58$

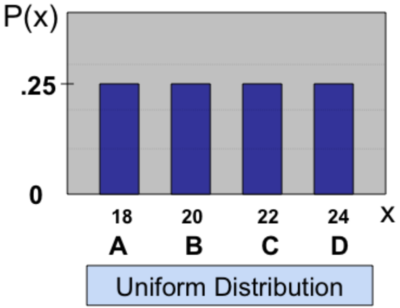




Developing a Sampling Distribution

(continued)

Summary Measures for the Population Distribution:



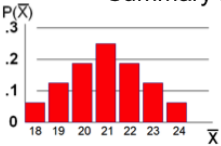
$$\mu = \frac{\sum X_i}{N} = \frac{18+20+22+24}{4} = 21$$

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

Developing a Sampling Distribution

(continued)

Summary Measures of the Sampling Distribution:



$$E(\bar{X}) = \frac{\sum \bar{X}_i}{N} = \frac{18 + 19 + 21 + \cdots + 24}{16} = 21 = \mu$$

$$\sigma_{\bar{X}} = \sqrt{\frac{\sum (\bar{X}_i - \mu)^2}{N}}$$

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

If the Population is Normal

- If a population is normal with mean μ and standard deviation σ, the sampling distribution of $\overline{X}$ is also normally distributed with

$$\boxed{\mu_{\overline{X}} = \mu} \quad \text{and} \quad \boxed{\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}}$$

Standard Normal Distribution for the Sample Means

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

$$\boxed{Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}}$$

where: $\overline{X}$ = sample mean

μ = population mean

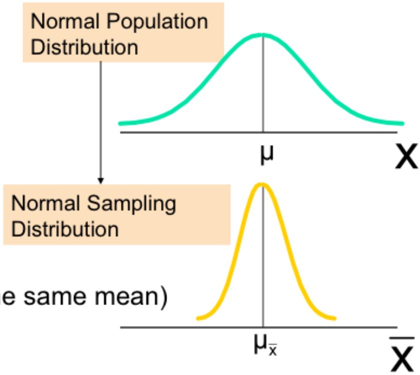
$\sigma_{\overline{X}}$ = standard error of the mean

Z is a standardized normal random variable with mean of 0 and a variance of 1

Sampling Distribution Properties

$$\boxed{E[\overline{X}] = \mu}$$

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



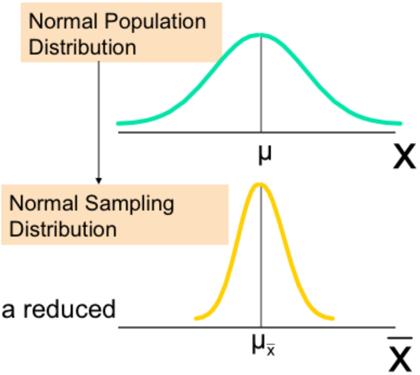
(both distributions have the same mean)

Sampling Distribution Properties

(continued)

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

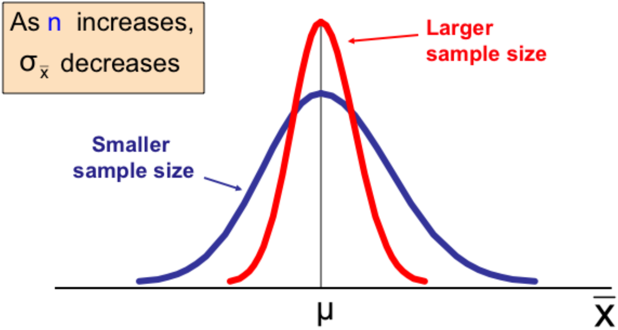
(the distribution of $\bar{X}$ has a reduced standard deviation)



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

Sampling Distribution Properties

(continued)



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

- Consider an experiment where *a random sample,* X_1, X_2, . . . , X_n , of size n drawn from a distributions with mean μ, variance σ^2, and $\overline{X}$ be the sample mean ($\overline{X}$ is a r.v.)

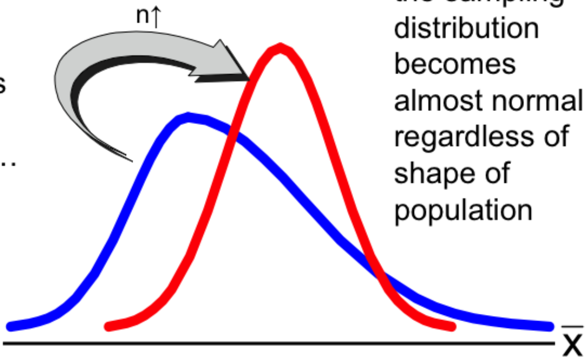
- Central Limit Theorem (CLT): as n becomes large, the distribution of $\overline{X}$ converges to N(μ, $\sigma_{\overline{x}}^{2}$), or we can say

$$\boxed{Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}}$$

approaches the standard normal distribution

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)

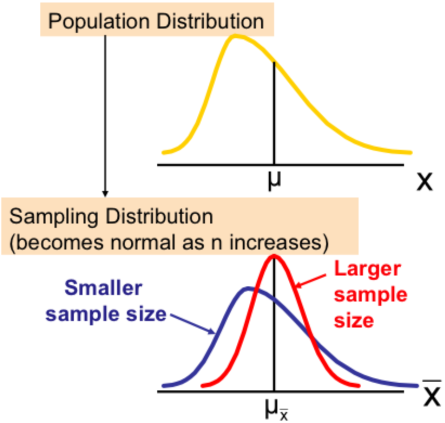
Sampling distribution properties:

Central Tendency

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

Variation

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



How Large is Large Enough?

- For most distributions, $n > 25$ will give a sampling distribution that is nearly normal
- For normal population distributions, the sampling distribution of the mean is always normally distributed

Example

- Suppose a large population has mean $\mu = 8$ and standard deviation $\sigma = 3$. Suppose a random sample of size $n = 36$ is selected.

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

Example

(continued)

Solution:

- Even if the population is not normally distributed, the central limit theorem can be used (n > 25)
- … so the sampling distribution of $\overline{X}$ is approximately normal
- … with mean $\mu_{\bar{x}} (= \mu) = 8$
- …and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$

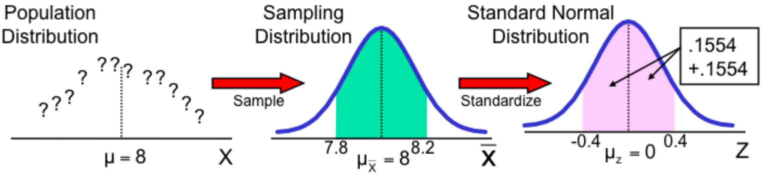
Example

(continued)

Solution (continued):

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

$$= P(-0.4 < Z < 0.4) = \boxed{0.3108}$$



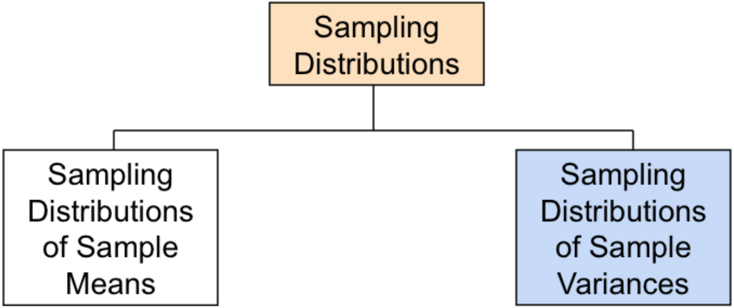
Acceptance Intervals

- Goal: determine a range within which sample means are likely to occur, given a population mean and variance
  - By the Central Limit Theorem, we know that the distribution of $\bar{X}$ is approximately normal if n is large enough, with mean μ and standard deviation $\sigma_{\bar{X}}$
  - Let $z_{\alpha/2}$ be the z-value that leaves area $\alpha/2$ in the upper tail of the normal distribution (i.e., the interval $-z_{\alpha/2}$ to $z_{\alpha/2}$ encloses probability $1 - \alpha$)
  - Then

$$\boxed{\mu \pm z_{\alpha/2}\sigma_{\bar{X}}}$$

is the interval that includes $\bar{X}$ with probability $1 - \alpha$

Sampling Distributions of Sample Variances



Sample Variance

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

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

- 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

$$\boxed{E[s^2] = \sigma^2}$$

- If the population distribution is normal, then

$$\boxed{Var(s^2) = \frac{2\sigma^4}{n-1}}$$

Chi-Square Distribution of Sample and Population Variances

- If the population distribution is normal then

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

has a chi-square (χ^2) distribution with $n - 1$ degrees of freedom

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?

→

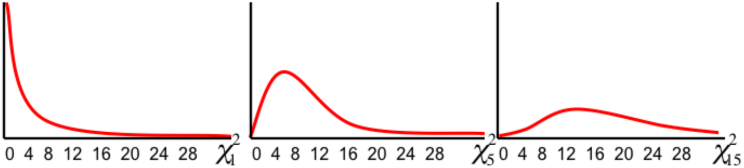
If the mean of these three values is 8.0,
then X_3 must be 9
(i.e., X_3 is not free to vary)

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)

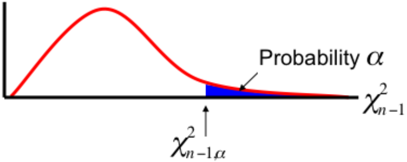
The Chi-square Distribution

- 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 Appendix Table 7 contains chi-square probabilities



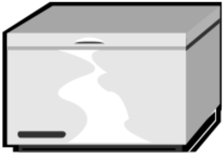
$\chi^2_{n-1,\alpha}$: the value of χ^2_{n-1} (the chi-square r.v. with (n-1) d.f.) with area α in the upper tail.

Note that this notation implies

$$P(\chi^2_{n-1} > \chi^2_{n-1,\alpha}) = \alpha$$

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
- What is the upper limit (K) for the sample variance such that the probability of exceeding this limit, given that the population standard deviation is 4, is less than 0.05?



Chi-square Example

We want to find K such that

$$P(s^2 > K) = 0.05$$

we don't know the distribution of s^2 directly! But we can do the following

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

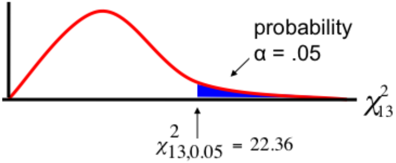
$$= P\left(\chi^2_{13} > \frac{(n-1)K}{16}\right)$$

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

Therefore, we can equivalently state the problem as: find K such that

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

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



$$\Rightarrow \frac{(n-1)K}{16} = 22.36$$

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

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.

χ^2_{13} : chi-square distributed with (n – 1) = 13 degrees of freedom

$\chi^2_{13,0.05}$: the value of χ^2_{13} (the chi-square r.v. with 13 d.f.) with area $0.05(=\alpha)$ in the upper tail, which is 22.36.

Problem

- Candidates for employment at city fire department are required to take a written aptitude test. Scores on this test are normally distributed with a mean of 280 and a standard deviation of 60. A random sample of nine test scores was taken.
- What is the standard error of the sample mean score?
- What is the probability that the sample mean score is more than 250?
- Suppose that the population standard deviation is, in fact, 40, rather than 60. Without doing the calculations, state how this would change your answer to part (a) and (b)?
- How large a sample is necessary to ensure that the probability that the sample mean differs from the population mean by more than 1.0 is less than 0.10?

Normal Distribution Approximation for Binomial Distribution

- Recall the binomial distribution:
  - n independent trials
  - probability of success on any given trial = P
- Random variable X:
  - $X_i = 1$ if the i^{th} trial is "success"
  - $X_i = 0$ if the i^{th} trial is "failure"

$$\boxed{E[X] = \mu = nP}$$

$$\boxed{Var(X) = \sigma^2 = nP(1-P)}$$

Normal Distribution Approximation for Binomial Distribution

(continued)

- The shape of the binomial distribution is approximately normal if n is large
- The normal is a good approximation to the binomial when $nP(1 - P) > 5$
- Standardize to Z from a binomial distribution:

$$\boxed{Z = \frac{X - E[X]}{\sqrt{Var(X)}} = \frac{X - np}{\sqrt{nP(1-P)}}}$$

Normal Distribution Approximation for Binomial Distribution

(continued)

- Let X be the number of successes from n independent trials, each with probability of success P.

- If $nP(1 - P) > 5$,

$$P(a < X < b) = P\left(\frac{a - nP}{\sqrt{nP(1-P)}} \le Z \le \frac{b - nP}{\sqrt{nP(1-P)}}\right)$$

Binomial Approximation Example

- 40% of all voters support ballot proposition A. What is the probability that between 76 and 80 voters indicate support in a sample of n = 200 ?
  - $E[X] = \mu = nP = 200(0.40) = 80$
  - $Var(X) = \sigma^2 = nP(1 - P) = 200(0.40)(1 - 0.40) = 48$

    (note: $nP(1 - P) = 48 > 5$)

$$
\begin{aligned}
P(76 < X < 80) &= P\left(\frac{76-80}{\sqrt{200(0.4)(1-0.4)}} \le Z \le \frac{80-80}{\sqrt{200(0.4)(1-0.4)}}\right) \\
&= P(-0.58 < Z < 0) \\
&= F(0) - F(-0.58) \\
&= 0.5000 - 0.2810 = 0.2190
\end{aligned}
$$