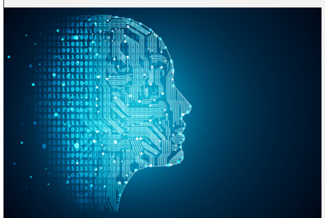


Data Science for Researchers and Scholars

Vasant G. Honavar
Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics, Public Health Sciences and Neuroscience
Director, Center for Artificial Intelligence Foundations and Scientific Applications
Associate Director, Institute for Computational and Data Sciences
Pennsylvania State University

vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

Sampling distributions

- Parameters are numerical descriptive measures for populations.
  - Two parameters for a normal distribution: mean μ and standard deviation σ.
  - One parameter for a binomial distribution: the success probability of each trial p.
- Often the values of parameters that specify the exact form of a distribution are unknown.
- You must rely on the sample to learn about these parameters.

Examples of Sampling

- A pollster is sure that the responses to his "agree/disagree" question will follow a binomial distribution, but p, the proportion of those who "agree" in the population, is unknown.
- An agronomist believes that the yield per acre of a variety of wheat is approximately normally distributed, but the mean μ and the standard deviation σ of the yields are unknown.
- If you want the sample to provide reliable information about the population, you must select your sample such that it is representative of the population!

Simple Random Sampling

- The sampling plan or experimental design determines
  - The amount of information you can extract, and
  - Often allows you to measure the reliability of your inference.
- Simple random sampling ensures that each possible sample of size n has an equal probability of being selected.

Sampling Distributions

- Any numerical descriptive measures calculated from the sample are called statistics.
- Statistics vary from sample to sample and hence are random variables. This variability is called sampling variability.
- The probability distributions of the statistics are called sampling distributions.
- In repeated sampling, they tell us what values of the statistics can occur and how often each value occurs.

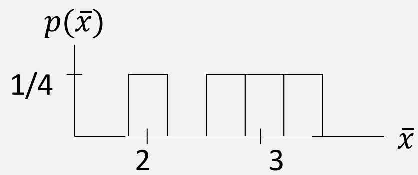
Example

<u>Population:</u> 3, 5, 2, 1

Draw samples of size *n* = 3 without replacement

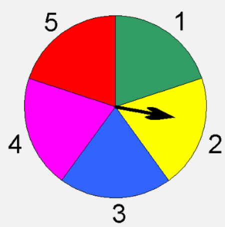
Possible samples	$\bar{x}$
3, 5, 2	$10/3 = 3.33$
3, 5, 1	$9/3 = 3$
3, 2, 1	$6/3 = 2$
5, 2, 1	$8/3 = 2.67$

Each value of $\bar{x}$ is equally likely, with probability 1/4



Example

- Consider a population that consists of the numbers 1, 2, 3, 4, 5 generated such that the probability of each of the values is 0.2 regardless of previous selections.
- This population could be described as the outcome associated with a roulette wheel shown with the distribution.



x	p(x)
1	0.2
2	0.2
3	0.2
4	0.2
5	0.2

Example

- If the sampling distribution for the means of samples of size two is analyzed, it looks like

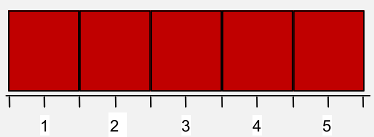
Sample	
1, 1	1
1, 2	1.5
1, 3	2
1, 4	2.5
1, 5	3
2, 1	1.5
2, 2	2
2, 3	2.5
2, 4	3
2, 5	3.5
3, 1	2
3, 2	2.5
3, 3	3

Sample	
3, 4	3.5
3, 5	4
4, 1	2.5
4, 2	3
4, 3	3.5
4, 4	4
4, 5	4.5
5, 1	3
5, 2	3.5
5, 3	4
5, 4	4.5
5, 5	5

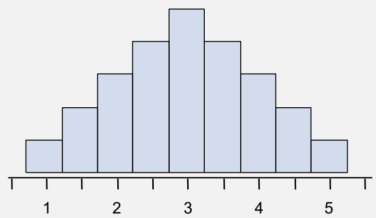
	frequency	p($\mathbf{x}$)
1	1	0.04
1.5	2	0.08
2	3	0.12
2.5	4	0.16
3	5	0.20
3.5	4	0.16
4	3	0.12
4.5	2	0.08
5	1	0.04
	25	

Example

The original distribution and the sampling distribution of means of samples with $n = 2$ are given below.



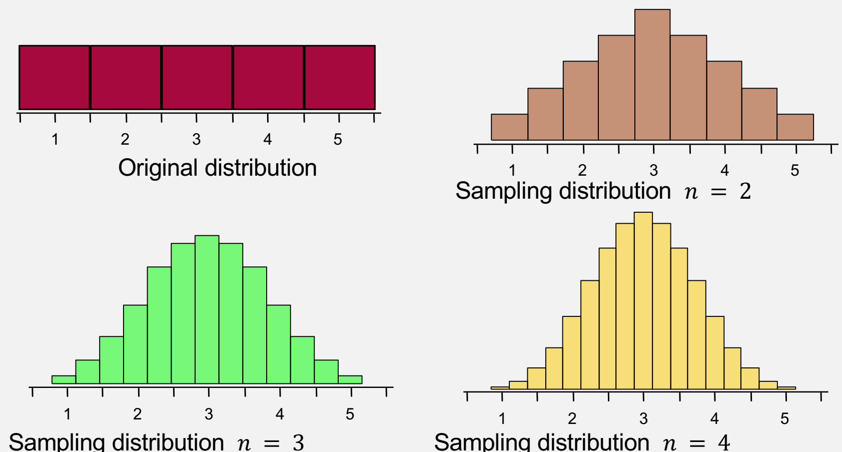
Original distribution



Sampling distribution for $n = 2$

Example

- Sampling distributions for n=3 and n=4 are shown below.
- What do you notice as n gets larger?
- The sampling distribution approaches normal distribution



Sampling Distribution of $\bar{x}$

If a random sample of *n* measurements is selected from a population with mean μ and standard deviation σ, the sampling distribution of the sample mean $\bar{x}$ will have a mean

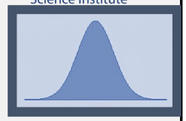
$$\mu_{\bar{x}} = \mu$$

and a standard deviation

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

Central Limit Theorem: If random samples of *n* observations are drawn from a nonnormal population with finite μ and standard deviation σ , then, when *n* is large, the sampling distribution of the sample mean $\bar{x}$ is approximately normally distributed, with mean μ and standard deviation $\sigma / \sqrt{n}$.
The approximation becomes more accurate as *n* becomes large.

Why is this Important?

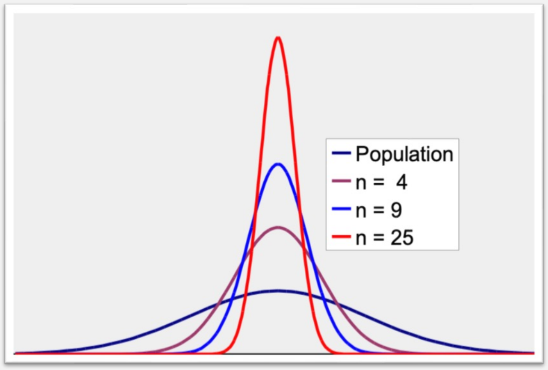


- The Central Limit Theorem also implies that the sum of n measurements is approximately normal with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$.
- Many statistics that are used for statistical inference are sums or averages of sample measurements.
- When n is large, these statistics will have approximately normal distributions.
- This will allow us to describe their behavior and evaluate the reliability of our inferences.

How Large is Large?

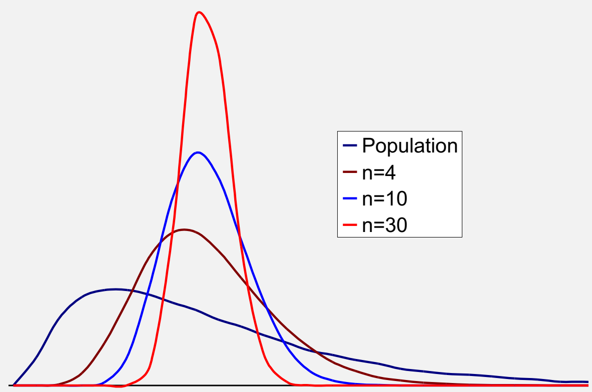
- If the sample is normal, then the sampling distribution of $\overline{x}$ will also be normal, no matter what the sample size.
- When the sample population is approximately symmetric, the distribution becomes approximately normal for relatively small values of n.
- When the sample population is skewed, the sample size must be at least 30 before the sampling distribution of $\overline{x}$ becomes approximately normal.

Sampling Distributions Illustrated



Symmetric distributions that resemble the normal distribution

Sampling distributions illustrated



Skewed population

Finding Probabilities for the Sample Mean

- If the sampling distribution of $\bar{x}$ is normal or approximately normal, *standardize or rescale* the interval of interest in terms of

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

- Find the appropriate area using the z distribution.

Example: A random sample of size *n* = 16 from a normal distribution with μ = 10 and σ = 8.

$$P(\bar{x} > 12) = P(\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < \frac{12 - 10}{8 / \sqrt{16}})$$

$$= P(z > 1) = .5 - .3413 = .1587$$

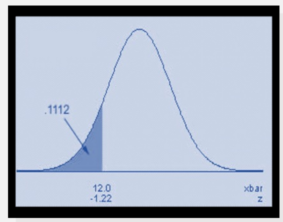
Example

- A soda filling machine is supposed to fill cans of soda with 12 fluid ounces.
- Suppose that the fills are actually normally distributed with a mean of 12.1 oz and a standard deviation of .2 oz.
- The probability of one can less than 12 is

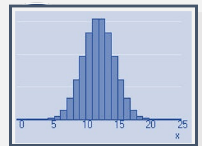
$$P(x < 12) = P(\frac{x-\mu}{\sigma} < \frac{12-12.1}{.2}) = P(z < -.5) = .5 - .1915 = .3085$$

What is the probability that the average fill for a 6-pack is less than 12 oz?

$$P(\bar{x} < 12) =$$
$$P(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < \frac{12-12.1}{.2/\sqrt{6}}) =$$
$$P(z < -1.22) = .1112$$

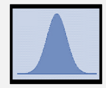


The Sampling Distribution of the Sample Proportion



- The Central Limit Theorem can be used to conclude that the binomial random variable *x* is approximately normal when *n* is large, with mean *np* and variance npq.
- The sample proportion, $\hat{p} = \frac{x}{n}$ is simply a *rescaling* of the binomial random variable *x*, dividing it by n.
- From the Central Limit Theorem, the sampling distribution of $\hat{p}$ will also be approximately normal, with a *rescaled* mean and standard deviation.

The Sampling Distribution of the Sample Proportion



✓A random sample of size *n* is selected from a binomial population with parameter *p*.

✓The sampling distribution of the sample proportion, $\hat{p} = \dfrac{x}{n}$

will have mean *p* and standard deviation $\sqrt{\dfrac{pq}{n}}$

✓If *n* is large, and *p* is not too close to zero or one, the sampling distribution of $\hat{p}$ will be approximately normal.

- The standard deviation of *p*-hat is sometimes called the STANDARD ERROR (SE) of *p*-hat.

Finding Probabilities for the Sample Proportion

✓If the sampling distribution of $\hat{p}$ is normal or approximately normal, *standardize or rescale* the interval of interest in terms of

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

✓Find the appropriate area using the normal table.

Example

- The soda bottler in the previous example claims that only 5% of the soda cans are underfilled.
- A quality control technician randomly samples 200 cans of soda.
- What is the probability that more than 10% of the cans are underfilled?

$n = 200$

S: underfilled can

p = P(S) = .05

$q = .95$

$np = 10$ $nq = 190$

OK to use the normal approximation

$$P(\hat{p} > .10)$$

$$= P(z > \frac{.10 - .05}{\sqrt{\frac{.05(.95)}{200}}}) = P(z > 3.24)$$

$$< .5 - .4990 = .001$$

This would be very unusual, if indeed $p = .05$!

Example

- Suppose 3% of the people contacted by phone are receptive to a certain sales pitch and buy your product. If your sales staff contacts 2000 people, what is the probability that more than 100 of the people contacted will purchase your product?

$$n = 2000, \qquad p = 0.03, np = 60, nq = 1940,$$

OK to use the normal approximation

$$P(\hat{p} > 100/2000) = P(z > \frac{.05 - .03}{\sqrt{\frac{.03(.97)}{2000}}}) = P(z > 5.24) \approx 0$$

Sampling - Summary

- Simplest sampling technique – random sampling
  - Each possible sample is equally likely
- Sampling distributions describe the possible values of a statistic and how often they occur in repeated sampling.
- If the underlying data are normally distributed, sampling distribution is a normal distribution
- The **Central Limit Theorem** states that sums and averages of measurements from an arbitrarily distributed population with finite mean μ and standard deviation σ have approximately normal distributions for large samples of size n.

Sampling - Sampling Distribution of the Sample Mean

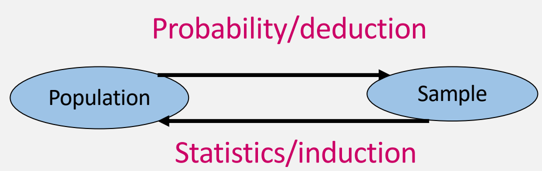
- When samples of size n are drawn from a normal population with mean μ and variance σ^2, the sample mean $\bar{x}$ has a normal distribution with mean μ and variance σ^2/n.
- When samples of size n are drawn from a nonnormal population with mean μ and variance σ^2, the Central Limit Theorem ensures that the sample mean $\bar{x}$ will have an approximately normal distribution with mean μ and variance σ^2/n when n is large (n ³ 30).
- Probabilities involving the sample mean μ can be calculated by standardizing the value of $\bar{x}$ using $z = \frac{x-\mu}{\sigma/\sqrt{n}}$

Summary – Sampling Distribution of the Sample Proportion

- When samples of size n are drawn from a binomial population with parameter p, the sample proportion $\hat{p}$ will have an approximately normal distribution with mean p and variance $pq\,/n$ as long as $np > 5$ and $nq > 5$.
- Probabilities involving the sample proportion $\hat{p}$ can be calculated by standardizing the value $\hat{p}$ using $z = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}$

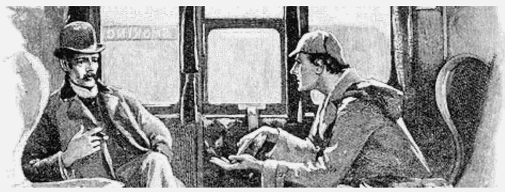
Probabilistic vs Statistical Reasoning

- In the last lecture, we looked at how the properties of a population govern what we see in sample(s) of the population
- Now we turn to going in the other direction: Given a sample, we try to understand the data generating process that could have generated the observed data
- This shift our mode of thinking from deductive reasoning to induction



Probabilistic vs Statistical Reasoning

- In many ways, science, or scholarly inquiry, is like detective work.
- We begin with a set of observations, we ask what can be said about the data generating process



- “Data! Data! Data!.. I can't make bricks without clay”

  — Sherlock Holmes, 1892

  — “ The Adventure of the Copper Beeches”

Shadows: Shadow Puppetry :: Data : Data Generating Process

Image source: Annie Katsura Rollins, Ballard Institute and Museum of Puppetry, photo by Kenneth Best

Parameters

- Populations are described by their probability distributions
  - If we assume a parametric form for the distribution, e.g., Normal, binomial, etc., then populations are described by the parameters of the respective distributions
    - Binomial populations are determined by a single parameter, p.
    - Normal distributions are described by the mean μ and the standard deviation σ.
- If the values of parameters are unknown, we have to make inferences about them using information provided by a sample from the underlying distribution
- Sample or data : distribution :: shadows : shadow puppetry
- The puppeteer whose machinations generate the shadows you see is hidden from you. Your goal is to learn his or her modus operandi.

Two types of statistical inference

- Estimation
  - Estimating or inferring the value of the parameter(s)
    - Maximum likelihood: What is the mean height of individuals of Asian descent given the sample of individuals of Asian descent you have observed?
    - Bayesian: What is the likely height of the next person of Asian descent you may encounter, given your prior belief about the heights of individuals of Asian descent, the heights of individuals of Asian descent that you have observed?
- Hypothesis testing
  - Deciding if the data support a preconceived idea or theory one has about a population
    - "Did the sample of individuals you have come from a population with mean height of 5.6" ?
    - "Was the newly discovered manuscript of unknown authorship written by Shakespeare?

Specify the type of statistical inference

- A consumer wants to estimate the average price of similar homes in her city before putting her home on the market.
- Estimation: Estimate the average price of similar homes in the city

- A manufacturer wants to know if a new type of steel is more resistant to high temperatures than an old type was.
- Hypothesis testing: Is the average efficacy of the new Covid vaccine μ_{New} greater than that of the old Covid vaccine μ_{Old}?

Methods of Statistical Inference

- Whether you are estimating parameters or testing hypotheses, statistical methods
  - Offer a sound basis for inference
  - A measure of the goodness or reliability of the inference

What is an estimator?

An estimator is a formula, that tells you how to calculate the estimate of a parameter of interest from the given sample.

- Point estimation yields a single value for the parameter
  - Example: The estimated probability of a coin coming up heads is 0.4
  - Underlying assumption:
    - The coin has a fixed parameter p
    - Our job is to estimate it.
    - How realistic is this assumption?
- Confidence interval is an interval such that for a chosen degree of confidence, expressed as a probability, the true value of the parameter is likely to fall inside the interval.
  - Example: 95% confidence interval for p is [0.3, 0.5]

Point Estimator of Population Mean

- Given a sample $S = \{x_1 \cdots x_n\}$, the point estimate of the population mean, μ, is the sample mean

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

  - Example: Suppose $S = \{4,5,6,3,4,6,3,5,8,1\}$ were the ratings given by viewers of the movie "Back to the future".
  - Suppose you believe that the viewers are a random sample of the viewers of "Back to the future".
  - What is the sample estimate of the mean rating of "Back to the future?
  - 4.5
- But why?

Point Estimation of Population Proportion

- A point estimate of p, population success rate of a binary experiment (e.g., coin tosses with outcomes H and T) is sample proportion of successes observed in the sample:

$$\hat{p} = \frac{n_H}{n_H + n_T}$$

- Example: Out of 100 people tested for Covid, 10 were positive.
  - What is the point estimate of p, the Covid positive rate in the population?

$$\hat{p} = \frac{10}{100} = 0.1$$

- But why?

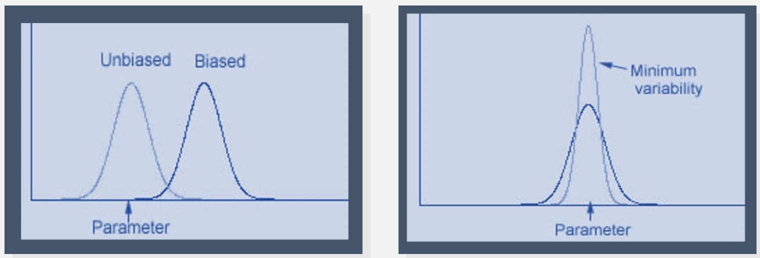
Properties of Point Estimators

- Since an estimator is calculated from sample values, it varies from sample to sample according to its sampling distribution.
- An estimator is unbiased if
  - The mean of its sampling distribution equals the parameter of interest.
  - It does not systematically overestimate or underestimate the target parameter.
- Both sample mean and sample proportion are unbiased estimators of population mean and proportion.
- Given n samples, the following sample variance is an unbiased estimator of population variance σ^2

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

Properties of Point Estimators

- Of all the unbiased estimators, we prefer the estimator whose sampling distribution has the smallest spread or variability.

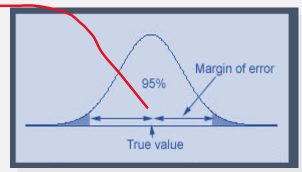


Confidence Intervals

- Confidence intervals depend on sampling distributions
- The shape of sampling distributions depend on sample sizes
- For large sample sizes, central limit theorem applies which allow us to use normal distributions
- For small sample sizes, we need to choose the right sampling distribution

Quantifying the error of Point Estimates

- Assumption: The sample sizes are large
- From the Central Limit Theorem, the sampling distributions of $\hat{\mu}$ and $\hat{p}$ will be approximately normal
- For unbiased estimators with normal sampling distributions, 95% of all point estimates will lie within 1.96 standard deviations of the parameter of interest.
- Margin of error: an upper bound on the difference between a particular estimate and the parameter that it estimates.
- Margin of error = 1.96 $\times$ standard deviation of the estimate



Estimating Means and Proportions

Point estimator of population mean $\mu : \bar{x}$

Margin of error $(n \geq 30) : \pm 1.96 \frac{s}{\sqrt{n}}$

For a binomial population,

Point estimator of population proportion $\mathrm{p} : \hat{\mathrm{p}} = \mathrm{x/n}$

Margin of error $: \pm 1.96 \sqrt{\frac{\hat{\mathrm{p}}\hat{\mathrm{q}}}{\mathrm{n}}}$

Assumption : $\mathrm{np} > 5$ and $\mathrm{nq} > 5$; or $0 < \mathrm{p} \pm 2\sqrt{\frac{\mathrm{pq}}{\mathrm{n}}} < 1$

Example

- A homeowner randomly samples 64 homes similar to her own and finds that the average selling price is $250,000 with a standard deviation of $15,000.
- Estimate the average selling price for all similar homes in the city.
- What is the margin of error?

$$\text{Point estimator of } \mu : \bar{x} = 250{,}000$$

$$\text{Margin of error} : \pm 1.96\frac{\hat{\sigma}\; s}{\sqrt{n}} = \pm 1.96\frac{15{,}000}{\sqrt{64}} = \pm 3675$$

Example

- A quality control technician wants to estimate the proportion of soda cans that are underfilled.
- He randomly samples 200 cans of soda and finds 10 underfilled cans.

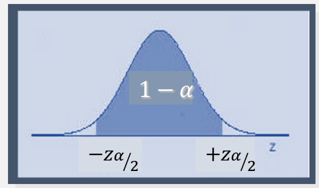
$n = 200 \qquad p = \text{proportion of underfilled cans}$

Point estimator of $p : \hat{p} = x/n = 10/200 = .05$

Margin of error : $\pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} = \pm 1.96\sqrt{\frac{(.05)(.95)}{200}} = \pm .03$

Confidence Interval

- Create an interval so that you are fairly sure that the parameter lies between these two values.
- "Fairly sure" means "with high probability", measured using the confidence coefficient, $1-\alpha$.
- Usually, $1-\alpha = 0.9, 0.95, 0.99$ …



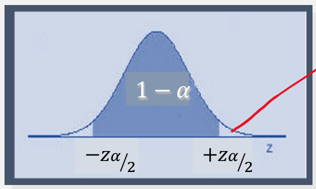
- For large-Sample size,

100($1-\alpha$)% confidence Interval:

Point Estimate $\pm z\alpha/_2$

Confidence Level

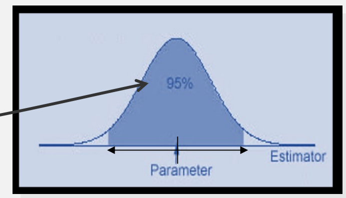
- To change to a general confidence level, $1 - \alpha$, pick a value of z that puts area $1 - \alpha$ in the center of the z distribution.



Tail area $\alpha/2$	$z_{\alpha/2}$
.05	1.645
.025	1.96
.005	2.58

- Suppose 1-α = .95

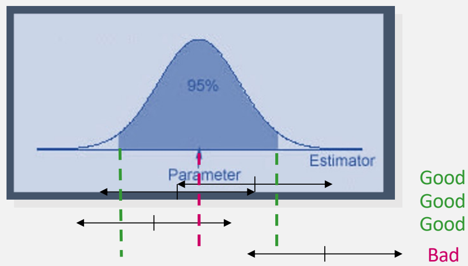
There is 95% probability that the interval constructed in this manner will contain the population mean



Confidence Interval

- Since we don't know the value of the parameter, consider which has a variable center.

Point Estimator $\pm$ 1.96 std error



- Only if the estimator falls in the tail areas will the interval fail to enclose the parameter. This happens only 5% of the time.

Interpretation of a Confidence Interval

- A confidence interval is calculated from one given sample.
- The interval either covers or misses the true parameter.
- Since the true parameter is unknown, you'll never know with certainty
- If independent samples are taken repeatedly from the same population, and a confidence interval calculated for each sample, then a certain percentage (confidence level) of the intervals will include the unknown population parameter.
- The confidence level associated with a confidence interval is the success rate of the confidence interval.

Confidence Intervals for Means and Proportions

Confidence interval for a population mean μ :

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

For a binomial population:

Confidence interval for a population proportion p :

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Example

A random sample of $n = 50$ males showed a mean average daily intake of dairy products equal to 756 grams with a standard deviation of 35 grams. Find a 95% confidence interval for the population average μ.

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}} \Rightarrow 756 \pm 1.96 \frac{35}{\sqrt{50}} \Rightarrow 756 \pm 9.70$$

$$\text{or } 746.30 < \mu < 765.70 \text{ grams.}$$

Example

Find a 99% confidence interval for μ, the population average daily intake of dairy products for men.

$$\bar{x} \pm 2.58 \frac{s}{\sqrt{n}} \Rightarrow 756 \pm 2.58 \frac{35}{\sqrt{50}} \Rightarrow 756 \pm 12.77$$

or $743.23 < \mu < 768.77$ grams.

Example

- Of a random sample of n = 150 college students, 104 of the students said that they had played on a soccer team during their K-12 years.
- Estimate the proportion of college students who played soccer in their youth with a 90% confidence interval.

$$\hat{p} \pm 1.645\sqrt{\frac{\hat{p}\hat{q}}{n}} \Rightarrow \frac{104}{150} \pm 1.645\sqrt{\frac{.69(.31)}{150}}$$

$$\Rightarrow \quad .69 \pm .06 \qquad \text{or } \; .63 < p < .75.$$

Estimating the Difference between Two Means

- Sometimes we are interested in comparing the means of two populations.
  - The average growth of plants fed using two different nutrients.
  - The average scores for students taught with two different teaching methods.
- To make this comparison

A random sample of size n_1 drawn from population 1 with mean μ_1 and variance σ_1^2.

A random sample of size n_2 drawn from population 2 with mean μ_2 and variance σ_2^2.

Comparing Two Means

	Mean	Variance	Standard Deviation
Population 1	μ_1	σ_1^2	σ_1
Population 2	μ_2	σ_2^2	σ_2

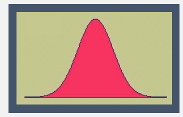
	Sample size	Mean	Variance	Standard Deviation
Sample from Population 1	n_1	$\bar{x}_1$	s_1^2	s_1
Sample from Population 2	n_2	$\bar{x}_2$	s_2^2	s_2

Estimating the Difference between Two Means

- We compare the two averages by making inferences about μ_1-μ_2, the difference in the two population averages.
  - If the two population averages are the same, then μ_1-μ_2 = 0.
  - The best estimate of μ_1-μ_2 is the difference in the two sample means

$$\bar{x}_1 - \bar{x}_2$$

The Sampling Distribution of $\bar{x}_1 - \bar{x}_2$



1. The mean of $\bar{x}_1 - \bar{x}_2$ is $\mu_1 - \mu_2$, the difference in the population means.

2. The standard deviation of $\bar{x}_1 - \bar{x}_2$ is $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

3. If the sample sizes (both n_1 and n_2) are large, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is approximately normal, and standard deviation can be estimated as $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

Estimating μ_1-μ_2

- For large samples, point estimates and their margin of error as well as confidence intervals are based on the standard normal (z) distribution.

Point estimate for $\mu_1 - \mu_2 : \bar{x}_1 - \bar{x}_2$

Margin of Error : $\pm 1.96\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$

Assumption :

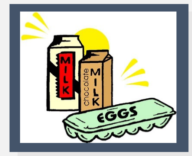
Both $n_1 \geq 30$ and $n_2 \geq 30$

Confidence interval for $\mu_1 - \mu_2$:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

Example

Avg Daily Intakes	Men	Women
Sample size	50	50
Sample mean	756	762
Sample Std Dev	35	30



- Compare the average daily intake of dairy products of men and women using a 95% confidence interval.

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\Rightarrow (756 - 762) \pm 1.96\sqrt{\frac{35}{50} + \frac{30}{50}} \quad \Rightarrow \quad -6 \pm 12.78$$

$$\text{or } -18.78 < \mu_1 - \mu_2 < 6.78.$$

Example, continued

$$-18.78 < \mu_1 - \mu_2 < 6.78$$

- Could you conclude, based on this confidence interval, that there is a difference in the average daily intake of dairy products for men and women?
- The confidence interval contains the value $\mu_1 - \mu_2 = 0$.
- Therefore, it is possible that $\mu_1 = \mu_2$.
- You would not want to conclude that there is a difference in average daily intake of dairy products for men and women.

Estimating the Difference between Two Proportions

- Sometimes we are interested in comparing the proportion of "successes" in two binomial populations.
  - The germination rates of untreated seeds and seeds treated with a fungicide.
  - The proportion of male and female voters who favor a particular candidate for governor.
- To make this comparison

A random sample of size n_1 drawn from binomial population 1 with parameter p_1.

A random sample of size n_2 drawn from binomial population 2 with parameter p_2.

Comparing Two Proportions

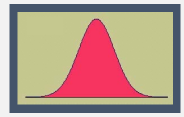
	Sample size	Sample Proportion	Sample Variance	Standard Deviation
Sample from Population 1	n_1	$\hat{p}_1 = \frac{x_1}{n_1}$	$\frac{\hat{p}_1\hat{q}_1}{n}$	$\sqrt{\frac{\hat{p}_1\hat{q}_1}{n}}$
Sample from Population 2	n_2	$\hat{p}_2 = \frac{x_2}{n_2}$	$\frac{\hat{p}_2\hat{q}_2}{n}$	$\sqrt{\frac{\hat{p}_2\hat{q}_2}{n}}$

Estimating the Difference between Two Means

- We compare the two proportions by making inferences about p_1-p_2, the difference in the two population proportions.
  - If the two population proportions are the same, then p_1-p_2 = 0.
  - The best estimate of p_1-p_2 is the difference in the two sample proportions,

$$\hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2}$$

The Sampling Distribution of $\hat{p}_1 - \hat{p}_2$



1. The mean of $\hat{p}_1 - \hat{p}_2$ is $p_1 - p_2$, the difference in the population proportions.

2. The standard deviation of $\hat{p}_1 - \hat{p}_2$ is $\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}$.

3. If the sample sizes (both n_1 and n_2) are large, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal, and stanard deviation can be estimated as

$$\text{SE} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}.$$

Estimating p_1-p_2

For large samples, point estimates and their margin of error as well as confidence intervals are based on the standard normal (*z*) distribution.

$$\text{Point estimate for } p_1\text{-}p_2 : \hat{p}_1 - \hat{p}_2$$

$$\text{Margin of Error} : \pm 1.96\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$\text{Confidence interval for } p_1 - p_2 :$$

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Assumption: both n_1 and n_2 are sufficiently large so that $-1 \le \hat{p}_1 - \hat{p}_2 \pm 2SE \le 1$

Example

Youth Soccer	Male	Female
Sample size	80	70
Played soccer	65	39

- Compare the proportion of male and female college students who said that they had played on a soccer team during their K-12 years using a 99% confidence interval.

$$(\hat{p}_1 - \hat{p}_2) \pm 2.58\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$\Rightarrow (\frac{65}{80} - \frac{39}{70}) \pm 2.58\sqrt{\frac{.81(.19)}{80} + \frac{.56(.44)}{70}} \quad \Rightarrow \quad .25 \pm .19$$

$$\text{or } .06 < p_1 - p_2 < .44.$$

Example, continued

$$.06 < p_1 - p_2 < .44$$

- Could you conclude, based on this confidence interval, that there is a difference in the proportion of male and female college students who said that they had played on a soccer team during their K-12 years?
- The confidence interval does not contains the value p_1-p_2 = 0. Therefore, it is not likely that p_1= p_2. You would conclude that there is a difference in the proportions for males and females.

A higher proportion of males than females played soccer in their youth.

Summary – Large Sample Point Estimators

To estimate one of four population parameters when the sample sizes are large, use the following point estimators with the appropriate margins of error.

Parameter	Point Estimator	Margin of Error
μ	$\bar{x}$	$\pm 1.96\left(\frac{s}{\sqrt{n}}\right)$
p	$\hat{p} = \frac{x}{n}$	$\pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$p_1 - p_2$	$(\hat{p}_1 - \hat{p}_2) = \left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right)$	$\pm 1.96\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$

Summary – Large Sample Confidence Intervals

To estimate one of four population parameters when the sample sizes are large, use the following interval estimators.

Parameter	$(1-\alpha)100\%$ Confidence Interval
μ	$\bar{x} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$
p	$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$
$\mu_1 - \mu_2$	$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$p_1 - p_2$	$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$

Summary: Large Sample Confidence intervals

- All values in the interval are possible values for the unknown population parameter.
- Any values outside the interval are unlikely to be the value of the unknown parameter.
  - To compare two population means or proportions, look for the value 0 in the confidence interval.
  - If 0 is in the interval, it is possible that the two population means or proportions are equal, and you should not declare a difference.
  - If 0 is not in the interval, it is unlikely that the two means or proportions are equal, and you can confidently declare a difference.