Lesson 7

Inferential Statistics

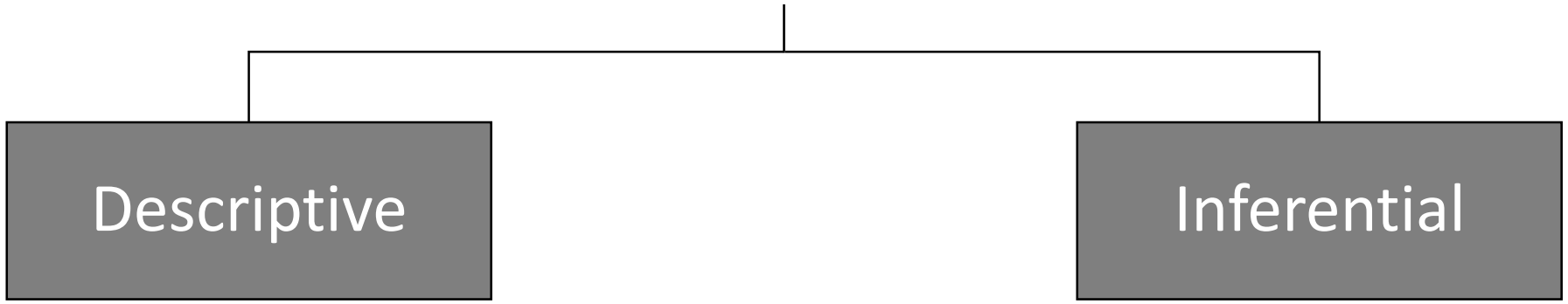
Kush Kulshrestha

Statistics

/stəˈtɪstɪks/

noun

the practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.



Describing, presenting, summarizing or organizing your data via numerical calculations or graphs or tables

Complex mathematics, allow us to infer trends, make assumptions & predictions about population based on samples

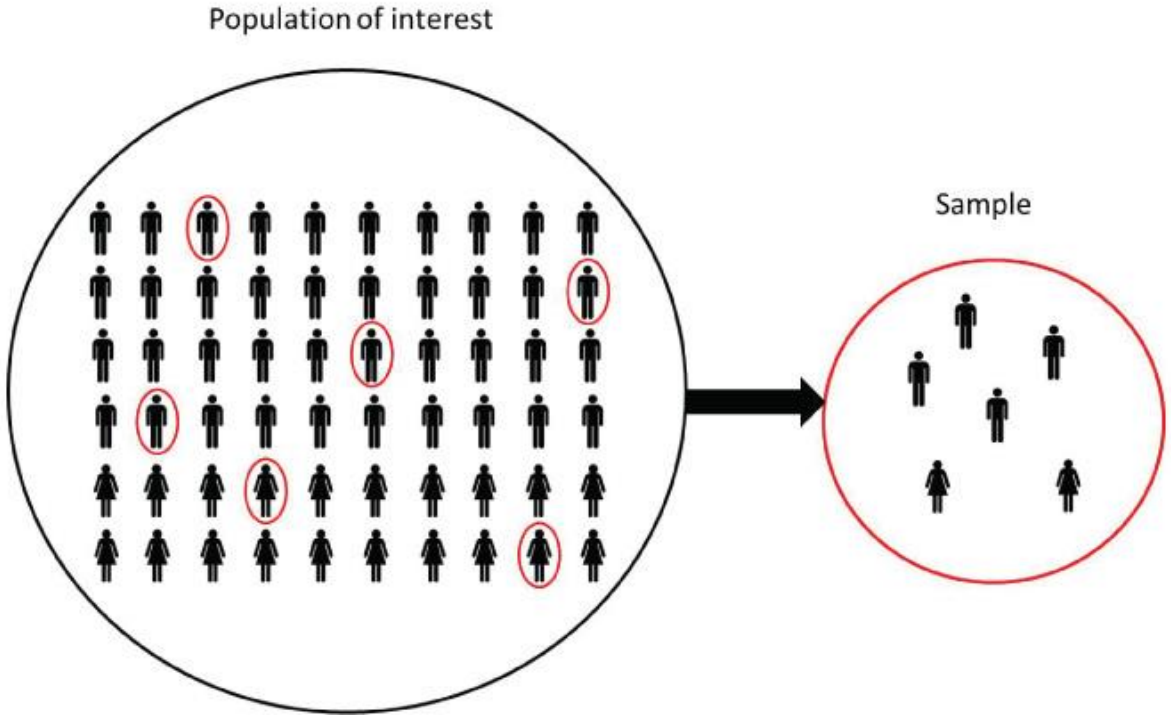
Population and Sample

Inferential Statistics is a part of statistics which uses various tools and techniques to derive some inference by looking at just a sample of the all of the values.

A **population** includes all of the elements from a set of data.

A **sample** consists one or more observations drawn from the population.

A sample can have fewer observations than the population, the same number of observations, or more observations.



Population and Sample

Key differences between Population and Sample:

A measurable characteristic of a population is called a parameter.	A measurable characteristic of a sample is called a statistic
Mean of the population is denoted by μ	Mean of a sample is represented by the symbol $\bar{x}$
Standard deviation of population is calculated in regular fashion and represented by σ	Two types of sample standard deviation: Biased and Unbiased represented by s

A pollster wants to know the percentage of voters that favour a flat-rate income tax.

The *actual* percentage of all the voters is a population parameter.
The *estimate* of that percentage, based on sample data, is a sample statistic.

Population and Sample

Key differences between Population and Sample:

Mean of Population:

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

N = number of items in the population

Mean of Sample:

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

n = number of items in the sample

Standard Deviation of Population:

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

Standard Deviation of Sample:

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

or

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$ (biased)

Population and Sample

Difference between biased and unbiased estimate of Sample Variance

Using (n-1) instead of (n) in Sample Standard Deviation and Sample Variance is called **Bessel's Correction**. This corrects the bias in the estimation of the population variance from sample variance.

Explanation:

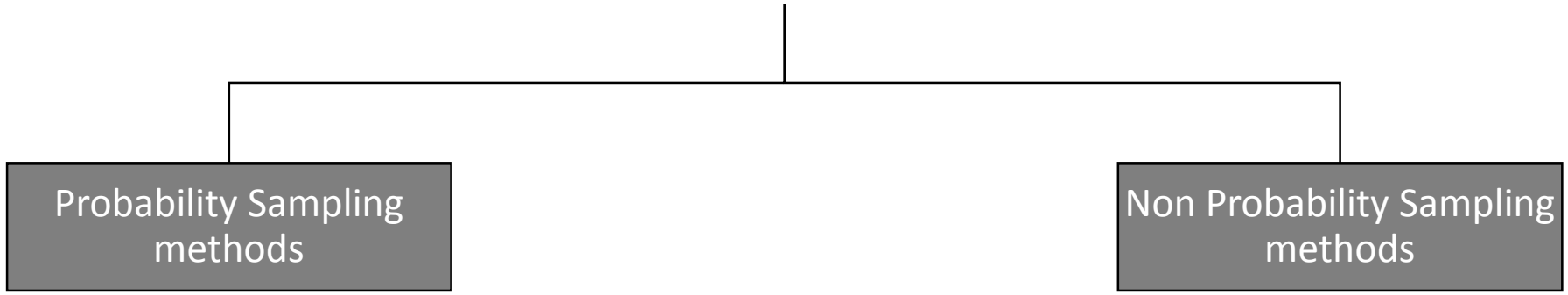
Population and Sample

Sampling

- Sampling is a procedure for selecting sample elements from the population.
- We often don’t have resources (time and money) needed to measure population parameters, so we try to estimate them.
- We try to measure the sample statistics and then try to estimate the population parameters from them.

The quality of a sample statistic (i.e., accuracy, precision, representativeness) is strongly affected by the way that sample observations are chosen; that is., by the sampling method.

Sampling methods fall into one of two different categories:



Population and Sample

Sampling

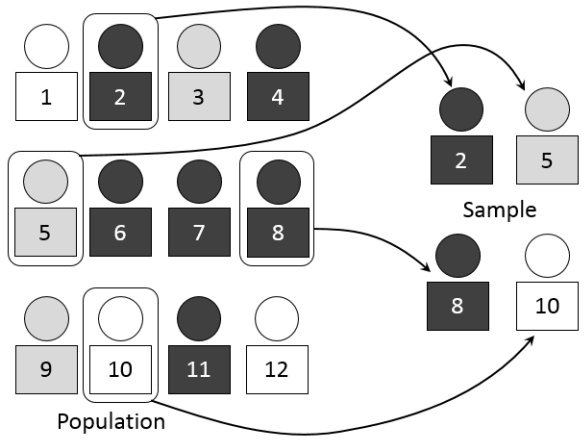
Probability samples:
With probability sampling methods, each population element has a known (non-zero) chance of being chosen for the sample.

Non probability samples:
With non-probability sampling methods, we do not know the probability that each population element will be chosen, and/or we cannot be sure that each population element has a non-zero chance of being chosen.

Non-probability sampling methods offer two potential advantages - convenience and cost.
The main disadvantage is that non-probability sampling methods do not allow you to estimate the extent to which sample statistics are likely to differ from population parameters. Only probability sampling methods permit that kind of analysis.

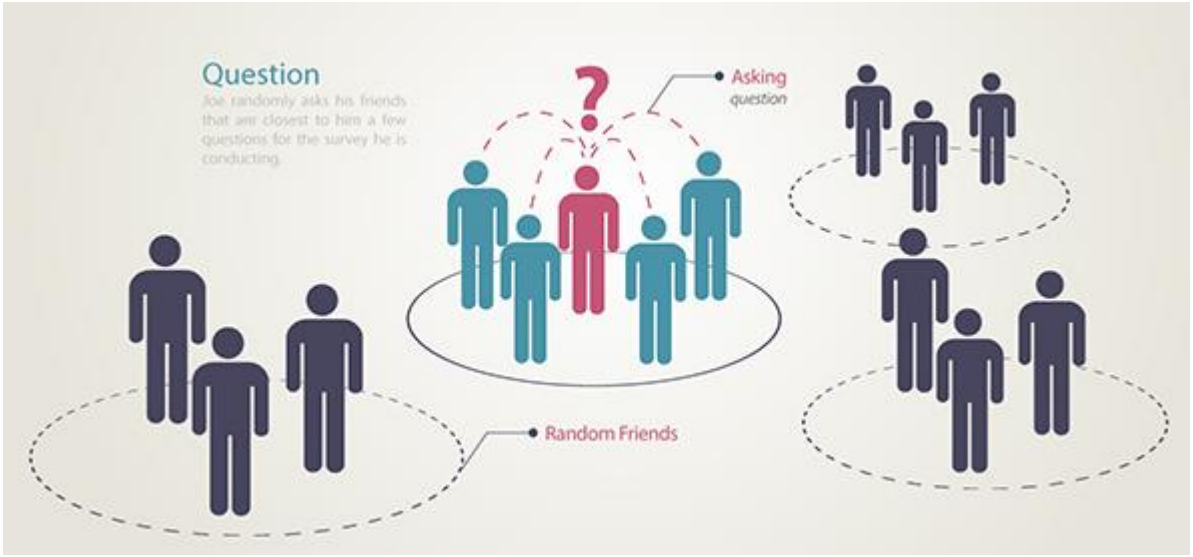
Population and Sample

Non Probability Sampling Methods:



Voluntary samples:
A voluntary sample is made up of people who self-select into the survey. Often, these folks have a strong interest in the main topic of the survey.

Convenience samples:
A convenience sample is made up of people who are easy to reach.



Population and Sample

Probability Sampling Methods:

The key benefit of probability sampling methods is that **they guarantee that the sample chosen is representative of the population**. This ensures that the statistical conclusions will be valid.

Simple random sampling: It has following properties:

a) Population consists of N objects
b) Sample consists of n objects
c) All possible samples of n objects are equally likely to occur.

Benefit of this method: It allows us to use statistical methods to analyze sample results and estimate population parameters. Non random sampling methods are not good for using statistical methods.

Sampling with and without replacement:
Lets say we have lottery method to select a simple random sample (picking a number from a bowl). After we pick a number from the bowl, we can put the number aside or we can put it back into the bowl. If we put the number back in the bowl, it may be selected more than once; if we put it aside, it can selected only one time.

When a population element can be selected more than one time, we are sampling with replacement. When a population element can be selected only one time, we are sampling without replacement.

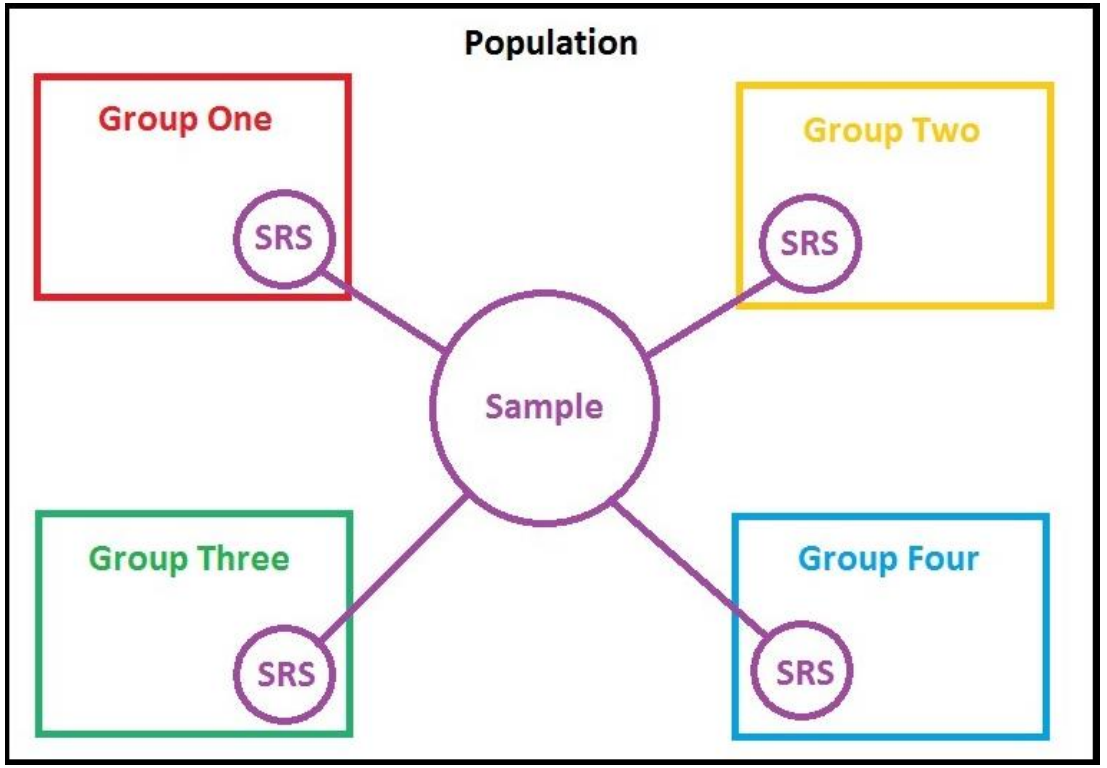
Population and Sample

Probability Sampling Methods:

Stratified Sampling

With stratified sampling, the population is divided into groups, based on some characteristic. Then, within each group, a probability sample (often a simple random sample) is selected. In stratified sampling, the groups are called **strata**.

As a example, suppose we conduct a national survey. We might divide the population into groups or strata, based on geography - north, east, south, and west. Then, within each stratum, we might randomly select survey respondents.



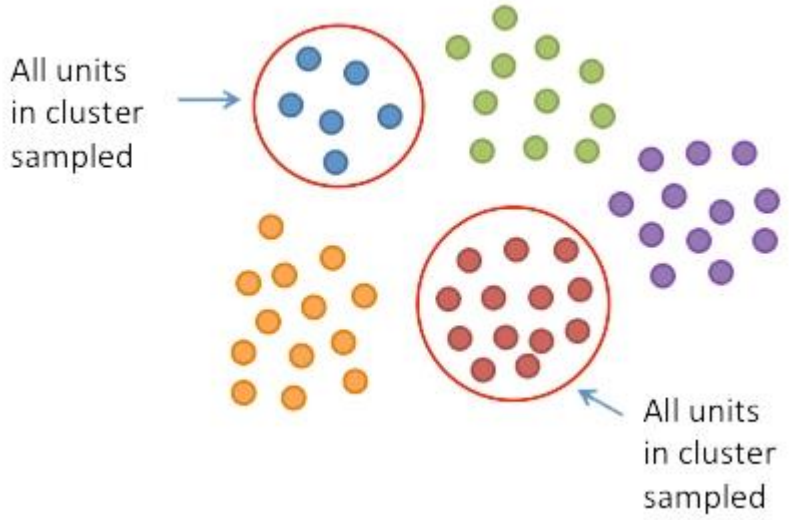
Population and Sample

Probability Sampling Methods:

Cluster Sampling

With cluster sampling, every member of the population is assigned to one, and only one, group. Each group is called a cluster. A sample of clusters is chosen, using a probability method (often simple random sampling). Only individuals within sampled clusters are surveyed.

Difference with Stratified Sampling:
With stratified sampling, the sample includes elements from each stratum. With cluster sampling, in contrast, the sample includes elements only from sampled clusters.



Population and Sample

Probability Sampling Methods:

Multistage Sampling

With multistage sampling, we select a sample by using combinations of different sampling methods.
For example, in Stage 1, we might use cluster sampling to choose clusters from a population. Then, in Stage 2, we might use simple random sampling to select a subset of elements from each chosen cluster for the final sample.

Systematic random Sampling

With systematic random sampling, we create a list of every member of the population. From the list, we randomly select the first sample element from the first *k* elements on the population list. Thereafter, we select every *kth* element on the list. This method is different from simple random sampling since every possible sample of *n* elements is not equally likely.

Problem:
An auto analyst is conducting a satisfaction survey, sampling from a list of 10,000 new car buyers. The list includes 2,500 Ford buyers, 2,500 GM buyers, 2,500 Honda buyers, and 2,500 Toyota buyers. The analyst selects a sample of 400 car buyers, by randomly sampling 100 buyers of each brand.
Is this an example of a simple random sample?

You can learn STATISTICS

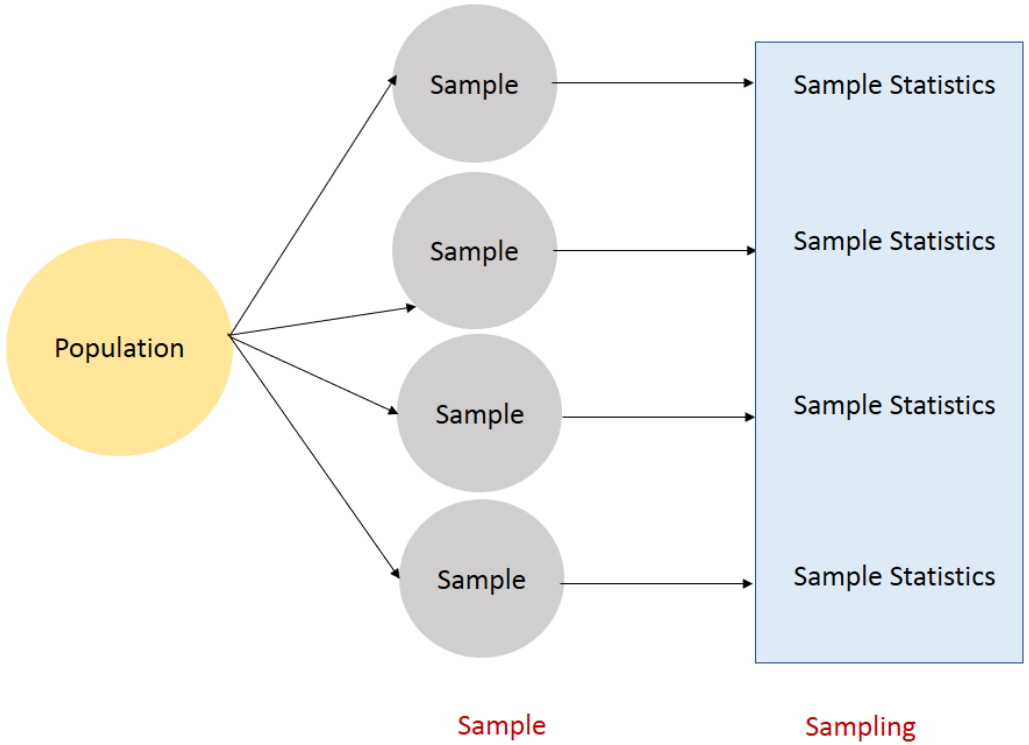
Sampling Distributions

Suppose that we draw all possible samples of size n from a given population. Suppose further that we compute a statistic (e.g., a mean, proportion, standard deviation) for each sample.

The probability distribution of this statistic is called a sampling distribution.
And the standard deviation of this statistic is called the standard error.

The variability of a sampling distribution is measured by its variance or its standard deviation. The variability of a sampling distribution depends on three factors:

N: The number of observations in the population.
n: The number of observations in the sample.
The way that the random sample is chosen.



Sampling Distribution of the Mean

Suppose we draw all possible samples of size *n* from a population of size *N*.
Suppose further that we compute a mean score for each sample. In this way, we create a sampling distribution of the mean.

We know the following about the sampling distribution of the mean:

1. The mean of the sampling distribution (μ_x) is equal to the mean of the population (μ).
2. And the standard error of the sampling distribution (σ_x) is determined by the standard deviation of the population (σ), the population size (N), and the sample size (n).

$$\mu_x = \mu$$
$$\sigma_x = [\ \sigma / \text{sqrt}(n)\ ] * \text{sqrt}[\ (N - n\ ) / (N - 1)\ ]$$

The factor sqrt[(N - n) / (N - 1)] is called the finite population correction.

When the population size is very large relative to the sample size, the fpc is approximately equal to one; and the standard error formula can be approximated by:

$$\sigma_x = \sigma / \text{sqrt}(n)$$

As a general rule, it is safe to use the approximate formula when the sample size is no bigger than 1/20 of the population size.

Sampling Distribution of the Proportion

Suppose we draw all possible samples of size *n* from a population of size *N*.

In a population of size *N*, suppose that the probability of the occurrence of an event (dubbed a "success") is P; and the probability of the event's non-occurrence (dubbed a "failure") is Q.

From this population, suppose that we draw all possible samples of size *n*. And finally, within each sample, suppose that we determine the proportion of successes *p* and failures *q*.

In this way, we create a sampling ***distribution of the proportion***.

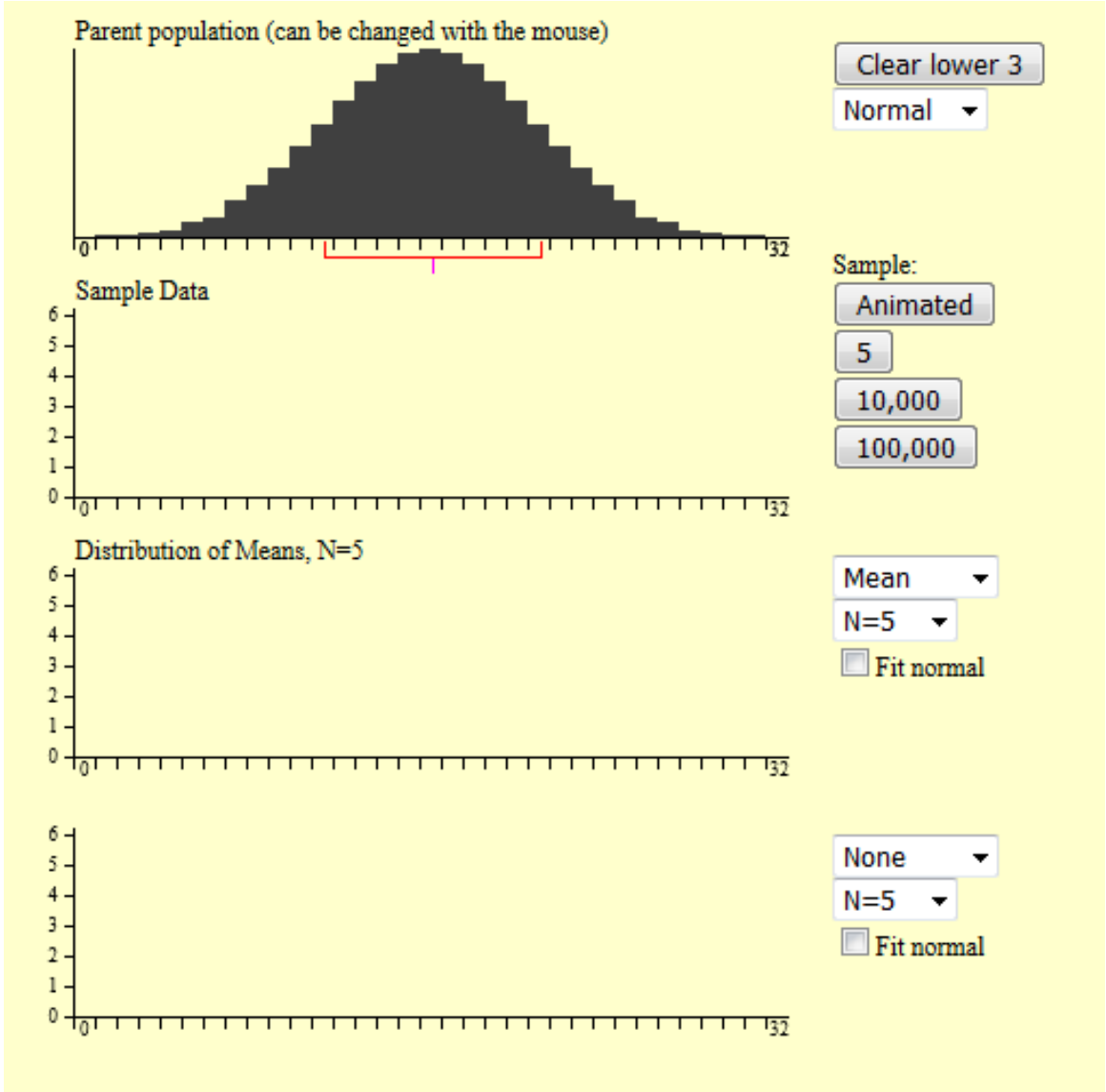
Results:
We find that the mean of the sampling distribution of the proportion (μ_p) is equal to the probability of success in the population (P).
And the standard error of the sampling distribution (σ_p) is determined by the standard deviation of the population (σ), the population size, and the sample size.

$$\mu_p = P$$
$$\sigma_p = [\ \sigma / \text{sqrt}(n)\ ] * \text{sqrt}[\ (N - n\ ) / (N - 1)\ ]$$
$$\sigma_p = \text{sqrt}[\ PQ/n\ ] * \text{sqrt}[\ (N - n\ ) / (N - 1)\ ]$$

Or $\sigma_p = \text{sqrt}[\ PQ/n\ ]$ if finite correction is not applied.

Demo time

Visit here: http://onlinestatbook.com/stat_sim/sampling_dist/



Sampling Distribution - Learnings

In the examples given so far, a population was specified and the sampling distribution of the mean and the range were determined.

In practice, the process proceeds the other way: you collect sample data and from these data you estimate parameters of the sampling distribution.
This knowledge of the sampling distribution can be very useful.

For example, knowing the degree to which means from different samples would differ from each other and from the population mean would give you a sense of how close your particular sample mean is likely to be to the population mean.

The most common measure of how much sample means differ from each other is the standard deviation of the sampling distribution of the mean. This standard deviation is called the **standard error** of the mean. If all the sample means were very close to the population mean, then the standard error of the mean would be small. On the other hand, if the sample means varied considerably, then the standard error of the mean would be large.

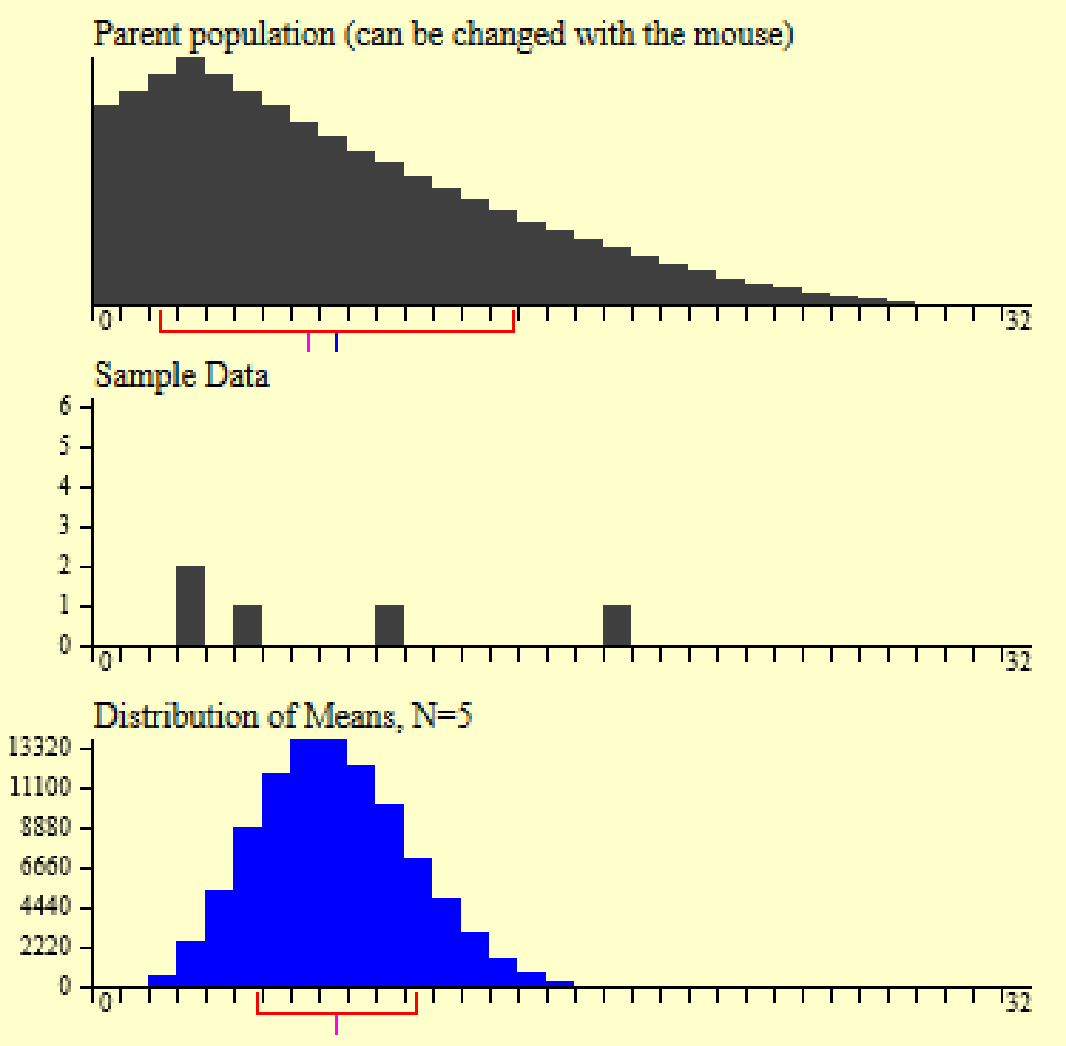
To be specific, assume your sample mean were 125 and you estimated that the standard error of the mean were 5. If you had a normal distribution, then it would be likely that your sample mean would be within 10 units of the population mean since most of a normal distribution is within two standard deviations of the mean.

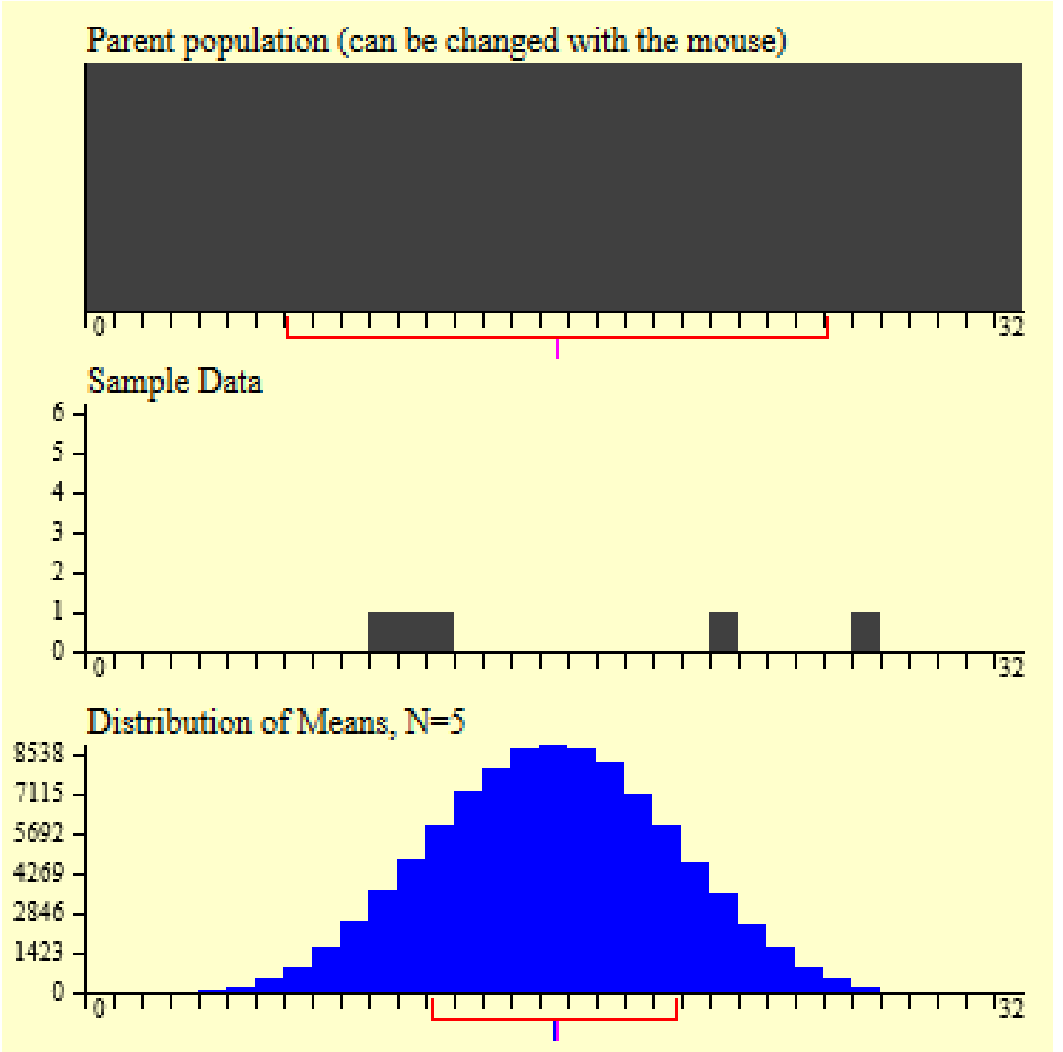
Central Limit Theorem

Given a population with a finite mean μ and a finite non-zero variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean of μ and a variance of σ^2/N as N, the sample size, increases.

Central Limit Theorem

What is remarkable is that regardless of the shape of the parent population, the sampling distribution of the mean approaches a normal distribution as N increases.





Sampling Distribution of Difference Between Means

Statistics problems often involve comparisons between two independent sample means.

Suppose we have **two populations with means equal to μ1 and μ2**. Suppose further that we **take all possible samples of size n1 and n2**.

Supposing following **assumptions are valid**:

- The size of each population is large relative to the sample drawn from the population. That is, N_1 is large relative to n_1, and N_2 is large relative to n_2.
- The samples are independent.
- The set of differences between sample means is normally distributed. This will be true if each population is normal or if the sample sizes are large.

As a result of this, we can infer:

1) The expected value of the difference between all possible sample means is equal to the difference between population means.

$$E(x_1 - x_2) = \mu_d = \mu_1 - \mu_2$$

2) The standard deviation of the difference between sample means (σ_d) is approximately:

$$\sigma_d = \text{sqrt}(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)$$

Sampling Distribution of Difference Between Means

Deriving second inference:

The variance of the difference between independent random variables is equal to the sum of the individual variances.

$$\sigma^2_d = \sigma^2_{(x1 - x2)} = \sigma^2_{x1} + \sigma^2_{x2}$$

If the populations N_1 and N_2 are both large relative to n_1 and n_2, respectively, then:

$$\sigma^2_{x1} = \sigma^2_1 / n_1$$

$$\sigma^2_{x2} = \sigma^2_2 / n_2$$

$$\sigma_d^2 = \sigma_1^2 / n_1 + \sigma_2^2 / n_2$$

$$\sigma_d = \text{sqrt}(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)$$

Sampling Distribution of Difference Between Means

Example:

For boys, the average number of absences in the first grade is 15 with a standard deviation of 7; for girls, the average number of absences is 10 with a standard deviation of 6.
In a nationwide survey, suppose 100 boys and 50 girls are sampled. What is the probability that the male sample will have *at most* three more days of absences than the female sample?

Solution:

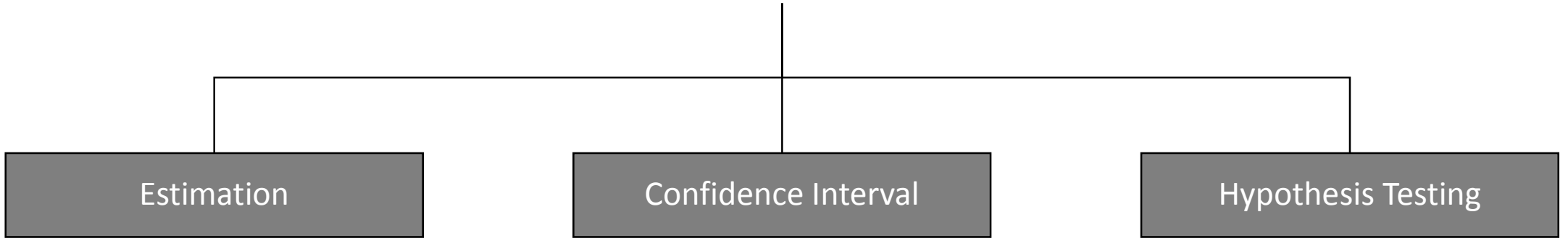
1. Find the mean difference (male absences minus female absences) in the population. $\mu_d = \mu_1 - \mu_2 = 15 - 10 = 5$
2. Find the standard deviation of the difference.

$$\sigma_d = \text{sqrt}(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)$$
$$\sigma_d = \text{sqrt}(7^2/100 + 6^2/50)$$
$$\sigma_d = \text{sqrt}(0.49 + .72) = \text{sqrt}(1.21) = 1.1$$

3. Find the probability. This problem requires us to find the probability that the average number of absences in the boy sample minus the average number of absences in the girl sample is less than 3.

Next Steps..

Now we have to learn about:



The process by which one makes inferences about a population, based on information obtained from a sample.

Describes the amount of uncertainty associated with a sample estimate of a population parameter.

Hypothesis testing refers to the formal procedures used by statisticians to accept or reject statistical hypotheses.

Some more distributions...

t Distribution

Chi-Square Distribution

T Distribution

It is a probability distribution that is used to estimate population parameters when:

1) The sample size is small or
2) When the population variance is unknown.

According to the central limit theorem, the sampling distribution of a statistic (like a sample mean) will follow a normal distribution, as long as the sample size is sufficiently large. Therefore, when we know the standard deviation of the population, we can compute a z-score, and use the normal distribution to evaluate probabilities with the sample mean.

But sample sizes are sometimes small, and often we do not know the standard deviation of the population. And we rely on the distribution of the t statistic (also known as the t score), whose values are given by:

$$t = [x - \mu] / [s / \sqrt{n}]$$

where x is the sample mean, μ is the population mean, s is the standard deviation of the sample, and n is the sample size.

The t distribution allows us to conduct statistical analyses on certain data sets that are not appropriate for analysis, using the normal distribution

Degrees of Freedom

Some estimates are based on more information than others. For example, an estimate of the variance based on a sample size of 100 is based on more information than an estimate of the variance based on a sample size of 5.

The degrees of freedom (dof) of an estimate is the number of independent pieces of information on which the estimate is based.

Case:
Assuming mean height of Martians is 6. We randomly sample one Martial and found that his height is 8.
Variance of this sample – (8-6)^2 – 4. So we estimate the variance of the population to be 4 based on this sample.

Now we sample one more Martian having height 5. His variance from the mean would be (6-5)^2 – 1.

Based on these two samples we would estimate the variance of the population to be 2.5 (avg of both)

Since this estimate is based on two independent pieces of information, it has two degrees of freedom. The two estimates are independent because they are based on two independently and randomly selected Martians.

Generally we don't have population mean, we have to estimate it as well.

Degrees of Freedom

We don't know the population mean and we have to estimate it from the samples now.

We have two samples with height 8 and 5. Estimation of population mean would be: avg(8,5) – 6.5
Computing variance of both of them:

Estimate 1 – (8 – 6.5)^2 = 2.25
Estimate 2 – (5 – 6.5)^2 = 2.25

Are these estimates independent?

No, because each height was used in estimation of the mean, so the first height influenced the mean and hence the estimate 2 of the variance. Changing the first height would change the estimate 2 of the variance.

The important point is that the two estimates are not independent and therefore we do not have two degrees of freedom.

In general, the degrees of freedom for an estimate is equal to the number of values minus the number of parameters estimated en route to the estimate in question. In the Martians example, there are two values (8 and 5) and we had to estimate one parameter (μ) on the way to estimating the parameter of interest (σ^2). Therefore, the estimate of variance has 2 - 1 = 1 degree of freedom. If we had sampled 12 Martians, then our estimate of variance would have had 11 degrees of freedom. Therefore, the degrees of freedom of an estimate of variance is equal to N - 1, where N is the number of observations.

T Distribution

There are actually many different t distributions. The particular form of the t distribution is determined by its **degrees of freedom**.

When estimating a mean score or a proportion from a single sample, the number of independent observations is equal to the sample size minus one. Hence, the distribution of the *t* statistic from samples of size 8 would be described by a t distribution having 8 - 1 or 7 degrees of freedom. Similarly, a t distribution having 15 degrees of freedom would be used with a sample of size 16.

Properties of the t – Distribution

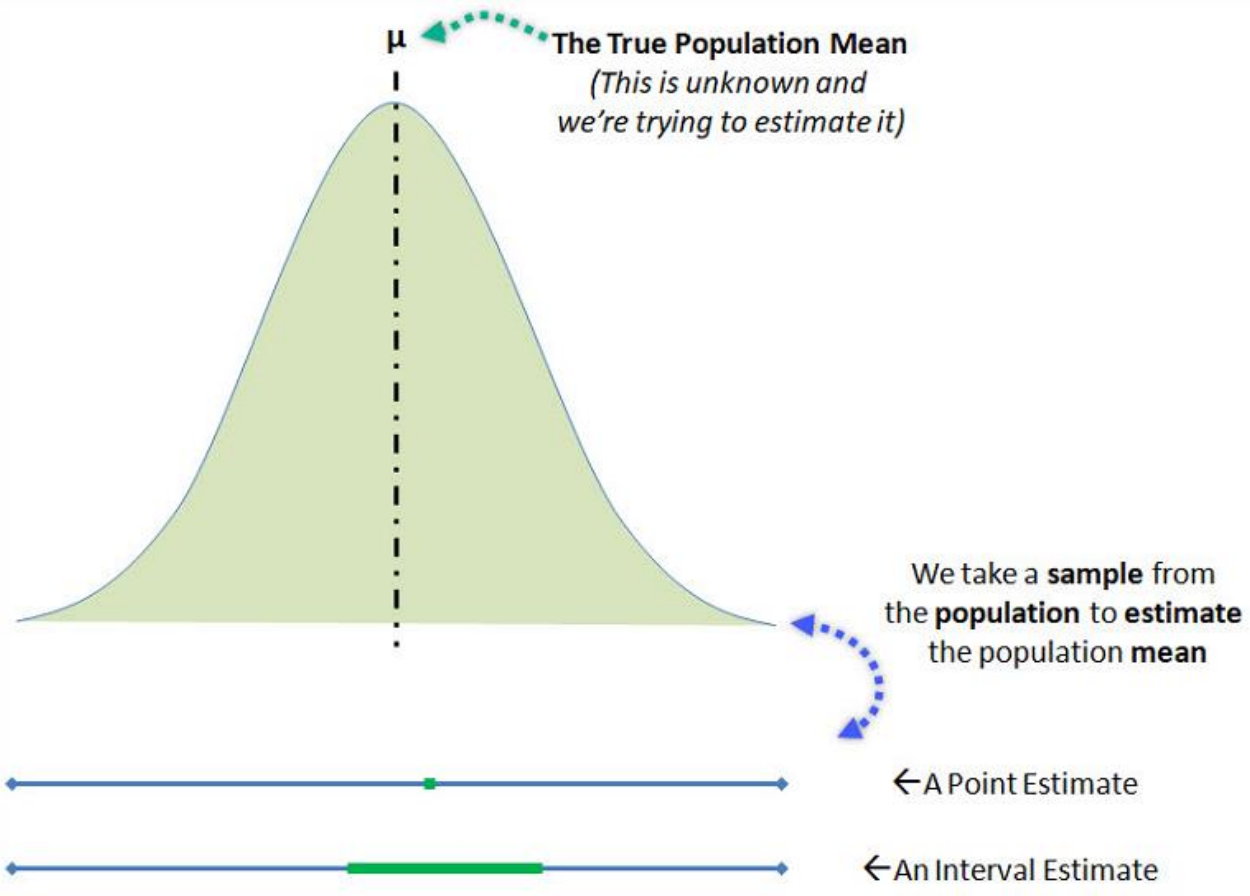
1. The mean of the distribution is equal to 0.
2. The variance is equal to v / (v - 2), where v is the degrees of freedom.
3. The variance is always greater than 1, although it is close to 1 when there are many degrees of freedom. With infinite degrees of freedom, the t distribution is the same as the standard normal distribution.

Estimation

In statistics, **estimation** refers to the process by which one makes inferences about a population, based on information obtained from a sample.
An estimate of a population parameter may be expressed in two ways:

Point Estimate:
A point estimate of a population parameter is a single value of a statistic. For example, the sample mean x is a point estimate of the population mean μ. Similarly, the sample proportion *p* is a point estimate of the population proportion *P*.

Interval Estimate:
An interval estimate is defined by two numbers, between which a population parameter is said to lie. For example, $a < x < b$ is an interval estimate of the population mean μ. It indicates that the population mean is greater than *a* but less than *b*.



Estimation

What is a Confidence Interval?

Statisticians use a confidence interval to express the precision and uncertainty associated with a particular sampling method. A confidence interval consists of three parts.

1. A confidence level
2. A statistic
3. A margin of error.

The confidence level describes the uncertainty of a sampling method. The statistic and the margin of error define an interval estimate that describes the precision of the method. The interval estimate of a confidence interval is defined by the *sample statistic* ± *margin of error*.

For example, suppose we compute an interval estimate of a population parameter. We might describe this interval estimate as a 95% confidence interval. This means that if we used the same sampling method to select different samples and compute different interval estimates, the true population parameter would fall within a range defined by the *sample statistic* ± *margin of error* 95% of the time.

Confidence intervals are preferred to point estimates, because confidence intervals indicate (a) the precision of the estimate and (b) the uncertainty of the estimate.

Estimation

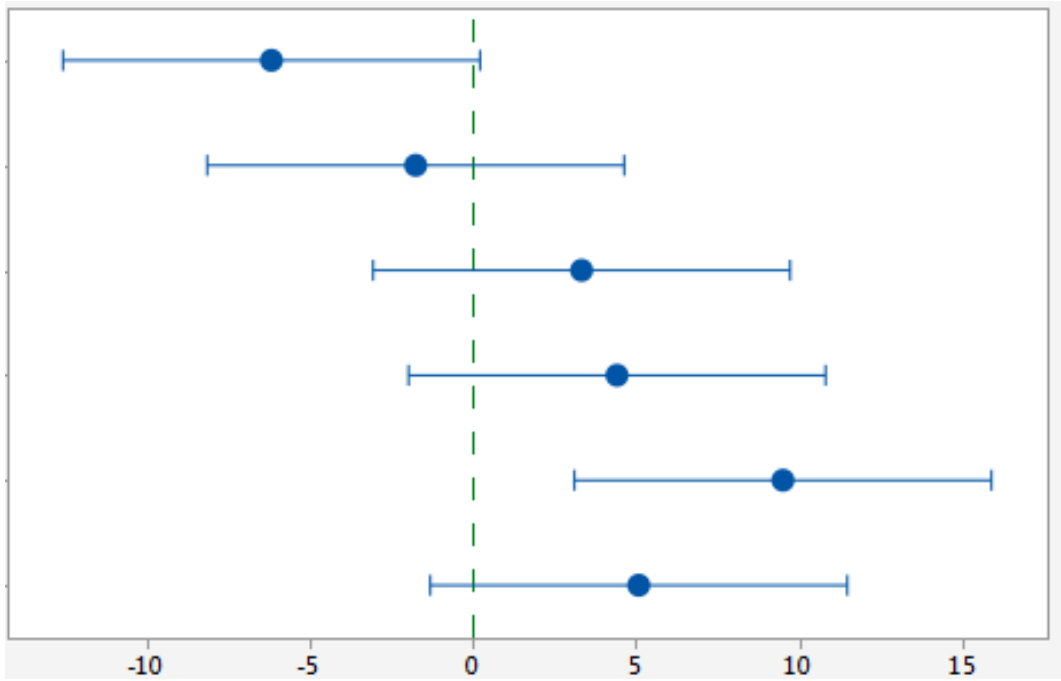
Confidence Interval?

A confidence level: The probability part of a confidence interval is called a **confidence level**.

- The confidence level describes the likelihood that a particular sampling method will produce a confidence interval that includes the true population parameter.

Interpretation of Confidence Level:

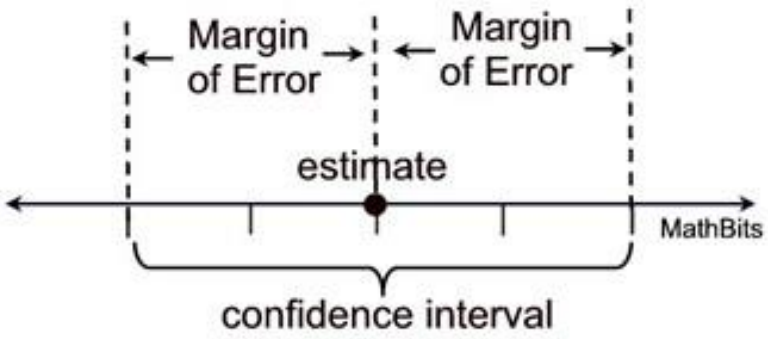
Suppose we collected all possible samples from a given population, and computed confidence intervals for each sample. Some confidence intervals would include the true population parameter; others would not. A 95% confidence level means that 95% of the intervals contain the true population parameter; a 90% confidence level means that 90% of the intervals contain the population parameter; and so on.



Estimation

Confidence Interval?

Margin of Error: In a confidence interval, the range of values above and below the sample statistic is called the margin of error.



For example, suppose the local newspaper conducts an election survey and reports that the independent candidate will receive 30% of the vote.

The newspaper states that the survey had a 5% margin of error and a confidence level of 95%.
These findings result in the following confidence interval: **We are 95% confident that the independent candidate will receive between 25% and 35% of the vote.**

- Many public opinion surveys report interval estimates, but not confidence intervals. They provide the margin of error, but not the confidence level. **To clearly interpret survey results you need to know both!** We are much more likely to accept survey findings if the confidence level is high (say, 95%) than if it is low (say, 50%).

Estimation

Understanding Confidence Interval:

Which of the following statements is true.
I. When the margin of error is small, the confidence level is high.
II. When the margin of error is small, the confidence level is low.
III. A confidence interval is a type of point estimate.
IV. A population mean is an example of a point estimate.

The confidence level is not affected by the margin of error. When the margin of error is small, the confidence level can low or high or anything in between. A confidence interval is a type of interval estimate, not a type of point estimate. A *population* mean is not an example of a point estimate; a *sample* mean is an example of a point estimate.

Estimation – Standard Error

The standard error is an estimate of the standard deviation of a statistic.

Standard Deviation of Sample Estimates: Statisticians use sample statistics to estimate population parameters. Naturally, the value of a statistic may vary from one sample to the next. The variability of a statistic is measured by its standard deviation.

The table below shows formulas for computing the standard deviation of statistics from simple random samples.

Sample mean, x	$\sigma_x = \sigma / \text{sqrt}(n)$
Sample proportion, p	$\sigma_p = \text{sqrt} [P(1 - P) / n]$
Difference between means, $x_1 - x_2$	$\sigma_{x1-x2} = \text{sqrt} [\sigma^2_1 / n_1 + \sigma^2_2 / n_2]$

Note: In order to compute the standard deviation of a sample statistic, you must know the value of one or more population parameters. For example, to compute the standard deviation of the sample mean (σ_x), you need to know the variance of the population (σ).

What if we don't know the population parameters?

Estimation – Standard Error

The standard error is an estimate of the standard deviation of a statistic.

Standard Error of Sample Estimates: Sadly, the values of population parameters are often unknown, making it impossible to compute the standard deviation of a statistic. When this occurs, use the standard error.

The standard error is computed from known sample statistics. The table below shows how to compute the standard error for simple random samples, assuming the population size is at least 20 times larger than the sample size.

Sample mean, x	$SE_x = s / sqrt(n)$
Sample proportion, p	$SE_p = sqrt [p(1 - p) / n]$
Difference between means, $x_1 - x_2$	$SE_{x1-x2} = sqrt [s^2_1 / n_1 + s^2_2 / n_2]$

The equations for the standard error are identical to the equations for the standard deviation, except for one thing - the standard error equations use statistics where the standard deviation equations use parameters.

Estimation – Margin of Error

In a confidence interval, the range of values above and below the sample statistic is called the margin of error.

For example, suppose we wanted to know the percentage of adults that exercise daily. We could devise a sample design to ensure that our sample estimate will not differ from the true population value by more than, say, 5 percent (the margin of error) 90 percent of the time (the confidence level).

The margin of error can be defined by either of the following equations:

Margin of error = Critical value x Standard deviation of the statistic
Margin of error = Critical value x Standard error of the statistic

How to find Critical value:

1. Compute alpha (α): α = 1 - (confidence level / 100)
2. Find the critical probability (p*): p* = 1 - α/2
3. To express the critical value as a z-score, find the z-score having a cumulative probability equal to the critical probability.
4. To express critical value as a t-statistics:
    1. find DOF.
    2. Critical t-statistics is the t statistics having degrees of freedom equal to DOF and cumulative probability equal to critical probability.

Estimation – Margin of Error

Should you express the critical value as a t statistic or as a z-score? One way to answer this question focuses on the population standard deviation.

- If the population standard deviation is known, use the z-score.
- If the population standard deviation is unknown, use the t statistic.

Another approach focuses on sample size.

- If the sample size is large, use the z-score.
- If the sample size is small, use the t statistic.

Estimation – Margin of Error

Example:
900 high school freshmen were randomly selected for a national survey. Among survey participants, the mean grade-point average (GPA) was 2.7, and the standard deviation was 0.4. What is the margin of error, assuming a 95% confidence level?
(TO ME – Explain single tailed experiments and double tailed experiments)

Margin of error = Critical value x Standard error of the statistic

We need – 1) Critical value and 2) SE of the statistic
Critical Value –

Alpha = 1 – (confidence level / 100) = 1 – 0.95 = 0.05
Critical probability = 1 – (alpha / 2) = 1 – (0.05 / 2) = 0.975
DOF – n-1 = 899
From t-dist table, value corresponding to critical probability 0.975 and 899 is 1.96

Standard Error = 0.4 / sqrt(900) = 0.013

ME = Critical value * Standard error = 1.96 * 0.013 = 0.025

Estimation – Margin of Error

How to Interpret Confidence Intervals

Suppose that a 90% confidence interval states that the population mean is greater than 100 and less than 200. How would you interpret this statement?

Some people think this means there is a 90% chance that the population mean falls between 100 and 200. This is incorrect. Like any population parameter, the population mean is a constant, not a random variable. It does not change. The probability that a constant falls within any given range is always 0.00 or 1.00.

The confidence level describes the uncertainty associated with a sampling method. Suppose we used the same sampling method to select different samples and to compute a different interval estimate for each sample. Some interval estimates would include the true population parameter and some would not. A 90% confidence level means that we would expect 90% of the interval estimates to include the population parameter; a 95% confidence level means that 95% of the intervals would include the parameter; and so on.

DEMO - http://www.rossmanchance.com/applets/ConfSim.html

Hypothesis Testing

A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true. Hypothesis testing refers to the formal procedures used by statisticians to accept or reject statistical hypotheses.

The best way to determine whether a statistical hypothesis is true would be to examine the entire population. Since that is often impractical, researchers typically examine a random sample from the population. If sample data are not consistent with the statistical hypothesis, the hypothesis is rejected.

There are two types of statistical hypotheses:
Null hypothesis. The null hypothesis, denoted by H_o, is usually the hypothesis that sample observations result purely from chance.
Alternative hypothesis. The alternative hypothesis, denoted by H_1 or H_a, is the hypothesis that sample observations are influenced by some non-random cause.

Example
Suppose we wanted to determine whether a coin was fair and balanced. A null hypothesis might be that half the flips would result in Heads and half, in Tails. The alternative hypothesis might be that the number of Heads and Tails would be very different. Symbolically, these hypotheses would be expressed as
H_o: $P = 0.5$
H_a: $P \neq 0.5$
Suppose we flipped the coin 50 times, resulting in 40 Heads and 10 Tails. Given this result, we would be inclined to reject the null hypothesis.

Hypothesis Testing

Some researchers say that a hypothesis test can have one of two outcomes: you accept the null hypothesis or you reject the null hypothesis.
Many statisticians, however, take issue with the notion of "accepting the null hypothesis." Instead, they say: you **reject the null hypothesis** or you **fail to reject the null hypothesis**.
Why the distinction between "acceptance" and "failure to reject?" - Acceptance implies that the null hypothesis is true. Failure to reject implies that the data are not sufficiently persuasive for us to prefer the alternative hypothesis over the null hypothesis.

Decision Errors : Two types of errors can result from a hypothesis test.
Type I error. A Type I error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type I error is called the **significance level**. This probability is also called **alpha**, and is often denoted by α.
Type II error. A Type II error occurs when the researcher fails to reject a null hypothesis that is false. The probability of committing a Type II error is called **Beta**, and is often denoted by β. The probability of *not* committing a Type II error is called the **Power** of the test.

Hypothesis Testing

P-value: The strength of evidence in support of a null hypothesis is measured by the **P-value**. Suppose the test statistic is equal to *S*. The P-value is the probability of observing a test statistic as extreme as *S*, assuming the null hypothesis is true. If the P-value is less than the significance level, we reject the null hypothesis.

Region of acceptance. The **region of acceptance** is a range of values. If the test statistic falls within the region of acceptance, the null hypothesis is not rejected. The region of acceptance is defined so that the chance of making a Type I error is equal to the significance level. The set of values outside the region of acceptance is called the **region of rejection**. If the test statistic falls within the region of rejection, the null hypothesis is rejected. In such cases, we say that the hypothesis has been rejected at the α level of significance.

Hypothesis Testing

One tailed and two tailed hypothesis testing:

A test of a statistical hypothesis, where the region of rejection is on only one side of the sampling distribution, is called a **one-tailed test**.
For example, suppose the null hypothesis states that the mean is less than or equal to 10. The alternative hypothesis would be that the mean is greater than 10. The region of rejection would consist of a range of numbers located on the right side of sampling distribution; that is, a set of numbers greater than 10.

A test of a statistical hypothesis, where the region of rejection is on both sides of the sampling distribution, is called a **two-tailed test**.
For example, suppose the null hypothesis states that the mean is equal to 10. The alternative hypothesis would be that the mean is less than 10 or greater than 10. The region of rejection would consist of a range of numbers located on both sides of sampling distribution; that is, the region of rejection would consist partly of numbers that were less than 10 and partly of numbers that were greater than 10.

Hypothesis Testing

Example 1:

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug, subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an affect on response time?

We write 2 hypothesizes first:
H_0 – The drug has no effect, i.e., the mean would be 1.2 s even with the drug
H_a – Drug has an effect – The mean does not equals 1.2 s when drug is given.

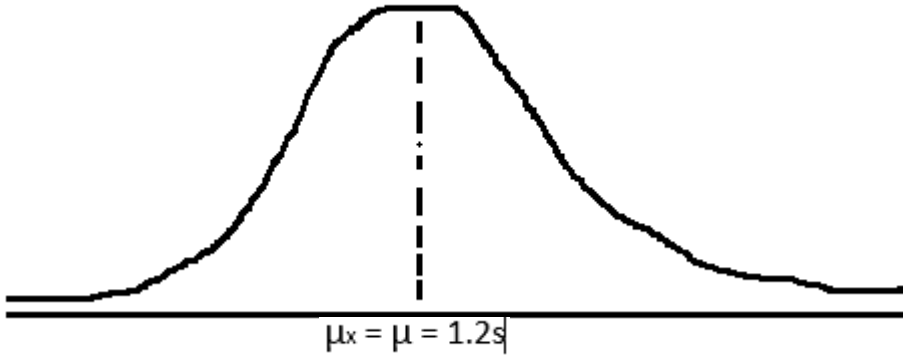
Should we accept alternate hypothesis or data is not convincing enough?
Assuming null hypothesis is true, we find the probability of this scenario. If that probability is really small, that means null hypothesis isn't true, because we have that value.

Hypothesis Testing

Example 1:

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug, subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an affect on response time?

Sampling Dist: $\mu_x = \mu = 1.2s$, s.d. of sampling dist = (population std. dev.) / sqrt(100)
= (sample std dev) / sqrt(100) (good approximator since sample size is big) = 0.05

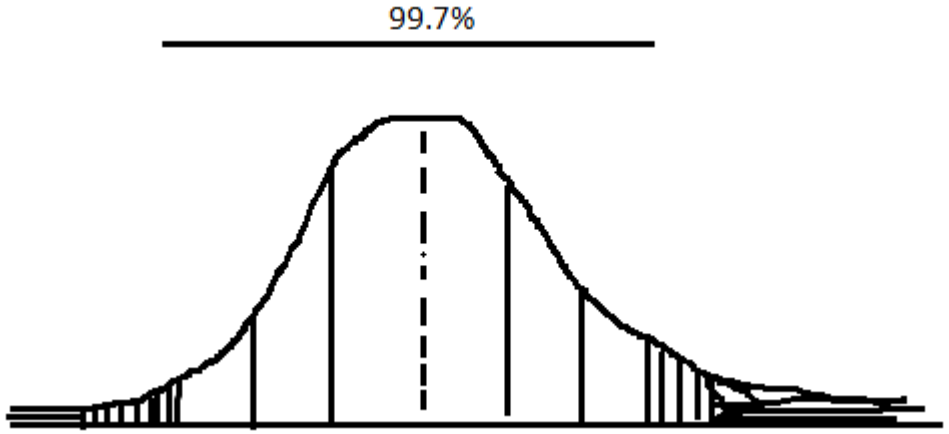


Now we have to find how many std dev away is 1.05 from the mean (z score). Or what is the probability of getting 1.05.

Z score – 1.2 – 1.05 / 0.05 - ~3.
So we are around 3 standard deviations away.

Hypothesis Testing

Example 1:



So probability of getting a value in the shaded area is ~0.3%.
Or 0.003
Since the probability is very less, we can reject the null hypothesis and go with the alternate hypothesis.

Probability of getting a result more extreme than the sample, given the null hypothesis, is called p-value.

And we reject the null hypothesis if the p value is less than a certain limit.
Conventionally that limit is 0.05.
The p value we got is 0.003. So we can reject the null hypothesis.

So there is a very small probability that we could have got that value in sample (1.05 s) if null hypothesis was true.
So we reject it.

Hypothesis Testing

Example 2:

The mean emission of all engines of a new design needs to be below 20 parts per million if the design is to meet new emission requirements. 10 engines are manufactured for testing purposes, and the emission level of each is determined. The emission data is:

15.6 16.2 22.5 20.5 16.4 19.4 16.6 17.9 12.7 13.9 (mean – 17.17 and stdev – 2.98)

Does the data supply sufficient evidence to conclude that this type of engine meets the new standard? Assuming we are willing to risk a Type I error with probability 0.01

H_0 : $\mu = 20$ ppm and H_1 : $\mu < 20$ ppm (Reject H_0 if p(mean = 17.17 | H_0 is true) < 1%

T = (17.17 – 20) / (sample std dev (2.98) / sqrt(10)) = -3.003 ~ -3.00

Now we need to figure out the probability of getting this t statistics in a t statistics with dof = 9.

(we care about 1 sided t dist.)

Or we can find out the t value at which the probability of getting any value <= to that value is 1%.

That is out threshold value. If the value we calculated is less than the threshold value, then we have less than 1 % chance of getting mean 17.17 if null hypothesis is true.

So we reject the null hypothesis.

Hypothesis Testing

Hypothesis Test for a Proportion:

- Compute the standard deviation (σ) of the sampling distribution: $\sigma = \sqrt{P * (1 - P) / n}$

where P is the hypothesized value of population proportion in the null hypothesis, and n is the sample size

- Test statistics: $z = (p - P) / \sigma$

where P is the hypothesized value of population proportion in the null hypothesis, p is the sample proportion, and σ is the standard deviation of the sampling distribution.

Example 1:

The CEO of a large electric utility claims that 80 percent of his 1,000,000 customers are very satisfied with the service they receive. To test this claim, the local newspaper surveyed 100 customers, using simple random sampling. Among the sampled customers, 73 percent say they are very satisified. Based on these findings, can we reject the CEO's hypothesis that 80% of the customers are very satisfied? Use a 0.05 level of significance.

(Example of Two tailed test)

Hypothesis Testing

Hypothesis Test for a Proportion:

Example 1:

Null hypothesis: $P = 0.80$

Alternative hypothesis: $P \neq 0.80$

Calculating standard error:

$$\sigma = \text{sqrt}[P * (1 - P) / n]$$
$$\sigma = \text{sqrt} [(0.8 * 0.2) / 100]$$
$$\sigma = \text{sqrt}(0.0016) = 0.04$$

Calculating z-statistic

$$z = (p - P) / \sigma = (.73 - .80)/0.04 = -1.75$$

For this z score, we find the probability: 0.04. Since two tailed, so total p value would be 0.04+0.04. = 0.08. Since the P-value (0.08) is greater than the significance level (0.05), we cannot reject the null hypothesis.

***Note**: The approach is appropriate because the sampling method was simple random sampling, the sample included at least 10 successes and 10 failures, and the population size was at least 10 times the sample size.*

Hypothesis Testing

Hypothesis Test for a Proportion:

Example 2:

Suppose the previous example is stated a little bit differently. Suppose the CEO claims that *at least* 80 percent of the company's 1,000,000 customers are very satisfied. Again, 100 customers are surveyed using simple random sampling. The result: 73 percent are very satisfied. Based on these results, should we accept or reject the CEO's hypothesis? Assume a significance level of 0.05. (One tailed test)

Null hypothesis: $P >= 0.80$
Alternative hypothesis: $P < 0.80$

Now we calculate the standard deviation (σ) and compute the z-score test statistic (z).

$$\sigma = \text{sqrt}[P * (1 - P) / n] = \text{sqrt} [(0.8 * 0.2) / 100]$$
$$\sigma = \text{sqrt}(0.0016) = 0.04$$
$$z = (p - P) / \sigma = (.73 - .80)/0.04 = -1.75$$

where P is the hypothesized value of population proportion in the null hypothesis, p is the sample proportion, and n is the sample size.

Since we have a one-tailed test, the P-value is the probability that the z-score is less than -1.75. We use the Normal Distribution Calculator to find $P(z < -1.75) = 0.04$. Thus, the P-value = 0.04.

Since the P-value (0.04) is less than the significance level (0.05), we cannot accept the null hypothesis.

Hypothesis Testing

Hypothesis Test for a Mean:

General Procedure:

- Compute the standard error (SE) of the sampling distribution: SE = s * sqrt{ (1/n) * [(N - n) / (N - 1)] }

where *s* is the standard deviation of the sample, N is the population size, and *n* is the sample size.

When the population size is much larger (at least 20 times larger) than the sample size, the standard error can be approximated by: SE = s / sqrt(n)

- The degrees of freedom (DF) is equal to the sample size (n) minus one. Thus, DF = n - 1.

- The test statistic is a t statistic (t) defined by the following equation. t = (x - μ) / SE

where x is the sample mean, μ is the hypothesized population mean in the null hypothesis, and SE is the standard error.

Example 1:

An inventor has developed a new, energy-efficient lawn mower engine. He claims that the engine will run continuously for 5 hours (300 minutes) on a single gallon of regular gasoline. From his stock of 2000 engines, the inventor selects a simple random sample of 50 engines for testing. The engines run for an average of 295 minutes, with a standard deviation of 20 minutes. Test the null hypothesis that the mean run time is 300 minutes against the alternative hypothesis that the mean run time is not 300 minutes. Use a 0.05 level of significance. (Assume that run times for the population of engines are normally distributed.)

Hypothesis Testing

Hypothesis Test for a Mean:

Example 1:

$$\text{Null hypothesis: } \mu = 300$$
$$\text{Alternative hypothesis: } \mu \neq 300$$

Calculating Standard error and t-statistic:

$$SE = s / \sqrt{n} = 20 / \sqrt{50} = 20/7.07 = 2.83$$
$$DF = n - 1 = 50 - 1 = 49$$
$$t = (x - \mu) / SE = (295 - 300)/2.83 = -1.77$$

where s is the standard deviation of the sample, x is the sample mean, μ is the hypothesized population mean, and n is the sample size.
Since we have a two-tailed test, the P-value is the probability that the t statistic having 49 degrees of freedom is less than -1.77 or greater than 1.77.
We use the t Distribution Calculator to find $P(t < -1.77) = 0.04$, and $P(t > 1.77) = 0.04$. Thus, the P-value = 0.04 + 0.04 = 0.08.

Since the P-value (0.08) is greater than the significance level (0.05), we cannot reject the null hypothesis.

Hypothesis Testing

Hypothesis Test for a Mean:

Example 2:

Bon Air Elementary School has 1000 students. The principal of the school thinks that the average IQ of students at Bon Air is at least 110. To prove her point, she administers an IQ test to 20 randomly selected students. Among the sampled students, the average IQ is 108 with a standard deviation of 10. Based on these results, should the principal accept or reject her original hypothesis? Assume a significance level of 0.01. (Assume that test scores in the population of engines are normally distributed.)

$$\text{Null hypothesis: } \mu >= 110$$

$$\text{Alternative hypothesis: } \mu < 110$$

Note that these hypotheses constitute a one-tailed test. The null hypothesis will be rejected if the sample mean is too small.

$$SE = s / sqrt(n) = 10 / sqrt(20) = 10/4.472 = 2.236$$

$$DF = n - 1 = 20 - 1 = 19$$

$$t = (x - \mu) / SE = (108 - 110)/2.236 = -0.894$$

where s is the standard deviation of the sample, x is the sample mean, μ is the hypothesized population mean, and n is the sample size.

The observed sample mean produced a t statistic test statistic of -0.894. We use the t Distribution Calculator to find P(t < -0.894) = 0.19. This means we would expect to find a sample mean of 108 or smaller in 19 percent of our samples, if the true population IQ were 110. Thus the P-value in this analysis is 0.19.

Since the P-value (0.19) is greater than the significance level (0.01), we cannot reject the null hypothesis.

Hypothesis Testing

Hypothesis Test difference between Means:

The test procedure, called the **two-sample t-test**,

When the null hypothesis states that there is no difference between the two population means (i.e., d = 0), the null and alternative hypothesis are often stated in the following form.

$$H_o: \mu_1 = \mu_2$$
$$H_a: \mu_1 \neq \mu_2$$

Standard Error SE = $\text{sqrt}[(s_1^2/n_1) + (s_2^2/n_2)]$

DF = $(s_1^2/n_1 + s_2^2/n_2)^2 / \{ [(s_1^2 / n_1)^2 / (n_1 - 1)] + [(s_2^2 / n_2)^2 / (n_2 - 1)] \}$

If DF does not compute to an integer, round it off to the nearest whole number. Some texts suggest that the degrees of freedom can be approximated by the smaller of $n_1 - 1$ and $n_2 - 1$; but the above formula gives better results.

The test statistic is a t statistic (t) defined by the following equation $t = [(x_1 - x_2) - d] / SE$

where x_1 is the mean of sample 1, x_2 is the mean of sample 2, d is the hypothesized difference between population means, and SE is the standard error.

Hypothesis Testing

Hypothesis Test difference between Means:

Example 1:
Within a school district, students were randomly assigned to one of two Math teachers - Mrs. Smith and Mrs. Jones. After the assignment, Mrs. Smith had 30 students, and Mrs. Jones had 25 students.
At the end of the year, each class took the same standardized test. Mrs. Smith's students had an average test score of 78, with a standard deviation of 10; and Mrs. Jones' students had an average test score of 85, with a standard deviation of 15.
Test the hypothesis that Mrs. Smith and Mrs. Jones are equally effective teachers. Use a 0.10 level of significance. (Assume that student performance is approximately normal.)

Null hypothesis: $\mu_1 - \mu_2 = 0$
Alternative hypothesis: $\mu_1 - \mu_2 \neq 0$

Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the difference between sample means is too big or if it is too small

$$SE = \text{sqrt}[(s_1^2/n_1) + (s_2^2/n_2)]$$
$$SE = \text{sqrt}[(10^2/30) + (15^2/25] = \text{sqrt}(3.33 + 9)$$
$$SE = \text{sqrt}(12.33) = 3.51$$

Hypothesis Testing

Hypothesis Test difference between Means:

Example 1:

Calculation of DOF and t-statistic:

$DF = (s_1^2/n_1 + s_2^2/n_2)^2 / \{ [(s_1^2 / n_1)^2 / (n_1 - 1)] + [(s_2^2 / n_2)^2 / (n_2 - 1)] \}$

$DF = (10^2/30 + 15^2/25)^2 / \{ [(10^2 / 30)^2 / (29)] + [(15^2 / 25)^2 / (24)] \}$

$DF = (3.33 + 9)^2 / \{ [(3.33)^2 / (29)] + [(9)^2 / (24)] \} = 152.03 / (0.382 + 3.375) = 152.03/3.757 = 40.47$

$t = [(x_1 - x_2) - d] / SE = [(78 - 85) - 0] / 3.51 = -7/3.51 = -1.99$

where s1 is the standard deviation of sample 1, s2 is the standard deviation of sample 2, n1 is the size of sample 1, n2 is the size of sample 2, x1 is the mean of sample 1, x2 is the mean of sample 2, d is the hypothesized difference between the population means, and SE is the standard error.

Since we have a two-tailed test, the P-value is the probability that a t statistic having 40 degrees of freedom is more extreme than -1.99; that is, less than -1.99 or greater than 1.99.

We use the t Distribution Calculator to find $P(t < -1.99) = 0.027$, and $P(t > 1.99) = 0.027$. Thus, the P-value = 0.027 + 0.027 = 0.054.

Since the P-value (0.054) is less than the significance level (0.10), we cannot accept the null hypothesis.