# Business Analytics

## Sampling & Hypothesis Testing

Proschool
An IMS Initiative

# Population and Sample



**Population:**

• An aggregate of objects or individuals under study is called population.

• Population may be a group of people suffering from a particular disease, collection of books, group of students.

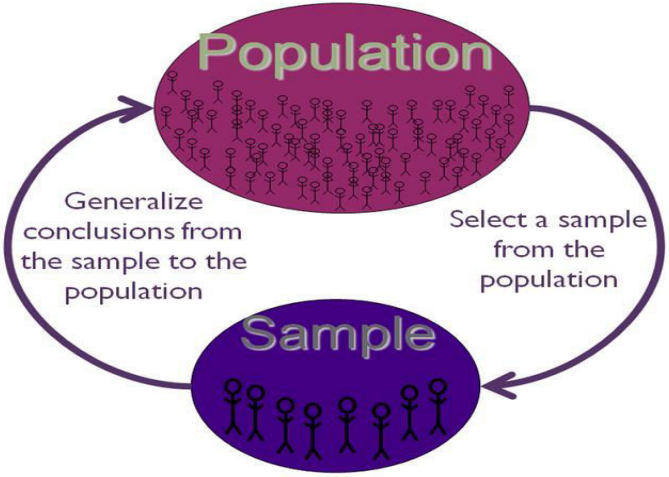
**Sample:**

• Any part of population under study is called a sample.

• While purchasing food grains, we inspect only a handful of grains and draw conclusions about the quality of the whole lot. In this case handful of grains is a sample and the whole lot is a population.

# Sampling

- “A sample is some portion of a population. Because many populations of interest are too large to work with directly, techniques of statistical sampling have been devised to obtain samples taken from larger populations.”
- Sampling is the process by which this part is chosen.



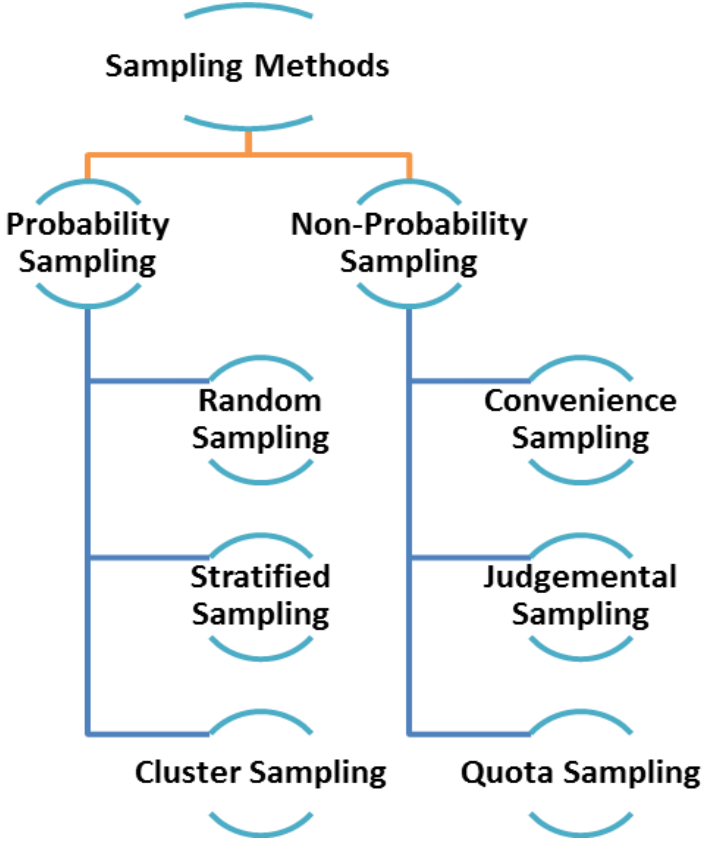
# Theory of Sampling



- The theory of sampling is as follows:
- Researchers want to gather information about a whole group of people (the **population).**
- Researchers can only observe a part of the population (the **sample).**
- The findings from the sample are **generalized,** or extended, back to the population.

# Sampling Methods





# Advantages of Sampling



- Makes the research of any type and size manageable.
- Significantly saves the costs of the research.
- Results in more accurate research findings.
- Provides an opportunity to process the information in a more efficient way.
- Accelerates the speed of primary data collection.

# Probability Overview



- Probability - *How **likely** something is to happen.*

$$\text{Probability} = \frac{\text{\# of ways a certain outcome can occur}}{\text{Total Possible Outcomes (Sample Space)}}$$

- As you can see, with this formula, we will write the probability of an event as a fraction. The numerator (in red) is the number of chances and the denominator (in blue) is the set of all possible outcomes. This is also known as the **sample space.**

- Uncertainty is all around us and we often come across real-life situations when we have to decide on making a choice from the available options.
- Questions like “Will it rain? Do I need to carry an umbrella today?” or “Will there be a rise in taxes? Which party will win the election this time?” All these situations demand a decision from us and this is the time when probability theory comes to our rescue.
- From weather forecasts, opinion polls to making business decisions, the concepts of probability come in handy at various aspects of our daily lives.
- What are the chances?

# Example 1

Given a standard die , determine the probability for the following events when rolling the die one time.

i)P(5)  ii) P(even number)  iii) P(7)

(Note: P(5) means probability of rolling a 5)

Sample space: {1, 2, 3, 4, 5, 6}
Number of sample points: 6

i) P(5) = # of ways outcome occurs / Total possible outcomes
= 1 / 6

# Example 1

ii) P(even number) = 3 / 6
(Since there are 3 even numbers in the sample space)

iii) P(7) = 0
(This is an impossible event because the die does not contain number 7. Whenever the probability is impossible, the answer is zero)

# Example 2

There are 4 blue marbles, 5 red marbles, 1 green marble and 2 black marbles in a bag. Suppose you select 1 marble at random. Find each of the following probability.

i) P( black )
ii) P( blue )
iii) P( blue or black)
iv) P( not green)
v) P( not purple)

Number of sample points: 12 ( There are 12 marbles in total )

i) P( black ) = 2 / 12 = 1 / 6
( There are 2 black marbles in the bag )

# Example 2

ii) P(blue) = 4 / 12 = 1 / 3
( There are 4 blue marbles in the bag )

iii) P( blue or black ) = 6 / 12 = 1 / 2
( There are 4 blue + 2 black marbles in the bag )

iv) P( not green ) = 11 / 12
( There's 1 green, so 12- 1 = 11 that are not green marbles )

v) P( not purple ) = 1
( no marble is purple )

# Interpreting Probability Result



- Mathematically, the probability that an event will occur is expressed as a number between 0 and 1.
- The probability of event A is represented by P(A).
- If P(A) equals zero, event A will almost definitely not occur.
- If P(A) is close to zero, there is only a small chance that event A will occur.
- If P(A) equals 0.5, there is a 50-50 chance that event A will occur.
- If P(A) is close to one, there is a strong chance that event A will occur.
- If P(A) equals one, event A will almost definitely occur.

# Interpreting Probability Result



- In a statistical experiment, the sum of probabilities for all possible outcomes is equal to one. This means, for example, that if an experiment can have three possible outcomes (A, B, and C), then P(A) + P(B) + P(C) = 1.

- Probability that the experiment results in a successful outcome (S) is:

P(S) = ( Number of successful outcomes ) / ( Total number of equally likely outcomes ) = r / n

**Consider the following experiment**.

An urn has 10 marbles. Two marbles are red, three are green, and five are blue. If an experimenter randomly selects 1 marble from the urn, what is the probability that it will be green?

In this experiment, there are 10 equally likely outcomes, three of which are green marbles. Therefore, the probability of choosing a green marble is 3/10 or 0.30.

Business Analytics

# Probability Sampling Methods

Proschool

An IMS Initiative

# Probability Sampling



**Probability sampling** is a **sampling** technique wherein the samples are gathered in a process that gives all the individuals in the population equal chances of being selected.

A researcher must identify specific sampling elements (e.g. persons) to include in the sample.

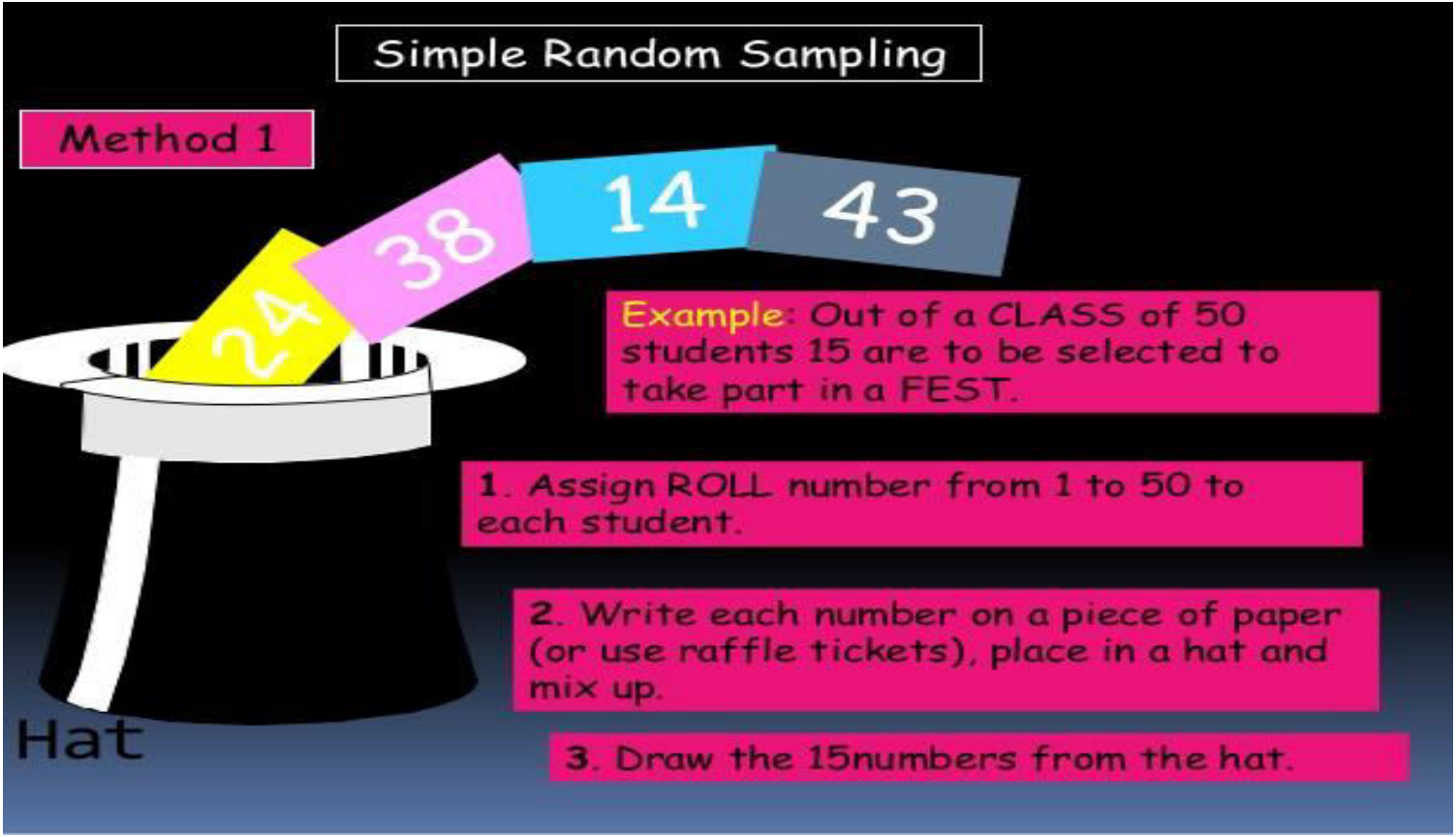
For example, if conducting a telephone survey, the researcher needs to try to reach the specific sampled person, by calling back several times, to get an accurate sample

# Random Sample



- The term random has a very precise meaning. **Each individual in the population of interest has an equal likelihood of selection.** This is a very strict meaning -- you can't just collect responses on the street and have a random sample.
- The assumption of an *equal chance of selection* means that sources such as a telephone book or voter registration lists are not adequate for providing a random sample of a community. In both these cases there will be a number of residents whose names are not listed. Telephone surveys get around this problem by random-digit dialing -- but that assumes that everyone in the population has a telephone.
- The key to random selection is that there is no bias involved in the selection of the sample. Any variation between the sample characteristics and the population characteristics is only a matter of chance.

# Simple Random Sampling

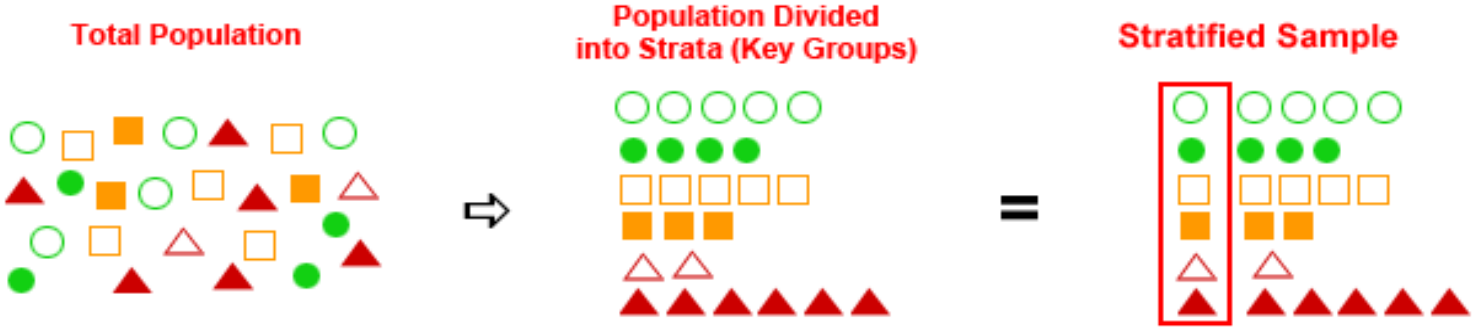


# Stratified Sample

- A stratified sample is a mini-reproduction of the population.
- If the population is not homogeneous, SRS is not very effective.
- Therefore, the entire population is divided into several homogeneous groups (strata) according to the characteristics of importance for the research (e.g. by gender, social class, blood group, education level, etc)
- Then the population is randomly sampled *within* each category or stratum. If 38% of the population is college-educated, then 38% of the sample is randomly selected from the college-educated population.
- Stratified samples are as good as or better than random samples, but they require a fairly detailed advance knowledge of the population characteristics, and therefore are more difficult to construct.

# Stratified Sample



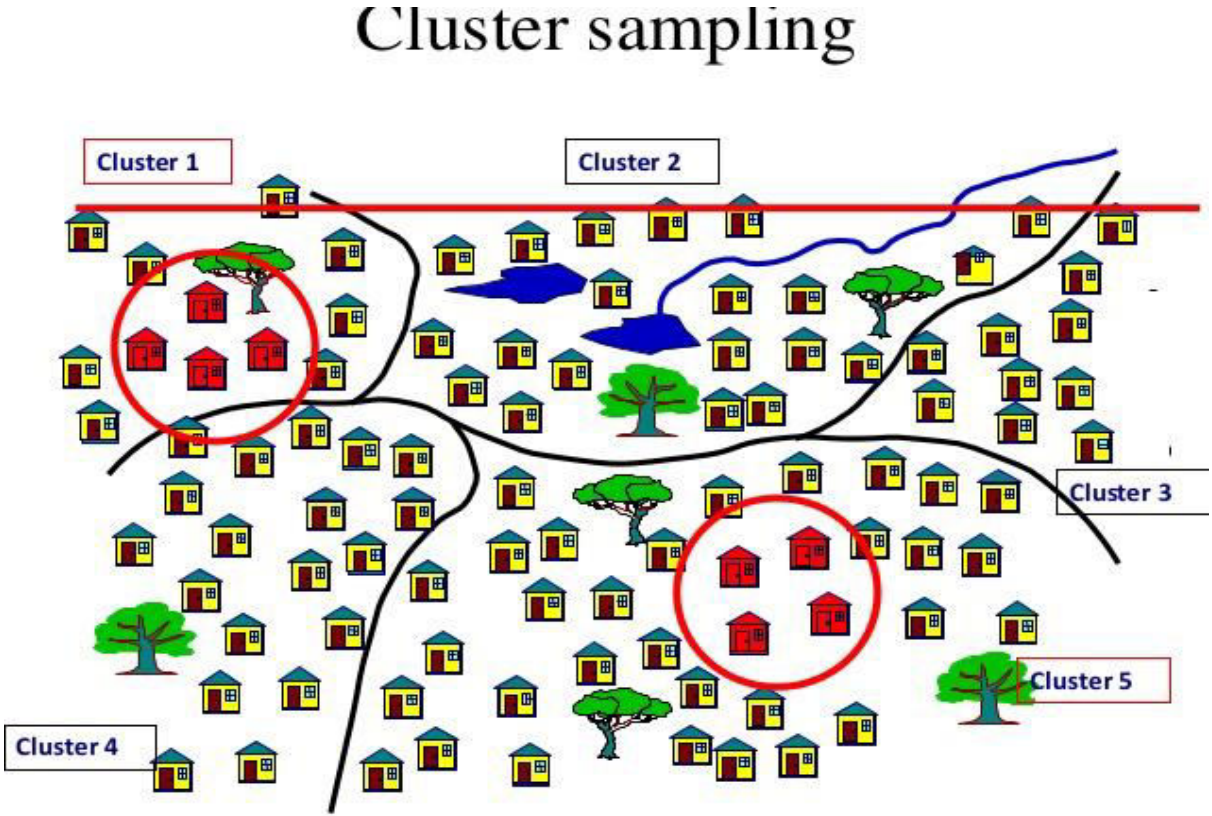


# Cluster Sampling



- **Cluster sampling** is a sampling technique used when "natural" but relatively homogeneous groupings are evident in a statistical population. It is often used in marketing research.
- In this technique, the total population is divided into these groups (or clusters) and a simple random sample of the groups is selected.

# Cluster Sampling



# Non Probability Sampling Methods

# Non Probability Sampling



- As they are not truly representative, non-probability samples are less desirable than probability samples. However, a researcher may not be able to obtain a random or stratified sample, or it may be too expensive.
- A researcher may not care about generalizing to a larger population.
- The validity of non-probability samples can be increased by trying to approximate random selection, and by eliminating as many sources of bias as possible.

# Quota Sample

- The defining characteristic of a quota sample is that the researcher deliberately sets the proportions of levels or strata within the sample. This is generally done to insure the inclusion of a particular segment of the population.
- The proportions may or may not differ dramatically from the actual proportion in the population. The researcher sets a quota, independent of population characteristics.

# Quota Sample Example



| | Chocolate Buyers | Respondent quota (sample size = 200) |
|---|---|---|
| Men | 40% | 80 |
| Women | 60% | 120 |

# Quota sample

*Example:*

- A researcher is interested in the attitudes of members of different religions towards the death penalty.
- In Iowa a random sample might miss Muslims (because there are not many in that state). To be sure of their inclusion, a researcher could set a quota of 3% Muslim for the sample.
- However, the sample will no longer be representative of the actual proportions in the population. This may limit generalizing to the state population. But the quota will guarantee that the views of Muslims are represented in the survey.

# Convenience sample



- **Convenience sampling** is a sample taken from a group you have easy access to. The idea is that anything learned from this study will be applicable to the larger population.
- By using a large, convenient size, you are able to more confidently say the sample represents the population.
- Furthermore, the convenient group you are testing should not be fundamentally different than if you had taken a sample from another area. If you are trying to say something about women, for example, then your convenient sample cannot be men.

# Convenience Sampling Example



- *Involves collecting information from members of the population who are conveniently available to provide this information.*

- *Example; 'Pepsi Challenge' contest with the purpose of determining whether people prefer one product over another, might be set up at a shopping mall visited by many shoppers.*

# Convenience Sample Example



- You are interested in the effects of caffeine on study habits of college students. To test the whole population you would need all current college students and a whole lot of time and soda.
- A sample would be a test of a few college students from all of the colleges in the India, requiring you to fly them in for the testing.
- A convenience sample would be a large group of college students from your local college or colleges. They are close by, are in college, and are not different than other college students.

# Judgment Sample



- **Judgment sample** is a type of nonrandom sample that is selected based on the opinion of an expert.
- Results obtained from a judgment sample are subject to some degree of bias, due to the frame and population not being identical.
- The frame is a list of all the units, items, people, etc., that define the population to be studied.
- Example:
- A TV researcher wants a quick sample of opinions about a political announcement. Taking views of people in the street.

Business Analytics

# Sampling & Estimation

Proschool

An IMS Initiative

# Sampling

- What is Sampling?

A shortcut method for investigating a whole population

Data is gathered on a small part of the whole parent population or sampling frame, and used to inform what the whole picture is like.

- Why sample?

In reality there is simply not enough; time, energy, money, labor/man power, equipment, access to suitable sites to measure every single item or site within the parent population or whole sampling frame.

Therefore an appropriate sampling strategy is adopted to obtain a representative, and statistically valid sample of the whole.

# Sampling

## Sampling considerations

- Larger sample sizes are more accurate representations of the whole.
- The sample size chosen is a balance between obtaining a statistically valid representation, and the time, energy, money, labor, equipment and access available.
- A sampling strategy made with the minimum of bias is the most statistically valid.

# Sampling



- Most approaches assume that the parent population has a normal distribution where most items or individuals clustered close to the mean, with few extremes
- A 95% probability or confidence level is usually assumed, for example 95% of items or individuals will be within plus or minus two standard deviations from the mean
- This also means that up to five per cent may lie outside of this - sampling, no matter how good can only ever be claimed to be a very close estimate

# Sampling

- Sampling techniques
- Three main types of sampling strategy:
- Random
- Systematic
- Stratified

# Random sampling

## Random sampling

- Least biased of all sampling techniques, there is no subjectivity each member of the total population has an equal chance of being selected
- Can be obtained using random number tables
- Microsoft Excel has a function to produce random number
- The function is simply:

  =rand()

# Random sampling



- Type rand() into a cell and it will produce a random number in that cell. Copy the formula throughout a selection of cells and it will produce random numbers.
- You can modify the formula to obtain whatever range you wish, for example if you wanted random numbers from one to 250, you could enter the following formula:
- =INT(250*RAND())+1
- Where INT eliminates the digits after the decimal, 250* creates the range to be covered, and +1 sets the lowest number in the range.

# Advantages & Disadvantages of Random Sampling



**Advantages:**

- Can be used with large sample populations
- Avoids bias

**Disadvantages:**

- Can lead to poor representation of the overall parent population or area if large areas are not hit by the random numbers generated. This is made worse if the study area is very large
- There may be practical constraints in terms of time available and access to certain parts of the study area

# Systematic sampling



- Samples are chosen in a systematic, or regular way.
- They are evenly/regularly distributed in a spatial context, for example every two meters along a transect line.
- They can be at equal/regular intervals in a temporal context, for example every half hour or at set times of the day.
- They can be regularly numbered, for example every 10th house or person

# Advantages & Disadvantages of Systematic Sampling

**Advantages:**

- It is more straight-forward than random sampling
- A grid doesn't necessarily have to be used, sampling just has to be at uniform intervals
- A good coverage of the study area can be more easily achieved than using random sampling

**Disadvantages:**

- It is more biased, as not all members or points have an equal chance of being selected
- It may therefore lead to over or under representation of a particular pattern

# Stratified Sampling



This method is used when the parent population or sampling frame is made up of sub-sets of known size. These sub-sets make up different proportions of the total, and therefore sampling should be stratified to ensure that results are proportional and representative of the whole.

# Advantages & Disadvantages of Stratified Sampling



**Advantages:**

- It can be used with random or systematic sampling, and with point, line or area techniques
- If the proportions of the sub-sets are known, it can generate results which are more representative of the whole population
- It is very flexible and applicable to many geographical enquiries
- Correlations and comparisons can be made between sub-sets

**Disadvantages:**

- The proportions of the sub-sets must be known and accurate if it is to work properly
- It can be hard to stratify questionnaire data collection, accurate up to date population data may not be available and it may be hard to identify people's age or social background effectively

# Estimation



- In statistics, estimation refers to the process by which one makes inferences about a population, based on information obtained from a sample.
- An estimator attempts to approximate the unknown parameters using the measurements.
- For example, it is desired to estimate the proportion of a population of voters who will vote for a particular candidate.
- That proportion is the parameter sought; the estimate is based on a small random sample of voters.

# Estimation



- Estimation is the process of determining a likely value for a population parameter (e.g. the true population mean or proportion) based on a random sample.
- In practice, a sample is drawn from the target population, and sample statistics (e.g. the sample mean or sample proportion) are used to generate estimates of the unknown parameter.
- The sample should be representative of the population, ideally with participants selected at random from the population.
- Because different samples can produce different results, it is necessary to quantify the sampling error or variation that exists among estimates from different samples.

Business Analytics

# Point & Interval Estimates

Proschool

An IMS Initiative

# Point and Internal estimates



- The point estimate is our best guess of the true value of the parameter, while the interval estimate gives a measure of accuracy of that point estimate by providing an interval that contains plausible values.
- 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 $\mu$. Similarly, the sample proportion $p$ is a point estimate of the population proportion $P$.

# Point and Internal estimates



- 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 $\mu$. It indicates that the population mean is greater than $a$ but less than $b$.

# R - Point Estimate of Population Mean

**Problem**

Find a point estimate of mean university student height with the sample data from survey.

**Solution**

For convenience, we begin with storing the survey data of student heights in the variable height. Survey.

```
> library(MASS)
> height.survey=survey$Height
```

It turns out not all students have answered the question, and we must filter out the missing values.

# R - Point Estimate of Population Mean



- It turns out not all students have answered the question, and we must filter out the missing values.
- Hence we apply the mean function with the "na.rm" argument as TRUE for skipping the missing values.

```
> mean(height.survey, na.rm=TRUE)
[1] 172.3809
```

- A point estimate of the mean student height is 172.38 centimeters.

# Interval Estimate of Population Mean-Known Variance



**Problem**

Assume the population standard deviation σ of the student height in survey is 9.48. Find the margin of error and interval estimate at 95% confidence level.

**Solution**

We first filter out missing values in survey$Height with the na.omit function, and save it in height.response.

```
> library(MASS)
> height.response=na.omit(survey$Height)
```

# Interval Estimate of Population Mean-Known Variance

- Then we compute the standard error of the mean.

```
> n=length(height.response)
> sigma=9.48                    # Population standard deviation
> SEmean=sigma/sqrt(n)          # standard error of mean
> SEmean
[1] 0.6557453
```

# Interval Estimate of Population Mean-Known Variance



Since there are two tails of the normal distribution, the 95% confidence level would imply the 97.5th percentile of the normal distribution at the upper tail. Therefore, in R, $z\alpha/2$ is given by qnorm(.975).

We multiply it with the standard error of the mean and get the margin of error.

# Interval Estimate of Population Mean-Known Variance



```
> E=qnorm(0.975)*SEmean
> E                          # margin of error
[1] 1.285237
```

We then add it up with the sample mean, and find the confidence interval as told.

```
> xbar=mean(height.response)  # sample mean
> xbar+ c(-E,E)
[1] 171.0956 173.6661
```

Assuming the population standard deviation σ being 9.48, the margin of error for the student height survey at 95% confidence level is 1.2852 centimeters. The confidence interval is between 171.10 and 173.67 centimeters.

# Interval Estimate of Population Mean-Unknown Variance

Problem

- Without assuming the population standard deviation of the student height in survey, find the margin of error and interval estimate at 95% confidence level.

Solution

- We first filter out missing values in survey$Height with the na.omit function, and save it in height.response.

```
> library(MASS)
> height.response=na.omit(survey$Height)
```

- Then we compute the sample standard deviation.

```
> n=length(height.response)
> s=sd(height.response)          # sample standard deviation
> SE=s/sqrt(n)                   # standard error estimate
> SE
[1] 0.6811677
```

# Interval Estimate of Population Mean-Unknown Variance

- Since there are two tails of the Student t distribution, the 95% confidence level would imply the97.5th percentile of the Student t distribution at the upper tail.
- Therefore, $t_{\alpha/2}$ is given by qt(.975, df=n-1).
- We multiply it with the standard error estimate SE and get the margin of error.

```
> E=qt(0.975, n-1)     # margin of error
> E
[1] 1.971435
```

- We then add it up with the sample mean, and find the confidence interval.

```
> xbar=mean(height.response)    # sample mean
> xbar + c(-E, E)
[1] 170.4094 174.3523
```

# Interval Estimate of Population Mean-Unknown Variance



- Without assumption on the population standard deviation, the margin of error for the student height survey at 95% confidence level is 1.3429 centimeters. The confidence interval is between 171.04 and 173.72 centimeters.

# Sampling using R



- R has the ability to sample with and without replacement.
- That is, choose at random from a collection of things such as the numbers 1 through 6 in the dice rolling example.
- The sampling can be done with replacement (like dice rolling) or without replacement (like a lottery).
- By default sample samples without replacement each object having equal chance of being picked.
- You need to specify replace=TRUE if you want to sample with replacement. Furthermore, you can specify separate probabilities for each if desired.

# R Codes

Here are some examples

```
> # Roll a die
> sample(1:6,10,replace=TRUE)
 [1] 4 2 6 1 2 6 4 2 6 2
>
> # Toss a coin
> sample(c('H','T'),10,replace=TRUE)
 [1] "T" "H" "H" "H" "T" "T" "T" "H" "H" "H"
>
> # Pick 6 of 54(a lottery)
> sample(1:54,6)              # no replacement
[1] 29  2  5 25 27  6
>
```

# R Codes



```
> # Pick a card
> cards=paste(rep(c("A",2:10,"J","Q","K"),4),c("H","D","S","C"))
> sample(cards,5)          # a pair of jacks without replacement
[1] "8 D" "9 D" "9 S" "5 D" "4 S"
>
> # Roll 2 dice
> dice=as.vector(outer(1:6,1:6,paste))
> sample(dice,5,replace=TRUE)   # replace when rolling die
[1] "3 5" "2 2" "2 1" "2 1" "2 4"
```

# Hypothesis Testing

- Develop null and alternative hypotheses to test for a given situation.
- Understand the difference between one- and two-tailed hypothesis tests.
- Understand Type I and Type II error

# Hypothesis Introduction



- In everyday life, we often have to make decisions based on incomplete information. These may be decisions that are important to us such as, "Will I improve my biology grades if I spend more time studying vocabulary?" or "Should I become a chemistry major to increase my chances of getting into med school?" This section is about the use of hypothesis testing to help us with these decisions.
- Hypothesis testing is a kind of statistical inference that involves asking a question, collecting data, and then examining what the data tells us about how to proceed.

# Hypothesis Introduction



- In a formal hypothesis test, hypotheses are always statements about the population. The hypothesis tests that we will examine in this chapter involve statements about the average values (means) of some variable in the population.
- For example, we may want to know if the average time that college freshmen spend studying each week is really 20 hours per week. We may want to compare this average time spent studying for freshmen that earned a GPA of 3.0 or higher and those that did not. In later chapters, we will be able to test if the average time studying differs for four groups: freshmen, sophomores, juniors and seniors.

# Developing Null and Alternative Hypotheses

- In statistical hypothesis testing, there are always two hypotheses. The hypothesis to be tested is called the null hypothesis and given the sym $H_0$.
- The null hypothesis states that there is no difference between a hypothesized population mean and a sample mean. It is the status quo hypothesis.
- For example, if we were to test the hypothesis that college freshmen study 20 hours per week, we would express our null hypothesis as:

$$H_0 : \mu = 20$$

# Alternative Hypothesis



- We test the null hypothesis against an alternative hypothesis, which is given the symbol $H_a$.
- The alternative hypothesis is often the hypothesis that you believe yourself! It includes the outcomes not covered by the null hypothesis. In this example, our alternative hypothesis would express that freshmen do not study 20 hours per week

$$H_a : \mu \neq 20$$

# Example A

We have a medicine that is being manufactured and each pill is supposed to have 14 milligrams of the active ingredient. What are our null and alternative hypotheses ?

Solution:-

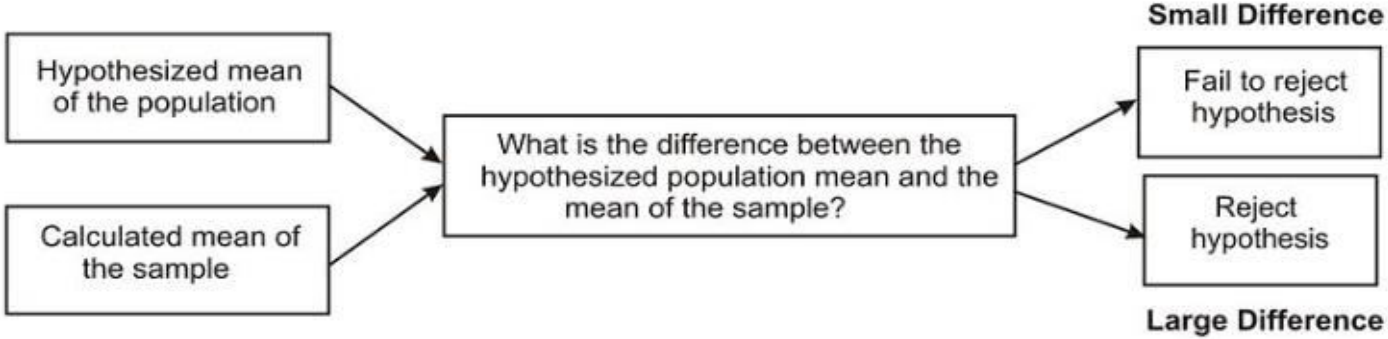
$$H_0 : \mu = 14$$
$$H_a : \mu \neq 14$$

Our null hypothesis states that the population has a mean equal to 14 milligrams. Our alternative hypothesis states that the population has a mean that is different than 14 milligrams.

# Decision Criterion: One & Two-Tailed Hypothesis Tests

- The alternative hypothesis can be supported only by rejecting the null hypothesis.
- To reject the null hypothesis means to find a large enough difference between your sample mean and the hypothesized (null) mean that it raises real doubt that the true population mean is 20.
- If the difference between the hypothesized mean and the sample mean isIf the difference is very small, we do not. In each hypothesis test, we have to decide in advance what the magnitude of that difference must be to allow us to reject the null hypothesis.
- Below is an overview of this process. Notice that if we fail to find a large enough difference to reject, we fail to reject the null hypothesis.

very large, we reject the null hypothesis.


Hypothesized mean of the population
Calculated mean of the sample
What is the difference between the hypothesized population mean and the mean of the sample?
Small Difference
Fail to reject hypothesis
Reject hypothesis
Large Difference

# Hypothesis Tests



- When a hypothesis is tested, a statistician must decide on how much of a difference between means is necessary in order to reject the null hypothesis.
- Statisticians first choose a level of significance or alpha ($\alpha$)/el for their hypothesis test.
- Similar to the significance level you used in constructing confidence intervals, this alpha level tells us how improbable a sample mean must be for it to be deemed "significantly different" from the hypothesized mean.

# Hypothesis Tests



- The most frequently used levels of significance are 0.05 and 0.01.

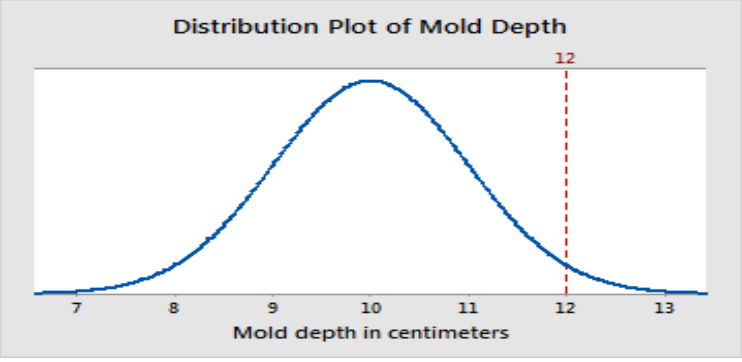
An alpha level of 0.05 means that we will consider our sample mean to be significantly different from the hypothesized mean if the chances of observing that sample mean are less than 5%.

- Similarly, an alpha level of 0.01 means that we will consider our sample mean to be significantly different from the hypothesized mean if the chances of observing that sample mean are less than 1%.

# What is a Z-value?

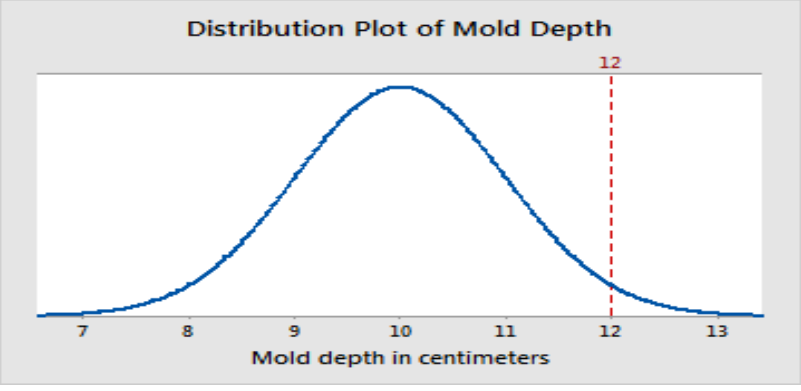


• The Z-value is a test statistic for Z-tests that measures the difference between an observed statistic and its hypothesized population parameter in units of standard error. For example, a selection of factory molds has a mean depth of 10cm and a standard deviation of 1 cm. A mold with a depth of 12 cm has a Z-value of 2, because its depth is two standard deviations greater than the mean. The vertical line represents this observation, and its location relative to the entire population:



# What is a Z-value?

Converting an observation to a Z-value is called standardization. To standardize a observation in a population, subtract the population mean from the observation of interest and divide the result by the population standard deviation. The result of these calculations is the Z-value associated with the observation of interest.



# What is a Z-value?

You can use the Z-value to determine whether to reject the null hypothesis. To determine whether to reject the null hypothesis compare the Z-value to your critical value, which can be found in a standard normal table in most statistics books. The critical value is $Z_{1-\alpha/2}$ for a two sided test and $Z_{1-\alpha}$ for a one sided test. If the absolute value of the Z-value is greater than the critical value, you reject the null hypothesis. If it is not, you fail to reject the null hypothesis.

For example, you want to know whether a second group of molds also has a mean depth of 10cm. You measure the depth of each mold in the second group, and calculate the group's mean depth. A 1-sample Z-test calculates a Z-value of -1.03. You choose an $\alpha$ of 0.05, which results in a critical value of 1.96. Because the absolute value of the Z-value is less than 1.96, you fail to reject the null hypothesis and cannot conclude that the mold's mean depth is different from 10cm.

# What is a t-value?



• The t-value is a test statistic for t-tests that measures the difference between an observed sample statistic and its hypothesized population parameter in units of standard error.

• A t-test compares the observed t-value to a critical value on the t-distribution with (n-1) degrees of freedom to determine whether the difference between the estimated and hypothesized values of the population parameter is statistically significant.

Applications of t-values include:

- comparing two sample means
- comparing the means of paired observations
- determining the significance of a regression coefficient
- comparing two regression coefficients

# What is a t-value?



You can also use t-values in a 1-sample t-test. For example, you want to determine whether the length of a manufactured part meets its target value of 10cm. You take a sample of 50 parts, conduct a two-sided 1-sample t-test on their mean length with the following hypotheses:

$H_0$: $\mu = 0$ (the mean length of all parts meets the target value )

$H_1$: $\mu \neq 0$ (the mean length of all parts does not meet the target value)

The test produces a t-value of 2.5. On the t-distribution with (n-1 = 49) degrees of freedom, this t-value corresponds to a p-value of 0.0158.

For most common significance levels, this result is statistically significant.

Therefore, you reject the null hypothesis that the mean length meets the target, and conclude that the process needs improvement.

# P-Value

- All hypothesis tests ultimately use a *p*-value to weigh the strength of the evidence (what the data are telling you about the population). The *p*-value is a number between 0 and 1 and interpreted in the following way:
- A small *p*-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
- A large *p*-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
- *p*-values very close to the cutoff (0.05) are considered to be marginal (could go either way). Always report the *p*-value so your readers can draw their own conclusions.

# P-Value



- For example, suppose a pizza place claims their delivery times are 30 minutes or less on average but you think it's more than that.
- You conduct a hypothesis test because you believe the null hypothesis, $H_o$, that the mean delivery time is 30 minutes max, is incorrect. Your alternative hypothesis ($H_a$) is that the mean time is greater than 30 minutes.
- You randomly sample some delivery times and run the data through the hypothesis test, and your *p*-value turns out to be 0.001, which is much less than 0.05. In real terms, there is a probability of 0.001 that you will mistakenly reject the pizza place's claim that their delivery time is less than or equal to 30 minutes.

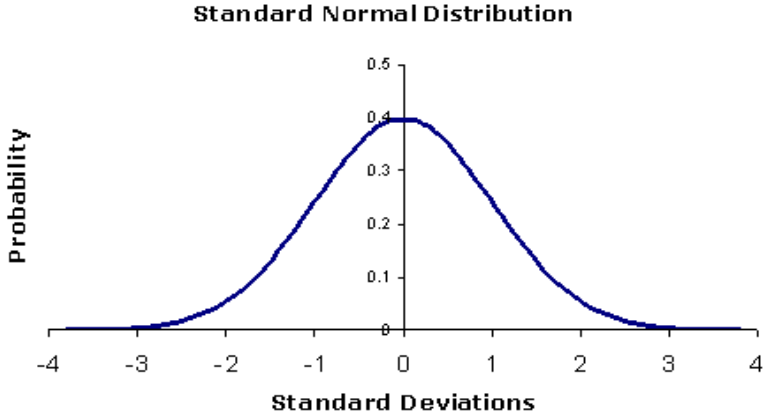
# P-Value

Since typically we are willing to reject the null hypothesis when this probability is less than 0.05, you conclude that the pizza place is wrong; their delivery times are in fact more than 30 minutes on average, and you want to know what they're gonna do about it!

(Of course, you could be wrong by having sampled an unusually high number of late pizza deliveries just by chance.)

# What is a Z score ? What is a P-value?

- The Z score is a test of statistical significance that helps you decide whether or not to reject the null hypothesis.
- The p-value is the probability that you have falsely rejected the null hypothesis.



# What is a Z score ? What is a P-value?



- Z scores are measures of standard deviation. For example, if a tool returns a Z score of +2.5 it is interpreted as "+2.5 standard deviations away from the mean".
- P-values are probabilities. Both statistics are associated with the standard normal distribution. This distribution relates standard deviations with probabilities and allows significance and confidence to be attached to Z scores and p-values.

# T-Score vs. Z-Score: What's the Difference?



- A **z-score** and a **t-score** are both used in hypothesis testing.
- Few topics in elementary statistics cause more confusion to students than deciding when to use the z-score and when to use the t-score.
- Generally, you'll use a z-score in testing more often than a t-score.
- The general rule of thumb for *when* to use a t-score is when your sample size meets the following two requirements:
- The sample size is below 30

# T-Score vs. Z-Score: What's the Difference?

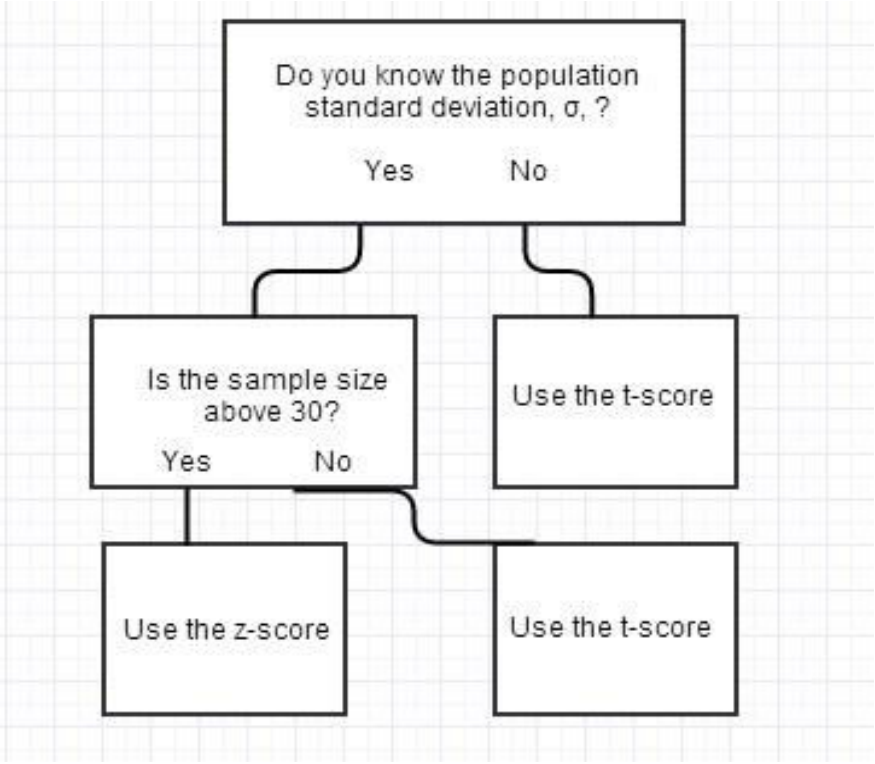


The population standard deviation is unknown (estimated from your sample data)

In other words, you **must** know the standard deviation of the **population** *and* your sample size **must** be above 30 in order for you to be able to use the z-score. Otherwise, use the t-score.

**Note:** You can estimate the population standard deviation, σ by using the standard deviation of the sample, s. However, you can only do this if your sample is 30 or above.

# T-Score vs. Z-Score: What's the Difference?



# Lower Tail Test of Population Mean-Known Variance



**Problem**

Suppose the manufacturer claims that the mean lifetime of a light bulb is more than 10,000 hours. In a sample of 30 light bulbs, it was found that they only last 9,900 hours on average. Assume the population standard deviation is 120 hours. At .05 significance level, can we reject the claim by the manufacturer?

# Lower Tail Test of Population Mean-Known Variance

## Solution

The null hypothesis is that $\mu \geq 10000$. We begin with computing the test statistic.

```
> xbar=9900          # sample mean
> mu0=10000          # hypothesized value
> sigma=120          # population standard deviation
> n=30               # sample size
> z=(xbar-mu0)/(sigma/sqrt(n))
> z                  # test statistic
[1] -4.564355
```

We then compute the critical value at .05 significance level.

```
> alpha=0.05
> z.alpha=qnorm(1-alpha)
> -z.alpha                  # critical value
[1] -1.644854
```

# Interpreting Output



**Answer**

The test statistic -4.5644 is less than the critical value of -1.6449. Hence, at .05 significance level, we reject the claim that mean lifetime of a light bulb is above 10,000 hours.

**Alternative Solution**

Instead of using the critical value, we apply the pnorm function to compute the lower tail **p-value** of the test statistic. As it turns out to be less than the .05 significance level, we reject the null hypothesis that $\mu \geq 10000$.

```
> pval=pnorm(z)
> pval                # lower tail p-value
[1] 2.505166e-06
```

# Upper Tail Test of Population Mean-Known Variance



**Problem**

Suppose the food label on a cookie bag states that there is at most 2 grams of saturated fat in a single cookie. In a sample of 35 cookies, it is found that the mean amount of saturated fat per cookie is 2.1 grams. Assume that the population standard deviation is 0.25 grams. At .05 significance level, can we reject the claim on food label?

# Upper Tail Test of Population Mean-Known Variance



**Solution**

The null hypothesis is that $\mu \leq 2$. We begin with computing the test statistic.

```
> xbar=2.1          # sample mean
> mu0=2             # hypothesized value
> sigma=0.25        # population standard deviation
> n=35              # sample size
> z=(xbar-mu0)/(sigma/sqrt(n))
> z                 # test statistic
[1] 2.366432
```

We then compute the critical value at .05 significance level.

```
> alpha=0.05
> z.alpha=qnorm(1-alpha)
> z.alpha
[1] 1.644854
```

# Interpreting Output

**Answer**

The test statistic 2.3664 is greater than the critical value of 1.6449. Hence, at .05 significance level, we reject the claim that there is at most 2 grams of saturated fat in a cookie.

**Alternative Solution**

Instead of using the critical value, we apply the pnorm function to compute the upper tail **p-value** of the test statistic. As it turns out to be less than the .05 significance level, we reject the null hypothesis that $\mu \leq 2$.

```
> pval=pnorm(z, lower.tail=FALSE)
> pval
[1] 0.008980239
```

# Two-Tailed Test of Population Mean-Known Variance



**Problem**

Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 35 penguins same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the population standard deviation is 2.5 kg. At .05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

# Two-Tailed Test of Population Mean-Known Variance

## Solution

The null hypothesis is that $\mu$ = 15.4. We begin with computing the test statistic.

```
> xbar=14.6          # sample mean
> mu0=15.4           # hypothesized value
> sigma=2.5          # population standard deviation
> n=35               # sample size
> z=(xbar-mu0)/(sigma/sqrt(n))
> z                  # test statistic
[1] -1.893146
```

We then compute the critical values at .05 significance level.

```
> alpha=0.05
> z.half.alpha=qnorm(1-alpha/2)
> c(-z.half.alpha, z.half.alpha)
[1] -1.959964  1.959964
```

# Interpreting Output



**Answer**

The test statistic -1.8931 lies between the critical values -1.9600 and 1.9600. Hence, at .05 significance level, we do *not* reject the null hypothesis that the mean penguin weight does not differ from last year.

**Alternative Solution**

Instead of using the critical value, we apply the pnorm function to compute the two-tailed **p-value** of the test statistic. It doubles the *lower* tail p-value as the sample mean is *less* than the hypothesized value. Since it turns out to be greater than the .05 significance level, we do *not*reject the null hypothesis that $\mu = 15.4$.

```
> pval=2*pnorm(z)      # lower tail
> pval                 # two tailed p-value
[1] 0.05833852
```

# Lower Tail Test of Population Mean-Unknown Variance



**Problem**

Suppose the manufacturer claims that the mean lifetime of a light bulb is more than 10,000 hours. In a sample of 30 light bulbs, it was found that they only last 9,900 hours on average. Assume the sample standard deviation is 125 hours. At .05 significance level, can we reject the claim by the manufacturer?

# Lower Tail Test of Population Mean-Unknown Variance

## Solution

The null hypothesis is that $\mu \geq 10000$. We begin with computing the test statistic.

```
> xbar=9900          # sample mean
> mu0=10000          # hypothesized value
> s=125              # sample standard deviation
> n=30               # sample size
> t=(xbar-mu0)/(s/sqrt(n))
> t                  # test statistic
[1] -4.38178
```

We then compute the critical value at .05 significance level.

```
> alpha=0.05
> t.alpha=qt(1-alpha, df=n-1)
> -t.alpha                  # critical value
[1] -1.699127
```

The test statistic -4.3818 is less than the critical value of -1.6991. Hence, at .05 significance level, we can reject the claim that mean lifetime of a light bulb is above 10,000 hours.

# Interpreting Output

## Alternative Solution

Instead of using the critical value, we apply the pt function to compute the lower tail p-value of the test statistic. As it turns out to be less than the .05 significance level, we reject the null hypothesis that $\mu \geq 10000$.

```
> pval=pt(t,df=n-1)
> pval          # lower tail p-value
[1] 7.035026e-05
```

# Lower Tail Test of Population Proportion



**Problem**

Suppose 60% of citizens voted in last election. 85 out of 148 people in a telephone survey said that they voted in current election. At 0.5 significance level, can we reject the null hypothesis that the proportion of voters in the population is above 60% this year?

# Lower Tail Test of Population Proportion



**Solution**

The null hypothesis is that $p \geq 0.6$. We begin with computing the test statistic.

```
> pbar=85/148          # sample proportion
> p0=0.6               # hypothesized value
> n=148                # sample size
> z=(pbar-p0)/(sqrt(p0*(1-p0)/n))
> z                    # test statistic
[1] -0.6375983
```

We then compute the critical value at .05 significance level.

```
> alpha=0.05
> z.alpha=qnorm(1-alpha)
> -z.alpha
[1] -1.644854
```

# Interpreting Output



**Answer**

The test statistic -0.6376 is *not* less than the critical value of -1.6449. Hence, at .05 significance level, we do *not* reject the null hypothesis that the proportion of voters in the population is above 60% this year.

**Alternative Solution 1**

Instead of using the critical value, we apply the pnorm function to compute the lower tail **p-value** of the test statistic. As it turns out to be greater than the .05 significance level, we do not reject the null hypothesis that $p \geq 0.6$.

```
> pval=pnorm(z)      # lower tail p-value
> pval
[1] 0.2618676
```

# Types of Error, Power of the Test

# Type I & Type II Error



**Type I Error:**

- A Type I error occurs if we reject the null hypothesis H0 (in favor of the alternative hypothesis HA) when the null hypothesis H0 is true.
- We denote $\alpha$ = P(Type I Error).

**Type II Error:**

- A Type II error occurs if we fail to reject the null hypothesis H0 when the alternative hypothesis HA is true.
- We denote $\beta$ = P(Type II Error).

# Type I & Type II Error



- When you do a hypothesis test, two types of errors are possible: type I and type II.
- The risks of these two errors are inversely related and determined by the level of significance and the power for the test.
- Therefore, you should determine which error has more severe consequences for your situation before you define their risks.
- No hypothesis test is 100% certain. Because the test is based on probabilities, there is always a chance of drawing an incorrect conclusion.

# Type I & Type II Error



**Type I error**

• When the null hypothesis is true and you reject it, you make a type I error. The probability of making a type I error is α, which is the level of significance you set for your hypothesis test. An α of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis.

• To lower this risk, you must use a lower value for α. However, using a lower value for alpha means that you will be less likely to detect a true difference if one really exists.

# Type I & Type II Error



**Type II error**

- When the null hypothesis is false and you fail to reject it, you make a type II error. The probability of making a type II error is β, which depends on the power of the test. You can decrease your risk of committing a type II error by ensuring your test has enough power.
- You can do this by ensuring your sample size is large enough to detect a practical difference when one truly exists. The probability of rejecting the null hypothesis when it is false is equal to 1–β. This value is the power of the test.

# Type I & Type II Error



| | True condition | |
|---|---|---|
| **Decision resulting from data analysis** | $H_0$ false ("Change has occurred") | $H_0$ true ("No change") |
| Reject $H_0$ ("Change has occurred") | Correct decision ($1-\beta$: Power of the test) | Error (Type I) ($\alpha$) |
| Fail to reject $H_0$ ("No change") | Error (Type II) ($\beta$) | Correct decision |

# Type I & Type II Error



**Examples of type I and type II error**

To understand the interrelationship between type I and type II error, and to determine which error has more severe consequences for your situation, consider the following example.

A medical researcher wants to compare the effectiveness of two medications. The null and alternative hypotheses are:

Null hypothesis (H0): $\mu 1 = \mu 2$

The two medications are equally effective.

Alternative hypothesis (H1): $\mu 1 \neq \mu 2$

The two medications are not equally effective.

# Type I & Type II Error



- A type I error occurs if the researcher rejects the null hypothesis and concludes that the two medications are different when, in fact, they are not.
- If the medications have the same effectiveness, the researcher may not consider this error too severe because the patients still benefit from the same level of effectiveness regardless of which medicine they take.
- However, if a type II error occurs, the researcher fails to reject the null hypothesis when it should be rejected. That is, the researcher concludes that the medications are the same when, in fact, they are different.
- This error is potentially life-threatening if the less-effective medication is sold to the public instead of the more effective one.

# Type I & Type II Error



• As you conduct your hypothesis tests, consider the risks of making type I and type II errors. If the consequences of making one type of error are more severe or costly than making the other type of error, then choose a level of significance and a power for the test that will reflect the relative severity of those consequences.

# Power of Test



What is the power of a test?

- The power of a statistical test gives the likelihood of rejecting the null hypothesis when the null hypothesis is false.
- Just as the significance level (alpha) of a test gives the probability that the null hypothesis will be rejected when it is actually true (a wrong decision), power quantifies the chance that the null hypothesis will be rejected when it is actually false (a correct decision).
- Thus, power is the ability of a test to correctly reject the null hypothesis.

# Power of Test

Why is it important?

- Although you can conduct a hypothesis test without it, calculating the power of a test beforehand will help you ensure that the sample size is large enough for the purpose of the test.
- Otherwise, the test may be inconclusive, leading to wasted resources.
- On rare occasions the power may be calculated after the test is performed, but this is not recommended except to determine an adequate sample size for a follow-up study

# Power of Test



How is it calculated?

- As an example, consider testing whether the average time per week spent watching TV is 4 hours versus the alternative that it is greater than 4 hours.
- We will calculate the power of the test for a specific value under the alternative hypothesis, say, 7 hours:
- The Null Hypothesis is H0: $\mu = 4$ hours
- The Alternative Hypothesis is H1: $\mu = 6$ hours
- Where $\mu$ = the average time per week spent watching TV.

# Power of Test

Under the null hypothesis μ is written as μ0 and under the alternative it is written as μ1. So here μ0 = 4 and μ1 = 6.

Suppose the standard deviation from past data is known to be 2 hours. To find the power of this test for a sample size of 4:

# Power of Test



At the 5% significance level, the decision criterion for the test is to reject H0 if Z > 1.645, where

$$Z = \frac{\bar{X} - \mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{\bar{X} - 4}{\left(2/\sqrt{4}\right)} = \frac{\bar{X} - 4}{1}$$

The 5% critical value from the standard normal distribution is 1.645. Equating the critical Z-value to the calculated Z gives the corresponding (hypothetical) sample mean value:

$$1.645 = \frac{\bar{X} - 4}{1} \Rightarrow \bar{X} = 5.645$$

# Power of Test



• Calculate the Z-statistic assuming the alternative hypothesis is true, i.e., μ1 = 6:

$$Z = \frac{\bar{X} - \mu_1}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{5.645 - 6}{\left(2/\sqrt{4}\right)} = -0.355$$

• P(Z > -0.355) = 0.6387. The power of the test is approximately 64%. In general, tests with 80% power and higher are considered to be statistically powerful.

# Thank You