CSC423: Data Analysis and Regression
CSC 324: Data Analysis & Statistical Software II

Instructor : Nandhini Gulasingam
Office Hours : Mon. 10:00 am– 11:30 am
Office : Lincoln Park
: 990 W. Fullerton Ave., 3rd Floor, Suite 3132
Email : mgulasin@cdm.depaul.edu
Phone : (773) 325-4917
Course Website : http://d2l.depaul.edu

Outline

- Review basic concepts
- Population vs. sample
- Inference for population mean
- Introduction to SAS and lab session

About the Course

- Statistical methods provide formal procedures to make *informed* decisions and predictions using data

- This course will discuss modeling approaches to analyze relationships among several variables of interest and to identify the effect of predictors on a variable of interest.

- Course topics include
  - Inference for a population mean.
  - Comparing two population means.
  - Multiple regression analysis. Model diagnostics
  - Modeling categorical variables
  - Logistic regression
  - ANOVA models

Statistical Software

Access instructions are posted in the syllabus

- SAS for Windows:
  - SAS 9.4 available in the computer labs in all DePaul campuses and for home PC (see course syllabus)
  - Using Virtual Lab – See instructions under D2L – SAS Resources
  - Online resource:
    - http://support.sas.com
    - http://www.ats.ucla.edu/stat/sas/

Textbook, Grading, Course Policy, and Schedule

Textbook:

- ***A Second Course in Statistics: Regression Analysis***, 7th ed., William Mendenhall, Terry L. Sincich, Prentice Hall, 2010 6th edition is ok
- Reading assignments are available under the syllabus
- Additional readings will be posted on D2L under the respective week

Grading:

- Homework and Programming assignments (40%)
- Late in-class Midterm Exam (30%) scheduled in Week 8
- Group project (30%)

Course Policy:

Course policy is listed under the syllabus and is available on D2L

Schedule:

Tentative schedule and due dates are available on the syllabus

Prerequisite Knowledge

- Simple descriptive statistics: mean, standard deviation, median, quartiles.
- Histograms, scatter plots, box plots, Normal distribution
- Inference on average: confidence intervals, hypothesis testing
- Correlation, simple linear regression, least squares estimates

RESOURCES:

- Chapters 1 reviews these concepts
- Online resource at http://onlinestatbook.com/2/

Review Basic Concepts (prerequisite)

Descriptives

- Variety of descriptive statistics - mean, median, mode, skewness, kurtosis, standard deviation, first quartile and third quartile, etc.

- Mean: μ
  - Mean of a set of data is the sum of the data values, divided by the number of data values
  - The mean is commonly known as the average
  - Is a measure of the center of a data set

- Median:
  - The median is one of the three primary ways to find the average of statistical data
  - It is the middle value after sorting the data ascending or descending
  - If odd # of rows → middle value
    if even # of rows → average of 2 middle values

Descriptives

- Mode:
  - The number which appears most often in a set of numbers
- Standard Deviation: σ
  - Standard deviation is a measure of the dispersion of a set of data from its mean
  - Its is calculated as the square root of variance
    - Variance is the average of the squared differences from the mean

$$\sigma = \sqrt{\frac{\Sigma(X - \bar{X})^2}{(n - 1)}}$$

where:

X = each score
$\bar{X}$ = the mean or average
n = the number of values
Σ means we sum across the values

Descriptives

- Quartiles vs Percentile:

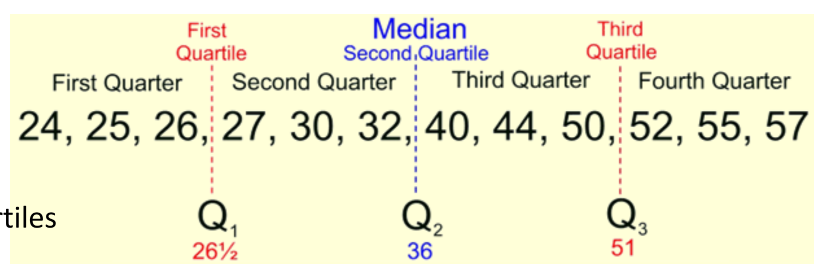
Quantiles

Quartiles divide the set of data into 4 equal parts, so that each part represents ¼ of the dataset (3 quartiles)

Percentiles

Percentiles divide the set of data into 100 equal parts, represented as percentages (0% to 100%)

Quartiles



Lower Quartile

Upper Quartile

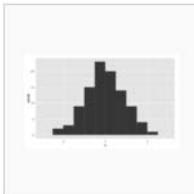
Inter Quartile Range (IQR)

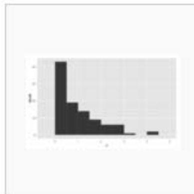
Percentile

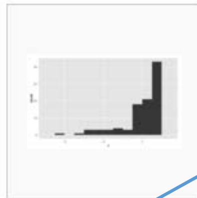
0% 25% 50% 75% 100%

Histogram

- Histogram is a graphical representation of the distribution of numerical data

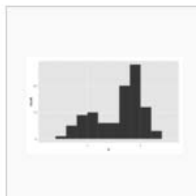

Symmetric, unimodal

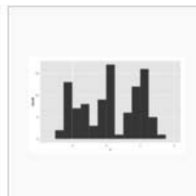

Skewed right


Skewed left

Left tail longer
Left skewed
Negatively skewed
mean < median< mode

Right tail longer
Right skewed
Positively skewed
mean > median> mode

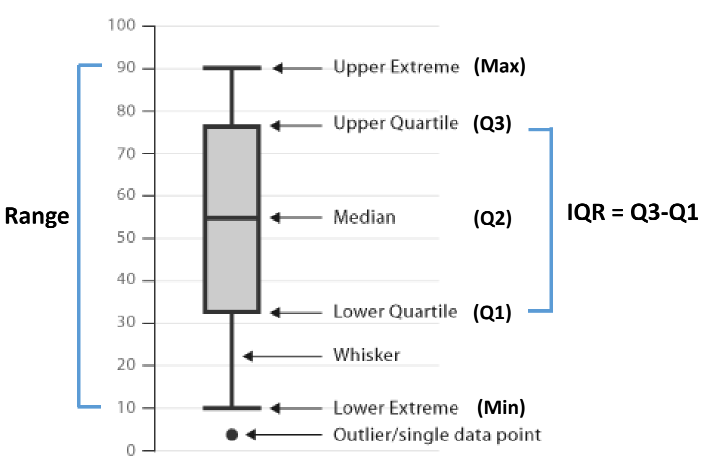

Bimodal

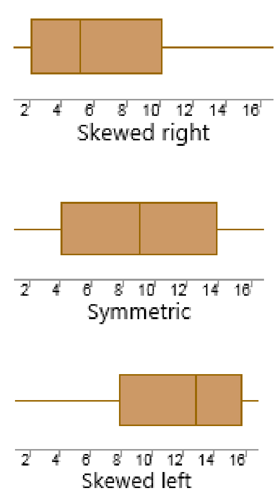

Multimodal


Symmetric

Boxplot

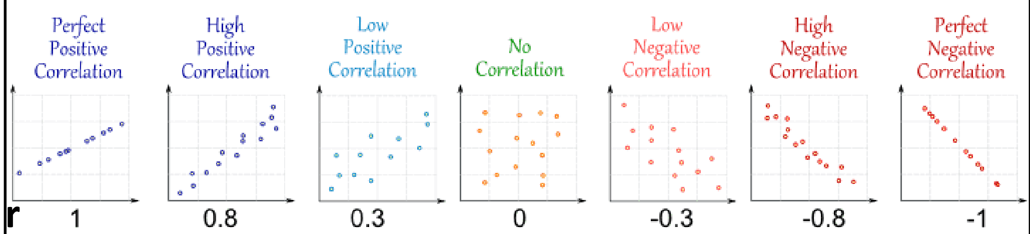
- The box plot (box and whisker plot) is a standardized way of displaying the distribution of data based on the five number summary: minimum, first quartile, median, third quartile, and maximum
- It can also display outliers





Scatter Plot

- Scatter plots show how much one variable is affected by another (i.e. correlation)
- Correlation coefficient “r” ranges from -1 to + 1



When value of one variable increases, the other also increases

When value of one variable increases, the other decreases

Normal Distribution

- The normal distribution is the most widely known and used of all distributions. Because it approximates many natural phenomenon so well, it has developed into a standard of reference for many probability problems
- Normal distribution has a bell-shaped density curve described by its mean and standard deviation
- Notation

$$\mathbf{N(\mu, \sigma^2)}$$

i.e. normally distributed with mean μ (mu) and variance σ^2 (sigma squared)

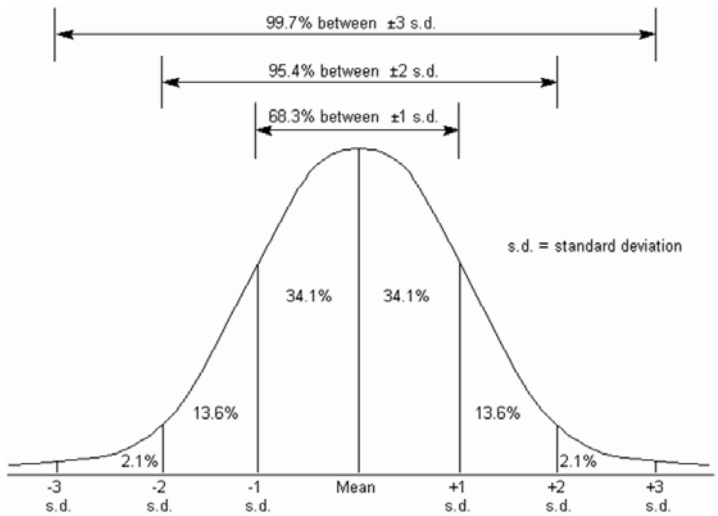
OR

$$\mathbf{X \sim N(\mu, \sigma^2)}$$

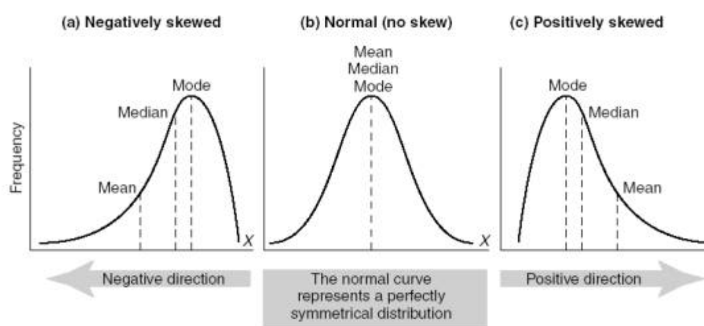
i.e. X is distributed $N(\mu, \sigma^2)$

Normal Distribution

- The curve is symmetrical, centered about its mean, with its spread determined by its standard deviation



Normal Distribution



negatively skewed
mean < median< mode
skewed to the left

Example:
Mean = 2165
Median = 3631

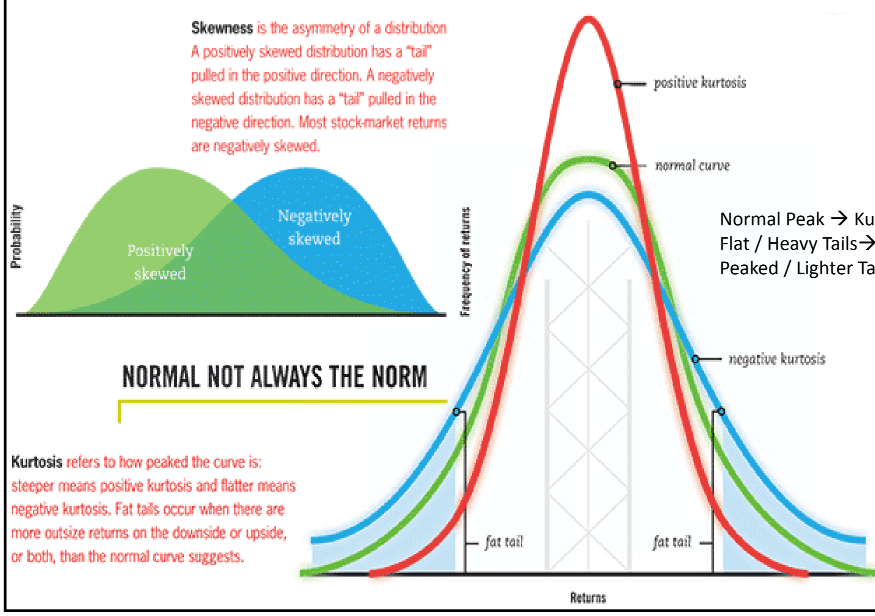
Mean < Median
→ Negative skew, skewed to the Left
→ Most values fall within the higher range

positively skewed
mean > median> mode
skewed to the right

Example:
Mean = 2165
Median = 1631

Mean > Median
→ Positive skew, skewed to the right
→ Most values fall within the lower range

Skewness vs Kurtosis



Normal Peak → Kurtosis = 3
Flat / Heavy Tails→ Kurtosis < 3
Peaked / Lighter Tails → Kurtosis > 3

Inferences about Population

Inferences about Population Central Values

What is the objective of statistics?

ANSWER: to make **inference about a population** based on information observed in a **sample of data**.

However data are often messy and hard to interpret

1. Predictions on returns of future investments
2. Health care data analytics: insurance claims, number of ER visits, patient treatments, etc…How can you use data to improve patient care and satisfaction?
3. Analyze online users' behavior on a certain website
4. Assessment of software reliability based on the number of failures of a certain software
5. Predicting likelihood of economic recovery based on various indicators such as unemployment rates, housing market, consumer confidence, etc…

Goals of Statistical Analyses

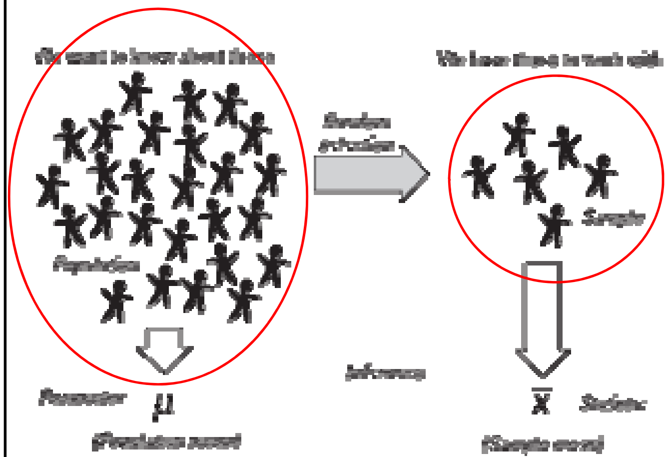
<u>Simple</u> statistical analyses often deal with **<u>one</u>** of these problems:

i. Estimate values of a population of interest

ii. Test an hypothesis

iii. Analyze the association among observed factors

First Ingredient: Observed Sample



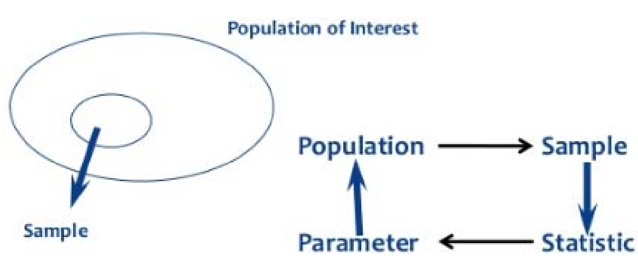
The quality of the statistical analysis depends on the quality of the sample.

If the data sample is not representative, analyzing the data and drawing conclusions will be unproductive-at best.

Random Sampling: every unit in the population has an equal chance to be chosen. It is representative of the population

Definitions

- **Population**: is any large collection of objects or individuals, such as Americans, students, or trees about which information is desired
- **Parameter**: is any summary number, like an average or percentage, that describes the entire population
- **Sample**: is a representative group drawn from the population
- **Statistic**: is any summary number, like an average or percentage, that describes the sample



Population Parameter vs. Sample Statistic

Sample Statistics	Population Parameter
Proportion $\hat{p}$	p
Sample mean $\bar{x}$ $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$	μ
Sample variance s^2 $s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$	σ^2
Sample standard deviation s $s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}}$	σ

In-Class Exercise

Q1 : What is the prevalence of eating hamburgers at DePaul?

Let's say DePaul has a population of approximately 42,000 students. A research question is "what proportion of these students eat hamburgers regularly?" A survey was administered to a sample of 867 DePaul students. Thirty-eight percent (38%) of the sampled students reported that they had hamburgers regularly. How confident can we be that 38% is close to the actual proportion of all DePaul students who ate hamburgers?

- What is the population?
- What is the sample?
- What is the parameter?
- What is the sample proportion?

Population: is any large collection of objects or individuals
Parameter: is any summary describing the entire population
Sample: is a representative group drawn from the population
Statistic: is any summary number describing the sample

In-Class Exercise

Q2 : The International Dairy Foods Association (IDFA) wants to estimate the average amount of calcium male teenagers consume. From a random sample of 50 male teenagers, the IDFA obtained a sample mean of 1081 milligrams of calcium consumed

- What is the population?
- What is the sample?
- What is the parameter?
- What is the sample statistic?

Population: is any large collection of objects or individuals
Parameter: is any summary describing the entire population
Sample: is a representative group drawn from the population
Statistic: is any summary number describing the sample

In-Class Exercise

Q3 : A sociologist wants to know the proportion of adults with children under the age of 18 that eat dinner together 7 nights a week. A simple random sample of 1122 adults with children under the age of 18 was obtained, and 337 of those adults reported eating dinner together with their families 7 nights a week

- What is the parameter?
- What is the sample statistic?

Population: is any large collection of objects or individuals
Parameter: is any summary describing the entire population
Sample: is a representative group drawn from the population
Statistic: is any summary number describing the sample

Inference on Means

Supercomputer Systems Data

A software engineer is trying to optimize system performance, and wants to collect data on the time in milliseconds between requests for a particular process service

- **Step I:** Design the data collection
- **Step II:** "Explore" the data
- **Step III**: Analyze the data

Question

- What is the population? → **time taken for all inter-requests in milliseconds**
- What is the parameter of interest? → **average time taken for all inter-requests in milliseconds**

Step I: Design the Data Collection

1) Sample: Take a **simple random sample** of 100 requests for the particular process service and record the inter-request time values

2) Recorded Data: Record **time in milliseconds**

   e.g : 2,808 4,201 3,848 12,345 31,556 … 4,236 7,432 7,940

Remark: Use a sample that is large enough to see the true nature of any effects, and obtain the sample choosing the *subjects* at random to eliminate unwanted bias

Step II: “Explore” the data

Summarize the data: display plots and compute summary statistics

<u>Descriptive Statistics:</u>

N	Mean	Std. Dev.	First Quartile	Median	Third Quartile
100	1161.63	1142.54	279.6022	736.2358	1737.2834

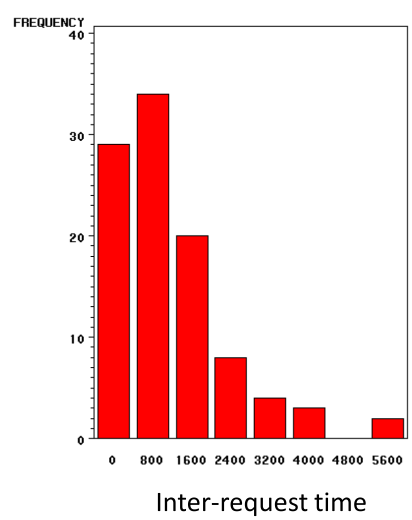
Step II: “Explore” the data

Summarize the data: display plots and compute summary statistics

<u>Visualize Data:</u> Histogram

What can you say about the data?

- Min value?
- Max value?
- Range
- Peak at?
- Outliers?
- Tail?
- Skewness?



Step III: Analyze the Data

Analyze the data and make inferences about the population

What we know…

- In our example we want to estimate the **<u>average time</u>** between requests to the processing system
- The population average μ is estimated using the **sample mean (a.k.a. sample average)**
- The average time for between requests is $\bar{\mathbf{x}}$ =1161.63 for the sample of 100 inter-requests

What we need to estimate…

- How close is this estimate to the actual population value?

→The accuracy of the sample estimate is measured by the

Standard Error and C.I.

Central Limit Theorem

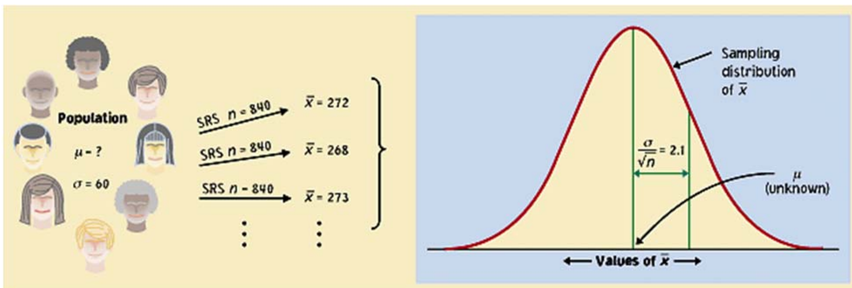
If the sample size n is large, **the sample average is approximately normal** with mean equal to the population mean and standard deviation equal to the standard deviation of the population average

$$\bar{X} \text{ is approximately } N(\mu, \frac{\sigma}{\sqrt{n}})$$

- The sample is a **<u>simple random sample</u>**
- The <u>larger the sample</u>, the more accurate the normal approximation is
- If the distribution of the population is <u>not symmetric</u>, the normal approximation is less accurate, and you need a larger sample

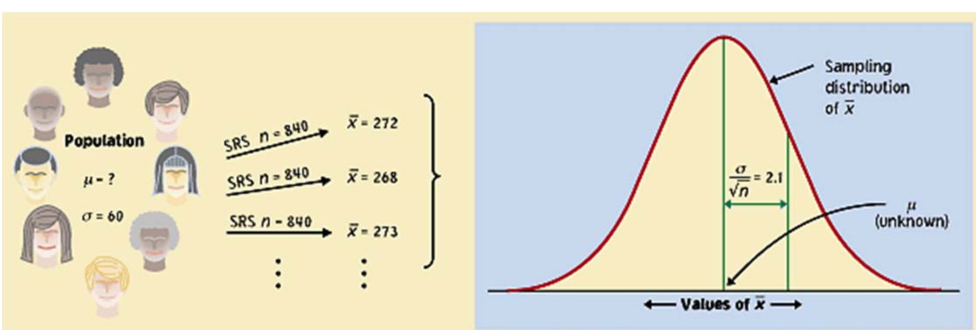
Statistical Confidence

- Although the sample mean, $\bar{x}$, is a unique number for any particular sample, if you pick a different sample you will probably get a different sample mean
- In fact, you could get many different values for the sample mean, and virtually none of them would actually equal the true population mean, μ
- We could then calculate an average of all of our sample means, this mean would equal the true population mean



Standard Deviation vs. Standard Error

- We can also calculate the Standard Deviation of the distribution of sample means
- The Standard Deviation of this distribution of sample means is the Standard Error of each individual sample mean
- Put another way, Standard Error is the Standard Deviation of the population mean



Standard Error

Given a sample of size n, the accuracy of the sample average as an estimate of the population average is measured by the **standard error,** defined as

$$S.E.(\overline{X}) = \frac{s}{\sqrt{n}}$$

Where

s is the sample std. dev. for the observations in the sample

The larger the sample, the more accurate the average is as an estimate of the population average

How accurately does a sample mean estimate the population mean μ?

It depends on the variation of the population of interest and on the sample size

Example

Suppose a bank manager wants to estimate the average call length for a bank service center

In a sample of 80 calls, the sample mean length is 196.6 sec and the sample standard deviation is 184.81

How accurate is the sample estimate?

Answer

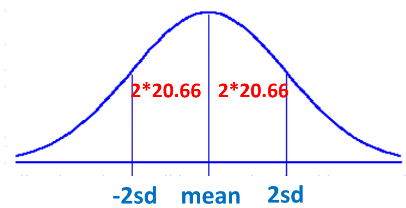
The standard error of the mean call length is computed as

$$S.E.(\overline{X}) = \frac{s}{\sqrt{n}}$$

Sample n = 80
Sample mean length = 196.6 sec
Sample Std. dev. s =184.81

$$S.E(\bar{x}) = 184.81/\sqrt{80} = 20.66$$ ← **1 S.E**

In fact, S.E tells us that we can be 95% confident that our observed sample mean is plus or minus 2 Standard Errors from the population mean



Sample Distribution	196.6-2(20.66) **(155.28)**	196.6 **(196.6)**	196.6 +2(20.66) **(237.92)**

Thus the true mean to be estimated is within **41.32** (i.e. 2* S.E) seconds away from the sample mean of 196.6 seconds

Estimating the Population Average and Confidence Intervals

Estimating the Population Average

- If we want to estimate the population average, we collect observations using a <u>simple random sample</u> and compute the sample average

The sample average is an estimate of the population average

The standard deviation of the sample average measures how accurate the estimate is

Confidence Intervals

- From the normal approximation, there is about 95% chance that the sample average is roughly within two standard deviations from the population average
- We can then construct a confidence interval for the population average
- The C.I. will give us a plausible range of values for the "true" average

Confidence Intervals for Large Samples

- Given a simple random sample of large size (n>50), the confidence interval for the mean μ of the population is computed as:

$$\bar{x} \pm m \qquad m = z^* \frac{s}{\sqrt{n}}$$

- Where $\bar{x}$ is the sample mean, **s** is the sample standard deviation, and **n** is the sample size

- Z* depends on the confidence level
- m is the margin of error that measures the accuracy of the sample estimate

Some Confidence Intervals for the Population Average

$\bar{x}$ is the sample average of n observations in a simple random sample of size n, where n is large (>50)

s is the sample standard deviation of the n observations.

The **confidence level C** says how confident we are that the procedure will "catch" the true population average μ

The 90% C.I. for the population mean: $\bar{x} \pm 1.64 * \frac{s}{\sqrt{n}}$

The 95% C.I. for the population mean: $\bar{x} \pm 1.96 * \frac{s}{\sqrt{n}}$

The 99% C.I. for the population mean: $\bar{x} \pm 2.57 * \frac{s}{\sqrt{n}}$

In-Class Exercise: Confidence Interval Mistakes

A sample of 400 students was asked to evaluate university's counseling services on a 1 to 10 scale. The sample mean was 8.6 with a sample standard deviation of 2.0.

a) An analyst computes the 95% confidence interval as 8.6 +/- 1.96* 2. 0. *What's the mistake?*

b) She corrects her mistake and states that "she is 95% confident that the sample mean falls between 8.404 and 8.796". *What's wrong with the statement?*

c) She corrects her mistake in part b) and states that "there is 95% probability that the true mean is between 8.404 and 8.796". *What's wrong with the statement? How would you correct it?*

General Remarks on C.I.'s

- The purpose of a C.I. is to estimate an unknown parameter with an indication of how accurate the estimate is and of how confident we are that the result is correct

- The methods used here rely on two assumptions:
  - simple random sample and
  - large sample size n

- The confidence level states the probability that the confidence interval contains the "true" value of the parameter

- The margin of error of a confidence interval decreases if either
  - The confidence level is smaller OR
  - The sample size n increases

$$\bar{x} \pm 2.57 * \frac{s}{\sqrt{n}}$$

Example: Server Upgrade – Confidence Interval

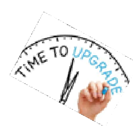
- A computer system goes through an expensive upgrade. To determine whether the new server is faster than the previous one, a certain process that ran for an average of 7.5 minutes on the old server is executed 30 times with the following results:

- **Observed results**: Average time is 6.7 min. with standard deviation of 1.2 min

What is the Confidence Interval for a process running on the new server?

Example: Confidence Interval – Server Upgrade

Compute a 95% confidence interval



- **Problem settings:**
  - Process ran for an average of 7.5 minutes on old system
  - Process is run 30 times on new server

    Observed results: Average time is 6.7 min. with standard deviation of 1.2 min

The 95% confidence interval for the average processing time is computed using the formula

$$\bar{x} \pm 1.96 * \frac{s}{\sqrt{n}}$$

$\bar{x}$ = 6.7
μ = 7.5
s = 1.2
n = 30

computed as 6.7 ± 1.96*1.2/$\sqrt{(30)}$ = (6.27, 7.13) minutes

Conclusion: Probability that the average time the process ran on the new server ranged between 6.27 min and 7.13 min

Hypothesis (Inference) Testing

Hypothesis Testing

A second goal of a statistical analysis is to **verify some claim** about the population on the basis of the data

- **A test of significance** is a procedure to assess the truth about a hypothesis using the observed data
- The results of the test are expressed in terms of a probability that measures how well the data supports the hypothesis

Example: Is the Server Upgrade worth while?

- A computer system goes through an expensive upgrade. To determine whether the new server is faster than the previous one, a certain process that ran for an average of 7.5 minutes on the old server is executed 30 times with the following results:

- **Observed results**: Average time is 6.7 min. with standard deviation of 1.2 min

Is the new server faster?

- Let's use **Significance Testing**

Computing a Test of Significance

Set up hypotheses
- Null hypothesis Ho
- Alternative hypothesis Ha

Compute Test Statistic

Compute p-value under the null hypothesis

State a conclusion

Step-1: Stating an Hypotheses – Server Upgrade

The null hypothesis H_0 expresses the idea that the observed difference is due to chance. It is a statement of "no effect" or "no difference", and is expressed in terms of the population parameter

Let μ denote the "true" execution time on the new server

H_0: μ =7.5min

The **alternative hypothesis H_a** represents the idea that the difference is real

The alternative hypothesis states that the new server execution time is ≠ 7.5 min (i.e. faster or slower) that is:

H_a: μ <7.5min or μ >7.5min

Step-2: Test Statistics and Significance - Server Upgrade

A **test statistic** is used to measure the difference between the data and the null hypothesis

Consider the statistical test

Null hypothesis Ho: μ = 7.5

Alternative hypothesis Ha: μ <7.5 or μ >7.5

The test statistic for **significance tests on averages** is called z-statistic and its general form is

Sample average

Average defined in Ho

$$z = \frac{\overline{x} - \mu}{s / \sqrt{n}}$$

Sample standard deviation

Step-2: Test Statistics and Significance - Server Upgrade

In the example the z-statistic is

$$z = \frac{6.7 - 7.5}{1.2/\sqrt{30}} = -3.65$$

$$z = \frac{\overline{x} - \mu}{s / \sqrt{n}}$$

$\overline{x}$ = 6.7
μ = 7.5
s = 1.2
n = 30

The distribution of the z-statistic is the **t_{n-1} distribution** with n-1 degrees of freedom, where n is the sample size

The t-distribution is symmetric around zero and has tails that are thicker than the standard normal distribution N(0,1)

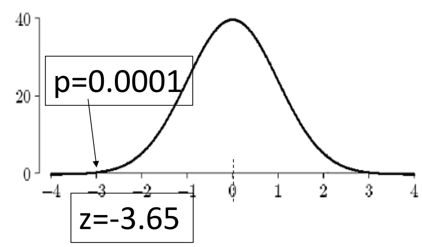
P-value

Confidence Levels

z-score (Standard Deviations)	p-value (Probability)	Confidence level
< -1.65 or > +1.65	< 0.10	90%
< -1.96 or > +1.96	< 0.05	95%
< -2.58 or > +2.58	< 0.01	99%

Step-2: Test Statistics and Significance - Server Upgrade

T-distribution with n-1=29 degrees of freedom



The probability of getting a sample average less than 3.65 S.D.'s below the null hypothesis value is extremely small

H_0: μ =7.5min
H_a: μ <7.5min or μ >7.5min

P-value = <0.01 → The probability that the new server executed the program for 7.5 minutes is very small (~ 0.0001)

Definitions of P-value

- The p-value is the probability of observing the value of the test statistic or a more extreme value than the observed one, assuming that the null hypothesis is true
- Hence a small p-value is strong evidence against the null hypothesis
- If the p-value is small, the null hypothesis does not provide a "good explanation" for the observed data

Significance Levels and P-values

- If **the p-value is small**, then **the null hypothesis should be rejected**
- In common statistical terminology:

✸ If **p-value <** **α=0.05,** the null hypothesis is rejected at 5% significance level. The test result is called "***statistically significant***".

✸ If **p-value <** **α=0.01,** the null hypothesis is rejected at 1% significance level. The test result is called "***highly significant***".

✸ If **p-value > 0.05**, the null hypothesis cannot be rejected.

The test is "***not significant***".

- It is better practice to summarize the test result reporting what test was used, the P-value and whether the test was "statistically significant", "highly significant" or "not significant"

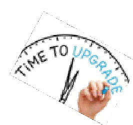
Alpha and P-Value w.r.t Null Hypothesis

P-value > alpha → Null will fly
P-value < alpha → Null must go (reject null hypothesis)

Goal is to reject the null hypothesis and accept the alternative hypothesis

Step-3: Test Statistics and Significance – Conclusion - Server Upgrade

Is the new server faster?



Given:

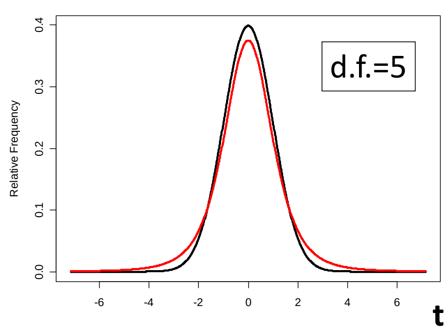
- A computer system goes through an expensive upgrade
- $\bar{x}$ = 6.7 z = -3.65
  μ = 7.5 p-value < 0.01
  s = 1.2
  n = 30
- H_0 : μ =7.5min
  H_a : μ <7.5min or μ >7.5min

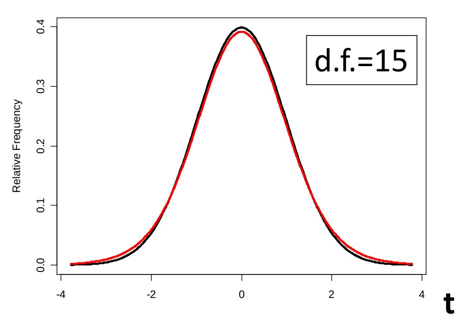
Conclusion: Since p-value is < 0.01, reject H_0 and accept the H_a . i.e. you cannot conclusively say if the server is faster or slower.

True Distribution of Test Statistic

- For small samples ($n<50$), the distribution of the z-statistic is the t_{n-1} distribution with n-1 degrees of freedom, where n is the sample size
- There are many t-distribution curves. Each curve is specified by its ***degrees of freedom***
- In the system upgrade example, we have $n=30$ observations, therefore the degrees of freedom are d.f. = $30–1=29$
- *The p-value is found using a table of values for the student's curves or a statistical package such as SAS*

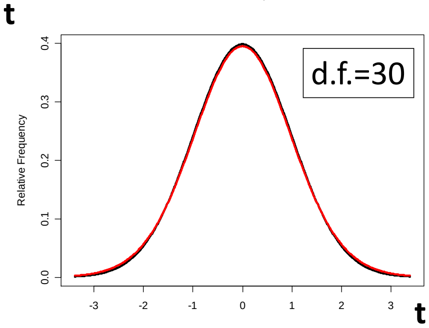
Comparing Student's Curve and Standard Normal Curve





Student's curve
Standard normal curve

Student's curve has "fatter" tails. For d.f. around 30, the student's curve is very similar to the standard normal curve



The t-distribution (a.k.a. Student's T distribution)

What is it?

This distribution was discovered by W. S. Gosset (born on 13 June 1876 in Canterbury, England)

He discovered the *t*-distribution in order to deal with small samples arising in statistical quality control.

The brewery had a policy against employees publishing under their own names, thus he published his research on the *t*-distribution under the pen name "Student".

W.S. Gosset
Chief statistician of Guinness Brewery (Dublin, Ireland)

When to Use t-test

- A test on averages that uses the t-distribution is called t-test
- When should we use it? Each of the following conditions should hold:

1. For computing a statistical test on averages
2. The sample is a simple random sample
3. Data are assumed to come from a symmetric distribution that is not too different from the normal distribution. (Not easy to check, typically true for measurements)
4. For larger samples, the t-test is equivalent to a z-test using the normal distribution

T-tests for a Population Average

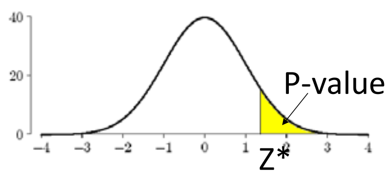
The sample average is $\bar{X}$ and the standard deviation in the sample is *s*

1. Compute the one-sided test: (upper-tailed test)

$$Ho: \mu = \mu_0$$
$$Ha: \mu > \mu_0$$

- The test z-statistic:

$$z^* = \frac{\bar{x} - \mu_0}{S.E.(\bar{x})} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$



The test p-value is equal to the area under the **t-distribution with n-1 degrees of freedom** to the right of z*

T-tests for a Population Average

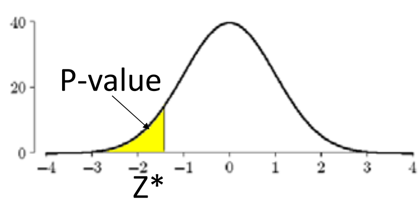
The sample average is $\bar{X}$ and the standard deviation in the sample is *s*

2. Compute the one-sided test: (lower-tailed test)

$$Ho: \mu = \mu_0$$
$$Ha: \mu < \mu_0$$

- The test z-statistic:

$$z^* = \frac{\bar{x} - \mu_0}{S.E.(\bar{x})} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$



The test p-value is equal to the area under **t-distribution with n-1 degrees of freedom** to the left of z*

T-tests for a Population Average

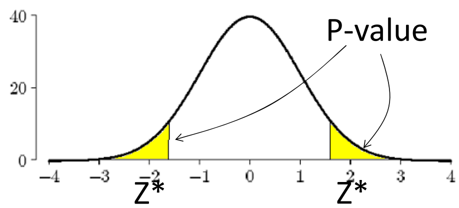
The sample average is $\overline{X}$ and the standard deviation in the sample is *s*

3. Compute the two-sided test: (2-tailed test)

$$Ho: \mu = \mu_0$$
$$Ha: \mu \neq \mu_0$$

- The test z-statistic:

$$z^* = \frac{\overline{x} - \mu_0}{S.E.(\overline{x})} = \frac{\overline{x} - \mu_0}{s / \sqrt{n}}$$



The test p-value is computed as the area under **t-distribution with n-1 degrees of freedom** to the left of -|z*| and to the right of +|z*|

Since the curve is symmetric: P-value=2 P(Z<-|z*|)

Type I Error

- Notice that the significance level α is very popular for reporting the test result
- The significance level α is the so-called Type I error, and represents the probability of incorrectly rejecting Ho when it is true
- However, it is better practice to summarize the test result reporting what test was used, the P-value and whether the test was “statistically significant”, “highly significant” or “not significant”

Summary of Hypothesis Test

- **Set up the null hypothesis** H_0 **and the alternative hypothesis** H_a
  Remember the test is designed to assess the strength against H_0, typically researchers are interested in proving Ha
- **Compute the test statistic value,** to measure the difference between the data and the null hypothesis
- **Compute the P-value.** This is the probability, calculated assuming that H_0 is true, of how strongly the data support H_0
- **State a conclusion.** You could choose a *significance level* α.
  If the P-value is less or equal than α, you conclude that the null hypothesis can be rejected at level α, and Ha is true. Otherwise you conclude that the data do not provide enough evidence to reject H_0