Sampling Distributions

Course Taught at SUAD

Dr. Tanujit Chakraborty

Faculty @ Sorbonne

tanujitisi@gmail.com

Quote of the day..

‘I avoid looking forward or backward, and try to keep looking upward.’

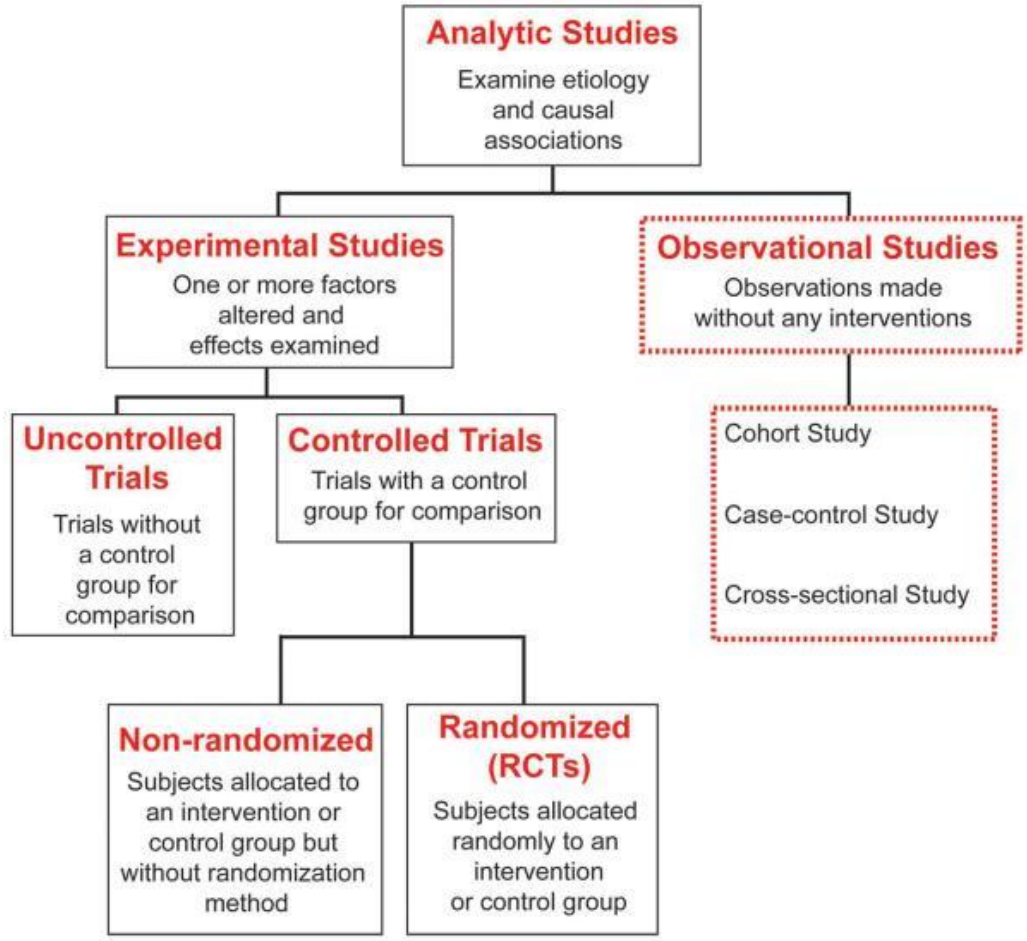
Charlotte Brontë

Today's Topics…

- Survey Sampling
- How to collect data?
- Basic concept of sampling distribution
- Central Limit Theorem
- Application of Central Limit Theorem
- Standard Sampling Distributions
  - Chi-square distribution
  - t distribution
  - F distribution
- Usage of sampling distributions

Introduction

- A primary goal of statistical studies is to collect data that can be used to make informed decisions.
- The ability to make good decisions depends on the quality of the information available.
- The data collection step is critical to obtaining reliable information.
- The conclusions that can be drawn depend on how the data are collected.



Observational Study

- A study is an **observational study** if the investigator observes characteristics of a sample selected from one or more existing populations.
- The goal of an observational study is usually to draw conclusions about the corresponding population or about differences between two or more populations.
- In a well designed observational study, the sample is selected in a way that is designed to produce a sample that is representative of the population.
- Example: an ecologist might be interested in estimating the average shell thickness of bald eagle eggs.
- In an observational study, it is impossible to draw clear cause-and-effect conclusions because we cannot rule out the possibility that the observed effect is due to some variable other than the explanatory variable being studied. Such variables are called confounding variables.

Experimental Study

- A study is an **experiment** if the investigator observes how a response variable behaves when one or more explanatory variables, also called factors, are manipulated.

- The usual goal of an experiment is to determine the effect of the manipulated explanatory variables (factors) on the response variable.

  - **Example:** An educator may wonder what would happen to test scores if the required lab time for a data science course were increased from 3 hours to 6 hours per week. To answer such questions, the researcher conducts an experiment to collect relevant data. The value of some response variable (test score) is recorded under different experimental conditions (3-hour lab and 6-hour lab). In an experiment, the researcher manipulates one or more explanatory variables to create the experimental conditions.

Examples:

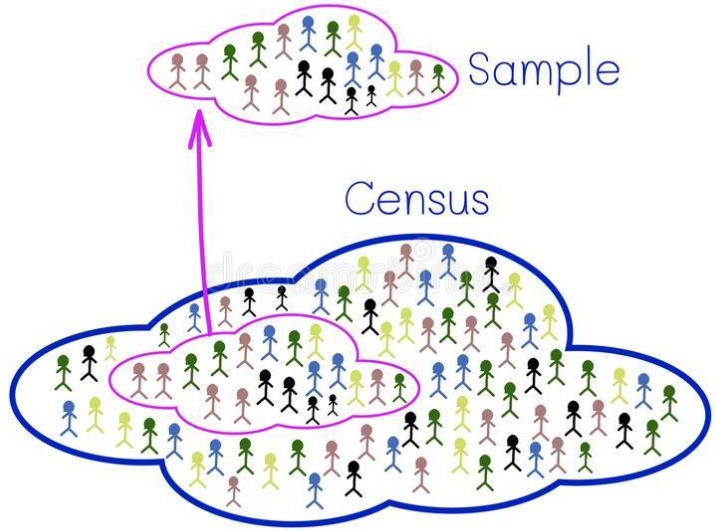
- The article "**Television's Value to Kids: It's All in How They Use It**" (***Seattle Times*, July 6, 2005**) described a study in which researchers analysed standardized test results and television viewing habits of 1700 children. They found that children who averaged more than 2 hours of television viewing per day when they were younger than 3 tended to score lower on measures of reading ability and short-term memory.

a. Is the study described an observational study or an experiment?

b. Is it reasonable to conclude that watching two or more hours of television is the cause of lower reading scores? Explain.

Sampling

- Many studies are conducted in order to generalize from a sample to the corresponding population.
- It is important that the sample be representative of the population.
- We must carefully consider the way in which the sample is selected.
- There are many reasons for selecting a sample rather than obtaining information from an entire population (a **census**).



Bias in Sampling

- **Selection bias** is introduced when the way the sample is selected systematically excludes some part of the population of interest.

Example:

- A researcher may wish to generalize from the results of a study to the population consisting of all residents of a particular city, but the method of selecting individuals may exclude the homeless or those without telephones.
- If those who are excluded from the sampling process differ in some systematic way from those who are included, the sample is virtually guaranteed to be unrepresentative of the population.

- If this difference between the included and the excluded occurs on a variable that is important to the study, conclusions based on the sample data may not be valid for the population of interest.

Bias in Sampling

- Response Bias: Tendency for samples to differ from the corresponding population because the method of observation tends to produce values that differ from the true value.

  Example:

  - In a town, 25% of car accidents among 18-20 year olds were alcohol-related. Do you support lowering the legal drinking age to 18?

- Non-Response Bias: Tendency for samples to differ from the corresponding population because data are not obtained from all individuals selected for inclusion in the sample.

  Example:

  - A polling company is conducting a study in a certain city on people's attitudes to ward their occupation. The company calls random phone numbers each day between the hours of 6.00 pm and 9.00 pm. Those work during the evening hours will be unable to take part in the study.

Random Sampling

- A **simple random sample of size *n*** is a sample that is selected from a population in a way that ensures that every different possible sample of the desired size has the same chance of being selected.

- **Simple random sampling methods**: The following steps are involved in selecting simple random sampling:

  - A list of all the members of the population is prepared initially and then each member is marked with a specific number (for example, there are nth members then they will be numbered from 1 to N).

  - From this population, random samples are chosen using two ways: random number tables and random number generator software. A random number generator software is preferred more as the sample numbers can be generated randomly without human interference.

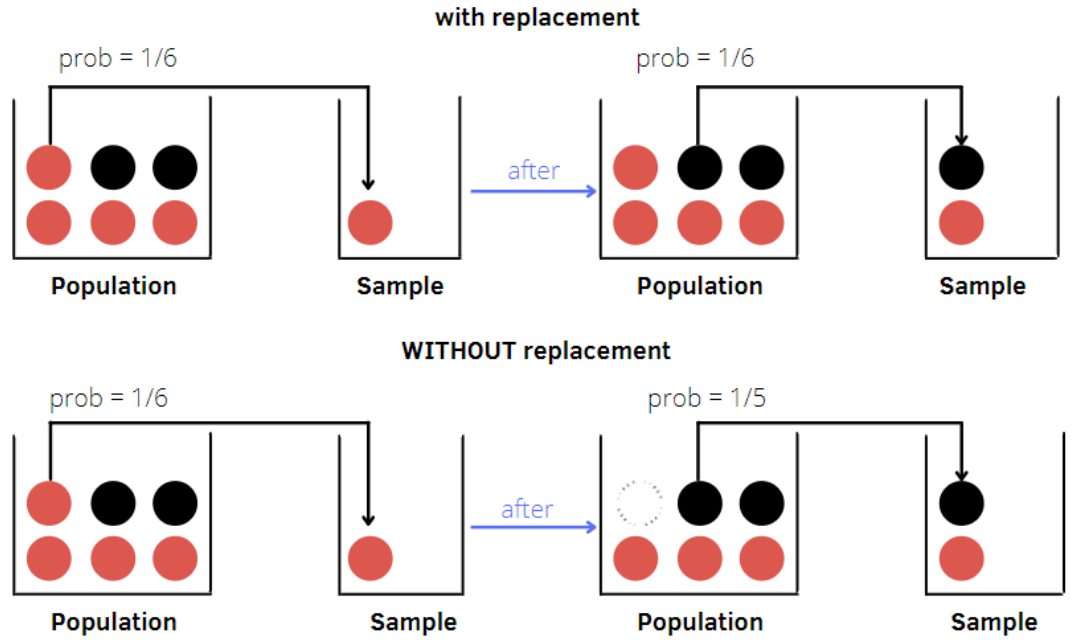
Random Sampling: Example

- Suppose a list containing the names of the 427 customers who purchased a new car during 2022 at a large dealership is available.
- The owner of the dealership wants to interview a sample of these customers to learn about customer satisfaction.
- She plans to select a simple random sample of 20 customers.
- Because it would be tedious to write all 427 names on slips of paper, random numbers can be used to select the sample.
- To do this, we can use three-digit numbers, starting with 001 and ending with 427, to represent the individuals on the list.

Random Sampling

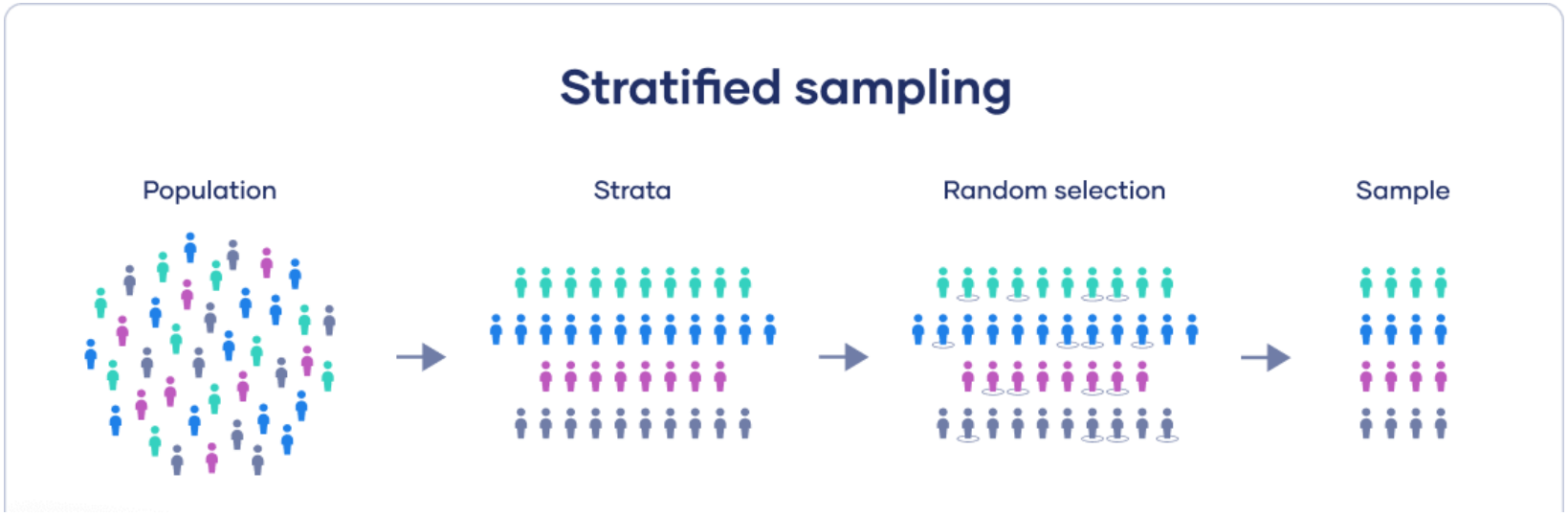
- **Sampling without replacement:** Once an individual from the population is selected for inclusion in the sample, it may not be selected again in the sampling process. A sample selected without replacement includes *n* distinct individuals from the population.

- **Sampling with replacement:** After an individual from the population is selected for inclusion in the sample and the corresponding data are recorded, the individual is placed back in the population and can be selected again in the sampling process. A sample selected with replacement might include any particular individual from the population more than once.



Stratified Sampling

- When the entire population can be divided into a set of non-overlapping subgroups, a method known as **stratified sampling** often proves easier to implement and more cost-effective than simple random sampling.
- In stratified random sampling, separate simple random samples are independently selected from each subgroup.
- the subgroups are called **strata** and each individual subgroup is called a **stratum.**



Stratified Sampling: Example

- To estimate the average cost of malpractice insurance, a researcher might find it convenient to view the population of all doctors practicing in a particular metropolitan area as being made up of four subpopulations: (1) surgeons, (2) family practitioners, (3) obstetricians, and (4) a group that includes all other areas of specialization.

- Rather than taking a random simple sample from the population of all doctors, the researcher could take four separate simple random samples — one from the group of surgeons, another from the family practitioners, and so on.

- These four samples would provide information about the four subgroups as well as information about the overall population of doctors.

Examples

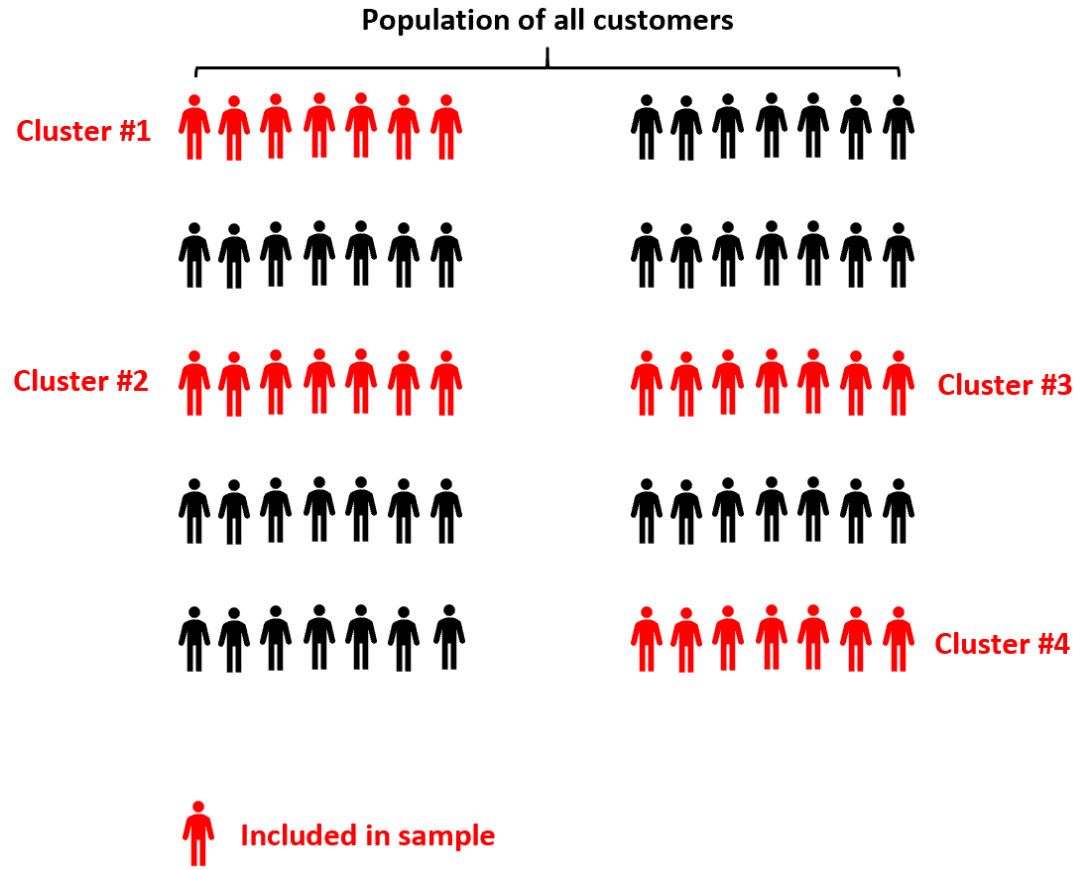
- During the previous calendar year, a county's small claims court processed 870 cases. Describe how a simple random sample of size $n = 20$ might be selected from the case files to obtain information regarding the average award in such cases.

- Suppose that you were asked to help design a survey of adult city residents in order to estimate the proportion that would support a sales tax increase. The plan is to use a stratified random sample, and three stratification schemes have been proposed.
  - Scheme 1: Stratify adult residents into four strata based on the first letter of their last name (A–G, H–N, O–T, U–Z).
  - Scheme 2: Stratify adult residents into three strata: college students, nonstudents who work full time, nonstudents who do not work full time.
  - Scheme 3: Stratify adult residents into five strata by randomly assigning residents into one of the five strata.

Which of the three stratification schemes would be best in this situation?

Cluster Sampling

- Sometimes it is easier to select groups of individuals from a population than it is to select individuals themselves.
- **Cluster sampling** involves dividing the population of interest into non-overlapping subgroups, called **clusters**.
- Clusters are then selected at random, and then *all* individuals in the selected clusters are included in the sample.

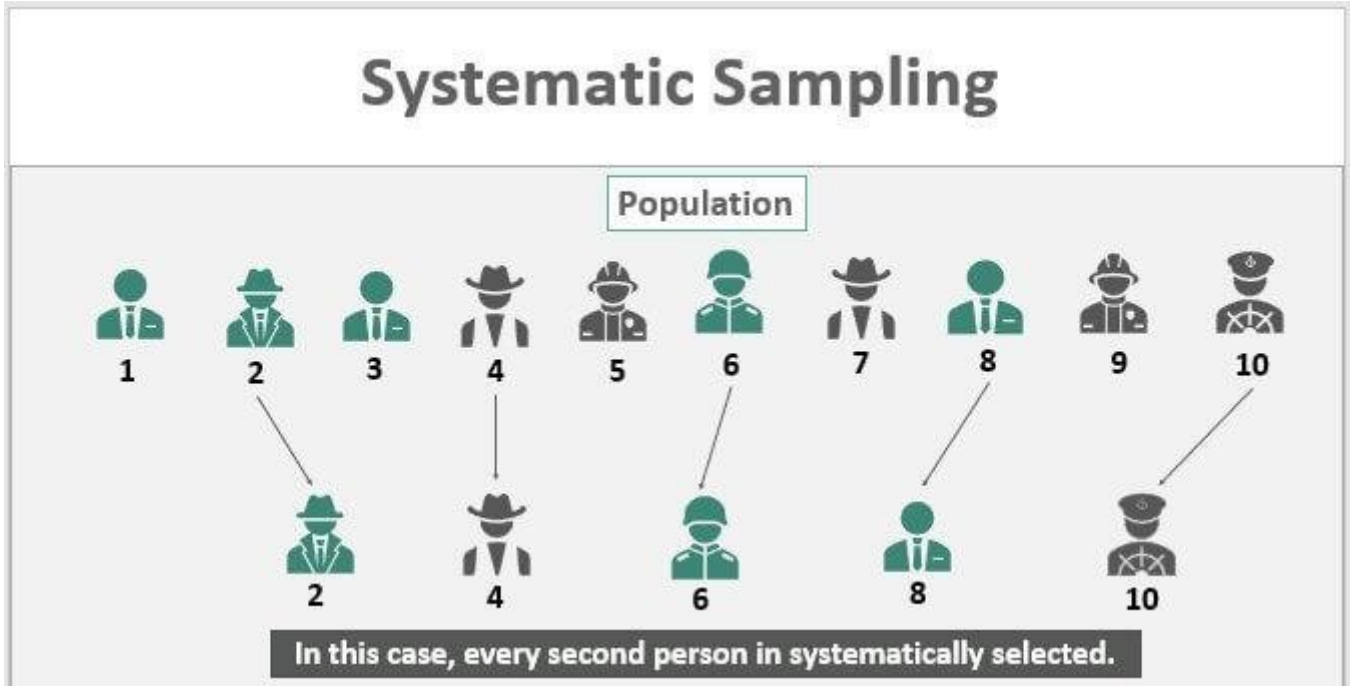


Cluster Sampling: Example

- Suppose that a large urban high school has 600 senior students, all of whom are enrolled in a first period homeroom.
- There are 24 senior homerooms, each with approximately 25 students.
- If school administrators wanted to select a sample of roughly 75 seniors to participate in an evaluation of the college and career placement advising available to students, they might find it much easier to select three of the senior homerooms at random and then include all the students in the selected homerooms in the sample.
- In this way, an evaluation survey could be administered to all students in the selected homerooms at the same time—certainly easier logistically than randomly selecting 75 students and then administering the survey to those individual seniors.

Systematic Sampling

- **Systematic sampling** is a procedure that can be used when it is possible to view the population of interest as consisting of a list or some other sequential arrangement.
- A value *k* is specified (for example, *k*=50 or *k*=200).
- Then one of the first *k* individuals is selected at random, after which every *k*th individual in the sequence is included in the sample.
- A sample selected in this way is called a **1 in *k* systematic sample**.



Exercise

- A sample of pages from a book is to be obtained, and the number of words on each selected page will be determined. For the purposes of this exercise, equations are not counted as words and a number is counted as a word only if it is spelled out—that is, *ten* is counted as a word, but *10* is not.

a. Describe a sampling procedure that would result in a simple random sample of pages from this book.

b. Describe a sampling procedure that would result in a stratified random sample. Explain why you chose the specific strata used in your sampling plan.

c. Describe a sampling procedure that would result in a systematic sample.

d. Describe a sampling procedure that would result in a cluster sample.

e. Using the process you gave in Part (a), select a simple random sample of at least 20 pages, and record the number of words on each of the selected pages. Construct a dotplot of the resulting sample values, and write a sentence or two commenting on what it reveals about the number of words on a page.

f. Using the process you gave in Part (b), select a stratified random sample that includes a total of at least 20 selected pages, and record the number of words on each of the selected pages. Construct a dotplot of the resulting sample values, and write a sentence or two commenting on what it reveals about the number of words on a page.

Simple Comparative Experiments

- An **experiment** is a study in which one or more explanatory variables are manipulated in order to observe the effect on a response variable.

- The **explanatory variables** are those variables that have values that are controlled by the experimenter. Explanatory variables are also called **factors.**

- The **response variable** is a variable that is not controlled by the experimenter and that is measured as part of the experiment.

- An **experimental condition** is any particular combination of values for the explanatory variables. Experimental conditions are also called **treatments.**

Simple Comparative Experiments

- An **extraneous variable** is one that is not one of the explanatory variables in the study but is thought to affect the response variable.
- Two variables are **confounded** if their effects on the response variable cannot be distinguished from one another.
- **Direct Control:** Holding extraneous variables constant so that their effects are not confounded with those of the experimental conditions (treatments).
- **Random Assignment** to ensure that the experiment does not systematically favour one experimental condition (treatment) over another.
- **Blocking:** Using extraneous variables to create groups (blocks) that are similar. All experimental conditions (treatments) are then tried in each block.
- **Replication:** Ensuring that there is an adequate number of observations for each experimental condition.
- **Experimental Units:** An experimental unit is the smallest unit to which a treatment is applied.

More on Experimental Design

- If the purpose of an experiment is to determine whether some treatment has an effect, it is important to include an experimental group that does not receive the treatment. Such a group is called a **control group**.
  - Not all experiments require the use of a control group.
- A **placebo** is something that is identical (in appearance, taste, feel, etc.) to the treatment received by the treatment group, except that it contains no active ingredients.
  - The placebo group would provide a better basis for comparison and would allow the researchers to determine whether the treatment had any real effect over and above the "placebo effect."

Examples

- The head of the quality control department at a printing company would like to carry out an experiment to determine which of three different glues results in the greatest binding strength. Although they are not of interest in the current investigation, other factors thought to affect binding strength are the number of pages in the book and whether the book is being bound as a paperback or a hardback.

a. What is the response variable in this experiment?

b. What explanatory variable will determine the experimental conditions?

c. What two extraneous variables are mentioned in the problem description? Are there other extraneous variables that should be considered?

Examples

- Red wine contains flavonol, an antioxidant thought to have beneficial health effects. But to have an effect, the antioxidant must be absorbed into the blood. The article "**Red Wine is a Poor Source of Bioavailable Flavonols in Men**" (***The Journal of Nutrition*** **[2001]: 745–748)** describes a study to investigate three sources of dietary flavonol—red wine, yellow onions, and black tea—to determine the effect of source on absorption. The article included the following statement:

- We recruited subjects via posters and local newspapers. To ensure that subjects could tolerate the alcohol in the wine, we only allowed men with a consumption of at least seven drinks per week to participate … Throughout the study, the subjects consumed a diet that was low in flavonols.

  a. What are the three treatments in this experiment?

  b. What is the response variable?

  c. What are three extraneous variables that the researchers chose to control in the experiment?

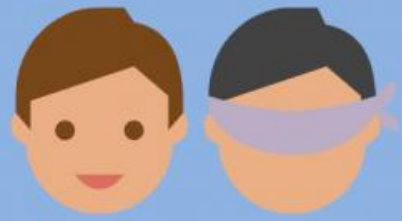
Single-Blind and Double-Blind Experiments

- A **single-blind** experiment is one in which the subjects do not know which treatment was received but the individuals measuring the response do know which treatment was received, or one in which the subjects do know which treatment was received but the individuals measuring the response do not know which treatment was received.

- A **double-blind** experiment is one in which neither the subjects nor the individuals who measure the response know which treatment was received.

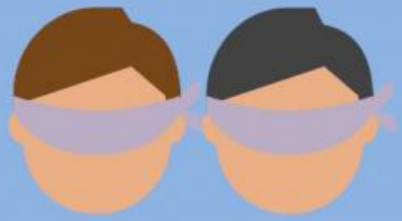
Single Blind vs Double Blind Study

Single Blind



Researcher knows treatment groups.

Double Blind



Neither researcher nor participants know treatment groups.

sciencenotes.org

Sampling Distributions

Random Sampling:

- The outcome of a statistical experiment may be recorded either as a numerical value or as a descriptive representation.

- Here we focus on sampling from distributions or populations and study such important quantities as the *sample mean* and *sample variance*.

Population

- A **population** consists of the totality of the observations with which we are concerned.
- The number of observations in the population is defined to be the size of the population.
- Each observation in a population is a value of a random variable X having some probability distribution $f(x)$.
- Hence, the mean and variance of a random variable or probability distribution are also referred to as the mean and variance of the corresponding population.

Sample

- In the field of statistical inference, statisticians are interested in arriving at conclusions concerning a population when it is impossible to observe the entire set of observations that make up the population.

- This brings us to consider the notion of sampling.

- A **sample** is a subset of a population.

- All too often we are tempted to choose a sample by selecting the most convenient members of the population.

- Any sampling procedure that produces inferences that consistently overestimate or consistently underestimate some characteristic of the population is said to be **biased**.

Random Sampling

- To eliminate any possibility of bias in the sampling procedure, it is desirable to choose a **random sample** in the sense that the observations are made independently and at random.
- In selecting a random sample of size n from a population $f(x)$, let us define the random variable $X_i, i = 1, 2, \ldots, n$, to represent the ith measurement or sample value that we observe.
- The random variables $X_1, X_2, \ldots, X_n$ will then constitute a random sample from the population $f(x)$ with numerical values $x_1, x_2, \ldots, x_n$ if the measurements are obtained by repeating the experiment n independent times under essentially the same conditions.
- Because of the identical conditions under which the elements of the sample are selected, it is reasonable to assume that the n random variables $X_1, X_2, \ldots, X_n$ are independent and that each has the same probability distribution $f(x)$.
- The probability distributions of $X_1, X_2, \ldots, X_n$ are, respectively, $f(x_1), f(x_2), \ldots, f(x_n)$, and their joint probability distribution is $f(x_1, x_2, \ldots, x_n) = f(x_1) f(x_2) \cdots f(x_n)$.

Some important Statistic(s)

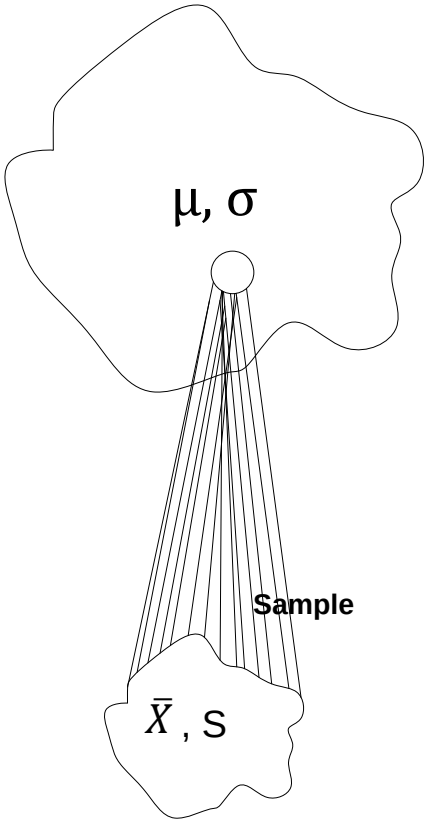
- Our main purpose in selecting random samples is to elicit information about the unknown population parameters.

- Any function of the random variables constituting a random sample is called a **statistic**.

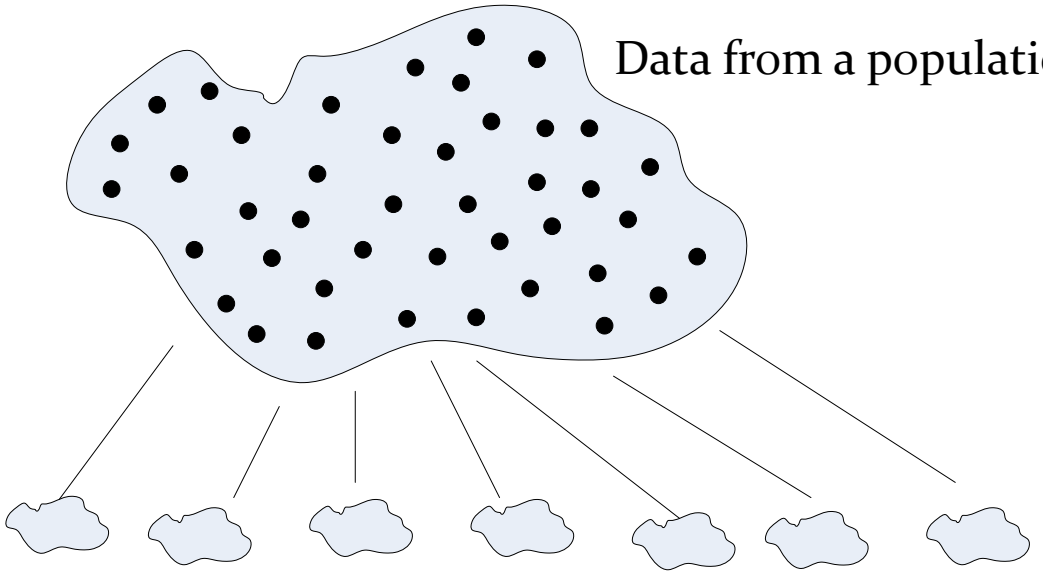
- Let $X_1, X_2, \ldots, X_n$ represent *n* random variables.
  - Sample Mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Note that the statistic $\bar{X}$ assumes the value $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, when X_1 assumes the value x_1, X_2 assumes the value x_2, and so forth. The term *sample mean* is applied to both the statistic $\bar{X}$ and its computed value $\bar{x}$.
  - Sample Variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. The computed value of S^2 for a given sample is denoted by s^2.

Introduction to Statistical Inference

The primary objective of statistical analysis is to use data from a sample to make inferences about the population statistics from which the sample was drawn.

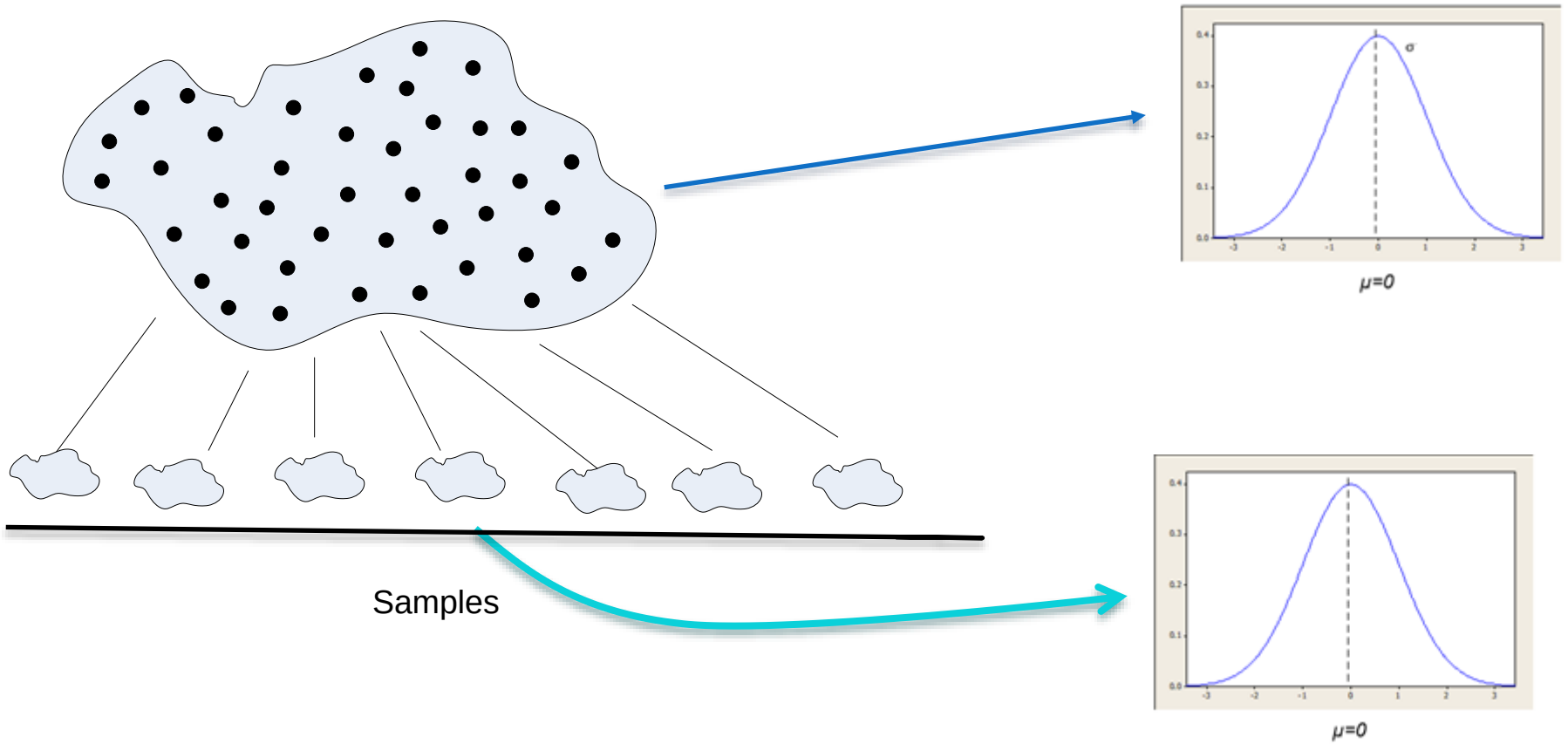


This lecture aims to learn the basic procedures for making such inferences.



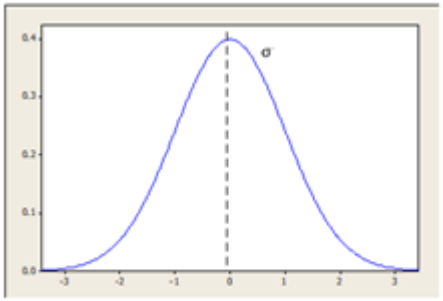
… Starting point

- Distribution of a sample's data



… Starting point

- Central Limit theorem…(Distribution of samples' statistics).

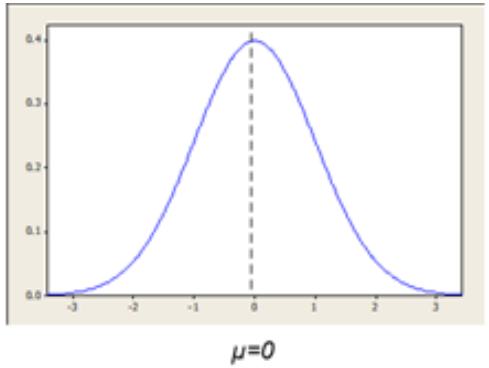


Population's statistics

$f(z: 0, \sigma)$

z-estimation $z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$

Samples' statistics

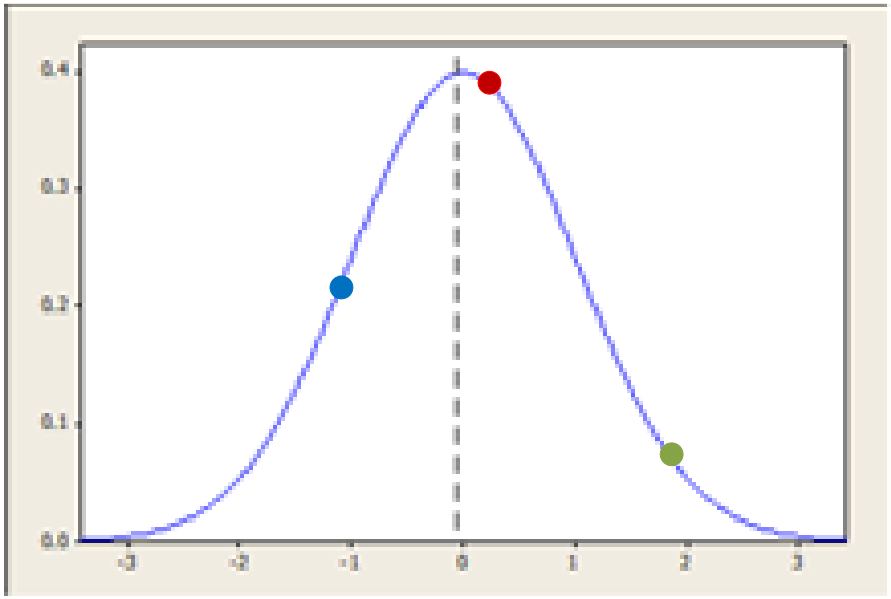


$f(z: 0, \sigma/\sqrt{n})$

… Starting point

- The interpretation of z estimation

Samples



$\mu=0$

Sample statistics

Sampling Distribution

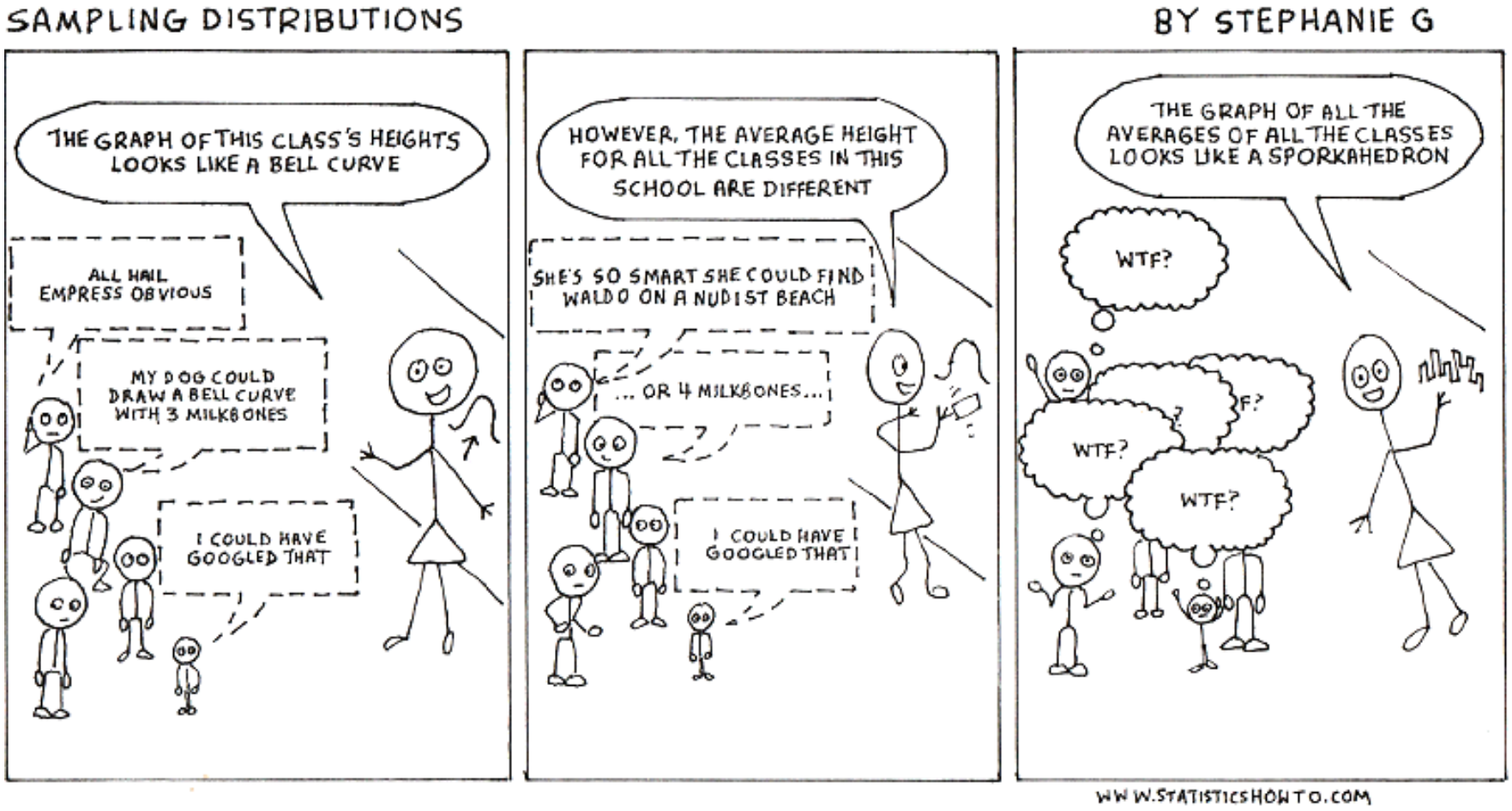
More precisely, sampling distributions are probability distributions and used to describe the variability of sample statistic.

Definition : Sampling distribution

The sampling distribution of a statistic is the probability distribution of that statistic.

- The probability distribution of sample mean (hereafter, will be denoted as $\bar{X}$) is called the sampling distribution of the mean (also, referred to as the distribution of sample mean).
- Like $\bar{X}$, we call sampling distribution of variance (denoted as S^2).
- Using the values of $\bar{X}$ and S^2 for different random samples of a population, we are to make inference on the parameters μ and σ^2 (of the population).

Sampling Distribution



https://www.statisticshowto.com/probability-and-statistics/sampling-in-statistics/sampling-distribution/

Sampling Distribution

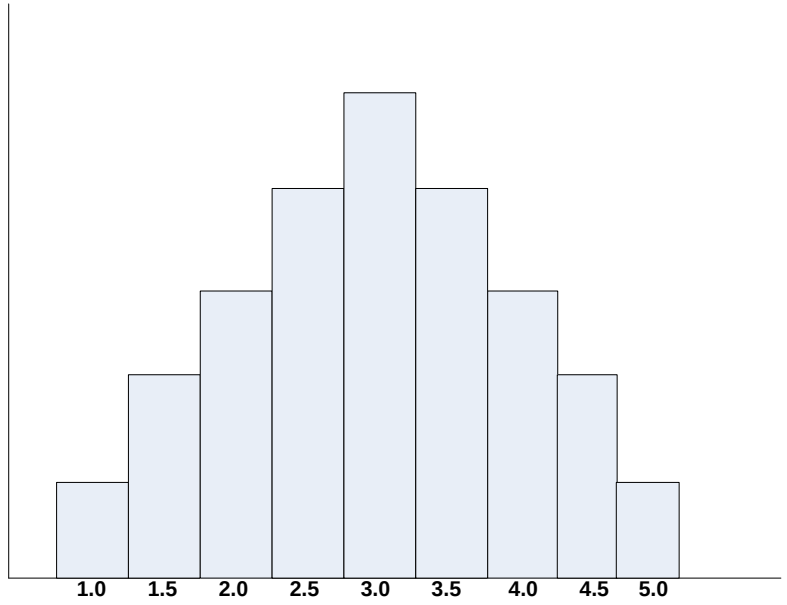
Example 1: Consider five identical balls numbered and weighting as 1, 2, 3, 4 and 5. Consider an experiment consisting of drawing two balls, replacing the first before drawing the second, and then computing the mean of the values of the two balls. Following table lists all possible samples and their mean.

Sample (X)	Mean ($\bar{X}$)	Sample (X)	Mean ($\bar{X}$)	Sample (X)	Mean ($\bar{X}$)
[1,1]	1.0	[2,4]	3.0	[4,2]	3.0
[1,2]	1.5	[2,5]	3.5	[4,3]	3.5
[1,3]	2.0	[3,1]	2.0	[4,4]	4.0
[1,4]	2.5	[3,2]	2.5	[4,5]	4.5
[1,5]	3.0	[3,3]	3.0	[5,1]	3.0
[2,1]	1.5	[3,4]	3.5	[5,2]	3.5
[2,2]	2.0	[3,5]	4.0	[5,3]	4.0
[2,3]	2.5	[4,1]	2.5	[5,4]	4.5
				[5,5]	5.0

Sampling Distribution

Sampling distribution of means

$\bar{X}$	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
$f(\bar{X})$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{4}{25}$	$\frac{5}{25}$	$\frac{4}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	$\frac{1}{25}$



Issues with Sampling Distribution

1. In practical situation, for a large population, it is infeasible to have all possible samples and hence probability distribution of sample statistics.

2. The sampling distribution of a statistic depends on

   - the size of the population
   - the size of the samples and
   - the method of choosing the samples.

Theorem on Sampling Distribution

Theorem 1: **Sampling distribution of mean and variance**

The sampling distribution of a random sample of size n drawn from a population with mean μ and variance σ^2 will have mean $\bar{X} = \mu$ and variance $\frac{\sigma^2}{n} = V(\bar{X})$.

Example 2: With reference to data in Example 1

For the population, $\mu = \frac{1+2+3+4+5}{5} = 3$

$$\sigma^2 = \frac{(25-1)}{12} = 2$$

Applying the theorem, we have $\bar{X} = 3$ *and* $\mathrm{V}(\bar{X}) = 1$.

Hence, the theorem is verified!

Central Limit Theorem

Theorem 1 is an amazing result and in fact, also verified that if we sampling from a population with unknown distribution, the sampling distribution of $\bar{X}$ will still be approximately normal with mean μ and variance $\frac{\sigma^2}{n}$ provided that the sample size is large.

This further, can be established with the famous "central limit theorem", which is stated below.

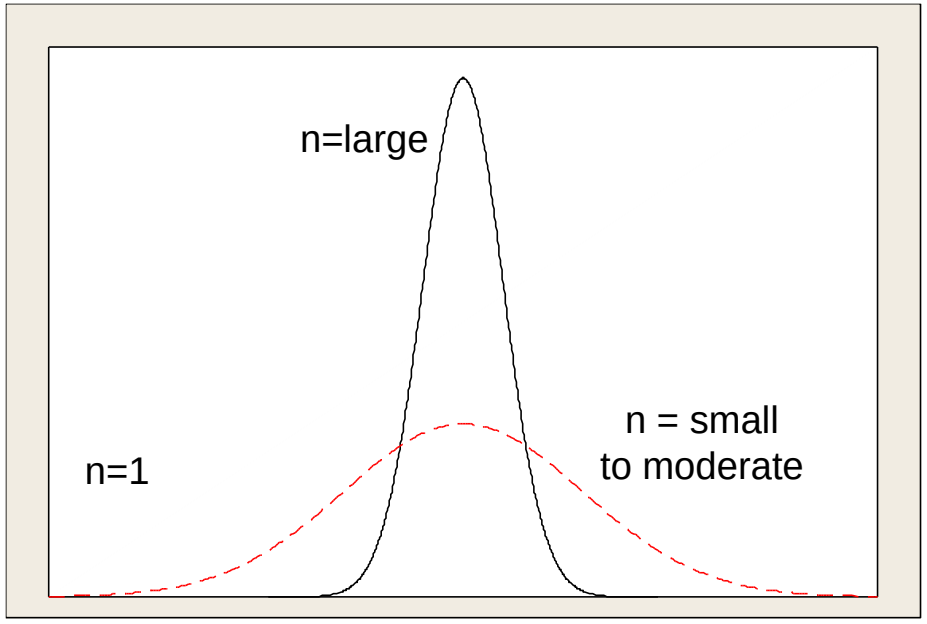
Theorem 2: Central Limit Theorem

If random samples each of size n are taken from any distribution with mean μ and variance σ^2, the sample mean $\bar{X}$ will have a distribution approximately normal with mean μ and variance $\frac{\sigma^2}{n}$; ***i.e., $E(\bar{X}) = \mu$ and $V(\bar{X}) = \frac{\sigma^2}{n}$.***

The approximation becomes better as n increases.

Applicability of Central Limit Theorem

- The normal approximation of $\overline{X}$ will generally be good if $n \geq 30$
- The sample size $n = 30$ is, hence, a guideline for the central limit theorem.
- The normality on the distribution of $\overline{X}$ becomes more accurate as n grows larger.



- One very important application of the **Central Limit Theorem** is the determination of reasonable values of the population mean μ and variance σ^2.

- For standard normal distribution, we have the z-transformation

$$Z = \frac{\overline{X} - \mu}{S} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Usefulness of the sampling distribution

- The mean of the sampling distribution of the mean is the population mean.
  - This implies that "on the average" the sample mean is the same as the population mean.
  - We therefore say that the sample mean is an unbiased estimate of the population mean.
- The variance of the distribution of the sample means is σ^2/n.
  - The standard deviation of the sampling distribution (i.e., $\frac{\sigma}{\sqrt{n}}$) of the mean, often called the standard error of the mean.
    - If σ is high then the sample are not reliable, for a very large sample size ($n \to \infty$), standard error tends to zero

Applicability of Central Limit Theorem

- One very important application of the **Central Limit Theorem** is the determination of reasonable values of the population mean μ and variance σ^2 having a sample, that is, a subset of a population.

- One very important deduction

For standard normal distribution, we have the z-transformation

$$z = \frac{x-\mu}{\sigma}$$ (discussed earlier)

Thus, for a sample statistics

$$Z = \frac{\overline{X}-\mu}{S} = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$$

Applicability of Central Limit Theorem

Example:

- A quiz test for the course CS61061 was conducted and it is found that mean of the scores $\mu = 90$ with standard deviation $\sigma = 20$.
- Now, all students enrolled in the course are randomly assigned to various sections of 100 students in each. A section (X) was checked, and the mean score was found as $\bar{X} = 86$.

- **What is the standard error rate?**

  The standard error rate (Central Limit Theorem) = $\frac{\sigma}{\sqrt{n}} = \frac{20}{100} = 2.0$

- **What is the probability of getting a mean of 86 or lower on the quiz test?**

  For standard normal distribution, we have the z-transformation

$$z = \frac{x-\mu}{\sigma}$$

  Thus, for a sample statistics

$$Z = \frac{\bar{X}-\mu}{S} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{86-90}{2} = -2. \qquad P(Z<-2)?$$

Exercise:

An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

Sampling Distribution of the Difference between Two Means

- Suppose that we have two populations, the first with mean μ_1 and variance σ_1^2, and the second with mean μ_2 and variance σ_2^2.

- Let the statistic $\bar{X}_1$ represent the mean of a random sample of size n_1 selected from the first population, and the statistic $\bar{X}_2$ represent the mean of a random sample of size n_2 selected from the second population, independent of the sample from the first population.

Sampling Distribution of the Difference between Two Means

Distribution of $\bar{X}_1 - \bar{X}_2$:

If independent samples of size n_1 and n_2 are drawn at random from two populations, discrete or continuous, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, then the sampling distribution of the differences of means, $\bar{X}_1 - \bar{X}_2$, is approximately normal distributed with mean and variance given by

$$\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2 \text{ and } \sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Hence,

$$Z = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

is approximately a standard normal variable.

Sampling Distribution of the Difference between Two Means

- If both n_1 and n_2 are greater than or equal to 30, the normal approximation for the distribution of $\bar{X}_1 - \bar{X}_2$ is very good when the underlying distributions are not too far away from normal.

- When n_1 and n_2 are less than 30, the normal approximation is reasonably good except when the populations are decidedly nonnormal.

- If both populations are normal, then $\bar{X}_1 - \bar{X}_2$ has a normal distribution no matter what the sizes of n_1 and n_2 are.

Exercise

Two independent experiments are run in which two different types of paint are compared. Eighteen specimens are painted using type *A*, and the drying time, in hours, is recorded for each. The same is done with type *B*. The population standard deviations are both known to be 1.0. Assuming that the mean drying time is equal for the two types of paint, find $P(\bar{X}_A - \bar{X}_B > 1.0)$, where $\bar{X}_A \; and \; \bar{X}_B$ are average drying times for samples of size $n_A = n_B = 18$.

Sampling Distribution of S^2

- If an engineer is interested in the population mean resistance of a certain type of resistor, the sampling distribution of $\bar{X}$ will be exploited once the sample information is gathered.

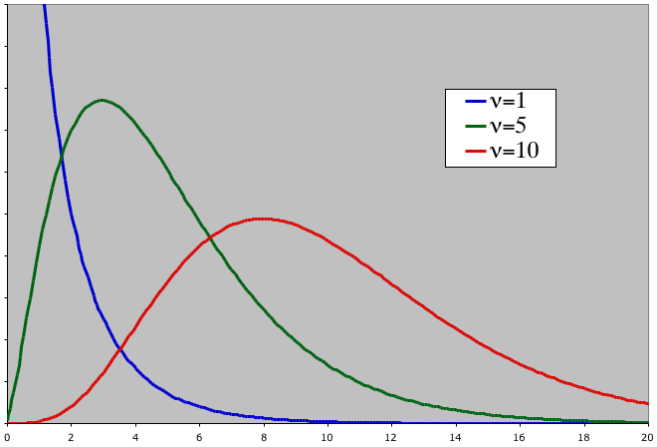
- On the other hand, if the variability in resistance is to be studied, clearly the sampling distribution of S^2 will be used in learning about the parametric counterpart, the population variance σ^2.

Standard Sampling Distributions

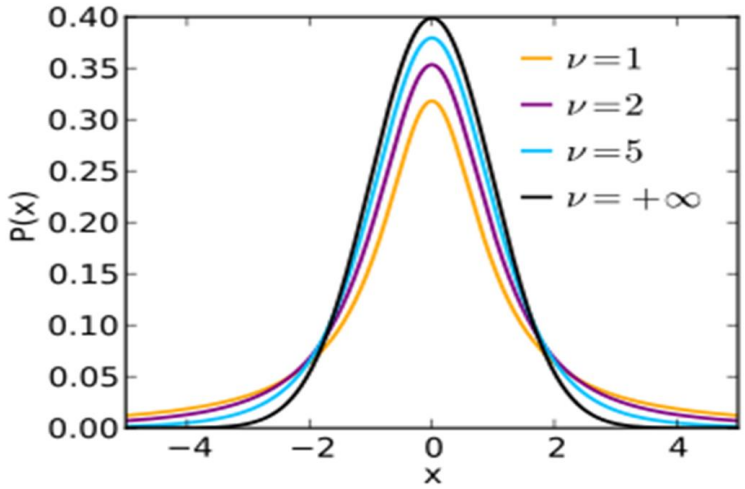
- Apart from the standard normal distribution to describe sampling distribution, there are some other quite different sampling, which are extensively referred in the study of statistical inference.

  - χ^2: Describes the distribution of variance.
  - t: Describes the distribution of normally distributed random variable standardized by an estimate of the standard deviation.
  - F: Describes the distribution of the ratio of two variables.

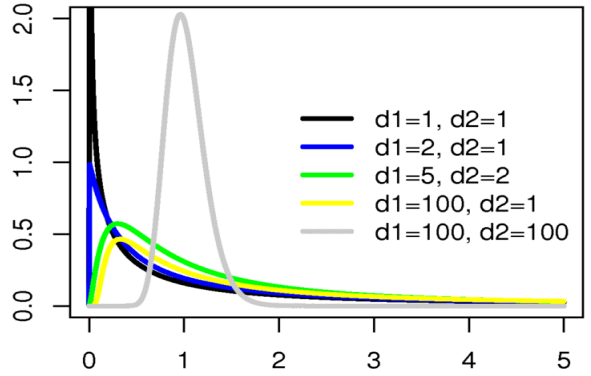
Standard Sampling Distributions



χ^2 - (Chi-Square) distribution curve



t- distribution curve



F – distribution curve

Chi-square Distribution

χ^2 Distribution

- A common use of the χ^2 distribution is to describe the distribution of the sample variance.

- Let X_1, X_2, . . . , X_n be an independent random variables from a normally distributed population with mean = μ and variance = σ^2.
- The χ^2 distribution can be written as

$$\chi^2 = (Z_1)^2 + (Z_2)^2 + (Z_3)^2 + \cdots + (Z_n)^2$$

where $Z_i = \frac{(X_i - \mu)}{S}$.

- This χ^2 is also a random variable of a distribution and is called χ^2-distribution (pronounced as Chi-square distribution).

The χ^2 Distribution

A common use of the χ^2 distribution is to describe the distribution of the sample variance.

Definition 1: χ^2 distribution

If $x_1, x_2 \ldots\ldots\ldots\ldots x_n$ are independent random variables having identical normal distribution with mean μ and variance σ^2, then the random variable

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

has a Chi squared distribution with *n* degrees of freedom. (How?)

The χ^2 Distribution

Note: Each of the n independent random variable $\left(\frac{x_i-\mu}{\sigma}\right)^2, i = 1, 2, 3, \ldots\ldots n$ has Chi-squared distribution with 1 degree of freedom.

Now we can derive χ^2- distribution for sample variance.
We can write

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}[(x_i - \overline{x}) + (\overline{x} - \mu)]^2$$

$$= \sum_{i=1}^{n}(x_i - \overline{x})^2 + n.(\overline{x} - \mu)^2$$

or, $$\frac{1}{\sigma^2}\sum(x_i - \mu)^2 = \frac{(n-1)S^2}{\sigma^2} + \frac{(\overline{x}-\mu)^2}{\sigma^2/n}$$

Chi-square distribution with n-degree

Chi-square distribution with (n-1) degree of freedom

Chi-square distribution with 1 degree of freedom $[= Z^2]$

Note: To **calculate degrees of freedom**, subtract the number of relations from the number of observations.

The χ^2 Distribution

Definition 2: χ^2-distribution for Sampling Variance

If S^2 is the variance of a random sample of size *n* taken from a normal population having the variance σ^2, then the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{\sigma}\right)^2$$

has a chi-squared distribution with $v = n - 1$ degrees of freedom.

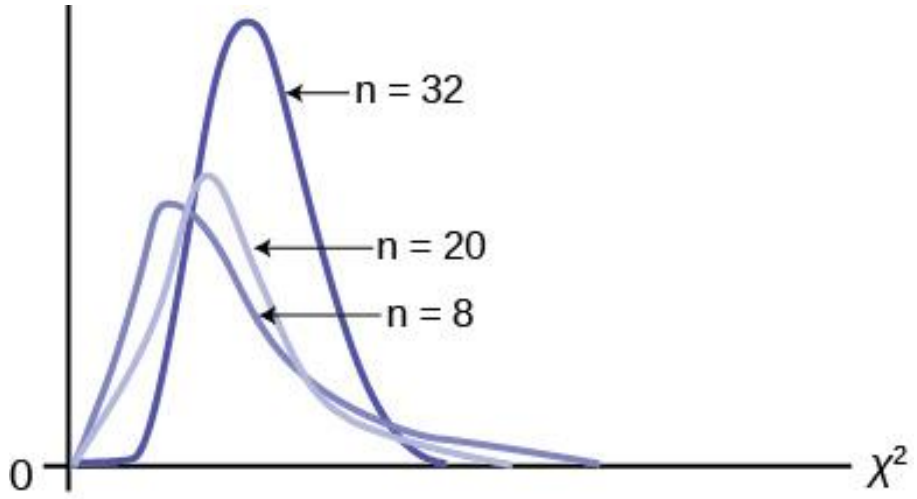
This way χ^2- distribution is used to describe the sampling distribution of S^2.

The χ^2 Distribution

- The χ^2 distribution can be written as

$$\chi^2 = \sum Z^2 = \sum \left(\frac{X-\bar{X}}{\sigma}\right)^2 = \frac{SS}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$$

- This expression χ^2 describes the distribution (of n samples) and thus having degrees of freedom $\upsilon = n\text{-}1$ and often written as $\chi^2(\upsilon)$, where υ is the only parameter in it.



Chi-Squared Distribution

Definition 3: Chi-squared distribution

The continuous random variable x has a Chi-squared distribution with v degrees of freedom, is given by

$$f(x:v) = \begin{cases} \dfrac{1}{2^{\frac{v}{2}}\,\Gamma(v/2)} x^{v/2-1} e^{-\frac{x}{2}}, & x > 0 \\ 0 & Otherwise \end{cases}$$

where v is a positive integer and

$$\Gamma(x) = \int_0^{\infty} e^{-t} t^{x-1}\, dt.$$

- The Chi-squared distribution plays an important role in statistical inference .
- The mean and variance of Chi-squared distribution are:

$$\mu = v \text{ and } \sigma^2 = 2v \text{ (Prove !)}$$

Some facts about χ^2 distribution

- The curves are non symmetrical and skewed to the right.
- χ^2 values cannot be negative since they are sums of squares.
- The mean of the χ^2 distribution is ν, and the variance is 2ν.
- When $\upsilon > 30$, the Chi-square curve approximates the normal distribution. Then, you may write the following

$$Z = \frac{\chi^2 - \upsilon}{\sqrt{2\upsilon}}$$

χ^2 is distribution of sample variances

- A common use of the χ^2 distribution is to describe the distribution of the sample variance. Let $X_1, X_2, \ldots, X_n$ be a random sample from a normally distributed population with mean = μ and variance = σ^2. Then the quantity $(n-1)S^2/\sigma^2$ is a random variable whose distribution is described by a χ^2 distribution with $(n-1)$ degrees of freedom, where S^2 is the usual sample estimate of the population variance. That is

$$S^2 = \frac{(X-\bar{X})^2}{n-1}$$

- In other words, the χ^2 distribution is used to describe the sampling distribution of S^2. Since we divide the sum of squares by degrees of freedom to obtain the variance estimate, the expression for the random variable having a $\chi 2$ distribution can be written

$$\chi^2 = \sum Z^2 = \sum \left(\frac{X-\bar{X}}{\sigma}\right)^2 = \frac{SS}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$$

Application χ^2 of values

Example: Judging the quality of a machine

A machine is to produce a ball of 100gm. It is desirable to have maximum deviation of 0.01gm (this is the desirable value of σ).

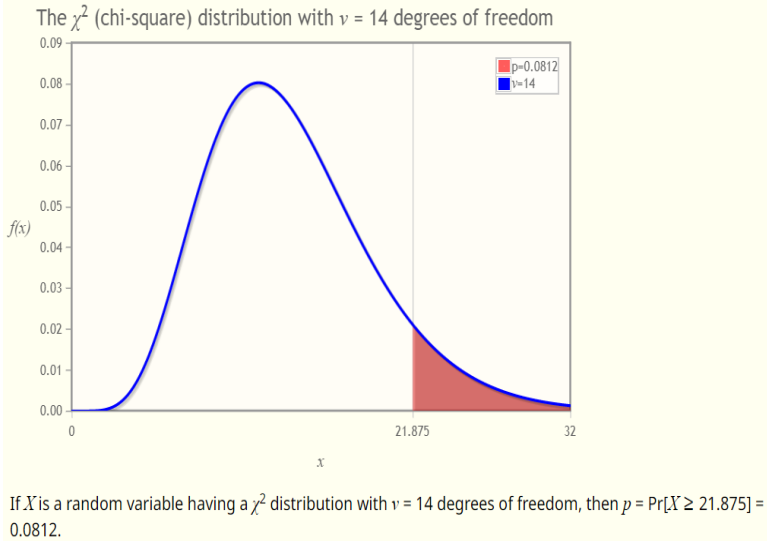
Suppose, 15 balls produced by the machine are select at random and it shows S = 0.0125gm.

What is the probability that the machine will produces an accurate ball?

χ^2 calculation can help us to know this value.

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{14\times(0.0125)^2}{(0.01)^2} = 21.875$$

This is the χ^2 value with 14 degrees of freedom. The value can be tested with χ^2 table to know the desired probability value.



Exercise

A manufacturer of car batteries guarantees that the batteries will last, on average, 3 years with a standard deviation of 1 year. If five of these batteries have lifetimes of 1.9, 2.4, 3.0, 3.5, and 4.2 years, should the manufacturer still be convinced that the batteries have a standard deviation of 1 year? Assume that the battery lifetime follows a normal distribution.

t Distribution

The t Distribution

- **The t Distribution**

1. To know the sampling distribution of mean we make use of Central Limit Theorem with $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$
2. This require the known value of σ a priori.
3. However, in many situation, σ is certainly no more reasonable than the knowledge of the population mean μ.
4. In such situation, only measure of the standard deviation available may be the sample standard deviation S.
5. It is natural then to substitute S for σ. The problem is that the resulting statistics is not normally distributed!
6. The t distribution is to alleviate this problem. This distribution is called $student's\ t$ or simply $t - distribution$.

The t Distribution

Definition: t −distribution

The t −distribution with v degrees of freedom actually takes the form

$$t(v) = \frac{Z}{\sqrt{\frac{\chi^2(v)}{v}}}$$

where Z is a standard normal random variable, and $\chi^2(v)$ is χ^2 random variable with v degrees of freedom.

The probability density function :

$$f(t) = \frac{\Gamma[(\vartheta+1)/2]}{\Gamma(\vartheta/2)\sqrt{\pi\vartheta}}\left(1+\frac{t^2}{\vartheta}\right)^{-(\vartheta+1)/2}, \qquad -\infty < t < \infty$$

This is known as t distribution with $\vartheta = n - 1$ degrees of freedom.

t Distribution

Corollary: Let $X_1, X_2, \ldots\ldots, X_n$ be independent random variables that are all normal with mean μ and standard deviation σ.

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Using this definition, we can develop the sampling distribution of the sample mean when the population variance, σ^2 is unknown.

That is,

$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ has the standard normal distribution.

$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ has the χ^2 distribution with $(n-1)$ degrees of freedom.

Thus, $T = \dfrac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}}$ or $\boxed{T = \frac{\bar{X}-\mu}{S/\sqrt{n}}}$

This is the $t - distribution$ with $(n-1)$ degrees of freedom.

t Distribution

- If the sample size is small, the values of S^2 fluctuate considerably from sample to sample.
- The distribution of T deviates appreciably from that of a standard normal distribution.
- If the sample size is large enough, say $n \geq 30$, the distribution of T does not differ considerably from the standard normal.
- For $n < 30$, it is useful to deal with the exact distribution of T.
- In developing the sampling distribution of T, we shall assume that our random sample was selected from a normal population.

Exercise

A chemical engineer claims that the population mean yield of a certain batch process is 500 grams per milliliter of raw material. To check this claim he samples 25 batches each month. If the computed *t*-value falls between $-t_{0.05}$ and $t_{0.05}$, he is satisfied with this claim. What conclusion should he draw from a sample that has a mean $\bar{x} = 518$ grams per milliliter and a sample standard deviation $s = 40$ grams? Assume the distribution of yields to be approximately normal.

F Distribution

F Distribution

- While it is of interest to let sample information shed light on two population means, it is often the case that a comparison of variability is equally important, if not more so.
- The F-distribution finds enormous application in comparing sample variances.
- Applications of the F-distribution are found in problems involving two or more samples.
- The statistic F is defined to be the ratio of two independent chi-squared random variables, each divided by its number of degrees of freedom.
- Hence, we can write

$$F = \frac{\chi_1^2/\vartheta_1}{\chi_2^2/\vartheta_2}$$

where χ_1^2 and χ_2^2 are independent random variables having chi-squared distributions with $\vartheta_1 = n_1 - 1$ and $\vartheta_2 = n_2 - 1$ degrees of freedom, respectively.

F Distribution

- The curve of the F-distribution depends not only on the two parameters ϑ_1 and ϑ_2 but also on the order in which we state them.
- Let f_α be the f-value above which we find an area equal to α.
- Writing $f_\alpha(\vartheta_1, \vartheta_2)$ for f_α with ϑ_1 and ϑ_2 degrees of freedom, then

$$f_{1-\alpha}(\vartheta_1, \vartheta_2) = \frac{1}{f_\alpha(\vartheta_2, \vartheta_1)}.$$

- Thus the f-value with 6 and 10 degrees of freedom, leaving an area of 0.95 to the right is

$$f_{0.95}(6,10) = \frac{1}{f_{0.05}(10,6)} = \frac{1}{4.06} = 0.246$$

F Distribution

- Probability density function:

$$h(x) = \frac{\Gamma[(\vartheta_1 + \vartheta_2)/2]}{\Gamma(\vartheta_1/2)\Gamma(\vartheta_2/2)} \left(\frac{\vartheta_1}{\vartheta_2}\right)^{\vartheta_1/2} \frac{x^{(\vartheta_1/2)-1}}{\left[1 + \left(\frac{\vartheta_1}{\vartheta_2}\right)x\right]^{(\vartheta_1+\vartheta_2)/2}}, 0 < x < \infty$$

with ϑ_1 and ϑ_2 degrees of freedom.

- If S_1^2 and S_2^2 are the variances of independent random samples of size n_1 and n_2 taken from normal populations with variances σ_1^2 and σ_2^2 respectively, then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

has an F-distribution with $\vartheta_1 = n_1 - 1$ and $\vartheta_2 = n_2 - 1$ degrees of freedom.

The F Distribution

Definition: F distribution

The statistics F is defined to be the ratio of two independent Chi-Squared random variables, each divided by its number of degrees of freedom. Hence,

$$\mathrm{F}(v_1, v_2) = \frac{\chi^2(v_1)/v_1}{\chi^2(v_2)/v_2}$$

Corollary : Recall that $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ is the Chi-squared distribution with $(n-1)$ degrees of freedom.

Therefore, if we assume that we have sample of size n_1 from a population with variance ${\sigma_1}^2$ and an independent sample of size n_2 from another population with variance ${\sigma_2}^2$, then the statistics

$$F = \frac{{S_1}^2/{\sigma_1}^2}{{S_2}^2/{\sigma_2}^2}$$

Note: The F distribution finds enormous applications in comparing sample variances.

Exercise

Pumpkins were grown under two experimental conditions. Two random samples of 11 and 9 pumpkins show the sample standard deviations of their weights as 0.8 and 0.5 respectively. Assuming that the weight distributions are normal, test the hypothesis that the true variances are equal, against the alternative that they are not, at the 10% level.

Summary of sampling distributions

Z – distribution:

- Typically it is used for comparing the mean of a sample to some hypothesized mean for the population in case of large sample, or when **population variance is known**.

t – distribution:

- **population variance is not known.** In this case, we use the variance of the sample as an estimate of the population variance.

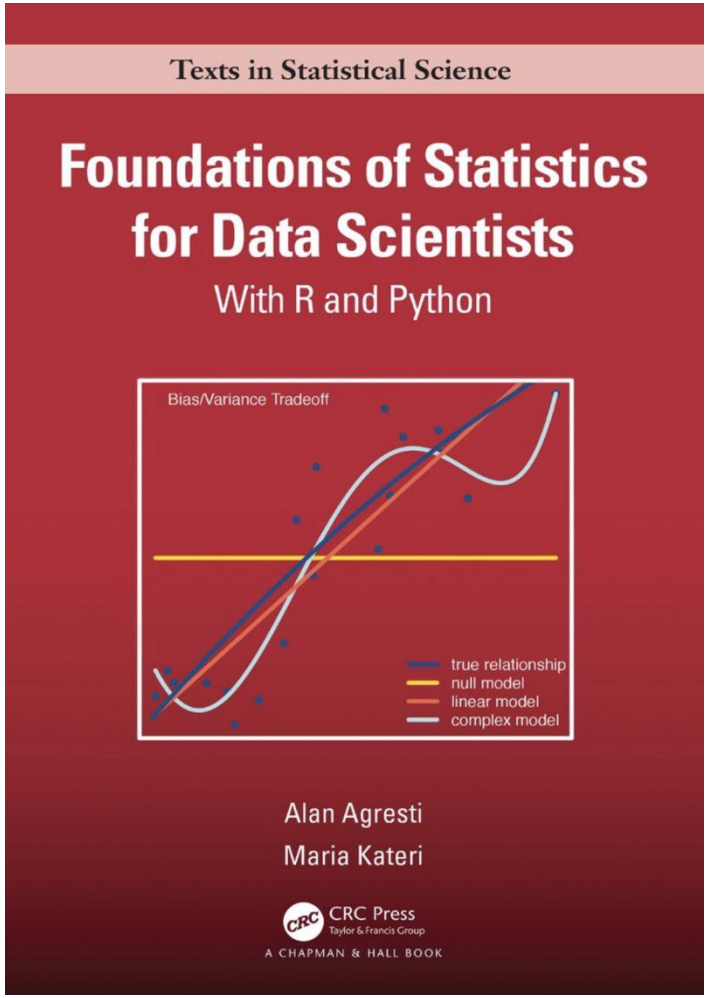
χ^2 – distribution:

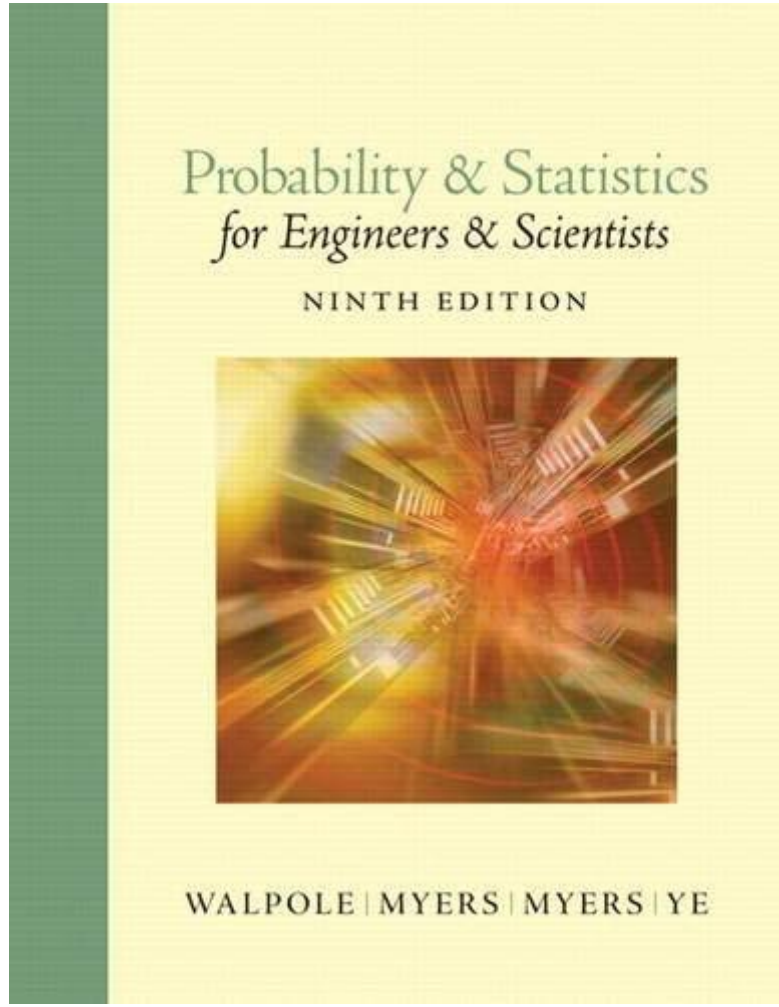
- It is used for comparing a sample variance to a theoretical population variance.

F – distribution:

- It is used for comparing the variance of two or more populations.

References:


Texts in Statistical Science
Foundations of Statistics for Data Scientists
With R and Python
Bias/Variance Tradeoff
true relationship
null model
linear model
complex model
Alan Agresti
Maria Kateri
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK


Probability & Statistics
for Engineers & Scientists
NINTH EDITION
WALPOLE | MYERS | MYERS | YE