SMA 2272/STA 2270 STATISTICS

PURPOSE

By the end of the course the student should be proficient in representing data graphically and handling summary statistics, simple correlation and best fitting line, and handling probability and probability distributions including expectation and variance of a discrete random variable.

DESCRIPTION

Classical and axiomatic approaches to probability. Compound and conditional probability, including Bayes' theorem. Concept of discrete random variable: expectation and variance. Data: sources, collection, classification and processing. Frequency distributions and graphical representation of data, including bar diagrams, histograms and stem-and-leaf diagrams. Measures of central tendency and dispersion. Skewness and kurtosis. Correlation. Fitting data to a best straight line.

Pre-Requisites: STA 2104 Calculus for statistics I, SMA 2104 Mathematics for Science.

COURSE TEXT BOOKS

1. Uppal, S. M., Odhiambo, R. O. & Humphreys, H. M. *Introduction to Probability and Statistics.* JKUAT Press, 2005. ISBN 9966-923-95-0
2. J Crawshaw & J Chambers *A concise course in A-Level statistics, with worked examples*, 3rd ed. Stanley Thornes, 1994 ISBN 0-534- 42362-0.

COURSE JOURNALS

1. Journal of Applied Statistics (J. Appl. Stat.) [0266-4763; 1360-0532]
2. Statistics (Statistics) [0233-1888]

FURTHER REFERENCE TEXT BOOKS AND JOURNALS

1. GM Clarke & D Cooke *A Basic Course in Statistics.* 5th ed. Arnold, 2004 ISBN13: 978-0-340-81406-2 ISBN10: 0-340-81406-3.
2. S Ross *A first course in Probability* 4th ed. Prentice Hall, 1994 ISBN-10: 0131856626 ISBN-13: 9780131856622.
3. P.S. Mann. *Introductory Statistics*. John Wiley & Sons Ltd, 2001 ISBN 13: 9780471395119.
4. Statistical Science (Stat. Sci.) [0883-4237]
5. Journal of Mathematical Sciences
6. Journal of Teaching Statistics

Introduction

What is statistics?

The Word statistics has been derived from Latin word "**Status**" or the Italian word "**Statista**", the meaning of these words is "**Political State**" or a Government. Early applications of statistical thinking revolved around the needs of states to base policy on demographic and economic data.

Definition

Statistics: *a branch of science that deals with collection presentation, analysis, and interpretation of data*. The definition points out 4 key aspects of statistics namely

(i) Data collection

(ii) Data presentation,

(iii) Data analysis, and

(iv) Data interpretation

Statistics is divided into 2 broad categories namely descriptive and inferential statistics.

Descriptive Statistics: summary values and presentations which gives some information about the data Eg the mean height of a 1st year student in JKUAT is170cm. 170cm is a statistics which describes the central point of the heights data.

Inferential Statistics: summary values calculated from the sample in order to make conclusions about the target population.

Types of Variables

Qualitative Variables: Variables whose values fall into groups or categories. They are called categorical variables and are further divided into 2 classes namely nominal and ordinal variables

a) Nominal variables: variables whose categories are just names with no natural ordering. Eg gender marital status, skin colour, district of birth etc

b) Ordinal variables: variables whose categories have a natural ordering. Eg education level, performance category, degree classifications etc

Quantitative Variables: these are numeric variables and are further divided into 2 classes namely discrete and continuous variables

a) Discrete variables: can only assume certain values and there are gaps between them. Eg the number of calls one makes in a day, the number of vehicles passing through a certain point etc

b) Continuous variables: can assume any value in a specified range. Eg length of a telephone call, height of a 1st year student in JKUAT etc

1. Data Collection:

1.1 Sources of Data

There are 2 sources for data collection namely Primary, and Secondary data

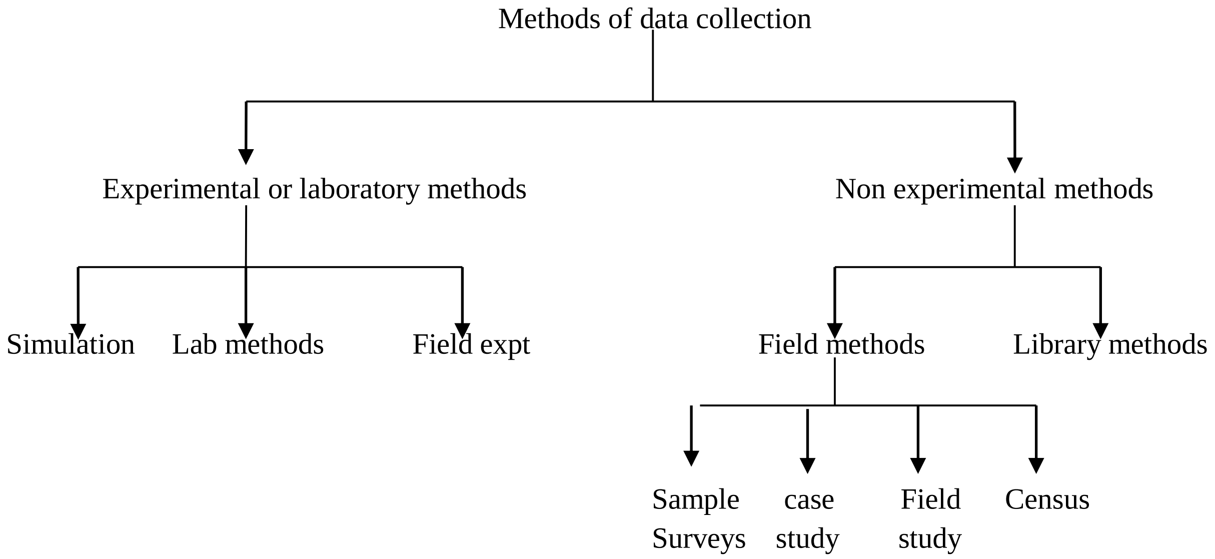
Primary data:- freshly collected ie for the first time. They are original in character ie they are the first hand information collected, compiled and published for some purpose. They haven't undergone any statistical treatment

***Secondary Data*:-** 2nd hand information mainly obtained from published sources such as statistical abstracts books encyclopaedias periodicals, media reports eg census report CD-roms and other electronic devices, internet. They are not original in character and have undergone some statistical treatment at least once.

1.2 Data Collection Methods

The 1st step in any investigation (inquiry) is data collection. Information can either be collected directly or indirectly from the entire population or a sample.

There are many methods of collecting data which includes the ones illustrated in the flow chart below



Experimental methods are so called because in them the investigator in a laboratory tests the hypothesis about the cause and effect relationship by manipulating the independent variables under controlled conditions.

Non-Experimental methods are so called because in them the investigator does not control or change any aspect of the situation under study but simply describes what naturally occurs at a certain point or period of time.

Non-Experimental methods are widely used in social sciences. Some of the Non-Experimental methods used for data collection are outlined below.

a) **Field study**:- aims at testing hypothesis in natural life situations. It differs from field experiment in that the researcher does not control or manipulate the independent variables but both of them are carried out in natural conditions

Merits*:*

(i) The method is realistic as it is carried out in natural conditions

(ii) It's easy to obtain data with large number of variables.

Demerits

(iii) Independent variables are not manipulated.

(iv) Co-operation of the organization is often difficult to obtain.

(v) Data is likely to contain unknown sampling biasness.

(vi) The dross rate (proportion of irrelevant data) may be high in such studies.

(vii) Measurement is not precise as in laboratory because of influence of confounding variables.

b) **Census.** A census is a study that obtains data from every member of a population (totality of individuals /items pertaining to certain characteristics). In most studies, a census is not practical, because of the cost and/or time required.

c) **Sample survey**. A sample survey is a study that obtains data from a subset of a population, in order to estimate population attributes/ characteristics. Surveys of human populations and institutions are common in government, health, social science and marketing research.

d) **Case study –**It's a method of intensively exploring and analyzing the life of a single social unit be it a family, person, an institution, cultural group or even an entire community. In this method no attempt is made to exercise experimental or statistical control and phenomena related to the unit are studied in natural. The researcher has several discretion in gathering information from a variety of sources such as diaries, letters, autobiographies, records in office, files or personal interviews.

Merits:

(i) The method is less expensive than other methods.
(ii) Very intensive in nature –aims at studying a few units rather than several
(iii) Data collection is flexible since the researcher is free to approach the problem from any angle.
(iv) Data is collected from natural settings.

Demerits

(i) It lacks internal validity which is basic to scientific evidence.
(ii) Only one unit of the defined population is studied. Hence the findings of case study cannot be used as abase for generalization about a large population. They lack external validity.
(iii) Case studies are more time consuming than other methods.

e) **Experiment.** An experiment is a controlled study in which the researcher attempts to understand cause-and-effect relationships. In experiments actual experiment is carried out on certain individuals / units about whom information is drawn. The study is "controlled" in the sense that the researcher controls how subjects are assigned to groups and which treatments each group receives.

f) **Observational study.** Like experiments, observational studies attempt to understand cause-and-effect relationships. However, unlike experiments, the researcher is not able to control how subjects are assigned to groups and/or which treatments each group receives. Under this method information, is sought by direct observation by the investigator.

1.3 Population and Sample

Population: The entire set of individuals about which findings of a survey refer to.

Sample: A subset of population selected for a study.

Sample Design: The scheme by which items are chosen for the sample.

Sample unit: The element of the sample selected from the population.

Unit of analysis: Unit at which analysis will be done for inferring about the population. Consider that you want to examine the effect of health care facilities in a community on prenatal care. What is the unit of analysis: health facility or the individual woman?.

Sampling Frames

For probability sampling, we must have a list of all the individuals (units) in the population. This list or sampling frame is the basis for the selection process of the sample. “A [sampling] frame is a clear and concise description of the population under study, by virtue of which the population units can be identified unambiguously and contacted, if desired, for the purpose of the survey” - Hedayet and Sinha, 1991

Based on the sampling frame, the sampling design could also be classified as:

Individual Surveys if List of individuals is available or when the size of population is small Special population

Household Surveys; If it’s Based on the census of the households and if the individual level information is unlikely to be available In practice, it’s limited to small geographical areas and know as “area sampling frame” Example: Demographic and Health Surveys (DHS)

Institutional Surveys If it’s Based on the census of say Hospital/clinic lists eg

i) 1990 National Hospital Discharge Survey
ii) National Ambulatory Medical Care Survey

Problems of Sampling Frame

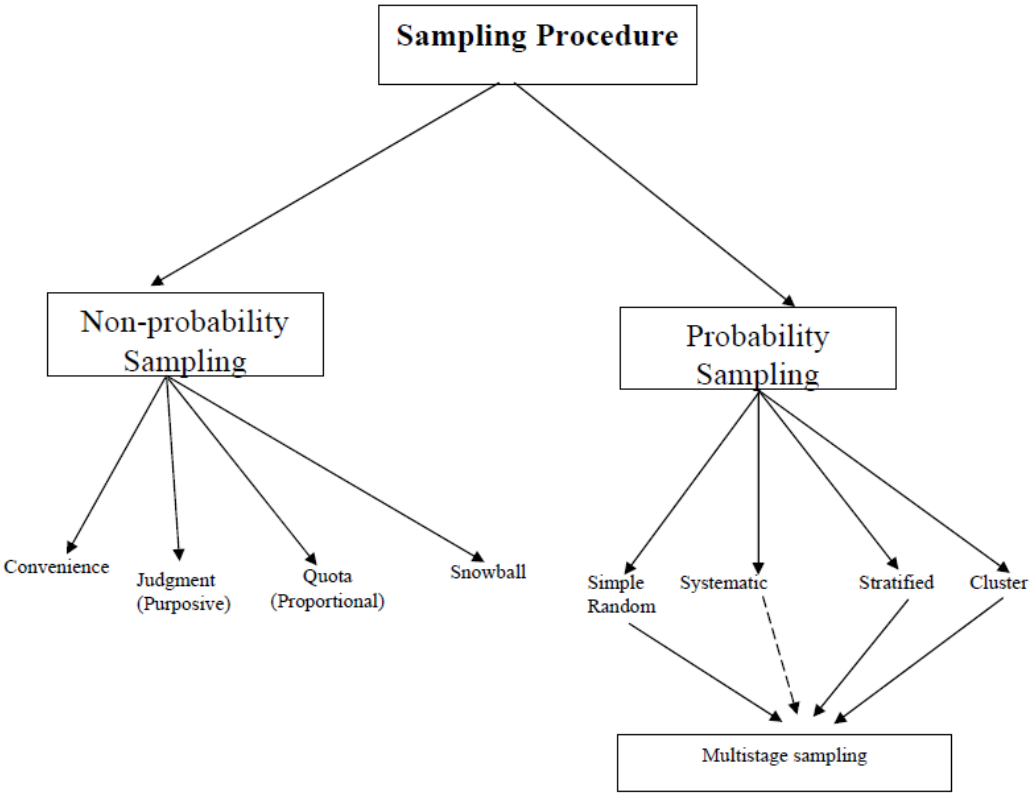
(i) Missing elements
(ii) Noncoverage
(iii) Incomplete frame
(iv) Old list
(v) Undercoverage
(vi) May not be readily available
(vii) Expensive to gather

1.4 Sampling

Sampling is a statistical process of selecting a representative sample. We have probability sampling and non-probability sampling **Probability Samples** involves a mathematical chance of selecting the respondent. Every unit in the population has a chance, greater than zero, of being selected in the sample. Thus producing unbiased estimates. They include;

(i) Simple random sampling
(ii) Systematic sampling
(iii) Stratified sampling
(iv) Cluster sampling
(v) multi-stage sampling

Non-probability sampling is any sampling method where some elements of the population have *no* chance of selection (also referred to as "out of coverage"/"undercovered"), or where the probability of selection can't be accurately determined. It yields a non-random sample therefore making it difficult to extrapolate from the sample to the population. They include; Judgement sample, purposive sample, convenience sample: **subjective** Snow-ball sampling: **rare group/disease study**



1.4.1 Sampling Procedure

Sampling involves two tasks

- How to select the elements?
- How to estimate the population characteristics – from the sampling units?

We employ some *randomization* process for sample selection so that there is no preferential treatment in selection which may introduce selectivity bias

1.4.2 Reasons Behind sampling

(i) Cost; the sample can furnish data 0f sufficient accuracy at much lower cost.
(ii) Time; the sample provides information faster than census thus ensuring timely decision making.
(iii) Accuracy; it is easier to control data collection errors in a sample survey as opposed to census.
(iv) Risky or destructive test call for sample survey not census eg testing a new drug.

1.4.3 *Probability Sampling Techniques*

a)...**Simple Random Sampling** (SRS)

In this design, each element has an equal probability of being selected from a list of all population units (sample of n from N population). Though it's attractive for its simplicity, the design is not usually used in the sample survey in practice for several reasons:

(i) Lack of listing frame: the method requires that a list of population elements be available, which is not the case for many populations.

(ii) Problem of small area estimation or domain analysis: For a small sample from a large population, all the areas may not have enough sample size for making small area estimation or for *domain* analysis by variables of interest.

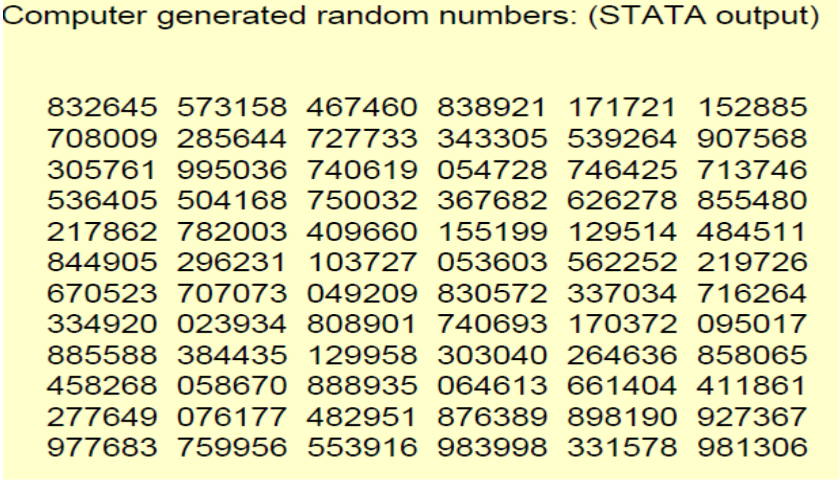
(iii) Not cost effective: SRS requires covering of whole population which may reside in a large geographic area; interviewing few samples spread sparsely over a large area would be very costly.

Implementation of SRS sampling:

(i) Listing (sampling) Frame

(ii) Random number table (from published table or computer generated)

(iii) Selection of sample



```
Computer generated random numbers: (STATA output)


  832645 573158 467460 838921 171721 152885
  708009 285644 727733 343305 539264 907568
  305761 995036 740619 054728 746425 713746
  536405 504168 750032 367682 626278 855480
  217862 782003 409660 155199 129514 484511
  844905 296231 103727 053603 562252 219726
  670523 707073 049209 830572 337034 716264
  334920 023934 808901 740693 170372 095017
  885588 384435 129958 303040 264636 858065
  458268 058670 888935 064613 661404 411861
  277649 076177 482951 876389 898190 927367
  977683 759956 553916 983998 331578 981306
```

b)..Systematic Sampling

Systematic sampling, either by itself or in combination with some other method, may be the most widely used method of sampling." In systematic sampling we select samples "evenly" from the list (sampling frame): First, let us consider that we are dividing the list evenly into some "blocks". Then, we select a sample element from each block.

In systematic sampling, only the first unit is selected at random, the rest being selected according to a predetermined pattern. To select a systematic sample of *n* units, the first unit is selected with a random start *r* from 1 to k sample, where *k=N/n* sample intervals, and after the selection of first sample, every k^{th} unit is included where $1 \leq r \leq k$.

An example:

Let N=100, n=10, then k=100/10. Then the random start *r* is selected between 1 and 10 (say, r=7). So, the sample will be selected from the population with serial indexes of: 7, 17, 27, ...,97. i.e., r, r+k, r+2k,......., r+(n-1)k

What could be done if k=N/n is not an integer?

Selection of systematic sampling when sampling interval (k) is not an integer

Consider, n=175 and N=1000. So, k=1000/175 = 5.71

One of the solution is to make k rounded to an integer, i.e., k=5 or k=6.Now, if k=5, then n=1000/5=200; or, If k=6, then n=1000/6 = 166.67 ~ 167. Which n should be chosen?

Solution
if k=5 is considered, stop the selection of samples when n=175 achieved.
if k=6 is considered, treat the sampling frame as a circular list and continue the selection of samples from the beginning of the list after exhausting the list during the first cycle.

An alternative procedure is to keep k non-integer and continue the sample selection as follows: Let us consider, k=5.71, and r=4. So, the first sample is 4th in the list. The second = (4+5.71) =9.71 ~9th in the list, the third =(4+2*5.71) =15.42 ~ 15th in the list, and so on. (The last sample is: 4+5.71*(175-1) = 997.54 ~ 997th in the list). Note that, k is switching between 5 and 6

Advantages:
Systematic sampling has many attractiveness:
(i) Provides a better random distribution than SRS
(ii) Simple to implement
(iii)May be started without a complete listing frame (say, interview of every 9th patient coming to a clinic).
(iv)With ordered list, the variance may be smaller than SRS (see below for exceptions

Disadvantages:
(i) Periodicity (cyclic variation)
(ii) linear trend

i.

When to use systematic sampling?
i) Even preferred over SRS
ii) When no list of population exists
iii) When the list is roughly of random order
iv) Small area/population

c)..Stratified Sampling

In stratified sampling the population is partitioned into groups, called *strata*, and sampling is performed separately within each *stratum*.
This sampling technique is used when;
i) Population groups may have different values for the responses of interest.
ii) we want to improve our estimation for each group separately.
iii) To ensure adequate sample size for each group.

In stratified sampling designs:
i) Stratum variables are mutually exclusive (no over lapping), e.g., urban/rural areas, economic categories, geographic regions, race, sex, etc. The principal objective of stratification is to reduce sampling errors.
ii) The population (elements) should be *homogenous* within-stratum, and the population (elements) should be *heterogeneous* between the strata.

Advantages
(i) Provides opportunity to study the stratum; variations - estimation could be made for each stratum

(ii) Disproportionate sample may be selected from each stratum
(iii) The precision is likely to increase as variance may be smaller than simple random case with same sample size
(iv) Field works can be organized using the strata (e.g., by geographical areas or regions)
(v) Reduce survey costs.

Disadvantages

(i) Sampling frame is needed for each stratum
(ii) Analysis method is complex
(iii) Correct variance estimation
(iv) Data analysis should take sampling "weight" into account for disproportionate sampling of strata
(v) Sample size estimation is difficult in practice

Allocation of Stratified Sampling

The major task of stratified sampling design is the appropriate allocation of samples to different strata.

Types of allocation methods:

(i) Equal allocation
(ii) Proportional to stratum size
(iii) Cost based sample allocation

Equal Allocation

Divide the number of sample units *n* equally among the *K* strata. ie $n_i = \frac{n}{k}$ Example: n = 100 and k= 4 strata $n_i = \frac{100}{4} = 25$ *units* in each stratum.

Disadvantages of equal allocation:

May need to use weighting to have unbiased estimates

Proportional allocation

Make the proportion of each stratum sampled identical to the proportion of the population. Ie Let the sample fraction be f= n/N. So, $n_i = fN_i = n\frac{N_i}{N}$, Where $\frac{N_i}{N}$ is the stratum weight.
Example: N = 1000, n = 100 $f = \frac{100}{1000} = 0.1$ now suppose $N_1 = 700$ and $N_2 = 300$ then $n_1 = 700 * 0.1 = 70$ and $n_2 = 300 * 0.3 = 30$

Disadvantage of proportional allocation:

Sample size in a stratum may be low thus providing unreliable stratum-specific results.

d)..Cluster Sampling

In many practical situations the population elements are grouped into a number of clusters. A list of clusters can be constructed as the sampling frame but a complete list of elements is often unavailable, or too expensive to construct. In this case it is necessary to use cluster sampling where a random sample of clusters is taken and some or all elements in the selected clusters are observed. Cluster sampling is also preferable in terms of cost, because it is much cheaper, easier and quicker to collect data from adjoining elements than elements chosen at random. On the other hand, cluster sampling is less informative and less efficient per elements in the sample, due to similarities of elements within the same cluster. The loss of efficiency, however, can often be compensated by increasing the overall sample size. Thus, in terms of unit cost, the cluster sampling plan is efficient.

e)..Multi-Stage Samples

Here the respondents are chosen through a process of defined stages. Eg residents within Kibera (Nairobi) may have been chosen for a survey through the following process:
Throughout the country (Kenya) the Nairobi may have been selected at random, (stage 1), within Nairobi, Langata (constituency) is selected again at random (stage 2), Kibera is then selected within Langata (stage 3), then polling stations from Kibera (stage 4) and then individuals from the electoral voters' register (stage 5)! As demonstrated five stages were gone through before the final selection of respondents were selected from the electoral voters' register.

Advantages of probability sample

(i) Provides a quantitative measure of the extent of variation due to random effects
(ii) Provides data of known quality
(iii)Provides data in timely fashion
(iv)Provides acceptable data at minimum cost
(v) Better control over nonsampling sources of errors
(vi)Mathematical statistics and probability can be applied to analyze and interpret the data

1.4.4 Non-probability Sampling

Social research is often conducted in situations where a researcher cannot select the kinds of probability samples used in large-scale social surveys. For example, say you wanted to study homelessness - there is no list of homeless individuals nor are you likely to create such a list. However, you need to get some kind of a sample of respondents in order to conduct your research. To gather such a sample, you would likely use some form of non-probability sampling.
There are four primary types of non-probability sampling methods:

a)..Convinience Sampling

It's a method of choosing subjects who are available or easy to find. This method is also sometimes referred to as haphazard, accidental, or availability sampling. The primary advantage of the method is that it is very easy to carry out, relative to other methods.

Demerit

- One can never be certain what population the participants in the study represent. The population is unknown.
- The method is haphazard, and the cases studied probably don't represent any population you could come up with. However, it's very useful for pilot studies

Advantages of convenience sample

(i) It's *very easy* to carry out with few rules governing how the *sample* should be collected.
(ii) The *relative cost* and *time* required to carry out a convenience sample are *small* in comparison to probability sampling techniques. This enables you to achieve the *sample size* you want in a *relatively fast* and *inexpensive* way.
(iii) The convenience sample may help you gather useful data and information that would not have been possible using *probability sampling techniques*, which require more formal access to *lists of populations* [see, for example, the article on simple random sampling].

For example, imagine you were interested in understanding more about employee satisfaction in a single, large organisation in the United States. You intended to collect your data using a questionnaire. The manager who has kindly given you access to conduct your research is unable to get permission to get a *list* of all employees in the organisation, which you would need to use a *probability sampling technique* such as simple random sampling or systematic random sampling.

However, the manager has managed to secure permission for you to spend two days in the organisation to collect as many questionnaire responses as possible. You decide to spend the two days at the entrance of the organisation where all employees have to pass through to get to their desks. Whilst a *probability sampling technique* would have been preferred, the convenience sample was the only sampling technique that you could use to collect data. Irrespective of the disadvantages of convenience sampling, discussed below, without the use of this sampling technique, you may not have been able to get access to any data on employee satisfaction in the organisation.

Disadvantages of convenience sampling

- The convenience sample often suffers from a number of *biases*. This can be seen in both of our examples, whether the 10,000 students we were studying, or the employees at the large organisation. In both cases, a convenience sample can lead to the *under-representation* or *over-representation* of particular *groups* within the *sample*. If we take the large organisation:

  It may be that the organisation has multiple sites, with employee satisfaction varying considerably between these sites. By conducting the survey at the headquarters of the organisation, we may have missed the differences in employee satisfaction amongst those at different sites, including non-office workers. We also do not know why some employees agreed to take part in the survey, whilst others did not. Was it because some employees were simply too busy? Did they not trust the intentions of the survey? Did others take part out of kindness or because they had a particular grievance with the organisation? These types of *biases* are quite typical in convenience sampling.
- Since the *sampling frame* is *not know*, and the *sample* is *not chosen at random*, the *inherent bias* in convenience sampling means that the sample is *unlikely* to be *representative* of the *population* being studied. This undermines your ability to make *generalisations* from your *sample* to the *population* you are studying.
  If you are an undergraduate or master's level dissertation student considering using *convenience sampling*, you may also want to read more about how to put together your *sampling strategy* [see the section: Sampling Strategy

b)..Quota Sampling

Quota sampling is designed to overcome the most obvious flaw of availability sampling. Rather than taking just anyone, you set quotas to ensure that the sample you get represents certain characteristics in proportion to their prevalence in the population. Note that for this method, you have to know something about the characteristics of the population ahead of time. Say you want to make sure you have a sample proportional to the population in terms of gender - you have to know what percentage of the population is male and female, then collect sample until yours matches. Marketing studies are particularly fond of this form of research design.

The primary problem with this form of sampling is that even when we know that a quota sample is representative of the particular characteristics for which quotas have been set, we have no way of knowing if sample is representative in terms of any other characteristics. If we set quotas for gender and age, we are likely to attain a sample with good representativeness on age and gender, but one that may not be very representative in terms of income and education or other factors.

Moreover, because researchers can set quotas for only a small fraction of the characteristics relevant to a study quota sampling is really not much better than availability sampling. To reiterate, you must know the characteristics of the entire population to set quotas; otherwise there's not much point to setting up quotas. Finally, interviewers often introduce bias when allowed to self-select respondents, which is usually the case in this form of research. In choosing males 18-25, interviewers are more likely to choose those that are better-dressed, seem more approachable or less threatening. That may be understandable from a practical point of view, but it introduces bias into research findings.

Imagine that a researcher wants to understand more about the career goals of students at a single university. Let's say that the university has roughly 10,000 students. suppose we were interested in *comparing the differences* in career goals between *male* and *female* students at the single university. If this was the case, we would want to ensure that the *sample* we selected had a *proportional* number of *male* and *female* students relative to the *population*.
To create a quota sample, there are three steps:
Choose the relevant grouping chsr and divide the population accordingly *gender*
Calculate a quota (number of *units* that should be included in *each* for group
Continue to invite units until the quota for each group is met

Advantages of quota sampling

i) It particularly useful when you are unable to obtain a probability sample, but you are still trying to create a sample that is as representative as possible of the population being studied. In this respect, it is the non-probability based equivalent of the stratified random sample.
ii) Unlike probability sampling techniques, especially stratified random sampling, quota sampling is much *quicker* and *easier* to carry out because it does not require a *sampling frame* and the strict use of random sampling techniques.
iii) The quota sample improves the *representation* of particular *strata* (*groups*) within the *population*, as well as ensuring that these *strata* are *not over-represented*. For example, it would ensure that we have sufficient male students taking part in the research (60% of our *sample size* of 100; hence, 60 male students). It would also make sure we did not have more than 60 male students, which would result in an *over-representation* of male students in our research.
iv) It allows *comparison* of *groups*.

Disadvantages of quota sampling

i) In quota sampling, the *sample* has not been chosen using *random selection*, which makes it impossible to determine the possible *sampling error*.
ii) this *sampling bias*. Thus no*statistical inferences* from the *sample* to the *population*. This can lead to problems of *external validity*.
iii) Also, with quota sampling it must be possible to clearly divide the *population* into *strata*; that is, *each unit* from the population must only belong to *one stratum*. In our example, this would be fairly simple, since our *strata* are *male* and *female* students. Clearly, a student could only be classified as either male or female. No student could fit into both categories (ignoring transgender issues).

c)..Purposive Sampling

Purposive sampling is a sampling method in which elements are chosen based on purpose of the study. Purposive sampling may involve studying the entire population of some limited group or a subset of a population. As with other non-probability sampling methods, purposive sampling does not produce a sample that is representative of a larger population, but it can be exactly what is needed in some cases - study of organization, community, or some other clearly defined and relatively limited group.

Advantages of purposive sampling

i) There are a wide range of *qualitative research designs* that researchers can draw on. Achieving the goals of such qualitative research designs requires different types of *sampling strategy* and *sampling technique*. One of the major benefits of purposive sampling is the wide range of sampling techniques that can be used across such qualitative research designs; purposive sampling techniques that range from *homogeneous sampling* through to *critical case sampling, expert sampling,* and more.

ii) Whilst the various purposive sampling techniques each have different goals, they can provide researchers with the justification to make *generalisations* from the sample that is being studied, whether such generalisations are *theoretical*, *analytic* and/or *logical* in nature. However, since each of these types of purposive sampling differs in terms of the nature and ability to make generalisations, you should read the articles on each of these purposive sampling techniques to understand their relative advantages.

iii) Qualitative research designs can involve multiple phases, with each phase building on the previous one. In such instances, different types of sampling technique may be required at each phase. Purposive sampling is useful in these instances because it provides a wide range of non-probability sampling techniques for the researcher to draw on. For example, *critical case sampling* may be used to investigate whether a phenomenon is worth investigating further, before adopting an *expert sampling* approach to examine specific issues further.

Disadvantages of purposive sampling

i) Purposive samples, irrespective of the type of purposive sampling used, *can be* highly prone to *researcher bias*. The idea that a purposive sample has been created based on the *judgement* of the researcher is not a good defence when it comes to alleviating possible researcher biases,

ii) specially when compared with *probability sampling* techniques that are designed to reduce such biases. However, this judgemental, subjective component of purpose sampling is only a major disadvantage when such judgements are *ill-conceived* or *poorly considered*; that is, where judgements have not been based on clear criteria, whether a theoretical framework, expert elicitation, or some other accepted criteria.

iii) The subjectivity and non-probability based nature of *unit* selection (i.e. selecting people, cases/organisations, etc.) in purposive sampling means that it can be difficult to defend the representativeness of the sample. In other words, it can be difficult to convince the reader that the judgement you used to select units to study was appropriate. For this reason, it can also be difficult to convince the reader that research using purposive sampling achieved *theoretical*/*analytic*/*logical generalisation*. After all, if different units had been selected, would the results and any generalisations have been the same?

d)..Snowball Sampling

Snowball sampling is a method in which a researcher identifies one member of some population of interest, speaks to him/her, and then asks that person to identify others in the population that the researcher might speak to. This person is then asked to refer the researcher to yet another person, and so on.

Snowball sampling is very good for cases where members of a special population are difficult to locate. For example,.*populations* that are subject to social stigma and marginalisation, such as suffers of AIDS/HIV, as well as individuals engaged in illicit or illegal activities, including prostitution and drug use. Snowball sampling is useful in such scenarios because:

The method creates a sample with questionable representativeness. A researcher is not sure who is in the sample. In effect snowball sampling often leads the researcher into a realm he/she knows little about. It can be difficult to determine how a sample compares to a larger population. Also, there's an issue of who respondents refer you to - friends refer to friends, less likely to refer to ones they don't like, fear, etc.

Snowball sampling is a useful choice of *sampling strategy* when the *population* you are interested in studying is *hidden* or *hard-to-reach*.

Advantages of Snowball Sampling

(i) The chain referral process allows the researcher to reach populations that are difficult to sample when using other sampling methods.

(ii) The process is cheap, simple and cost-efficient.

(iii) This sampling technique needs little planning and fewer workforce compared to other sampling techniques.

Disadvantages of Snowball Sampling

(i) The researcher has little control over the sampling method. The subjects that the researcher can obtain rely mainly on the previous subjects that were observed.

(ii) Representativeness of the sample is not guaranteed. The researcher has no idea of the true distribution of the population and of the sample.

(iii) Sampling bias is also a fear of researchers when using this sampling technique. Initial subjects tend to nominate people that they know well. Because of this, it is highly possible that the subjects share the same traits and characteristics, thus, it is possible that the sample that the researcher will obtain is only a small subgroup of the entire populatio

1.4.5 Limitations of Sampling

a) Sampling frame: may need complete enumeration
b) Errors of sampling may be high in small areas
c) May not be appropriate for the study objectives/questions
d) Representativeness may be vague, controversial

1.4.6 Characteristics of Good sampling

A good sample should;

a) Meet the requirements of the study objectives
b) Provides reliable results
c) Clearly understandable
d) Manageable/realistic: could be implemented
e) Time consideration: reasonable and timely
f) Cost consideration: economical
g) Interpretation: accurate, representative
h) Acceptability

1.5 Survey Administration

1.5.1 Steps in Survey

1. Setting the study objectives; What are the objectives of the study? Is survey the best procedure to collect data? Why other study design (experimental, quasi-experimental, community randomized trials, epidemiologic designs,,e.g., case-control study) is not appropriate for the study? What information/data need to be collected?

2. **Defining the study *population;*** Representativeness Sampling frame

3. Decide sample design: alternative considerations

4. Questionnaire design; Appropriateness, acceptability, culturally appropriate, understandable Pre-test: Appropriate, acceptable, culturally appropriate, will answer

5. Fieldwork; Training/Supervision Quality monitoring Timing: seasonality

6. Quality assurance Every steps Minimizing errors/bias/cheating

7. Data entry/compilation Validation Feedback

8. Analysis: Design consideration

9. Dissemination

10. Plans for next survey: what did you learn, what did you miss?

1.5.2 Modes of Survey Administration

a) Self-Administered Surveys
b) Personal interview
c) Telephone
d) Mail
e) Computer assisted self-interviewing(CASI)Variants: CAPI (personal interview); CATI (telephone interview) – Replaces the papers
f) Combination of methods

a)..Self-Administered Surveys

Self-administered surveys have special strengths and weaknesses.

They are useful in describing the characteristics of a large population and make large samples feasible.

Advantages:

i) ***Low cost.*** Extensive training is not required to administer the survey. Processing and analysis are usually simpler and cheaper than for other methods.

ii) ***Reduction in biasing error.*** The questionnaire reduces the bias that might result from personal characteristics of interviewers and/or their interviewing skills.

iii) ***Greater anonymity***. Absence of an interviewer provides greater anonymity for the respondent. This is especially helpful when the survey deals with sensitive issues.

iv) Convenience to the respondents (may complete any time at his/her own convenient time)

v) Accessibility (greater coverage, even in the remote areas)

vi) May provide more reliable information (e.g., may consult with others or check records to avoid recall bias)

Disadvantages:

i) ***Requires simple questions.*** The questions must be straightforward enough to be comprehended solely on the basis of printed instructions and definitions.

ii) ***No opportunity for probing.*** The answers must be accepted as final. Researchers have no opportunity to clarify ambiguous answers.

iii) ***Low response rate;*** respondents may not respond to all questions and/or may not return questionnaire

iv) The respondent must be literate to read and understand the questionnaire

v) Introduce self selection bias

vi) Not suitable for complex questionnaire

b). . .Interview Surveys

Unlike questionnaires interviewers ask questions orally and record respondents' answers. This type of survey generally decreases the number of **—**"do not know" and **—**"no answer" responses, compared with self-administered surveys. Interviewers also provide a guard against confusing items. If a respondent has misunderstood a question, the interviewer can clarify, thereby obtaining relevant responses.

Interviewer selection: background characteristics (race, sex, education, culture) listening skill recording skill experience unbiased observation/recording

Interviewer training: be familiar with the study objectives and significance thorough familiarity with the questionnaire contextual and cultural issues privacy and confidentiality informed consent and ethical issues unbiased view mock interview session

Supervision of the interviewer: Spot check Questionnaire check Reinterview (reliability check)

Advantages

i) ***Flexibility.*** Allows flexibility in the questioning process and allows the interviewer to clarify terms that are unclear.

ii) ***Control of the interview situation.*** Can ensure that the interview is conducted in private, and respondents do not have the opportunity to consult one another before giving their answers.

iii) ***High response rate.*** Respondents who would not normally respond to a mail questionnaire will often respond to a request for a personal interview.

iv) May record non-verbal behaviour, activities, facilities, contexts

v) Complex questionnaire may be used

vi) Illiterate respondents may participate

Disadvantages

i) ***Higher cost.*** Costs are involved in selecting, training, and supervising interviewers; perhaps in paying them; and in the travel and time required to conduct interviews.

ii) ***Interviewer bias.*** The advantage of flexibility leaves room for the interviewer's personal influence and bias, making an interview subject to interviewer bias.
iii) ***Lack of anonymity.*** Often the interviewer knows all or many of the respondents. Respondents may feel threatened or intimidated by the interviewer, especially if a respondent is sensitive to the topic or to some of the questions.
iv) Less accessibility
v) Inconvenience
vi) Often no opportunity to consult records, families, relatives

c).. Telephone Interview

Advantages:

(i) Less expensive
(ii) Shorter data collection period than personal interviews
(iii) Better response than mail surveys

Disadvantages:

(i) Biased against households without telephone, unlisted number
(ii) Nonresponse
(iii) Difficult for sensitive issues or complex topics
(iv) Limited to verbal responses

d)… Focus Groups

Focus groups are useful in obtaining a particular kind of information that would be difficult to obtain using other methodologies. A focus group typically can be defined as a group of people who possess certain characteristics and provide information of a qualitative nature in a focused discussion

Focus groups generally are composed of six to twelve people. Size is conditioned by two factors: the group must be small enough for everyone to participate, yet large enough to provide diversity. This group is special in terms of purpose, size, composition, and procedures. Participants are selected because they have certain characteristics in common that relate to the topic at hand, such as parents of gang members, and, generally, the participants are unfamiliar with each other. Typically, more than one focus group should be convened, since a group of seven to twelve people could be too atypical to offer any general insights on the gang problem.

A trained moderator probes for different perceptions and points of view, without pressure to reach consensus. Focus groups have been found helpful in assessing needs, developing plans, testing new ideas, or improving existing programs

Advantages:

i) Flexibility allows the moderator to probe for more in-depth analysis and ask participants to elaborate on their responses.
ii) Outcomes are quickly known.
iii) They may cost less in terms of planning and conducting than large surveys and personal interviews.

Limitations

i) A skilled moderator is essential.
ii) Differences between groups can be troublesome to analyze because of the qualitative nature of the data.
iii) Groups are difficult to assemble. People must take the time to come to a designated place at a particular time.
iv) Participants may be less candid in their responses in front of peers.

1.6 Sample Size Determination

Sample Size Determination is influenced factors like the purpose of the study, population size, the risk of selecting a "bad" sample, and the allowable sampling error.

There are several approaches to determining the sample size. These include using a census for small populations, imitating a sample size of similar studies, using published tables, and applying formulas to calculate a sample size.

Using a census for small populations
One approach is to use the entire population as the sample. It's impractical for large populations. A census eliminates sampling error and provides data on all the individuals in the population. Finally, virtually the entire population would have to be sampled in small populations to achieve a desirable level of precision

Using a sample size of a similar study
Another approach is to use the same sample size as those of studies similar to the one you plan. Without reviewing the procedures employed in these studies you may run the risk of repeating errors that were made in determining the sample size for another study. However, a review of the literature in your discipline can provide guidance about "typical" sample sizes which are used.

Using published tables
One can also rely on published tables which provide the sample size for a given set of criteria. Yamane, 1967 Table 2.1 and Table 2.2 present sample sizes that would be necessary for given combinations of precision, confidence levels, and variability.
NB, i) these sample sizes reflect the number of *obtained* responses, and not necessarily the number of surveys mailed or interviews planned (this number is often increased to compensate for non-response). Ii) the sample sizes in Table 2.2 presume that the attributes being measured are distributed normally or nearly so. If this assumption cannot be met, then the entire population may need to be surveyed.

Using formulas to calculate a sample size
Yamane (1967) provides a simplified formula to calculate sample sizes. A 95% confidence level and P = .5 are assumed for this Equation. $n = \frac{N}{1+Ne^2}$ Where n is the sample size, *N* is the population size, and e is the level of precision

2. Data Presentation

2.1 Frequency Distributions Tables

Definitions

Raw Data: unprocessed data ie data in its original form.
Frequency Distribution: The organization of raw data in table form with classes and frequencies. Rather it's a list of values and the number of times they appear in the data set. We have grouped and ungrouped frequency distribution tables for large and small data sets respectively.

2.1.1 Construction of Ungrouped Frequency Distributions

- Note the largest and smallest observations in the data
- Starting with the smallest value, tally the observations of each quantity.
- Count the number of tallies for each quantity and record it as frequency.

Example: Construct an ungrouped frequency table for the data below
16 14 15 13 12 14 16 15 15 14 17 16 13 16 15 14 18 13 15 17

Solution

Value	12	13	14	15	16	17	18
Tally	/	///	////	~~////~~	////	//	/
Frequency	1	3	4	5	4	2	1

Note ; when tallying ~~////~~ is used for 5 counts and not /////

2.1.2 Construction of Grouped Frequency Distributions

When the number of observations is too large and/or when the variable of interest is continuous, it's cumbersome to consider the repetition of each observation. A quick and more convenient way is to group the range of values into a number of exclusive groups or classes and count the class frequency. The resulting table is called a grouped frequency distribution table.

A grouped frequency distribution consists of *classes* and their corresponding *frequencies.* Each raw data value is placed into a quantitative or qualitative category called a **class.** The **frequency** of a class then is the number of data values contained in a specific class.

Steps in construction

- Select the number of classes k. Choose the smallest integer k such that

  $2^k > n \Rightarrow k > \dfrac{\log n}{\log 2}$ Eg if n = 30 $k > \dfrac{\log 30}{\log 2} \Rightarrow k = 5$
- Identify the largest and smallest observation and compute the ranrge R=largest –smallest value.
- Identify the smallest unit of measurement (u) used in the data collection (ie the accuracy of the measurement.)

Eg for the data 10 30 20 50 30 60 u=10 For the data 12 15 11 17 13 u=1

For the data 1.6 3.2 2.8 5.6 3.5 1.6 u=0.1

- Estimate the class width/interval. i = round up $\left(\frac{R}{k}\right)$ to the nearest u

  Eg round up 3.13 t0 the nearest 0.1 is 3.2
- Pick the starting value (lower class limit of the 1st class (LCL$_1$)) as the smallest value used in the computation of R above. Successive LCL$_s$ are got by adding I to the previous LCL
- Find the upper class limit of the 1st class (UCL$_1$)) by subtracting u from LCL$_2$. Successive UCL$_s$ are got by adding I to the previous UCL
- If necessary find the class boundaries as follows $LCB = LCL - \frac{1}{2}u \quad and \quad UCB = UCL + \frac{1}{2}u$
- Tally the number of observations falling in each class and record the frequency.

  NB a value x fall into a class $LCL - UCL \quad if \quad LCB \le x \le UCB$

Example 1

Organize the data below into a grouped frequency table.

15.0 17.4 10.3 9.2 20.7 18.9 16.6 22.4 23.7 18.6 26.1 16.5 19.7 12.9 15.7
30.8 15.4 20.3 24.0 29.6 18.3 23.7 17.8 24.6 23.0 21.4 32.8 12.5 17.5 18.3
23.2 21.6 20.8 29.8 24.5 28.4 13.5 17.1 27.1 27.9

Solution

$u = 0.1 \quad n = 40 \quad \Rightarrow \quad k > \frac{\log 40}{\log 2} \Rightarrow k = 6 \qquad \text{Range} = 32.8 - 9.2 = 23.6$

$\text{i} = \text{round up } \left(\frac{23.6}{6}\right) \text{to the nearest } 0.1 \; = 4.0$

$\text{Now } \text{LCL}_1 = 9.2 \;\Rightarrow\; \text{LCL}_2 = 13.2 \quad \text{LCL}_3 = 17.2 \text{ etc}$

$\Rightarrow \; \text{UCL}_1 = \text{LCL}_2 - u = 13.2 - 0.1 = 13.1, \quad \text{UCL}_2 = 17.1 \quad \text{UCL}_3 = 21.1 \text{ etc}$

The frequency table is as shown belo

Class	**Boundaries**	**tally**	**Freq**	**C.F**
9.2 - 13-1	9.15 - 13-15	////	4	4
13.2 - 17.1	13.15 - 17.15	~~////~~ //	7	11
17.2 - 21-1	17.15 - 21-15	~~////~~ ~~////~~ /	11	22
21.2 - 25.1	21.15 - 25.15	~~////~~ ~~////~~	10	32
25.2 - 29.1	25.15 - 29.15	////	4	36
29.2 - 33-	29.15 - 33-15	////	4	40

Sometimes due to convenience, it may be necessary to slightly lower the starting value. Eg in the above case we may use 9.0 in place of 9.2.

Example 2

These data represent the record high temperatures in degrees Fahrenheit (F) for each of the 50 states. Construct a grouped frequency distribution for the data.

112 100 127 120 134 118 105 110 109 112 110 118 117 116 118 122 114 114 105 109 107 112 114 115 118 117 118 122 106 110 116 108 110 121 113 120 119 111 104 111 120 113 120 117 105 110 118 112 114 114

Solution

$u = 1 \quad n = 50 \;\Rightarrow\; k > \frac{\log 50}{\log 2} \;\Rightarrow\; k = 6 \qquad \text{Range} = 134 - 100 = 34$

i = round up $\left(\frac{34}{6}\right)$ to the nearest1 $= 6$

Now $\text{LCL}_1 = 100 \;\Rightarrow\; \text{LCL}_2 = 106 \quad \text{LCL}_3 = 112$ etc

$\Rightarrow\; U\text{CL}_1 = \text{LCL}_2 - u = 106 - 1 = 105, \quad \text{UCL}_2 = 111 \quad \text{UCL}_3 = 117$ etc

The frequency table is as shown below

Class	**Boundaries**	**tally**	**Freq**	**C.F**
100 - 105	99.5 – 105.5	~~////~~	5	5
106 - 111	105.5 - 111.5	~~////~~ ~~////~~ ///	13	18
112 - 117	111.5 – 117.5	~~////~~ ~~////~~ ~~////~~ //	16	34
118 - 123	117.5 -123.5	~~////~~ ~~////~~ ////	14	48
124 - 129	124.5 - 129.5	/	1	49
130 - 135	129.5 - 135-5	/	1	50

Exercise

1) The data shown here represent the number of miles per gallon (mpg) that 30 selected four-wheel-drive sports utility vehicles obtained in city driving. Construct a frequency distribution, and analyze the distribution. 12 17 12 14 16 18 16 18 12 16 17 15 15 16 12 15 16 16 12 14 15 12 15 15 19 13 16 18 16 14
2) Suppose a researcher wished to do a study on the ages of the top 50 wealthiest people in the world. The researcher first would have to get the data on the ages of the people. In this case, these ages are listed in *Forbes Magazine*. 49 57 38 73 81 74 59 76 65 69 54 56 69 68 78 65 85 49 69 61 48 81 68 37 43 78 82 43 64 67 52 56 81 77 79 85 40 85 59 80 60 71 57 61 69 61 83 90 87 74 Organize the data into a grouped frequency table
3) The data represent the ages of our Presidents at the time they were first inaugurated. 57 61 57 57 58 57 61 54 68 51 49 64 50 48 65 52 56 46 54 49 50 47 55 55 54 42 51 56 55 54 51 60 62 43 55 56 61 52 69 64 46 54
   a) Were the data obtained from a population or a sample? Explain your answer.
   b) What was the age of the oldest and youngest President?
   c) Construct a frequency distribution for the data.

d) Are there any peaks in the distribution?

e) identify any possible outliers.

4) The state gas tax in cents per gallon for 25 states is given below. Construct a grouped frequency distribution for the data. 7.5 16 23.5 17 22 21.5 19 20 27.1 20 22 20.7 17 28 20 23 18.5 25.3 24 31 14.5 25.9 18 30 31.5

5) Listed are the weights of the NBA's top 50 players. Construct a grouped frequency distribution and analyze the results in terms of peaks, extreme values, etc.
240 210 220 260 250 195 230 270 325 225 165 295 205 230 250 210 220 210
230 202 250 265 230 210 240 245 225 180 175 215 215 235 245 250 215 210
195 240 240 225 260 210 190 260 230 190 210 230 185 260

6) The number of stories in each of the world's 30 tallest buildings is listed below. Construct a grouped frequency distribution and analyze the results in terms of peaks, extreme values, etc. 88 88 110 88 80 69 102 78 70 55 79 85 80 100 60 90 77 55 75 55 54 60 75 64 105 56 71 70 65 72

7) The average quantitative GRE scores for the top 30 graduate schools of engineering are listed. Construct a grouped frequency distribution and analyze the results in terms of peaks, extreme values, etc. 767 770 761 760 771 768 776 771 756 770 763 760 747 766 754 771 771 778 766 762 780 750 746 764 769 759 757 753 758 746

8) The number of passengers (in thousands) for the leading U.S. passenger airlines in 2004 is indicated below. Use the data to construct a grouped frequency distribution with a reasonable number of classes and comment on the shape of the distribution.

9) 91.570 86. 755 81.066 70.786 55.373 42.400 40.551 21.119 16.280 14.869 13.659 13.417 3.170 12.632 11.731 10.420 10.024 9.122 7.041 6.954 6.406 6.362 5. 930 5.585

10) The heights (in feet above sea level) of the major active volcanoes in Alaska are given here. Construct a grouped frequency distribution for the data. 4,265 3,545 4,025 7,050 11,413 3,490 5,370 4,885 5,030 6,830 4,450 5,775 3,945 7,545 8,450 3,995 10,140 6,050 10,265 6,965 150 8,185 7,295 2,015 5,055 5,315 2,945 6,720 3,465 1,980 2,560 4,450 2,759 9,430 7,985 7,540 3,540 11,070 5,710 885 8,960 7,015

2.2 Graphical Displays

After you have organized the data into a frequency distribution, you can present them in graphical form. The purpose of graphs in statistics is to convey the data to the viewers in pictorial form. It is easier for most people to comprehend the meaning of data presented graphically than data presented numerically in tables or frequency distributions. This is especially true if the users have little or no statistical knowledge.

Statistical graphs can be used to describe the data set or to analyze it. Graphs are also useful in getting the audience's attention in a publication or a speaking presentation. They can be used to discuss an issue, reinforce a critical point, or summarize a data set. They can also be used to discover a trend or pattern in a situation over a period of time.

The commonly used graphs in research are; the pie chart, bar chart, histogram, frequency polygon and the cumulative frequency curve (Ogive).

2.2.1 Pie Chart

It's a circular graph having radii divide a circle into sectors proportional in angle to the relative size of the quantities in the category being represented.

How to Draw

(i) Add up the given quantities and let s be the sum of the values

(ii) For each quantity x, calculate the representative angle and percentage as $\frac{x}{s}(360^0)$ and $\frac{x}{s}(100\%)$ respectively

(iii)Draw a circle and divide it into sectors using the angles calculated in step ii above

(iv)Label the sector by the group represented and indicate the corresponding percentage.

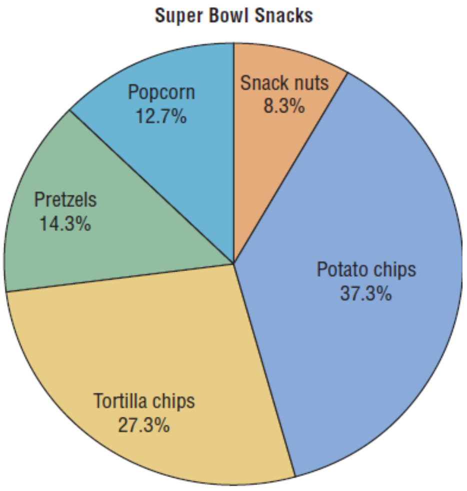
Example

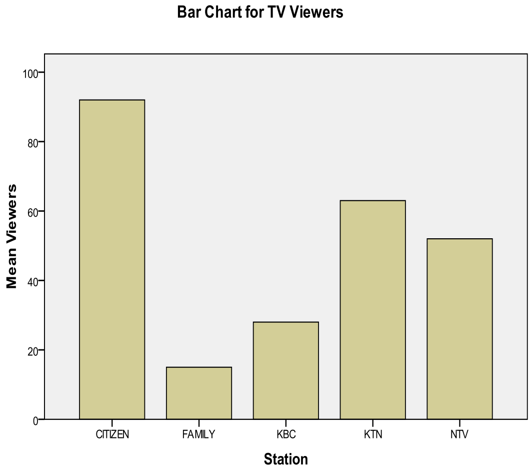
This frequency distribution shows the number of pounds of each snack food eaten during the Super Bowl. Construct a pie graph for the data.

Snack	Potato chips	Tortilla chips	Pretzels	Popcorn	Snack nuts
Pounds (in millions)	11.2	8.2	4.3	3.8	2.5

Solution

Snack	Potato chips	Tortilla chips	Pretzels	Popcorn	Snack nuts	Total
Pounds (in millions)	11.2	8.2	4.3	3.8	2.5	30.0
Representative Angle	134	98	52	46	30	360
Representative %age	37.3	27.3	14.3	12.7	8.3	99.9





2.2.2 Bar chart

A bar chart consists of a set of equal spaced rectangles whose heights are proportional to the frequency of the category /item being considered. The X axis in a bar chart can represent the number of categories.

Note: Bars are of uniform width and there is equal spacing between the bars.

Example

A sample of 250 students was asked to indicate their favourite TV channels and their responses were as follows.

TV station	KBC	NTV	CITIZEN	KTN	FAMILY
No. of viewers	28	52	92	63	15

Draw a bar chart to represent this information.

Solution see the bar chart above and on the right.

2.23 Pareto Charts

It consist of a set of continuous rectangles where the variable displayed on the horizontal axis is qualitative or categorical and the frequencies are displayed by the heights of vertical bars, which are arranged in order from highest to lowest. A **Pareto chart** is used to represent a frequency distribution for a categorical variable,

Points to note when drawing a Pareto Chart

i) Make the bars the same width.

ii) Arrange the data from largest to smallest according to frequency.

iii) Make the units that are used for the frequency equal in size.

When you analyze a Pareto chart, make comparisons by looking at the heights of the bars.

Example

The table shown here is the average cost per mile for passenger vehicles on state turnpikes. Construct and analyze a Pareto chart for the data.

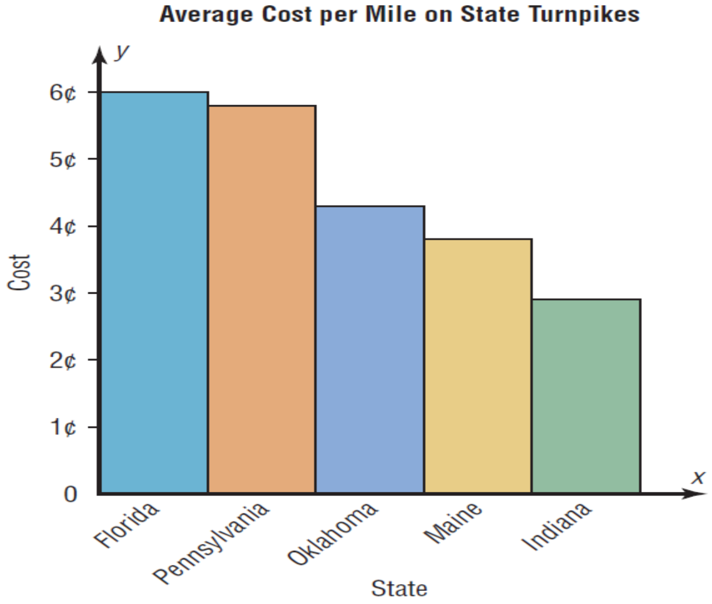
State	Indiana	Oklahoma	Florida	Maine	Pennsylvania
Number	2.9	4.3	6.0	3.8	5.8

Solution

Arrange the data from the largest to smallest according to frequency.

State	Florida	Pennsylvania	Oklahoma	Maine	Indiana
Number	6.0	5.8	4.3	3.8	2.9

Draw and label the *x* and *y* axes and then the bars corresponding to the frequencies. The Pareto chart shows that Florida has the highest cost per mile. The cost is more than twice as high as the cost for Indiana.



2.2.4 Histogram

It consists of a set of continuous rectangles such that the areas of the rectangles are proportional to the frequency. For ungrouped data, the heights of each bar is proportional to frequency. For grouped data, the height of each rectangle is the relative frequency *h* and is given by $\text{h} = \dfrac{\text{frequency}}{\text{IClass Interval}}$. The width of the bars is determined by the class boundaries.

Example

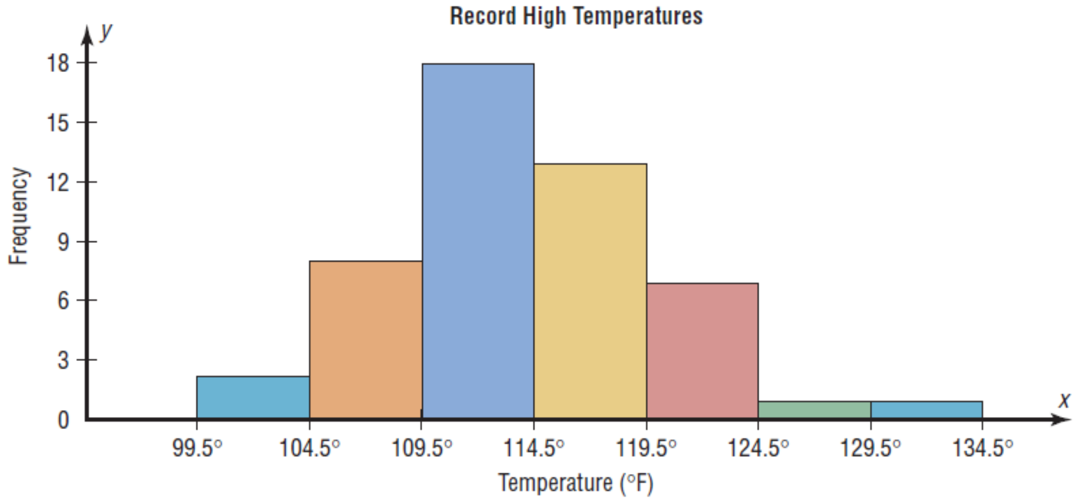
Construct a histogram to represent the data shown below

Class	100-104	105-109	110 -114	115-119	120 - 24	125-129	130 -134

Freq	2	8	18	13	7	1	1

Solution

Boundaries	99.5-104.5	104.5-109.5	109.5 - 114.5	114.5-119.5	*119.5 – 124.5*	*124.5-129.5*	129.5 - 134.5
Heights	2	8	18	13	7	1	1



2.2.5 Frequency polygon

It's a plot of frequency against mid points joined with straight line segments between consecutive points. It can also be obtained by joining the mid point of the tops of the bars in a histogram. The gaps at both ends are filled by extending to the next lower and upper imaginary classes assuming frequency zero.

Example: Consider the following frequency distribution.

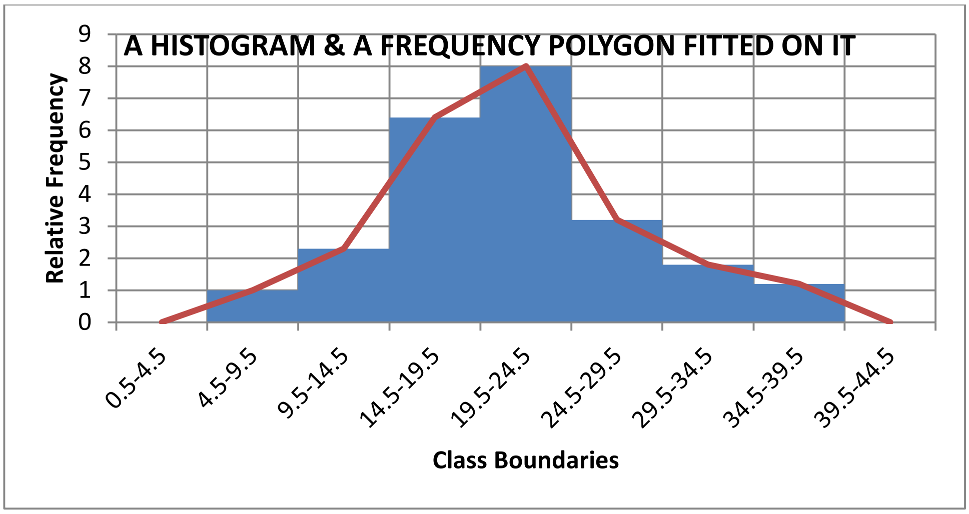
Class	5-9	10-14	15-19	20-24	25-29	30-34	35-39
freq	5	12	32	40	16	9	6

Draw a histogram to represent this information and fit a frequency polygon on it.

Solution

Boundaries	4.5-9.5	9.5-14.5	14.5-19.5	19.5-24.5	24.5-29.5	29.5-34.5	34.5-39.5
heights	1	2.4	6.4	8	3.2	1.8	1.2

The corresponding histogram is as shown below.



2.2.6 Cumulative frequency curve (ogive)

It is a plot of cumulative frequency against upper boundaries joined with a smooth curve. The gap on the lower end is filled by extending to the next lower imaginary class assuming frequency zero. This graph is useful in estimating median and other measures of location.

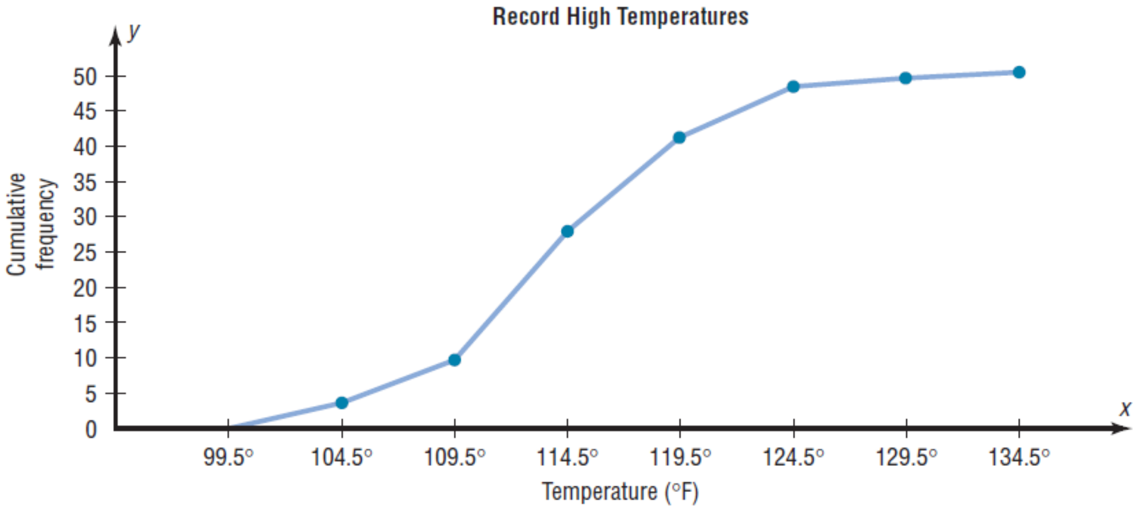
Example:

Construct an ogive to represent the data shown below

Class	100-104	105-109	110 -114	115-119	120 - 24	125-129	130 -134
Freq	2	8	18	13	7	1	1

Solution

Upper Boundaries	**99.5**	104.5	109.5	114.5	*119.5*	*124.5*	*129.5*	134.5
CF	**0**	2	10	28	41	48	49	50



Exercise

1) Construct a pie chart and a bar graph showing the blood types of the army inductees described in the frequency distribution is repeated here.

Blood group	A	B	AB	O
Frequency	5	7	4	9

2) The table below shows the average money spent by first-year college students. Draw a pie chart and a bar graph for the data.

Nature of Expense	Electronics	Dorm decor	Clothing	Shoes
Amount(in $)	728	344	141	72

3) The table shown here is the average cost per mile for passenger vehicles on state turnpikes. Draw a pie chart and a bar graph for the data.

State	Indiana	Oklahoma	Florida	Maine	Pennsylvania
Number	2.9	4.3	6.0	3.8	5.8

4) The following data are based on a survey from American Travel Survey on why people travel. Construct a pie chart a bar graph and a Pareto chart for the data and analyze the results.

Purpose	Personal business	Visit friends or relatives	Work-related	Leisure
Number	146	330	225	299

5) The following percentages indicate the source of energy used worldwide. Construct a Pareto chart and a vertical pie chart, a bar graph and a Pareto graph for the energy used.

Energy Type	Petroleum	Coal	Dry natural gas	Hydroelectric	Nuclear	Others
percentage	39.8	23.2	22.4	7.0	6.4	1.2

6) The following elements comprise the earth's crust, the outermost solid layer. Illustrate the composition of the earth's crust with a pie chart and a bargraph for this data.

Element	Oxygen	Silicon	Aluminum	Iron	Calcium	Others
percentage	45.6	27.3	8.4	6.2	4.7	7.8

7) The sales of recorded music in 2004 by genre are listed below. Represent the data with an appropriate graph.

Rock	Country	Rap/hip-hop	R&B/urban	Pop	Religious	Children's	Jazz	Classical	Oldies	Soundtracks	New age	Others
23.9	13.0	12.1	11.3	10.0	6.0	2.8	2.7	2.0	1.4	1.1	1.0	8.9

8) The top 10 airlines with the most aircraft are listed. Represent these data with an appropriate graph.

American	Continental	United	Southwest	Northwest	American Eagle	U.S. Airways	Lufthansa (Ger.)
714	364	603	327	424	245	384	233

9) The top prize-winning countries for Nobel Prizes in Physiology or Medicine are listed here. Represent the data with an appropriate graph.

United States	Denmark	United Kingdom	Austria	Germany	Belgium	Sweden	Italy	France	Australia	Switzerland
80	5	24	4	16	4	8	3	7	3	6

10) Construct a histogram, frequency polygon, and an ogive for the distribution (shown here) of the miles that 20 randomly selected runners ran during a given week.

Class	6-10	11-15	16 -20	21-25	26 - 30	31-35	36 -40
Freq	1	2	3	5	4	3	2

11) For 108 randomly selected college applicants, the following frequency distribution for entrance exam scores was obtained. Construct a histogram, frequency polygon, and ogive for the data.

Class	90-98	99-107	108-116	117125	126-134
Freq	6	22	43	28	9

Applicants who score above 107 need not enrol in a summer developmental program. In this group, how many students do not have to enroll in the developmental program?

12) Thirty automobiles were tested for fuel efficiency, in miles per gallon (mpg). The following frequency distribution was obtained. Construct a histogram, a frequency polygon, and an ogive for the data.

Class	8-12	13-17	18-22	23-27	28-32
Freq	3	5	15	5	2

13) The salaries (in millions of dollars) for 31 NFL teams for a specific season are given in this frequency distribution.

Class	39.9-42.8	42.9-45.8	45.9-48.8	48.9-51.8	51.9-54.8	54.9-57.8
Freq	2	2	5	5	12	5

Construct a histogram, a frequency polygon, and an ogive for the data; and comment on the shape of the distribution.

14) In a study of reaction times of dogs to a specific stimulus, an animal trainer obtained the following data, given in seconds. Construct a histogram, a frequency polygon, and an ogive for the data; analyze the results.

Class	2.3-2.9	3.0-3.6	3.7-4.3	4.4-5.0	5.1-5.7	5.8-6.4
Freq	10	12	6	8	4	2

15) The animal trainer in question above selected another group of dogs who were much older than the first group and measured their reaction times to the same stimulus. Construct a histogram, a frequency polygon, and an ogive for the data.

Class	2.3-2.9	3.0-3.6	3.7-4.3	4.4-5.0	5.1-5.7	5.8-6.4
Freq	1	3	4	16	14	4

Analyze the results and compare the histogram for this group with the one obtained in the above question. Are there any differences in the histograms?

16) The frequency distributions shown indicate the percentages of public school students in fourth-grade reading and mathematics who performed at or above the required proficiency levels for the 50 states in the United States. Draw histograms for each, and decide if there is any difference in the performance of the students in the subjects.

Class	18-22	23-27	28-32	33-37	38-42	43-48
Reading Freq	7	6	14	19	3	1
Math Freq	5	9	11	16	8	1

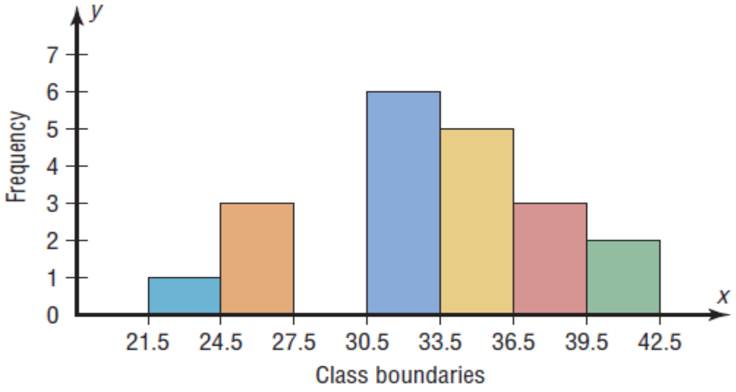
Using the histogram shown here, Construct a frequency distribution; include class limits, class frequencies, midpoints, and cumulative frequencies. Hence answer these questions.

a) How many values are in the class 27.5–30.5?

b) How many values fall between 24.5 and 36.5?

c) How many values are below 33.5?

d) How many values are above 30.5?



3 NUMERICAL SUMMARIES

A numerical summary for a set of data is referred to as a statistic if the data set is a sample and a parameter if the data set is the entire population.

Numerical summaries are categorized as measures of location and measures of spread. Measures of location can further be classified into measures of central tendancy and measures of relative positioning (quantiles).

3.1 Measures of Location

Before discussing the measures of location, its important to consider summation notation and indexing

Index (subscript) Notation: Let the symbol x_i (read 'x sub t'*i*) denote any of the n values $x_1, x_2, ..., x_n$ assumed by a variable X. The letter i in x_i (i=1,2,. . . ,n) is called an index or subscript. The letters j, k, p, q or s can also be used.

Summation Notation: $\sum_{i=1}^{n} x_i = x_1 + x_2 + ...+ x_n$

Example: $\sum_{i=1}^{n} X_i Y_i = X_1 Y_1 + X_2 Y_2 + ...+ X_N Y_N$ **and**

$$\sum_{i=1}^{n} aX_i = aX_1 + aX_2 + ...+ aX_N = a(X_1 + X_2 + ...+ X_N) = a\sum_{i=1}^{n} X_i$$

3.1.1 Measures of Central Tendency (Averages)

A Measures of Central Tendency of a set of numbers is a value which best represents it. There are three different types of Central Tendencies namely the mean, median and mode. Each has advantages and disadvantages depending on the data and intended purpose.

Arithmetic Mean

The arithmetic mean of a set of values $x_1, x_2, ..., x_n$, denoted $\bar{x}$ if the data set is a sample, is found by dividing the sum of the set of numbers with the actual number of values. Ie

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + ...+ x_n}{n}$$

Example 1 Find the mean of 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10.

Solution

Sum of values: 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 = 55

Number of values = 10

Mean of values $\bar{x} = \frac{55}{10} = 5.5$

Note: If the numbers $x_1, x_2, ..., x_n$ occur $f_1, f_2, ..., f_n$ times respectively, (occur with frequencies $f_1, f_2, ..., f_n$), the arithmetic mean is, given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} f_i x_i = \frac{f_1 x_1 + f_2 x_2 + ...+ f_n x_n}{f_1 + f_2 + ...+ f_n}$$

where n is the total frequency. This is the formular for the mean of a grouped data.

Example 2 The grades of a student on six examinations were 84, 91, 72, 68, 91 and 72. Find the arithmetic mean.

The arithmetic mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} f_i x_i == \frac{1(84)+2(91)+2(72)+1(68)}{1+2+2+1} = 79.67$

Example 3 If 5, 8, 6 and 2 occur with frequencies 3, 2, 4 and 1 respectively, the arithmetic mean is $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} f_i x_i = \frac{3(5)+2(8)+4(6)+1(2)}{3+2+4+1} = 5.7$

Exercise

1 Find the mean of 9, 3, 4, 2, 1, 5, 8, 4, 7, 3

2 A sample of 5 executives received the following amount of bonus last year: sh 14,000, sh 15,000, sh 17,000, sh 16,000 and sh y. Find the value of y if the average bonus for the 5executives is sh 15,400

Properties of the Arithmetic Mean

(1) The algebraic sum of the deviations of a set of numbers from their arithmetic mean is zero, that is $\sum_{i=1}^{n}\left(x_i - \overline{x}\right) = 0$.

(2) $\sum_{i=1}^{n}(x_i - a)^2$ is minimum if and only if $a = \overline{x}$.

(3) If n_1 numbers have mean $\overline{x}_1$, n_2 numbers have mean $\overline{x}_2$,…, n_k numbers have mean $\overline{x}_k$, then the mean of all the numbers called the combined mean is given by

$$\overline{x}_c = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + ... + n_k\overline{x}_k}{n_{11} + n_2 + ... + n_k} = \frac{\sum n_i \overline{x}_i}{\sum n_i}$$

Median

It's the value below which and above which half of the observations fall when ranked in order of size. The position of the median term is given by $\left(\frac{n+1}{2}\right)^{th}$ Value.

NB if n is even we average the middle 2 terms

For grouped data median is estimated using the formular

$$\text{Median} = \text{LCB} + \left(\frac{\left(\frac{n+1}{2}\right) - Cf_a}{f}\right) \times i$$

where LCB, f and i are the lower class boundary, frequency and class interval of the median class. Cf$_a$ is the cumuilative frequency of the class above the median class.

Remark: The disadvantage of median is that it is not sensitive against changes in the data.

Mode

It’s the value occurring most frequently in a data set. If each observation occurs the same number of times, then there is no mode. When 2 or more observation occurs mos frequently in a data then the data is said to be multimodal.

For ungrouped data it’s very easy to pick out the mode. However If the data is grouped, mode is estimated using the formular

$$\text{Mode} = \text{LCB} + \left(\frac{f - f_a}{2f - f_a - f_b} \right) \times i$$

where LCB, f and i are the lower class boundary, frequency and class interval of the modal class. f_a and f_b are frequencies of the class above and below the modal class. respectively

Example 1

Find the median and mode of the following data: 19, 13, 18, 14, 12, 25, 11, 10, 17, 23, 19.

Solution

Sorted data: 10, 11, 12, 13, 14, 17, 18, 19, 19, 23, 25.

$\text{n} = 11 \text{ thus Median} = \left(\tfrac{11+1}{2}\right)^{th} \text{Value} = 6^{th} \text{ Value} = 17$

Mode=19 since it appears most frequently in this data set as compared to other observations.

Example 2 Find the median and mode of the data: 2, 4, 8, 7, 9, 4, 6, 10, 8, and 5.

Solution

Array: 2, 4, 4, 5, 6, 7, 8, 8, 9, 10

$\text{n} = 10 \text{ thus Median} = \left(\tfrac{10+1}{2}\right)^{th} \text{Value} = 5.5^{th} \text{ Value} = \frac{6+7}{2} = 6.5$

Mode 4 and 8 ie the data is bimodal.

Example 3

Estimate the mean, median and mode for the following frequency distribution:

Class	5-9	10-14	15-19	20-24	25-29	30-34	35-39
Freq	5	12	32	40	16	9	6

Solution

Boundaries	4.5-9.5	9.5-14.5	14.5-19.5	19.5-24.5	24.5-29.5	29.5-34.5	34.5-39.5
Mid pts (x)	7	12	17	22	27	32	37
Frequency	5	12	32	40	16	9	6
Xf	35	144	544	880	432	288	222
CF	5	17	49	89	105	114	120

Mean $\bar{x} = \frac{\sum fx}{n} = \frac{35+144+...+222}{120} = \frac{2545}{120} \approx 21.2083$

$\text{n} = 120 \text{ thus Median} = 60.5^{th} \text{ Value} \Rightarrow$ Median class is 19.5-24.5 thus

$$\text{Median} = \text{LCB} + \left(\frac{\left(\frac{n+1}{2}\right) - Cf_a}{f} \right) \times i = 19.5 + \left(\frac{60.5 - 49}{40} \right) \times 5 = 20.9375$$

The modal class (class with the highest frequency) is 19.5-24.5 therefore

$$\text{Mode} = \text{LCB} + \left(\frac{f - f_a}{2f - f_a - f_b} \right) \times i = 19.5 + 5\left(\frac{40-32}{80-32-16} \right) = 20.75$$

Exercise

1. Find the mean median and mode for the following data: 9, 3, 4, 2, 9, 5, 8, 4, 7, 4
2. Find the mean median and mode of 1, 2, 2, 3, 4, 4, 5, 5, 5, 5, 7, 8, 8 and 9
3. The number of goals scored in 15 hockey matches is shown in the table.

No of goals	1	2	3	4	5
No of matches	2	1	5	3	4

Calculate the mean number of goals cored

4. The table shows the heights of 30 students in a class calculate an estimate of the mean mode and median height.

Height (cm)	140<x<144	144<x<148	148<x<152	152<x<156	156<x<160	160<x<164
No of students	4	5	8	7	5	1

5. Estimate the mean, median and mode for the following frequency distribution:

Class	1-4	5-8	9-12	13-16	17-20	21-24
frequency	10	14	20	16	12	8

Class	40-59	60-79	80-99	100-119	120-139	140-159	160-179	180-199
freq	5	12	28	40	16	9	6	4

The Empirical Relation between the Mean, Median and Mode

$$MEAN - MODE = 3(MEAN - MEDIAN)$$

The above relation is true for unimodal frequency curves which are asymmetrical.

3.1.2 Other Types of Means

These will include weighted, harmonic and geometric means.

The Weighted Arithmetic Mean

The weighted arithmetic mean of a set of n numbers $x_1, x_2, ..., x_n$ having corresponding weights $w_1, w_2, ..., w_n$ is defined as

$$\bar{x}_w = \frac{w_1 x_1 + w_2 x_2 + ... + w_n x_n}{w_{11} + w_2 + ... + w_n} = \frac{\sum w_i x_i}{\sum w_i}$$

Example1 Consider the following table with marks obtained by two students James (mark x) and John (mark y). The weights are to be used in determining who joins the engineering course whose requirement is a weighted mean of 58% on the four subjects below;

Subject	Maths	English	History	Physics	Total
Mark x	25	87	83	30	225
Mark y	70	45	35	75	225
Weight	3.6	2.3	1.5	2.6	10

Working the products of the marks and the weights we get

Subject	Maths	English	History	Physics	Total
Wx	90	200.1	124.5	78	492.6
Wy	252	103.5	52.5	195	603

Now $\overline{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{492.6}{10} = 49.26$ and $\overline{y}_w = \frac{\sum w_i y_i}{\sum w_i} = \frac{603}{10} = 60.3$

Clearly John qualifies but James does not.

Example 2 If a final examination is weighted 4 times as much as a quiz, a midterm examination 3 times as much as a quiz**,** and a student has a final examination grade of 80, a midterm examination grade
of 95 and quiz grades of 90, 65 and 70, the mean grade is

$$\overline{X} = \frac{1(90)+1(65)+1(70)+3(95)+4(80)}{1+1+1+3+4} = \frac{830}{10} = 83.$$

Question A tycoom has 3 house girls who he pays Ksh 4,000 each per month, 2 watch men who he pays Ksh 5,000 each and some garden men who receives Ksh 7,000 each. If he pays out an average of Ksh 5,700 per month to these people, find the number of garden men.

Question A student's grades in laboratory, lecture, and recitation parts of a computer course were 71, 78, and 89, respectively.

(a) If the weights accorded these grades are 2,4, and 5, respectively, what is an average grade?

(b) What is the average grade if equal weights are used?

The Geometric and Harmonic Means

Let $x_1, x_2, ..., x_n$ be the sample values, the geometric mean GM is given by

$$GM = \sqrt[n]{x_1 \times x_2 \times .... \times x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

and the harmonic mean is given by

$$HM = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}} = \frac{n}{\sum\left(\frac{1}{x_i}\right)}.$$

The Relation between the Arithmetic, Geometric and Harmonic Means:

$$HM \leq GM \leq \overline{x}.$$

The formulas for geometric and harmonic means of a frequency distribution are respectively given by;

$$GM = \sqrt[n]{x_1^{f_1} \times x_2^{f_2} \times .... \times x_n^{f_n}} = \sqrt[n]{\prod_{i=1}^{n} x_i^{f_i}} \Rightarrow \log(GM) = \frac{1}{n}\sum_{i-1}^{n} f_i \log x_i$$

and

$$HM = \frac{n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + ... + \frac{f_n}{x_n}} = \frac{n}{\sum\left(\frac{f_i}{x_i}\right)} = \left[\frac{1}{n}\sum\left(\frac{f_i}{x_i}\right)\right]^{-1}$$

where $n = \sum f_i$ and x_i are the midpoints

Example 1: Find the harmonic and the geometric mean of the numbers 2,4 and 8

Solution The geometric mean $GM = \sqrt[3]{2\times4\times8} = \sqrt[3]{64} = 4$ and

the harmonic mean $HM = \dfrac{3}{\frac{1}{2}+\frac{1}{4}+\frac{1}{8}} = \dfrac{3}{\frac{7}{8}} = \dfrac{34}{7} \approx 3.43$

Example 2:

Find the harmonic and geometric mean of the frequency table below

x	13	14	15	16	17
f	2	5	13	7	3

Solution

The harmonic mean $HM = \dfrac{30}{\frac{2}{13}+\frac{5}{14}+\frac{13}{15}+\frac{7}{16}+\frac{3}{17}} \approx 15.$ and

The geometric mean $GM = \sqrt[30]{13^2\times14^5\times15^{13}\times16^7\times17^3} \approx 15.09837$

Exercise:

1. Find the harmonic and the geometric mean of the numbers 10, 12, 15, 5 and 8
2. The number of goals scored in 15 hockey matches is shown in the table below . Calculate the harmonic and geometric mean number of goals scored.

No of goals	1	3	5	6	9
No of matches	2	1	5	3	4

3. Find the harmonic and geometric mean of the frequency table below

Class	0-29	30-49	50-79	80-99
Frequency	20	30	40	10

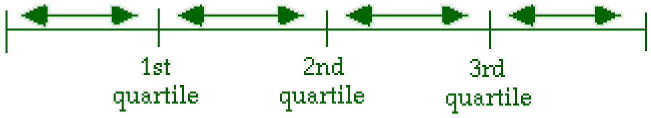
3.1.3 Measures of Relative Positioning (Quantiles)

These are values which divide a sorted data set into N equal parts. They are also known as quantiles or N-tiles. The commonly used quantiles are; **Quartiles, Deciles and Percentiles** These 3 divides a sorted data set into four, ten and hundred divisions, respectively. These measures of position are useful for comparing scores within one set of data. You probably all took some type of college placement exam at some point. If your composite math score was say 28, it might have been reported that this score was in the 94th percentile. What does this mean? This does not mean you received a 94% on the test. It does mean that of all the students who took that exam, 94% of them scored lower than you did (and 6% higher).

Remark For a set of data you can divide the data into three quartiles (Q_1, Q_2, Q_3), nine deciles ($D_1, D_2, ...D_9$) and 99 percentiles ($P_1, P_2, ..., P_{99}$).To work with percentiles, deciles and quartiles - you need to learn to do two different tasks. First you should learn how to find the percentile that corresponds to a particular score and then how to find the score in a set of data that corresponds to a given percentile.

Quartiles

They divide a sorted data set into 4 equal parts and we have lower, middle and upper quartiles denoted Q_1, Q_2 and Q_3 respectively. The lower quartile Q_1 separates the bottom 25% from the top 75%, Q_2 is the median and Q_3 separates the top 25% from the bottom 75% as illustrated below .



The Kth quartile is given by: $Q_k = \frac{k}{4}(n+1)^{th}$ value where k=1,2,3

Deciles and Percentiles

Similarly the Kth Deciles D_k and the Kth Percentiles P_k are respectively given by;

$$D_k = \tfrac{k}{10}(n+1)^{th} \text{ Value and } P_k = \tfrac{k}{100}(n+1)^{th} \text{ value}$$

NB For ungrouped data we may be forced to use linear interpolation for us to get the required Kth quantile. However for grouped data the Kth Value is given by

$$K^{th}Value = \text{LCB} + \left(\frac{K - Cf_a}{f}\right)\times i$$

where LCB, i and f are the lower class boundary. class interval and frequency of the class containing the K^{th} value. Cf$_a$ is the cumuilative frequency of the class above this particular class

Example 1: Find the lower and upper quartiles, the 7th decile and the 85th percentile of the following data. 3, 6, 9, 10, 7, 12, 13, 15, 6, 5, 13

Solution

Sorted data: 3, 5, 6, 6, 7, 9, 10, 12, 13, 13, 15 Here n=11

$$Q_1 = \tfrac{1}{4}(11+1)^{th} = 3^{rd}\, value = 6 \quad \text{Similarly } Q_3 = \tfrac{3}{4}(11+1)^{th} = 9^{th}\, value = 13$$

$$D_7 = \tfrac{7}{10}(11+1)^{th} = 7.7^{th}\, value = \underbrace{7^{th}\, value + 0.7(8^{th}\, value - 7^{th}\, value)}_{\text{linear interpolat ion}} = 10 + 0.7(12-10) = 11.4$$

$$P_{85} = \tfrac{85}{100}(11+1)^{th} = 10.2^{th}\, value = \underbrace{10^{th}\, value + 0.2(11^{th}\, value - 10^{th}\, value)}_{\text{linear interpolat ion}} = 13 + 0.2(15-13) = 13.4$$

Example 2:

Estimate the lower quartile, 4th decile and the 72nd percentile for the frequency table below

Class	1-4	5-8	9-12	13-16	17-20	21-24
frequency	10	14	20	16	12	8

Solution

Boundaries	0.5-4.5	4.5-8.5	8.5-12.5	12.5-16.5	16.5-20.5	20.5-24.5
C.F	10	24	44	60	72	80

For this data n=80

$$Q_1 = \tfrac{1}{4}(80+1)^{th} = 20.25^{th}\, value = 4.5 + \left(\frac{20.25-10}{14}\right)\times 4 \approx 7.428571$$

$$D_4 = \tfrac{4}{10}(80+1)^{th} = 32.4^{th}\, value = 8.5 + \left(\frac{32.4-24}{20}\right)\times 4 \approx 10.18$$

$$P_{72} = \tfrac{72}{100}(80+1)^{th} = 58.32^{th}\ value = 12.5 + \left(\frac{58.32-44}{16}\right)\times 4 \approx 16.08$$

Exercise

a) Find the lower and upper quartiles, the 7th decile and the 85th percentile of the data.

a) 9, 3, 4, 2, 9, 5, 8, 4, 7, 4 b) 1, 2, 2, 3, 4, 4, 5, 5, 5, 5, 7, 8, 8 and 9

2) The number of goals scored in 15 hockey matches is shown in the table.

No of goals	1	2	3	4	5
No of matches	2	1	5	3	4

Estimate the lower quartile, 4th decile and the 72nd percentile of the number of goals cored

4) The table shows the heights of 30 students in a class calculate an estimate of the upper and lower quartile of the height.

Height (cm)	140<x<144	144<x<148	148<x<152	152<x<156	156<x<160	160<x<164
No of students	4	5	8	7	5	1

5) The grouped frequency table gives information about the distance each of 150 people travel to work.

Height (cm)	0<d<5	5<d<10	10<d<15	15<d<20	20<d<25	25<d<30
No of students	4	5	8	7	5	1

a) Work out what percentage of the 150 people travel more than 20 km to work

(b) Calculate an estimate for the median distance travelled to work by the people?

Properties of measures of Location

(i) They are affected by change of origin. Adding or subtracting a constant from each and every observation in a data set causes all the measures of location to shift by the same magnitude. That is $\text{New measure} = \text{old measure} \pm \text{k}$

(ii) They are affected by change of scale. Multiplying each and every observation in a data set by a constant value scales up all the measures of location by the same magnitude.. That is $\text{New measure} = K(\text{old measure})$

Example: Consider the three sets of data A, B and C below

Set A: 65, 53, 42, 52, 53 $\overline{x}_A = 53$ and $\text{Median}_A = 53$

Set B: 15, 3, -8, 2, 3 $\overline{x}_B = 3$ and $\text{Median}_B = 3$

Set C: 45, 9, -24, 6, 9 $\overline{x}_C = 9$ and $\text{Median}_C = 9$

- Notice that set B is obtained by subtracting 50 from each and every observation in set A and clearly $\overline{x}_B = \overline{x}_A - 50$ and $\text{Median}_B = \text{Median}_A - 50$ Therefore $\text{New measure} = \text{old measure} \pm \text{k}$. This is referred to as change of origin.
- Effectively set C is obtained by multiplying each and every observation in set B by 3 and clearly $\overline{x}_C = 3\overline{x}_B$ and $\text{Median}_C = 3\text{Median}_B$ Thus $\text{New measure} = K(\text{old measure})$ This is referred to as change of scale.

3.2 Measures of Spread/ Dispersion

Spread is the degree of scatter or variation of the variable about the central value. Examples of these measures includes: the range, Inter-Quartile range, Quartile

Deviation also called semi Inter-Quartile range, Mean Absolute Deviation, Variance and standard deviation.

Inter-Quartile range and Semi Inter-Quartile Range

Inter-Quartile range (IQR) is the difference between the upper and lower quartiles. Half of this difference is called Quartile Deviation or the semi Inter-Quartile range (SIQR) Ie

$$\text{IQR} = \text{Q}_3\text{-Q}_1 \text{ and } \text{SIQR} = \tfrac{1}{2}\left(\text{Q}_3\text{-Q}_1\right)$$

Mean Absolute Deviation (MAD)

It is the average of the absolute deviations from the mean and it's given by

$\text{MAD} = \dfrac{\sum\left|\text{x - }\overline{\text{x}}\right|}{n}$ for ungrouped data but for grouped data $\text{MAD} = \dfrac{\sum \text{f}\left|\text{x - }\overline{\text{x}}\right|}{n}$

Example 1:

Find the quartile deviation and the mean absolute deviation for the following data.

3, 6, 9, 10, 7, 12, 13, 15, 6, 5, 13

Solution

Sorted data: 3, 5, 6, 6, 7, 9, 10, 12, 13, 13, 15

Recall $Q_1 = 6$ and $Q_3 = 13$ ie from earlier calculations.

Thus $\text{SIQR} = \tfrac{1}{2}\left(\text{Q}_3\text{-Q}_1\right) = \tfrac{1}{2}(13-6) = 3.5$

$$\overline{\text{x}} = \frac{3+5+6+6+7+9+10+12+13+13+15}{11} = 9$$

$$\text{MAD} = \frac{\sum\left(\text{x - }\overline{\text{x}}\right)}{n} = \frac{|3-9|+|5-9|+|6-9|+...+|15-9|}{11} = \frac{6+4+3+...+6}{11} = \frac{36}{11} \approx 3.2727$$

Variance and Standard Deviation

Ignoring the negative sign in order to compute MAD is not the only option we have to deal with deviations. We can square the deviations and then average. The average of the squared deviations from the mean is called the variance denoted s^2 and its given by

$\text{s}^2 = \tfrac{1}{\text{n}}\sum(\text{x - }\overline{\text{x}})^2$ A little algebraic simplification of this formular gives $\text{s}^2 = \tfrac{1}{\text{n}}\sum x^2 - \overline{\text{x}}^2$

For grouped data $\text{s}^2 = \tfrac{1}{\text{n}}\sum \text{f}\,(\text{x - }\overline{\text{x}})^2 = \tfrac{1}{\text{n}}\sum \text{f}x^2 - \overline{\text{x}}^2$ where n is the sum of frequencies.

To reverse the squaring on the units we find the square root of the variance.Standard Deviation denoted s is the square root of variance.

Example 1: Find the variance and standard deviation for the data.

3, 6, 9, 10, 7, 12, 13, 15, 6, 5, 13

Solution

$$\overline{\text{x}} = \frac{3+5+6+6+7+9+10+12+13+13+15}{11} = 9$$

$$\text{S}^2 = \frac{\sum\left(\text{x - }\overline{\text{x}}\right)^2}{n} = \frac{(3-9)^2+(5-9)^2+(6-9)^2+...+(15-9)^2}{11} = \frac{36+16+9+...+36}{11} = \frac{143}{11} = 13$$

Standard deviation $\text{s} = \sqrt{\text{variance} =} = \sqrt{13} \approx 3.60555$.

Example 2 Find the standard deviation of the data: 2, 4, 8, 7, 9, 4, 6, 10, 8, and 5.

Solution

Mean $\bar{x} = \frac{\sum x}{n} = \frac{2+4+8+...+5}{10} = \frac{63}{10} = 6.3$ and $\sum x^2 = 2^2 + 4^2 + 8^2 + ... + 5^2 = 455$

Standard deviation $s = \sqrt{\frac{1}{n}\sum x^2 - \bar{x}^2} = \sqrt{45.5 - 6.3^2} \approx 2.4104$.

Example 3 Estimate the mean, and standard deviation for the frequency table below:

Class	5-9	10-14	15-19	20-24	25-29	30-34	35-39
freq	5	12	32	40	16	9	6

Solution

Mid pts (x)	7	12	17	22	27	32	37	sum
Freq (f)	5	12	32	40	16	9	6	120
xf	35	144	544	880	432	288	222	2545
fx^2	245	1728	9248	19360	11664	9216	8214	59675

Mean $\bar{x} = \frac{\sum fx}{n} = \frac{2545}{120} \approx 21.2083$ and $\sum fx^2 = 59675$

Standard deviation $s = \sqrt{\frac{1}{n}\sum fx^2 - \bar{x}^2} = \sqrt{\frac{59675}{120} - 21.2083^2} \approx 6.8919$.

Exercise

1) Find the quartile deviation, the mean absolute deviation and the standard deviation of the data: a) 9, 3, 4, 2, 9, 5, 8, 4, 7, 4 b) 1, 2, 2, 3, 4, 4, 5, 5, 5, 5, 7, 8, 8 and 9

2) 2) The number of goals scored in 20 hockey matches is shown in the table.

No of goals	1	2	3	4	5
No of matches	2	5	6	3	4

Estimate the quartile deviation, the mean absolute deviation and the standard deviation of the number of goals cored

3) consider the frequency table below and estimate quartile deviation, the mean absolute deviation and the standard deviation

Class	8-12	13-17	18-22	23-27	28-32	33-37
Freq	3	10	12	9	5	1

4) The table shows the heights of 30 students in a class calculate an estimate of the quartile deviation, the mean absolute deviation and the standard deviation of the height.

Height (cm)	140<x<144	144<x<148	148<x<152	152<x<156	156<x<160	160<x<164
No of students	4	5	8	7	5	1

5) The grouped frequency table gives information about the distance each of 150 people travel to work. Calculate an estimate for the quartile deviation and the standard deviation of the distance travelled to work by the people

Height (cm)	0<d<5	5<d<10	10<d<15	15<d<20	20<d<25	25<d<30
# of students	4	5	8	7	5	1

.

Properties of measures of Spread

i) They are not affected by change of origin. Adding or subtracting a constant from each and every observation in a data set does not affect any measures of spread. That is $\text{New measure} = \text{old measure}$

iii) They are affected by change of scale. Multiplying each and every observation in a data set by a constant value scales up all the measures of spread by the same value except in the case of variance which is scaled up by a square of the same constant.

ie $\text{New measure} = K(\text{old measure})$ but $\text{New variance} = \text{k}^2 \times \text{old variance}$

Example: Consider the three sets of data A, B and C below

Set A: 65, 53, 42, 52, 53 Range=23, $\text{MAD}_A = 4.8$ and $\text{Variance}_A = 66.5$

Set B: 15, 3, -8, 2, 3 Range=23, $\text{MAD}_B = 4.8$ and $\text{Variance}_B = 66.5$

Set C: 45, 9, -24, 6, 9 Range=69, $\text{MAD}_C = 14.4$ and $\text{Variance}_C = 598.5$

- Notice that set B is obtained by subtracting 50 from each and every observation in set A and clearly $\text{MAD}_B = \text{MAD}_A$ and $\text{Variance}_B = \text{Variance}_A$ Therefore there is no effect on the change of origin ie $\text{New measure} = \text{old measure}.$.
- Effectively set C is obtained by multiplying each and every observation in set B by 3 and clearly $\text{MAD}_C = 3 \times \text{MAD}_B$ and $\text{Variance}_C = 3^2 \times \text{Variance}_B$ Thus

  $\text{New measure} = K(\text{old measure})$ and $\text{New Variance}_C = k^2 \times \text{old Variance}_B$

Mean and Standard Deviation Using a Calculator

- When on, press mode key to get;

COMP	SD	REG
1	2	3

- Press 2 to select SD for statistical data.
- Enter data one by one pressing m+ after every value entered. The screen will be showing the number of observations that are fully entered.
- Pressing shift then 1 gives

$\sum x^2$	$\sum x$	n
1	2	3

Typing 1 then = gives the value of $\sum x^2$

Similarly typing 2 then = gives the value of $\sum x$

- Pressing shift then 2 gives

$\bar{x}$	$x\sigma_n$	$x\sigma_{n-1}$
1	2	3

Which are the mean uncorrected standard deviation and the corrected standard deviation

Example using your calculator, obtain the mean and standard deviation of the following data: 31, 52, 29, 60, 58

Solution

Entering data 31M+ 52 M+ 29 M+ 60 M+ 58 M+

$\bar{x} = 46$ and $\text{s} = 14.91643$

Question Redo the above example using the data: 235, 693, 484, 118, 470

3.3 Assumed Mean and the Coding Formular

If the observations are too large such that the natural computation of totals is tedious, we can take one of the observations as the working/assumed mean. Let A be any guessed or assumed arithmetic mean and let $d_i = x_i - A$ be the deviations of x_i from A, then

$$\text{mean } \bar{x} = A + \tfrac{1}{n}\sum \text{f}d = A + \bar{d}$$

and

$$\text{Variance } S^2 = \tfrac{1}{n}\sum fd^2 - \left(\tfrac{1}{n}\sum fd\right)^2 = \tfrac{1}{n}\sum fd^2 - \bar{d}^2$$

Respectably where A = Assumed mean which is generally taken as mid point of the middle class or the class where frequency is large

Remark:, in most cases deviations (d) of x_i from A is a multiple of the class interval ie

$$d_i = t_i \times i \;\; \Rightarrow t_i = \frac{d_i}{i} = \frac{x_i - A}{i}.$$

In these cases we can use t rather than d in computation. The above formulae reduces to

$$\bar{x} = A + \tfrac{i}{n}\sum \text{ft} = A + i\bar{t} \quad \text{and}$$

$$S^2 = i^2\left[\tfrac{1}{n}\sum ft^2 - \left(\tfrac{1}{n}\sum ft\right)^2\right] = i^2\left[\tfrac{1}{n}\sum ft^2 - \bar{t}^2\right]$$

respectably the latter formulae are referred to as coding formulae

Example

Using coding formulae, find the mean and standard deviation of the following data

Class	340-349	350-359	360-369	370-379	380-389
Freq	2	3	7	5	3

Solution

Class	Mid pts	Freq	$t = \frac{x - 364.5}{i}$	ft	ft^2
340-349	344.5	2	-2	-4	8
350-359	354.5	3	-1	-3	3
360-369	364.5	7	0	0	0
370-379	374.5	5	1	5	5
380-389	384.5	3	2	6	12
Total		**20**		**4**	**28**

$$\bar{x} = A + \tfrac{i}{n}\sum \text{ft} = 364.5 + \tfrac{10}{20}(4) = 366.5$$

$$S = i \times \sqrt{\tfrac{1}{n}\sum ft^2 - \left(\tfrac{1}{n}\sum ft\right)^2} = 10 \times \sqrt{\tfrac{28}{20} - \left(\tfrac{4}{20}\right)^2} \approx 11.6619$$

Exercise

1) Consider the following frequency distribution.

classes	10-14	15-19	20-24	25-29	30-34
frequency	7	11	14	13	5

Estimate the mean and standard deviation using coding formula

2) Using coding formular, find the mean and standard deviation of the frequency table below

Class	10-20-	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100	100-110
Freq	4	5	7	13	16	11	9	6	4	3

3) The table shows the speed distribution of vehicles on Thika Super high way on a typical day.

Speed (km/hr)	60-69	70-79	80-89	90-99	100-109	110-119	120-129	130-139	140-149
No of vehicles	138	163	325	541	427	214	110	52	30

Using coding formulae, find the mean speed and the standard deviation of the speeds.

4) The following table shows a frequency distribution of the weekly wages of 65 employees at the P&R Company.

Wages	$250.00 -259.99	$260.00 -269.99	$270.00- 279.99	$280.00- 289.99	$290.00- 299.99	$300.00- 309.99	$310.00- 319.99
No. of employees	8	10	16	14	10	5	2

Find the mean wage and the standard deviation of the wages using coding formular

3.4 Measures of Relative Dispersion:

These measures are used in comparing spreads of two or more sets of observations. These measures are independent of the units of measurement. These are a sort of ratio and are called coefficients.

Suppose that the two distributions to be compared are expressed in the same units and their means are equal or nearly equal. Then their variability can be compared directly by using their standard deviations. However, if their means are widely different or if they are expressed in different units of measurement, we can not use the standard deviations as such for comparing their variability. We have to use the relative measures of dispersion in such situations. Examples of these Measures of relative dispersion includes; Coefficient of quartile deviation, Coefficient of mean deviation and the Coefficient of variation

3.4.1 Coefficient of Quartile Deviation and Coefficient of Mean Deviation

The Coefficient of Quartile Deviation of x CQD(x) is given by $\text{CQD}(x) = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100\%$

The Coefficient of Mean Deviation CMD(x) is given by $\text{CMD}(x) = \frac{\text{MAD}}{\text{Mean}} \times 100\%$

3.4.2 Coefficient of Variation:

Coefficient of variation is the percentage ratio of standard deviation and the arithmetic mean. It is usually expressed in percentage. The coefficient of variation of x denoted C.V(x) is given by the formula

$$\text{C.V}(x) = \frac{S}{\bar{x}} \times 100\%$$

where $\bar{x}$ is the mean and S is the standard deviation of x.

The coefficient has no units ie it's independent of the units of measurements. It is useful in comparing spreads of two or more populations. The smaller the coefficient of variation, the higher the peak and the lower the spread.

Note: Standard deviation is absolute measure of dispersion while. Coefficient of variation is relative measure of dispersion.

Example 1: Consider the distribution of the yields (per plot) of two ground nut varieties. For the first variety, the mean and standard deviation are 82 kg and 16 kg respectively. For the second variety, the mean and standard deviation are 55 kg and 8 kg respectively. Then we have, for the first variety

$$\text{C.V}(x) = \tfrac{16}{82}\times 100 \approx 19.5\%$$

For the second variety

$$\text{C.V}(x) = \tfrac{8}{55}\times 100 \approx 14.5\%$$

It is apparent that the variability in second variety is less as compared to that in the first variety. But in terms of standard deviation the interpretation could be reverse.

Example 2: Below are the scores of two cricketers in 10 innings. Find who is more „consistent scorer‟ by Indirect method.

A	204	68	150	30	70	95	60	76	24	19
B	99	190	130	94	80	89	69	85	65	40

Solution:

From a calculator, $\bar{x}_A = 79.6, S_A = 58.2$ $\bar{x}_B = 94.1$ and $S_B = 41.1$

Coefficient of variation for player A is $\text{C.V}(x) = \tfrac{58.2}{79.6}\times 100 \approx 73.153\%$

Coefficient of variation for player B is $\text{C.V}(x) = \tfrac{41.1}{94.1}\times 100 \approx 43.7028\%$

Coefficient of variation of A is greater than coefficient of variation of B and hence we conclude that player B is more consistent

Exercise

1) Find the coefficient of quartile deviation, the coefficient of mean deviation and the Coefficient of variation n of x for the following data: a) 9, 3, 4, 2, 9, 5, 8, 4, 7, 4
   b) 1, 2, 2, 3, 4, 4, 5, 5, 6, 6, 7, 8, 8 and 9 c) 3, 6, 9, 10, 7, 12, 13, 15, 6, 5, 13
   d) data on marks given by the table below

Marks Obtained	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of Students	6	12	22	24	16	12	8

2) If the weights of 7 ear-heads of sorghum are 89, 94, 102, 107, 108, 115 and 126 g. Find the arithmetic mean and standard deviation using a calculator hence determine the coefficient of variation of the ear-heads of sorghum
3) The following are the 381soybean plant heights in Cms collected from a particular plot. Using coding formula, Find the mean and Standard deviation of the plants hence determine the coefficient of variation of the 1soybean plant heights:

Plant heights (Cms)	6.8-7.2	7.3-7.7	7.8-8.2	8.3-8.7	8.8-9.2	9.3-9.7	9.8-10.2	10.3-10.7	10.8-11.2	11.3-11.7	11.8-12.2	12.3-12.7
No. of Plants	9	10	11	32	42	58	65	55	37	31	24	7

3.5 Measures of Skewness and Kurtosis

3.5.1 Skewness

Before discussing the concept of skewness, an understanding of the concept of **symmetry** is essential. A plot of frequency against class mark joined with a smooth curve can help us to visually assess the symmetry of a distribution. Usually symmetry is about the central value. Symmetry is said to exist in a distribution if the smoothed frequency polygon of the distribution can be divided into two identical halves wherein each half is a mirror image of the other **Skewness** on the other hand means lack of symmetry and it can be positive or negative. Basically, if the distribution has a tail on the right, (See figure below), then the distribution is positively skewed Eg Most students having vey low marks in an examination. However if the distribution has a tail on the left, then the distribution is negatively skewed. (see figure below). Eg Most students having vey high marks in an examination

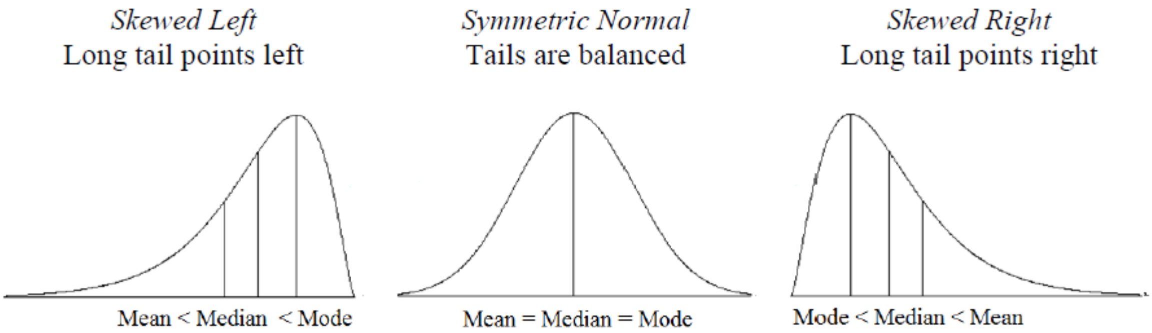


Figure 1. Sketches showing general position of mean, median, and mode in a population.

Measures of Skewness

Generally for any set of values $x_1, x_2, x_3, ...x_n,$ the **moment coefficient of skewness** α_3 is given by $\alpha_3 = \frac{\sum f\left(x - \bar{x}\right)^3}{nS^3}$ where S is the standard deviation of X. It's worth noting that if $\alpha_3 < 0,$ the distribution is negatively skewed, if $\alpha_3 > 0,$ the distribution is positively skewed and if $\alpha_3 = 0$ the distribution is normal

Other measures of Skewness includes the Karl Pearson coefficient of Skewness $\mathrm{SK_p}$, Bowley's coefficient of Skewness $\mathrm{SK_B}$ and Kelley's coefficient of Skewness $\mathrm{SK_k}$.

The **Karl Pearson's coefficient of Skewness** is based upon the ***divergence of mean from mob*** in a skewed distribution. Recall the empirical relation between mean, median and mode which states that, for a moderately symmetrical distribution, we have

$$\text{Mean - Mode} = 3\ (\text{Mean - Median})$$

Hence Karl Pearson's coefficient of skewness is defined by;

$$\mathrm{SK_p} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}},$$

The **Bowley's coefficient of Skewness** is based on quartiles. For a symmetrical distribution, it is seen that $\mathrm{Q_1}$ and $\mathrm{Q_3}$ ***are*** equidistant ftom median.

$$\mathrm{SK_B} = \frac{\mathrm{Q_3 - 2Q_2 + Q_1}}{\mathrm{Q_3 - Q_1}} \text{ where } \mathrm{Q_k} \text{ is the } \mathrm{K^{th}} \text{ quartile.}$$

The **Kelly's coefficient of Skewness** is based on $\mathrm{P_{90}}$ and, $\mathrm{P_{10}}$ so that only 10% of the observations on each extreme are ignored.. This is an improvement over the Bowley's coefficient which leaves 25% of the observatories on each extreme of the distribution.

$$\mathrm{SK_k} = \frac{\mathrm{P_{90} - 2P_{50} + P_{10}}}{\mathrm{P_{90} - P_{10}}} \text{ where } \mathrm{P_k} \text{ is the } \mathrm{K^{th}} \text{ percentile.}$$

Interpreting Skewness

If the coefficient of skewness is positive, the data are positively skewed or skewed right, meaning that the right tail of the distribution is longer than the left. If the coefficient of skewness is negative, the data are negatively skewed or skewed left, meaning that the left tail is longer. If the coefficient of skewness = 0, the data are perfectly symmetrical. But a skewness of exactly zero is quite unlikely for real-world data, so *how can you interpret the skewness number*? Bulmer, M. G., *Principles of Statistics* (Dover,1979) — a classic — suggests this rule of thumb: If the coefficient of skewness is:-

- less than −1 or greater than +1, the distribution is *highly skewed.*
- between −1 and $-\frac{1}{2}$. or between $+\frac{1}{2}$. and +1, the distribution is *moderately skewed.*
- between $-\frac{1}{2}$ and $+\frac{1}{2}$.., the distribution is *approximately symmetric.*

Example: The following figures relate to the size of capital of 285 companies :

Capital (in *Ks* lacs.)	1-5	6-10	11-15	16-20	21-25	26-30	31-35
No. of companies	20	27	29	38	48	53	70

Compute the Bowley's coefficients of skewness and interpret the results.

Solution

Boundaries	0.5-5.5	5.5-10.5	10.5-15.5	15.5-20.5	20.5-25.5	25.5-30.5	30.5-35.5
CF	20	47	76	114	162	215	285

$$Q_1 = \tfrac{1}{4}(286)^{\text{th}}\text{ value} = 71.5^{\text{th}}\text{ value} = 10.5 + \left(\frac{71.5-47}{29}\right)\times 5 \approx 14.7241$$

$$Q_2 = \tfrac{1}{2}(286)^{\text{th}}\text{ value} = 143^{\text{rd}}\text{ value} = 20.5 + \left(\frac{143-114}{48}\right)\times 5 \approx 23.5208$$

$$Q_3 = \tfrac{3}{4}(286)^{\text{th}}\text{ value} = 214.5^{\text{th}}\text{ value} = 25.5 + \left(\frac{214.5-162}{53}\right)\times 5 \approx 30.4528$$

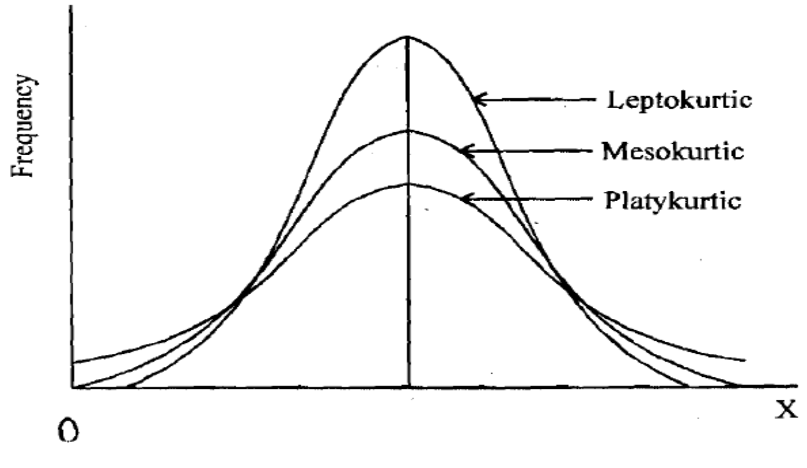
$$SK_p = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} = \frac{30.4528 - 2\times 23.5208 + 14.7241}{30.4528 - 14.7241_1} \approx -0.11855.$$

This value lies between $-\frac{1}{2}$ and $+\frac{1}{2}$, therefore the distribution is approximately symmetric.

Question: Compute the Karl Pearson's and the Kelly’s coefficient of skewness for the above data and interpret the results.

3.5.2 Kurtosis

It measures the peakedness of a distribution. If the values of x are very close to the mean, the peak is very high and the distribution is said to be **Leptokurtic**. On the other hand if the values of x are very far away from the mean, the peak is very low and the distribution is said to be **Pletykurtic.** Finally if x values are at a moderate distance from the mean then the peak is moderate and the distribution is said to be **mesokurtic.**.



Measures of Kurtosis

Generally for a set of values $x_1, x_2, x_3, ...x_n$, the moment coefficient of kurtosis α_4 is given by $\alpha_4 = \dfrac{\sum f\left(x - \bar{x}\right)^4}{nS^4}$ where $\bar{x}$ and S are the arithmetic mean and standard deviation of X.

Example: Calculate the coefficient of Skewness α_3 and the coefficient of kurtosis α_4 for the data 5, 6, 7, 6, 9, 4, 5

Solution

$\bar{x} = \frac{1}{n}\sum x = \frac{42}{7} = 6$ and Standard deviation $s = \sqrt{\frac{1}{n}\sum (x - \bar{x})^2} = \frac{4}{\sqrt{7}}$

x	5	6	7	6	9	4	5	Sum
$\left(x - \bar{x}\right)^2$	1	0	1	0	9	4	1	16
$\left(x - \bar{x}\right)^3$	-1	0	1	0	27	-8	-1	18
$\left(x - \bar{x}\right)^4$	1	0	1	0	81	16	1	100

Coefficient of Skewness $\alpha_3 = \dfrac{\sum \left(x - \bar{x}\right)^3}{nS^3} = \dfrac{18}{7} \times \left(\frac{\sqrt{7}}{4}\right)^3 \approx 0.744118$

Coefficient of kurtosis $\alpha_4 = \dfrac{\sum f\left(x - \bar{x}\right)^4}{nS^4} = \dfrac{100}{7} \times \left(\frac{\sqrt{7}}{4}\right)^4 \approx 2.73438$

Exercise

1. Find the moment coefficient of Skewness and kurtosis for the dat below. a) 9, 3, 4, 2, 9, 5, 8, 4, 7, 4 b) 1, 2, 2, 3, 4, 4, 5, 5, 6, 6, 7, 8, 8 and 9 c) 3, 6, 9, 10, 7, 12, 13, 15, 6, 5, 13
   d) data on marks given by the table below

Marks Obtained	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of Students	6	12	22	24	16	12	8

   e) Data given by the table below

Marks Obtained	0-10	10-20	20-30	30-40
No. of Students	1	3	4	2

2. Compute the Bowley's coefficient of skewness, the Kelly's coefficient of skewness and the Percentile coefficient of kurtosis for the following data and interpret the results.
   a) 9, 3, 4, 2, 9, 5, 8, 4, 7, 4 b) 1, 2, 2, 3, 4, 4, 5, 5, 6, 6, 7, 8, 8 and 9
   c) 3, 6, 9, 10, 7, 12, 13, 15, 6, 5, 13 d) data on heights given by the table below

Heightl (in *inches*.)	58	59	60	61	62	63	64	65
No. of persons	10	18	30	42	35	28	16	8

   e) data on daily expenditure of families given by the table below

Daily Expenditure (Rs)	0-20	20-40	40-60	60-80	80-100
No. of persons	13	25	27	19	16

   f) Data on marks given by the table below

Marks Obtained	0-20	20-40	40-60	60-80	80-100
No. of Students	8	28	35	17	12

3. The following measures were computed for a frequency distribution : Mean = 50, coefficient of Variation = 35% and Karl Pearson's Coefficient of Skewness $SK_p = -0.25$. Compute Standard Deviation, Mode and Median of the distribution.

4. Bivariate Data

4.1 Introduction

So far we have confined our discussion to the distributions involving only one variable. Sometimes, in practical applications, we might come across certain set of data, where each item of the set may comprise of the values of two or more variables.

A Bivariate Data is a a set of paired measurements which are of the form

$$(x_1 , y_1) , (x_2 , y_2), ...,(x_n , y_n)$$

Examples

i. Marks obtained in two subjects by 60 students in a class.
ii. The series of sales revenue and advertising expenditure of the various branches of a company in a particular year.
iii. The series of ages of husbands and wives in a sample of selected married couples.

In a bivariate data, each pair represents the values of the two variables. Our interest is to find a relationship (if it exists) between the two variables under study.

4.2 Scatter Diagrams and Correlation

A scatter diagram is a tool for analyzing relationships between two variables. One variable is plotted on the horizontal axis and the other is plotted on the vertical axis. The pattern of their intersecting points can graphically show relationship patterns. Most often a scatter diagram is used to prove or disprove cause-and-effect relationships. While the diagram shows relationships, it does not by itself prove that one variable *causes* the other. In brief, the easiest way to visualize Bivariate Data is through a Scatter Plot.

"Two variables are said to be correlated if the change in one of the variables results in a change in the other variable".

4.2.1: Positive and Negative Correlation

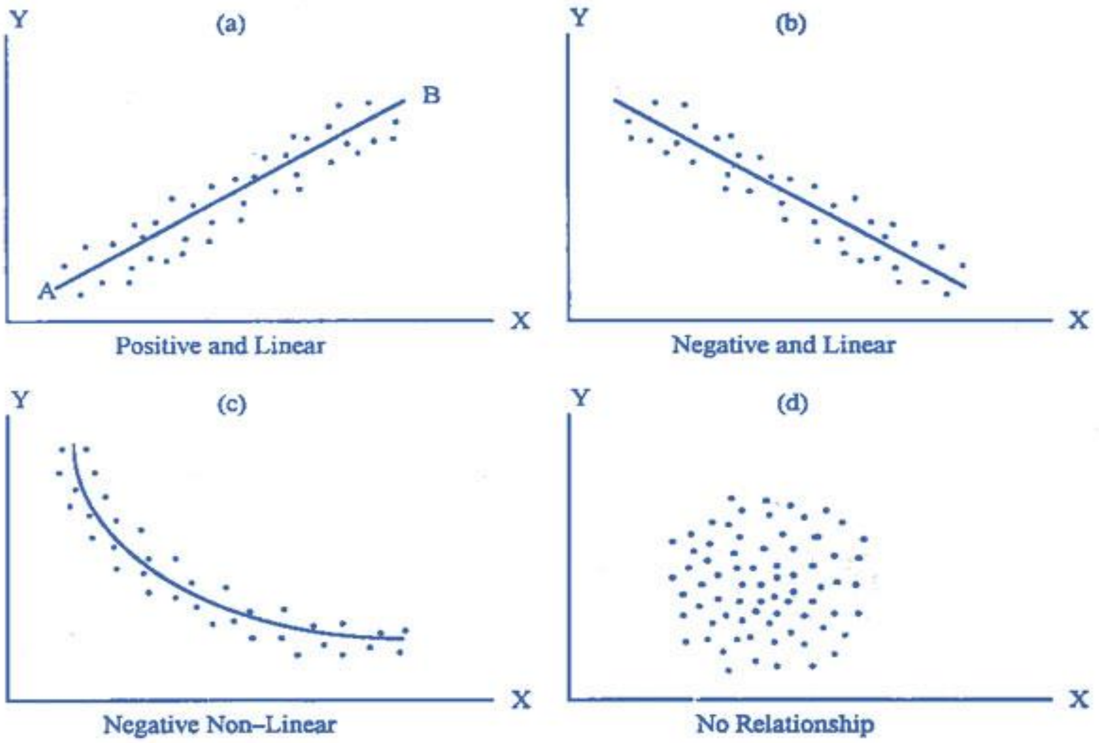
If the values of the two variables deviate in the same direction i.e. if an increase (or decrease) in the values of one variable results, on an average, in a corresponding increase (or decrease) in the values of the other variable the correlation is said to be positive.

Some examples of series of positive correlation are:

i. Heights and weights;
ii. Household income and expenditure;
iii. Price and supply of commodities;
iv. Amount of rainfall and yield of crops.

Correlation between two variables is said to be negative or inverse if the variables deviate in opposite direction. That is, if the increase in the variables deviate in opposite direction. That is, if increase (or decrease) in the values of one variable results on an average, in corresponding decrease (or increase) in the values of other variable.

Eg Price and demand of goods.



4.2.2 Interpreting a Scatter Plot

Scatter diagrams will generally show one of six possible correlations between the variables:

i. *Strong Positive Correlation* The value of Y clearly increases as the value of X increases.
ii. *Strong Negative Correlation* The value of Y clearly decreases as the value of X increases.
iii. *Weak Positive Correlation* The value of Y increases slightly as the value of X increases.
iv. *Weak Negative Correlation* The value of Y decreases slightly as the value of X increases.
v. *Complex Correlation* The value of Y seems to be related to the value of X, but the relationship is not easily determined.
vi. *No Correlation* There is no demonstrated connection between the two variables

4,3 Correlation Coefficient

Correlation coefficient measures the degree of linear association between 2 paired variables It takes values from + 1 to – 1.

i. If r = +1,we have **perfect positive** relationship
ii. If r = -1,we have **perfect negative** relationship
iii. If r = 0 there is **no** relationship ie the variables are **uncorrelated.**

4,3 .1 Pearson's Product Moment Correlation Coefficient

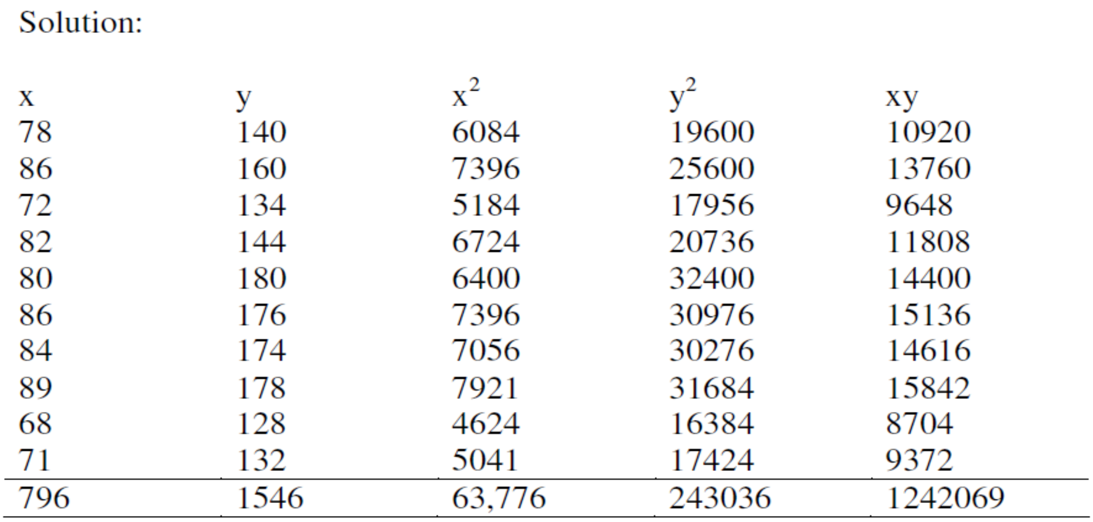
Pearson's product moment correlation coefficient, usually denoted by r, is one example of a correlation coefficient. It is a measure of the linear association between two variables that have been measured on interval or ratio scales, such as the relationship between height in inches and weight in pounds. However, it can be misleadingly small when there is a relationship between the variables but it is a non-linear one.

The correlation coefficient r is given by $r = \dfrac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$

Example:

: A study was conducted to find whether there is any relationship between the weight and blood pressure of an individual. The following set of data was arrived at from a clinical study. Let us determine the coefficient of correlation for this set of data. The first column represents the serial number and the second and third columns represent the weight and blood pressure of each patient.

Weight	78	86	72	822	80	86	84	89	68	71
Blood Pressure	140	160	134	144	180	176	174	178	128	132



Solution:

x	y	x^2	y^2	xy
78	140	6084	19600	10920
86	160	7396	25600	13760
72	134	5184	17956	9648
82	144	6724	20736	11808
80	180	6400	32400	14400
86	176	7396	30976	15136
84	174	7056	30276	14616
89	178	7921	31684	15842
68	128	4624	16384	8704
71	132	5041	17424	9372
796	1546	63,776	243036	1242069

Thus

$$r = \frac{10(124206) - (796)(1546)}{\sqrt{[\,(10)63776 - (796)^2\ (10)\,][(243036) - (1546)^2]}} = \frac{11444}{\sqrt{(1144)(40244)}} = 0.5966$$

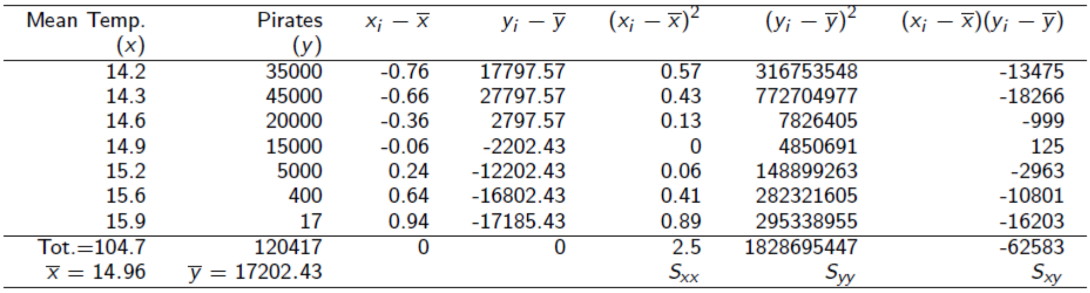
It can be shown that $r = \dfrac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}}$

Example:

Obtain the correlation coefficient of the following data

Mean Temp. (x)	14.2	14.3	14.6	14.9	15.2	15.6	15.9
Pirates (y)	35000	45000	20000	15000	5000	400	17

Solution



Mean Temp. (x)	Pirates (y)	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
14.2	35000	-0.76	17797.57	0.57	316753548	-13475
14.3	45000	-0.66	27797.57	0.43	772704977	-18266
14.6	20000	-0.36	2797.57	0.13	7826405	-999
14.9	15000	-0.06	-2202.43	0	4850691	125
15.2	5000	0.24	-12202.43	0.06	148899263	-2963
15.6	400	0.64	-16802.43	0.41	282321605	-10801
15.9	17	0.94	-17185.43	0.89	295338955	-16203
Tot.=104.7	120417	0	0	2.5	1828695447	-62583
$\overline{x} = 14.96$	$\overline{y} = 17202.43$			S_{xx}	S_{yy}	S_{xy}

We then have that $r = \dfrac{-62583}{\sqrt{2.5(1828695447)}} \approx -0:93$

4.3 .2 Spearman rank correlation coefficient

Data which are arranged in ascending order are said to be in **ranks** or **ranked data.**. The coefficient of correlation for such type of data is given by **Spearman rank difference correlation coefficient** and is denoted by R.

R is given by the formula $R = 1 - \dfrac{6\sum d^2}{n(n^2 - 1)}$

Example

The data given below are obtained from student records.(Grade Point Average (x) and Graduate Record exam score (y)) Calculate the rank correlation coefficient 'R' for the data.

Subject	1	2	3	4	5	6	7	8	9	10
X	8.3	8.6	9.2	9.8	8.0	7.8	9.4	9.0	7.2	8.6
y	2300	2250	2380	2400	2000	2100	2360	2350	2000	2260

Solution

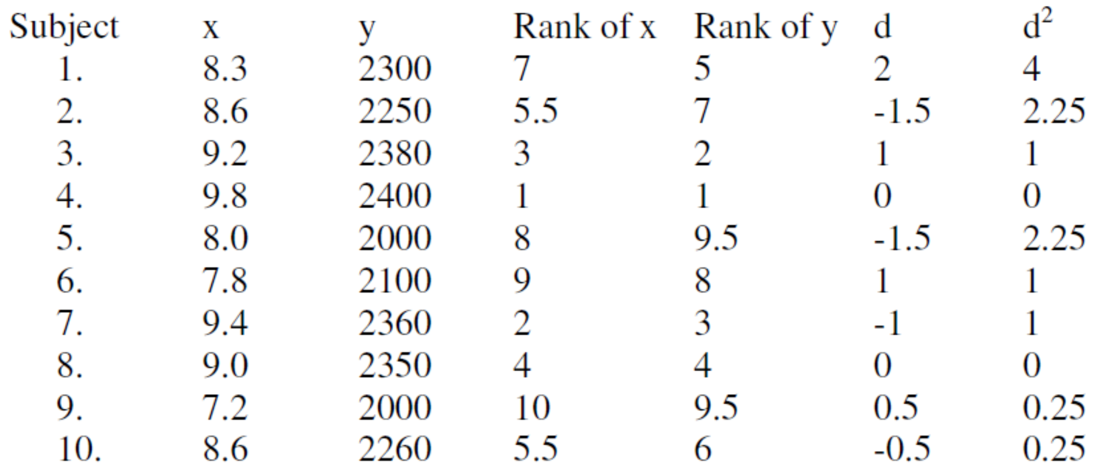
Note that in the x row, we have two students having a grade point average of 8.6 also in the y row; there is a tie for 2000.

Now we arrange the data in descending order and then rank 1,2,3,. . . . .10 accordingly. In case of a tie, the rank of each tied value is the mean of all positions they occupy. In x, for instance, 8.6 occupy ranks 5 and 6. So each has a rank $\dfrac{5+6}{2} = 5.5$

Similarly in 'y' 2000 occupies ranks 9 and 10, so each has rank 9.5

Now we come back to our formula $R = 1 - \dfrac{6\sum d^2}{n(n^2 - 1)}$

We compute d, square it and substitute its value in the formula



Subject	x	y	Rank of x	Rank of y	d	d^2
1.	8.3	2300	7	5	2	4
2.	8.6	2250	5.5	7	-1.5	2.25
3.	9.2	2380	3	2	1	1
4.	9.8	2400	1	1	0	0
5.	8.0	2000	8	9.5	-1.5	2.25
6.	7.8	2100	9	8	1	1
7.	9.4	2360	2	3	-1	1
8.	9.0	2350	4	4	0	0
9.	7.2	2000	10	9.5	0.5	0.25
10.	8.6	2260	5.5	6	-0.5	0.25

So here, n = 10 and $\sum d^2 = 12$. So

$$R = 1 - \frac{6(12)}{10(100-1)} = 1 - 0.0727 = 0.9273$$

Note: If we are provided with only ranks without giving the values of x and y we can still find Spearman rank difference correlation R by taking the difference of the ranks and proceeding in the above shown manner.

4,4 Regression

If two variables are significantly correlated, and if there is some theoretical basis for doing so, it is possible to predict values of one variable from the other.

Regression analysis, in general sense, means the estimation or prediction of the unknown value of one variable from the known value of the other variable. It is one of the most important statistical tools which is extensively used in almost all sciences – Natural, Social and Physical.

Regression analysis was explained by M. M. Blair as follows:

“Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.”

3.4.1 Regression Equation

Regression analysis can be thought of as being sort of like the flip side of correlation. It has to do with finding the equation for the kind of straight lines you were just looking at Suppose we have a sample of size n and it has two sets of measures, denoted by x and y. We can predict the values of y given the values of x by using the equation, $y^* = a + bx$

Where the coefficients ‘a’ and ‘b’ are real numbers given by

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \quad \text{and} \quad a = \frac{\sum y - b\sum x}{n}$$

The symbol y^* refers to the predicted value of y from a given value of x from the regression equation.

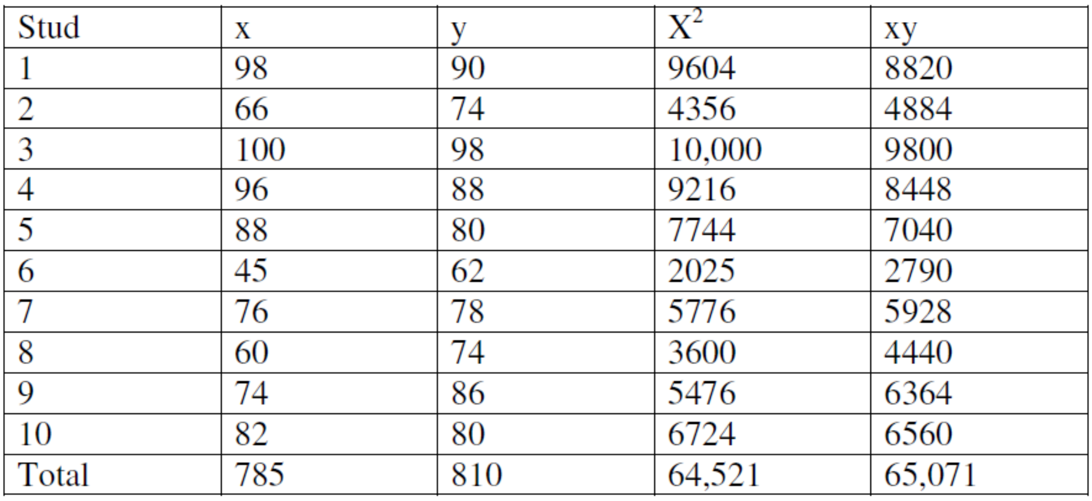
Example:

Scores made by students in a statistics class in the mid-term and final examination are given here. Develop a regression equation which may be used to predict final examination scores from the mid – term score.

Student	1	2	3	4	5	6	7	8	9	10
Mid term	98	66	100	96	88	45	76	60	74	82
Final	90	74	98	88	80	62	78	74	86	80

Solution:

We want to predict the final exam scores from the mid term scores. So let us designate 'y' for the final exam scores and 'x' for the mid term exam scores. We open the following table for the calculations.



Stud	x	y	X^2	xy
1	98	90	9604	8820
2	66	74	4356	4884
3	100	98	10,000	9800
4	96	88	9216	8448
5	88	80	7744	7040
6	45	62	2025	2790
7	76	78	5776	5928
8	60	74	3600	4440
9	74	86	5476	6364
10	82	80	6724	6560
Total	785	810	64,521	65,071

$$b = \frac{10\,(65{,}071) - \ 785\,(810)}{10(\,64,521) - \ (785)^2} = \frac{14{,}860}{28{,}985} = 0.5127 \ \text{ and } \ a = \frac{810 - 785(\,0.5127)}{10} = 40.7531$$

Thus, the regression equation is given by $y^* = 40.7531 + (0.5127)\,\text{x}$

We can use this to find the projected or estimated final scores of the students.
Eg for the midterm score of 50 the projected final score is

$y^* = 40.7531 + (0.5127)\,50 = 66.3881$, which is a quite a good estimation.

To give another example, consider the midterm score of 70. Then the projected final score is

$y^* = 40.7531 + (0.5127)\,70 = 76.6421$, which is again a very good estimation.

<u>Practice Problems</u>:

1. Consider the following data and draw a scatter plot

X	1.0	1.9	2.0	2.9	3.0	3.1	4.0	4.1	5
Y	10	99	100	999	1,000	1,001	10,000	10,001	100,000

2. . Let variable X is the number of hamburgers consumed at a cook-out, and variable Y is the number of beers consumed. Develop a regression equation to predict how many beers a person will consume given that we know how many hamburgers that person will consume.

Subject	1	2	3	4	5
Hamburgers	5	4	3	2	1

Beers	8	10	4	6	2

3. A horse owner is investigating the relationship between weight carried and the finish position of several horses in his stable. Calculate r and R for the data given

Weight carried	110	113	120	115	110	115	117	123	106	108	110	110
Position Finished	2	6	3	4	6	5	4	2	1	4	1	3

4. The top and bottom number which may appear on a die are as follows Calculate r and R for these values. Are the results surprising?

Top	1	2	3	4	5	6
Bottom	5	6	4	3	1	2

5. The ranks of two sets of variables (Heights and Weights) are given below. Calculate the Spearman rank difference correlation coefficient R.

	1	2	3	4	5	6	7	8	9	10
Heights	2	6	8	4	7	4	9.5	4	1	9.5
Weights	9	1	9	4	5	9	2	7	6	3

6. Researchers interested in determining if there is a relationship between death anxiety and religiosity conducted the following study. Subjects completed a death anxiety scale (high score = high anxiety) and also completed a checklist designed to measure an individuals degree of religiosity (belief in a particular religion, regular attendance at religious services, number of times per week they regularly pray, etc.) (high score = greater religiosity . A data sample is provided below:

X	38	42	29	31	28	15	24	17	19	11	8	19	3	14	6
y	4	3	11	5	9	6	14	9	10	15	19	17	10	14	18

a) What is your computed answer?
b) What does this statistic mean concerning the relationship between death anxiety and religiosity?
c) What percent of the variability is accounted for by the relation of these two variables?

7. The data given below are obtained from student records.(Grade Point Average (x) and Graduate Record exam score (y)) Calculate the regression equation and compute the estimated GRE scores for GPA = 7.5 and 8.5..

Subject	11	12	13	14	15	16	17	18	19	20
X	8.3	8.6	9.2	9.8	8.0	7.8	9.4	9.0	7.2	8.6
y	2300	2250	2380	2400	2000	2100	2360	2350	2000	2260

8. A horse was subject to the test of how many minutes it takes to reach a point from the starting point. The horse was made to carry luggage of various weights on 10 trials.. The data collected are presented below in the table. Find the regression equation between the load and the time taken to reach the goal. Estimate the time taken for the loads of 35 Kgs , 23 Kgs, and 9 Kgs. Are the answers in agreement with your intuitive feelings? Justify.

Trial Number	1	2	3	4	5	6	8	8	9	10
Weight (in Kgs)	11	23	16	32	12	28	29	19	25	20
Time taken (in mins)	13	22	16	47	13	39	43	21	32	22

9. A study was conducted to find whether there is any relationship between the weight and blood pressure of an individual. The following set of data was arrived at from a clinical study.

Serial Number	1	2	3	4	5	6	8	8	9	10
Weight	78	86	72	822	80	86	84	89	68	71
Blood Pressure	140	160	134	144	180	176	174	178	128	132

10. It is assumed that achievement test scores should be correlated with student's classroom performance. One would expect that students who consistently perform well in the classroom (tests, quizes, etc.) would also perform well on a standardized achievement test (0 - 100 with 100 indicating high achievement (x)). A teacher decides to examine this hypothesis. At the end of the academic year, she computes a correlation between the students achievement test scores (she purposefully did not look at this data until after she submitted students grades) and the overall

g.p.a.(y) for each student computed over the entire year. The data for her class are provided below.

X	98	96	94	88	01	77	86	71	59	63	84	79	75	72	86	85	71	93	90	62
y	3.6	2.7	3.1	4.0	3.2	3.0	3.8	2.6	3.0	2.2	1.7	3.1	2.6	2.9	2.4	3.4	2.8	3.7	3.2	1.6

a) Compute the correlation coefficient.
b) What does this statistic mean concerning the relationship between achievement test prformance and g.p.a.?
c) What percent of the variability is accounted for by the relationship between the two variables and what does this statistic mean?
d) What would be the slope and y-intercept for a regression line based on this data?
e) If a student scored a 93 on the achievement test, what would be their predicted G.P.A.? If they scored a 74? A 88?

11. With the growth of internet service providers, a researcher decides to examine whether there is a correlation between cost of internet service per month (rounded to the nearest dollar) and degree of customer satisfaction (on a scale of 1 - 10 with a 1 being not at all satisfied and a 10 being extremely satisfied). The researcher only includes programs with comparable types of services. A sample of the data is provided below.

Cost of internet (in $)	11	18	17	15	9	5	12	19	22	25
satisfaction	6	8	10	4	9	6	3	5	2	10

a) Compute the correlation coefficient.
b) What does this statistic mean concerning the relationship between amount of money spent per month on internet provider service and level of customer satisfaction?
c) What percent of the variability is accounted for by the relationship between the two variables and what does this statistic mean?

12. It is hypothesized that there are fluctuations in norepinephrine (NE) levels which accompany fluctuations in affect with bipolar affective disorder (manic-depressive illness). Thus, during depressive states, NE levels drop; during manic states, NE levels increase. To test this relationship, researchers measured the level of NE by measuring the metabolite 3-methoxy-4-hydroxyphenylglycol (MHPG in micro gram per 24 hour) in the patient's urine experiencing varying levels of mania/depression. Increased levels of MHPG are correlated with increased metabolism (thus higher levels) of central nervous system NE. Levels of mania/depression were also recorded on a scale with a low score indicating increased mania and a high score increased depression. The data is provided below.

MHPG	980	1209	1403	1950	1814	1280	1073	1066	880	776
Affect	22	26	8	10	5	19	26	12	23	28

a) Compute the correlation coefficient.
b) What does this statistic mean concerning the relationship between MHPG levels and affect?
c) What percent of the variability is accounted for by the relationship between the two variables?
d) What would be the slope and y-intercept for a regression line based on this data?
e) What would be the predicted affect score if the individual had an MHPG level of 1100? of 950? of 700?

13. The table below contains 25 cases -- the mother's weight in kilograms and the infant's birth weight in grams. Does this data suggest some relationship between the mother's weight and the infant's birth weight Why would such a relationship be important

. **Prepregnancy Weights of Mothers and Birthweights of their Infants**		
Case Number	**Mother's Weight (kg)**	**Infant's Birthweight (g)**
1	49.4	3515
2	63.5	3742
3	68.0	3629

4	52.2	2680
5	54.4	3006
6	70.3	4068
7	50.8	3373
8	73.9	4124
9	65.8	3572
10	54.4	3359
11	73.5	3230
12	59.0	3572
13	61.2	3062
14	52.2	3374
15	63.1	2722
16	65.8	3345
17	61.2	3714
18	55.8	2991
19	61.2	4026
20	56.7	2920
21	63.5	4152
22	59.0	2977
23	49.9	2764
24	65.8	2920
25	43.1	2693

5 PROBABILITY

5.1 What is Probability?

Probability theory is the branch of mathematics that studies the possible outcomes of given events together with the outcomes' relative likelihoods and distributions. In common usage, the word “probability" is used to mean the chance that a particular event (or set of events) will occur expressed on a linear scale from 0 (impossibility) to 1 (certainty). Factually, It is the study of random or indeterministic experiments eg tossing a coin or rolling a die. If we roll a die, we are certain it will come down but we are uncertain which face will show up. Ie the face showing up is indeterministic. Probability is a way of summarizing the uncertainty of statements or events. It gives a numerical measure for the degree of certainty (or degree of uncertainty) of the occurrence of an event.

We often use P to represent a probability Eg P(rain) would be the probability that it rains. In other cases Pr(.) is used instead of just P(.).

Definitions

- Experiment: A process by which an observation or measurement is obtained. Eg tossing a coin or rolling a die.
- Outcome: Possible result of a random experiment. Eg a 6 when a die is rolled once or a head when a coin is tossed.
- Sample space: Also called the probability space and it is a collection or set of all possible outcomes of a random experiment. Sample space is usually denoted by S or Ω or U
- Event: it's a subset of the sample space. Events are usually denoted by upper case letters.

Suppose the sample space S consists of n(S) equally likely outcomes and n(E) of those are favourable for an event E then probability of an event E is the ratio of the number of favourable outcomes n(E) to the total number of all possible outcomes n(S) ie

$$\mathrm{P(E)} = \frac{\text{number of favourable outcomes}}{\text{total number of possible outcomes}} = \frac{\mathrm{n(E)}}{\mathrm{n(S)}}$$

5.2 Approaches to Probability

There are three ways to define probability, namely classical, empirical and subjective probability.

5.2.1 Classical probability

Classical or theoretical probability is used when each outcome in a sample space is equally likely to occur. The underlying idea behind this view of probability is symmetry. Ie if the sample space contains n outcomes that are fairly likely then P(one outcome)=1/n.

The classical probability for an event A is given by

$$\mathrm{P(A)} = \frac{\text{Number of outcomes in A}}{\text{Total number of outcomes in S}} = \frac{\mathrm{n(A)}}{\mathrm{n(S)}}$$

Eg Roll a die and observe that $\mathrm{P(A)} = \text{P(rolling a 3)} = \frac{1}{6}$.

Example

A fair die, with faces numbered 1 to 6, is rolled once, write down the saple space S hence find the probability that the score showing up is ; a) a multiple of 3 b) a prime number.

Solution

$S = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$ Multiples of 3 are 3 and 6 while prime numbers are 2, 3 and 5

Thus $\text{P(Multiple of 3)} = \frac{2}{6} = \frac{1}{3}$ and $\text{P(prime number)} = \frac{3}{6} = \frac{1}{2}$

5.2.2 Frequentist or Empirical probability

When the outcomes of an experiment are not equally likely, we can conduct experiments to give us some idea of how likely the different outcomes are. For example, suppose we are interested in measuring the probability of producing a defective item in a manufacturing process. The probability could be measured by monitoring the process over a reasonably long period of time and calculating the proportion of defective items.

In a nut shell Empirical (or frequrntist or statistical) probability is based on observed data. The empirical probability of an event A is the relative frequency of event A, that is

$$P(A) = \frac{\text{Frequency of event A}}{\text{Total number of observations}}$$

Example 1

The following are the counts of fish of each type that you have caught before.

Fish Types	Blue gill	Red gill	Crappy	Total
No of times caught	13	17	10	40

Estimate the probability that the next fish you catch will be a Blue gill.

$P(\text{Blue gill}) = \frac{13}{40} = 0.325$

Example 2

A girl lists the number of male and female children her parent and her parent's brothers and sisters have. Her results were as tabulated below

	Males	Females
Her parents	2	5
Her mother's sisters	6	8
Her mother's brothers	4	8
Her father's sisters	5	8
Her father's brothers	7	7
Totals	24	36

d) Find the probability that, if the girl has children of her own, the 1st born will be a girl.

e) If the girl eventually has 10 children, how many are likely to be males?

Solution

a) Following the family pattern, $P(\text{1st born will be a girl}) = \frac{36}{60} = 0.6$

b) 60% of the children will be females $\Rightarrow 40\%$ will be males. Thus 4 out of 10 children are likely to be males.

Remark: The empirical probability definition has a weakness that it depends on the results of a particular experiment. The next time this experiment is repeated, you are likely to get a somewhat different result. However, as an experiment is repeated many times, the empirical probability of an event, based on the combined results, approaches the theoretical probability of the event.

5.2.3 Subjective Probability:

Subjective probabilities result from intuition, educated guesses, and estimates. For example: given a patient's health and extent of injuries a doctor may feel that the patient has a 90% chance of a full recovery.. Subjectivity means two people can assign different probabilities to the same event.

Regardless of the way probabilities are defined, they always follow the same laws, which we will explore in the following Section.

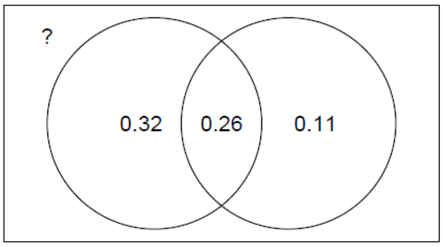
Exercise

1) What is the probability of getting a total of 7 or 11, when two dice are rolled?
2) Two cards are drawn from a pack, without replacement. What is the probability that both are greater than 2 and less than 8?
3) A permutation of the word "white" is chosen at random. Find the probability that it begins with a vowel. Also find the probability that it ends with a consonant.
4) Find the probability that a leap year will have 53 Sundays.

5) Two tetrahedral (4-sided) symmetrical dice are rolled, one after the other. Find the probability that;
   a) both dice will land on the same number.
   b) each die will land on a number less than 3.
   c) the two numbers will differ by at most 1.
   Will the answers change if we rolled the dice simultaneously?

Ways to represent probabilities:

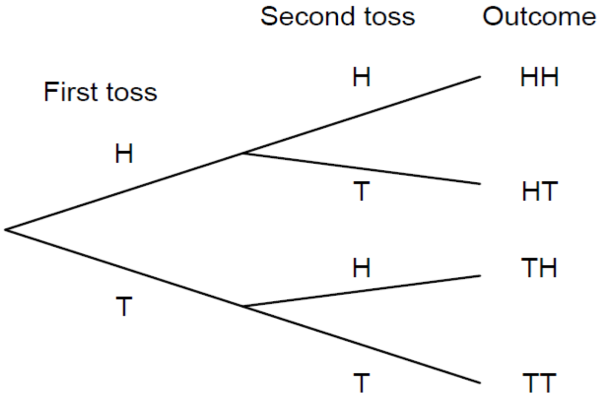
1) *Venn diagram*; We may write the probabilities inside the elementary pieces within a Venn diagram. For example, $P(AB') = 0.32$ and $P(A) = P(AB) + P(AB') = 0.58$ [why?] The relative sizes of the pieces do not have to match the numbers.



2) *Two-way table;* This is a popular way to represent statistical data. The cells of the table correspond to the intersections of row and column events. Note that the contents of the table add up accross rows and columns of the table. The bottom-right corner of the table contains P(S) = 1

	B	B'	Total
A	0.26	0.32	0.58
A'	0.11	?	0.42
Total	0.37	0.63	1

Tree diagram; Tree diagrams or probability trees are simper clear ways of representing probabilistic information. A tree diagram may be used to show the sequence of choices that lead to the complete description of outcomes. For example, when tossing two coins, we may represent this as follows



A tree diagram is also often useful for representing conditional probabilities

5.3 Review of set notation

Complement: The complement of event A, (denoted A'), is the set of all outcomes in a sample that are not included in the event A.

Intersection of events: The event $A \cap B$ (or simply AB) read as 'A intersection B' consists of outcomes that are contained within both events A and B. The probability of this event is the probability that both events A and B occur [but not necessarily at the same time]. Here after we will abbreviate intersection as AB.

Unions of Events: The event $A \cup B$ read as 'A union B' consists of the outcomes that are contained within at least one of the events A and B. The probability of this event $P(A \cup B)$; is the probability that events A and/or B occurs.

Set notation

Suppose a set S consists of points labelled 1, 2, 3 and 4. We denote this by $S = \{1, 2, 3, 4\}$. . If $A = \{1, 2\}$and $B = \{2, 3, 4\}$, then A and B are subsets of S, denoted by $A \subset S$ and$B \subset S$ (B is contained in S). We denote the fact that 2 is an element of A by $2 \in A$.

The union of A and B, $A \cup B = \{1, 2, 3, 4\}$. If $C = \{\ 4\}$, then $A \cup C = \{1, 2,\ 4\}$. The intersection$A \cap B = AB = \{2\}$: The complement of A, is$A' = \{3, 4\}$.

Distributive laws; $A \cap (B \cup C) = AB \cup AC$ and$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

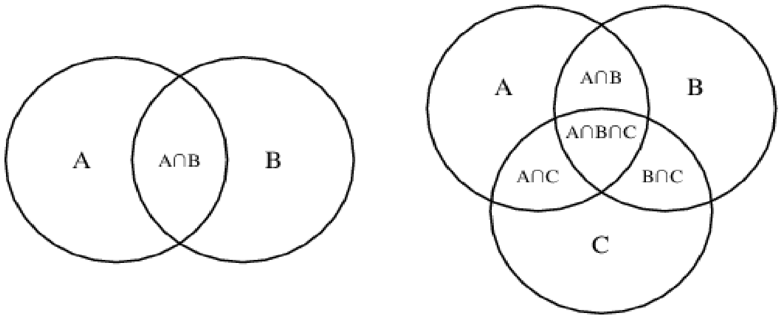
De Morgan's Law; $(A \cup B)' = A'B'$ and $(AB)' = A' \cup B'$

Venn diagram

Venn diagram is a diagram that shows all possible logical relations between a finite collection of sets.

Venn diagram is often used to illustrate the relations between sets (events).

The sets A and B are represented as circles; operations between them (intersections, unions and complements) can also be represented as parts of thevdiagram. The entire sample space S is the bounding box. See Figure 2.1



Exercise

1) Use the Venn diagrams to illustrate Distributive laws and De Morgan's law.
2) Simplify the following (Draw the Venn diagrams to visualize)
   a) $(A')'$ b) $(AB)' \cup A$ c) $AB \cup AB'$ d) $(A \cup B \cup) \cap B$
3) Represent by set notation and exhibit on a Venn diagram the following events
   a) both A and B occur
   b) exactly one of A, B occurs
   c) A and B but not C occur s
   d) at least one of A, B, C occurs
   e) at most one of A, B, C occurs
4) The sample space consists of eight capital letters (outcomes), A, B, C ,...,H. Let V be the event that the letter represents a vowel, and L be the event that the letter is made of straight lines. Describe the outcomes that comprise
   a) VL b) $V \cup L'$ c) $V'L'$

5) Out of all items sent for refurbishing, 40% had mechanical defects, 50% had electrical defects, and 25% had both. Denoting A = fan item has a mechanical defect and B = fan item has an electrical defect, fill the probabilities into the Venn diagram and determine the quantities listed below. a) $P(A)$ b) $P(AB)$ c) $P(A'B)$ d) $P(A'B')$ e) $P(A\cup B)$ f) $P(A'\cup B')$ g) $P[(A\cup B)']$

6) A sample of mutual funds was classified according to whether a fund was up or down last year (A and A') and whether it was investing in international stocks (B and B'). The probabilities of these events and their intersections are represented in the two-way table below. Fill in all the question marks hence find the probability of $A\cup B$

	B	B'	
A	0.33	?	?
A'	?	?	0.52
	0.64	?	1

5.4 Rules of Probability

1) For an experiment with a sample space $S=\{E_1, E_2, \dots, E_n\}$ we can assign probabilities

$$P(E_1), P(E_2), \dots , P(E_n) \text{ provided that } 0 \le P(E_i) \le 1 \text{ and } P(S)=\sum_{i=1}^{n} P(E_i)=1$$

Remark:

a) If a set (event) A consists of outcomes $E_1, E_2, \dots, E_k$, then $P(A)=\sum_{i=1}^{k} P(E_i)$

b) If $E = S$ then $P(E)=P(S)=1$ and If E is has no elements, (ie if E is empty $\Rightarrow$ $E=\phi$ or $\{\ \}$), then $P(E)=P(\phi)=0$

2) If E is an event in the sample space S, then E' (called the complement of E) is an event in S but outside E. $P(E)+P(E')=1 \Rightarrow P(E')=1-P(E)$

3) If the sample space S contains n disjoint events $E_1, E_2, \dots, E_n$, then

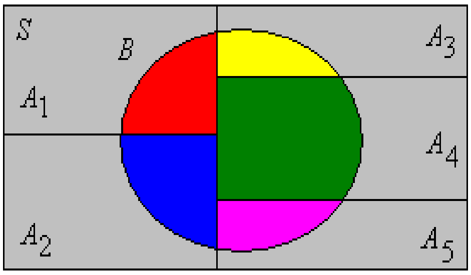
$$P(E_1)+P(E_2)+,\dots P(E_n)=\sum_{i=1}^{n} P(E_i)=1$$

4) Let A and B be two events such that $A\subseteq B$, then $P(A)\le P(B)$

5) For any two events A and B, $P(A\cup B)=P(A)+P(B)-P(AB)$ where $P(AB)=P(A\cap B)$. Extension of this rule leads to the **Inclusion–Exclusion Principle.** This principle is a way to extend the general addition rule to 3 or more events. Here we will limit it to 3 events. $P(A\cup B\cup C)=P(A)+P(B)-P(AB)-P(AC)-P(BC)+P(ABC)$

6) **Law of Partitions:** The law of partitions is a way to calculate the probability of an event. Let $A_1, A_2, \dots, A_k$ form a partition of the sample space Ω .then, for any events B,

$$P(B)=P(A_1B)+P(A_2B)+\dots.+P(A_kB)=\sum_{i=1}^{k} P(A_iB)$$



Example 1

The Probability that John passes a Maths exam is 4/5 and that he passes a Chemistry exam is 5/6. If the probability that he passes both exams is 3/4, find the probability that he will pass at least one exam.

Solution

Let M be the event thet John passes Math exam, and C be the event thet John passes Chemistry exam.

$$\text{P(John passes at least one exam)} = P(M \cup C) = P(M) + P(C) - P(MC) = \tfrac{4}{5} + \tfrac{5}{6} - \tfrac{3}{4} = \tfrac{53}{60}$$

Example 2

A fair die, with faces numbered 1 to 6, is rolled twice and the sum of the scores showing up noted. Let A be the event that the sum of the scores is greater than 7, B be the event that the sum of the scores is is a multiple of 3 and C be the event that the sum of the scores is a prime number. Show that $P(A \cup B) = P(A) + P(B) - P(AB)$ and also find $P(A \cup C)$, $P(BC)$ and $P(BC')$

Solution

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

$A \cap B$ means the set of all multiples of 3 that are greater than 7. Clearly $P(A \cap B) = \tfrac{5}{36}$

$A \cup B$ means the set of all values that are multiples of 3 and/or greater than 7. Clearly $P(A \cup B) = \tfrac{22}{36} = \tfrac{11}{18} = P(A) + P(B) - P(AB)$

1. $P(A) = \tfrac{15}{36} = \tfrac{5}{12}$ and $P(B) = \tfrac{12}{36} = \tfrac{1}{3}$
2. $P(C) = \tfrac{15}{36} = \tfrac{5}{12}$

$A \cap C$ means the set of all multiples of 3 that are prime number. Clearly $P(A \cap C) = \tfrac{2}{36} = \tfrac{1}{18}$

$P(A \cup C) = P(A) + P(C) - P(AC) = \tfrac{5}{12} + \tfrac{5}{12} - \tfrac{1}{18} = \tfrac{7}{9}$

$B \cap C$ means the set of all greater than 7 that are prime numbers. Clearly $P(BC) = \tfrac{2}{36} = \tfrac{1}{18}$

$B \cap C'$ means the set of all greater than 7 that are not prime numbers. Clearly $P(BC') = \tfrac{11}{36}$

Exercise

1) Which of the following is a probability function defined on $S = \{E_1, E_2,\ E_3\}$
   a) $P(E_1) = \tfrac{1}{4}, P(E_2) = \tfrac{1}{3}$ and $P(E_3) = \tfrac{1}{2}$ b) $P(E_1) = \tfrac{1}{3}, P(E_2) = \tfrac{1}{6}$ and $P(E_3) = \tfrac{1}{2}$
   c) $P(E_1) = \tfrac{2}{3}, P(E_2) = -\tfrac{1}{3}$ and $P(E_3) = \tfrac{2}{3}$ d) $P(E_1) = 0, P(E_2) = \tfrac{1}{3}$ and $P(E_3) = \tfrac{2}{3}$
2) As a foreign language, 40% of the students took Spanish and 30% took French, while 60% took at least one of these languages. What percent of students took both Spanish and French?
3) In a class of 100 students, 30 are in mathematics. Moreover, of the 40 females in the class, 10 are in Mathematics. If a student is selected at random from the class, what is the probability that the student will be a male or be in mathematics?
4) The probability that a car stopped at a road brook will have faulty breaks is 0.23, the probability that it will have badly worn out tyres is 0.24 and the probability that it will have faulty breaks and/or badly worn out tyres is 0.38. Find the probability that a car which has just been stopped will have both faulty breaks and badly worn out tyres.

5) Given two events A and B in the sane sample space such that $\mathrm{P(A)=0.59}$, $\mathrm{P(B)=0.3}$ and $\mathrm{P(AB)=0.21}$. Find; a) $\mathrm{P(A\cup B)}$ b) $\mathrm{P(A'B)}$ c) $\mathrm{P(AB')}$ d) $\mathrm{P(A\cup B')}$

6) Let A and B be two events in the sane sample space such that $\mathrm{P(A\cup B)}=\frac{3}{4}$, $\mathrm{P(B')}=\frac{2}{3}$ and $\mathrm{P(AB)}=\frac{1}{4}$.Find $\mathrm{P(B)}$, $\mathrm{P(A)}$ and $\mathrm{P(AB')}$

7) Suppose that $\mathrm{P(A)=0.4}$, $\mathrm{P(B)=0.5}$ and $\mathrm{P(A\cap B)=0.2}$ Find; a) $\mathrm{P(A\cup B)}$ b) $\mathrm{P(A'B')}$ c) $\mathrm{P}[\mathrm{A'}\cap(A\cup \mathrm{B})]$ d) $\mathrm{P}[\mathrm{A}\cup(\mathrm{A'B})]$

8) A die is loaded such that even numbers are twice as likely as odd numbers. Find the probability that for a single toss of this die the spot showing up is greater than 3

9) A point is selected at random inside an equilateral triangle of sides 3 units. Find the probability that its distance to any corner is greater than 1 unit.

Definition: (Odds of an event)

It's the ratio of the probability of an event occurring to that of the event not happening. If A is an event then the odds of A is given by $\frac{\mathrm{P(A)}}{\mathrm{P(A')}}=\frac{\mathrm{P(A)}}{\mathrm{1\text{-}P(A)}}$

Example

Find $\mathrm{P(A)}$ and $\mathrm{P(A')}$ if the odds of event A is $\frac{5}{4}$

Solution

$\frac{\mathrm{P(A)}}{\mathrm{1\text{-}P(A)}}=\frac{5}{4} \Rightarrow 4(1-\mathrm{P(A)})=5\mathrm{P(A)} \Rightarrow 4=9\mathrm{P(A)} \Rightarrow \mathrm{P(A)}=\frac{4}{9}$ and $\mathrm{P(A')}=\frac{5}{9}$

Question: Find $\mathrm{P(E)}$ and $\mathrm{P(E')}$ if the odds of event E is (i) $\frac{3}{4}$ (ii) $\frac{a}{b}$

5.5 Relationship between Events

- Compound event: Two or more events combined together. Eg AB is a compound event
- Mutually exclusive events: Two events A and B are said to be mutually exclusive if they cannot occur simultaneously. That is if the occurrence of A totally excludes the occurrence of B. Effectively events A and B are said to be mutually exclusive if they disjoint. ie $\mathrm{A\cap B}=\phi \Rightarrow \mathrm{P(AB)=0}$
- Exhaustive events: Events whose union equals the sample space.
- Independent events: Two events A and B are said to be independent if the occurrence of A does not affect the occurrence of B. If events A and B are independent then $\mathrm{P(AB)=P(A)\times P(B)}$

Remark: Three events A, B and C are said to be jointly independent if and only if

(i) $\mathrm{P(AB)=P(A)\times P(B)}$, $\mathrm{P(AC)=P(A)\times P(C)}$ and $\mathrm{P(BC)=P(B)\times P(C)}$ (ie they are pairwise independent) and

(ii) if $\mathrm{P(ABC)=P(A)\times P(B)\times P(C)}$

Note it does not necessarily mean that if events A,B and C are pairwise independent then they are jointly independent

Example 1

Roll a fair die twice and define A to be the event that the sum of the scores showing up is greater than 7, B be the event that the sum of the scores showing up is a multiple of 3 and C be the event that the sum of the scores showing up is a prime number. Which of the events A, B and C are independent? Are the 3 events jointly independent?

Solution

From the above example, $P(A) = P(C) = \frac{5}{12}$, $P(B) = \frac{1}{3}$, $P(AB) = \frac{5}{36}$ and $P(AC) = P(BC) = \frac{1}{18}$
Since $P(AB) \neq 0$ events A and B are not mutually exclusive. Similarly events A & C and B and C are not mutually exclusive

$P(A) \times P(B) = \frac{5}{12} \times \frac{1}{3} = \frac{5}{36} = P(AB) \Rightarrow$ A and B are independent events.

$P(A) \times P(C) = \frac{5}{12} \times \frac{5}{12} = \frac{25}{144} \neq P(AC) \Rightarrow$ A and C are dependent events.

$P(B) \times P(C) = \frac{1}{3} \times \frac{5}{12} = \frac{5}{36} \neq P(BC) \Rightarrow$ B and C are dependent events.

The 3 events are not jointly independent since pairwise independence is not satisfied.

Example 2

Three different machines in a factory have the following probabilities of breaking down during a shift.

Machine	A	B	C
probability	$\frac{4}{15}$	$\frac{3}{10}$	$\frac{2}{11}$

Find the probability that in a particular shift,;
a) All the machines will break down
b) None of the machines will break down.

Solution

Since the events of breaking down of machines are independent, probability that all the machines will break down is given by $P(ABC) = \frac{4}{15} \times \frac{3}{10} \times \frac{2}{11} = \frac{4}{275}$

The probability that none of the machines will break down is given by
$P(ABC) = \frac{11}{15} \times \frac{7}{10} \times \frac{9}{11} = \frac{21}{50}$

Exercises

1) In a game of archery the probability that A hits the target is $\frac{1}{3}$ and the probability that B hits the target is $\frac{2}{5}$. What is the probability that the target will be hit?
2) Toss a fair coin 3 times and let A be the event that two or more heads appears, B be the event that all outcomes are the same and C be the event that at most two tails appears. Which of the events A, B and C are independent? Are the 3 events jointly independent?
3) A fair coin and a fair die are rolled together once. Let A be the event that a head and an even number appears, B be the event that a prime number appears and C be the event that a tail and an odd number appears.
   a) Express explicitly the event that i) A and B occurs ii) Only B occurs iii) B and C occur
   b) Which of the events A, B and C are independent and which ones are mutually exclusive?
4) A die is loaded so that the probability of a face showing up is proportional to the face number. Write down the probability of each sample point. If A is the event that an even number appears, B is the event that a prime number appears and C is the event that an odd number appears.
   a) Find the probability that: i) A and/or B occurs ii) A but not B occurs iii) B and C occurs d) A and/or C occurs
   b) Which of the events A, B and C are independent and which ones are mutually exclusive?

Theorem 1: If events A and B are independent, then A and B' are also independent

Proof

Decomposing A into two disjoint events AB and AB'. We can write $P(A)=P(AB)+P(AB') \Rightarrow P(AB')=P(A)-P(AB)=P(A)-P(A)\times P(B)$ since events A and B are independent. Thus $P(AB')=P(A)[1-P(B)]=P(A)\times P(B') \Rightarrow$ A and B' are independent

Theorem 2: If events A and B are independent, then A' and B' are also independent

Proof

Decomposing B' into two disjoint events AB' and $A'B'$. We can write $P(B')=P(AB')+P(A'B') \Rightarrow P(A'B')=P(B')-P(AB')=P(B')-P(A)\times P(B')$ since events A and B' are independent. (From theorem 1 above) Thus $P(A'B')=[1-P(A)]P(B')=P(A')\times P(B') \Rightarrow$ A' and B' are also independent

5.6 Counting Rules useful in Probability

In some experiments it is helpful to list the elements of the sample space systematically by means of a tree diagram,. In many cases, we shall be able to solve a probability problem by counting the number of points in the sample space without actually listing each element.

Theorem (Multiplication principle)

If one operation can be performed in n_1 ways, and if for each of these a second operation can be performed in n_2 ways, then the two operations can be performed together in n_1n_2 ways.

Eg How large is the sample space when a pair of dice is thrown?

Solution; The first die can be thrown in $n_1 = 6$ ways and the second in $n_2 = 6$ Ways. Therefore, the pair of dice can land in $n_1n_2 = 36$ possible ways.

The above theorem can naturally be extended to more than two operations: if we have $n_1, n_2, ..., n_k$ consequent choices, then the total number of ways is $n_1 \times n_2 \times .... \times n_k$

Permutations

Permutations refer to an arrangement of objects when the order matters (for example, letters in a word).The number of permutations of n distinct objects taken r at a time is ${}_nP_r = \frac{n!}{(n-r)!}$

Example

From among ten employees, three are to be selected to travel to three out-of-town plants A, B, and C, one to each plant. Since the plants are located in different cities, the order in which the employees are assigned to the plants is an important consideration. In how many ways can the assignments be made?

Solution;

Because order is important, the number of possible distinct assignments is ${}_{10}P_3 = 720$

In other words, there are ten choices for plant A, but then only nine for plant B, and eight for plant C. This gives a total of 10(9)(8) ways of assigning employees to the plants.

Combinations

The term combination refers to the arrangement of objects when order does not matter. For example, choosing 4 books to buy at the store in any order will leave you with the same set of books. The number of distinct subsets or combinations of size r that can be selected from n distinct objects, (r _ n), is given by ${}_nC_r = \frac{n!}{r!(n-r)!}$

Example 1

In the previous example, suppose that three employees are to be selected from among the ten available to go to the same plant. In how many ways can this selection be made?

Solution

Here, order is not important; we want to know how many subsets of size r = 3 can be selected from n = 10 people. The result is ${}_{10}C_3 = 120$

Example 2

In a poker hand consisting of 5 cards, find the probability of holding 2 aces and 3 jacks.

Solution

The number of ways of being dealt 2 aces from 4 is ${}_4C_2 = 6$ and the number of ways of being dealt 3 jacks from 4 is ${}_4C_3 = 4$

The total number of 5-card poker hands, all of which are equally likely is ${}_{52}C_5 = 2,598,960$

Hence, the probability of getting 2 aces and 3 jacks in a 5-card poker hand is P(C) =

$$P(C) == \frac{6 \times 4}{2,598,960}$$

Example 3

A university warehouse has received a shipment of 25 printers, of which 10 are laser printers and 15 are inkjet models. If 6 of these 25 are selected at random to be checked by a particular technician, what is the probability that; a) exactly 3 of these selected are laser printers? b) at least 3 inkjet printers?

Solution

First choose 3 of the 15 inkjet and then 3 of the 10 laser printers.

There are ${}_{15}C_3$ and ${}_{10}C_3$ ways to do it, and therefore

$$P(\text{exactly } 3 \text{ of the } 6) = \frac{{}_{15}C_3 \times {}_{10}C_3}{{}_{25}C_6} = 0.3083$$

$$P(\text{at least } 3) = \frac{{}_{15}C_3 \times {}_{10}C_3}{{}_{25}C_6} + \frac{{}_{15}C_4 \times {}_{10}C_2}{{}_{25}C_6} + \frac{{}_{15}C_5 \times {}_{10}C_1}{{}_{25}C_6} + \frac{{}_{15}C_6 \times {}_{10}C_0}{{}_{25}C_6} = 0.8530$$

Exercises

1) An incoming lot of silicon wafers is to be inspected for defectives by an engineer in a microchip manufacturing plant. Suppose that, in a tray containing 20 wafers, 4 are defective. Two wafers are to be selected randomly for inspection. Find the probability that neither is defective.
2) A person draws 5 cards from a shuffled pack of cards. Find the probability that the person has at least 3 aces. Find the probability that the person has at least 4 cards of the same suit.
3) A California licence plate consists of a sequence of seven symbols: number, letter, letter, letter, number, number, number, where a letter is any one of 26 letters and a number is one among *0, 1,… 9*. Assume that all licence plates are equally likely. What is the probability that;
    a) all symbols are different?
    b) b) all symbols are different and the first number is the largest among the numbers?
4) A bag contains 80 balls numbered 1.... 80. Before the game starts, you choose 10 different numbers from amongst 1.... 80 and write them on a piece of paper. Then 20 balls are selected (without replacement) out of the bag at random. What is the probability that;
    a) all your numbers are selected?
    b) none of your numbers is selected?

    c) exactly 4 of your numbers are selected?

5) A full deck of 52 cards contains 13 hearts. Pick 8 cards from the deck at random (a) without replacement and (b) with replacement. In each case compute the probability that you get no hearts.
6) Three people enter the elevator on the basement level. The building has 7 floors. Find the probability that all three get off at different floors.
7) In a group of 7 people, each person shakes hands with every other person. How many handshakes did occur?
8) A marketing director considers that there's “overwhelming agreement" in a 5-member focus group when either 4 or 5 people like or dislike the product. If, in fact, the product's popularity is 50% (so that all outcomes are equally likely), what is the probability that the focus group will be in “overwhelming agreement" about it? Is the marketing director making a judgement error in declaring such agreement “overwhelming"?
9) A die is tossed 5 times. Find the probability that we will have 4 of a kind.
10) A tennis tournament has 2*n* participants, *n* Swedes and *n* Norwegians. First, *n* people are chosen at random from the 2*n* (with no regard to nationality) and then paired randomly with the other *n* people. Each pair proceeds to play one match. An outcome is a *set* of *n* (ordered) pairs, giving the winner and the loser in each of the *n* matches. (a) Determine the number of outcomes. (b) What do you need to assume to conclude that all outcomes are equally likely? (c) Under this assumption, compute the probability that all Swedes are the winners.
11) A group of 18 Scandinavians consists of 5 Norwegians, 6 Swedes, and 7 Finns. They are seated at random around a table. Compute the following probabilities: (a) that all the Norwegians sit together, (b) that all the Norwegians and all the Swedes sit together, and (c) that all theNorwegians, all the Swedes, and all the Finns sit together.
12) In a lottery, 6 numbers are drawn out of 45. You hit a jackpot if you guess all 6 numbers correctly, and get $400 if you guess 5 numbers out of 6. What are the probabilities of each of those events?
13) There are 21 Bachelor of Science programs at New Mexico Tech. Given 21 areas from which to choose, in how many ways can a student select:
    a) A major area and a minor area?
    b) A major area and first and second minor?
14) From a box containing 5 chocolates and 4 hard candies, a child takes ahandful of 4 (at random). What is the probability that exactly 3 of the 4 arechocolates?
15) If a group consist of 8 men and 6 women, in how many ways can a committee of 5 be selected if:
    a) The committee is to consist of 3 men and 3 women.
    b) There are no restrictions on the number of men and women on the committee.
    c) There must at least one man.
    d) There must be at least one of each sex.
16) Suppose we have a lot of 40 transistors of which 8 are defective. If we sample without replacement, what is the probability that we get 4 good transistors in the first 5 draws?
17) A housewife is asked to rank four brands A, B, C, and D of household cleaner according to her preference, number one being the one she prefers most, etc. she really has no preference among the four brands. Hence, any ordering is equally likely to occur.
    a) Find the probability that brand A is ranked number one.
    b) Find the probability that brand C is number one D is number 2 in the rankings.
    c) Find the probability that brand A is ranked number one or number 2.
18) How many ways can one arrange the letters of the word ADVANTAGE so
19) that the three As are adjacent to each other?

20) Eight tires of different brands are ranked 1 to 8 (best to worst) according to mileage performance. If four of these tires are chosen at random by a customer, find the probability that the best tire among the four selected by the customer is actually ranked third among the original eight.

5.7 Conditional Probability and Independence

Humans often have to act based on incomplete information. If your boss has looked at you gloomily, you might conclude that something's wrong with your job performance. However, if you know that she just suffered some losses in the stock market, this extra information may change your assessment of the situation. Conditional probability is a tool for dealing with additional information like this.

Conditional probability is the probability of an event occurring given the knowledge that another event has occurred. The conditional probability of event A occurring, given that event B has occurred is denoted by P(A/B) and is read "probability of A given B" and is given by

$$\mathrm{P(A/B)} = \frac{\mathrm{P(AB)}}{\mathrm{P(B)}} \quad \text{provided} \quad \mathrm{P(B)} > 0 \ \ \text{Similarly} \ \ \mathrm{P(B/A)} = \frac{\mathrm{P(AB)}}{\mathrm{P(A)}} \quad \text{provided} \ \ \mathrm{P(A)} > 0$$

$$\Rightarrow \ \mathrm{P(AB)} = \mathrm{P(A/B)} \times \mathrm{P(B)} = \mathrm{P(B/A)} \times \mathrm{P(A)}$$

Remark: Another way to express independence is to say that the knowledge of B occurring does not change our assessment of P(A). This means that if A and B are independent then $\mathrm{P(A/B)} = \mathrm{P(A)}$ and $\mathrm{P(B/A)} = \mathrm{P(B)}$

Example

In a large metropolitant area, the probability of a family owning a colour T.V , a computer or both 0.86, 0.35 and 0.29 respectively. What is the probability that a family chosen at random during a survey will own a colour T.V and/or a computer? Given that the family chosen at random during a survey owns a colour T.V, what is the probability that it will own a computer?

Solution

Let T and C be the event of owning a colour T.V and a computer respectively. Then

$$\mathrm{P(T \cup C)} = \mathrm{P(T)} + \mathrm{P(C)} - \mathrm{P(TC)} = 0.86 + 0.35 \text{-} 0.29 = 0.92$$

$$\mathrm{P(C/T)} - \frac{\mathrm{P(TC)}}{\mathrm{P(T)}} = \frac{0.29}{0.86} \approx 0.$$

Reduced sample space approach

In case when all the outcomes are equally likely, it is sometimes easier to find conditional probabilities directly, without having to apply the above equation. If we already know that B has happened, we need only to consider outcomes in B, thus reducing our sample space to B.

Then, $\mathrm{P(A/B)} = \dfrac{\text{Number of outcomes in AB}}{\text{Number of outcomes in B}}$

For example, $\text{P(a die is 3 / a die is odd)} = \frac{1}{3}$ and $\text{P(a die is 4 / a die is odd)} = 0$

Example

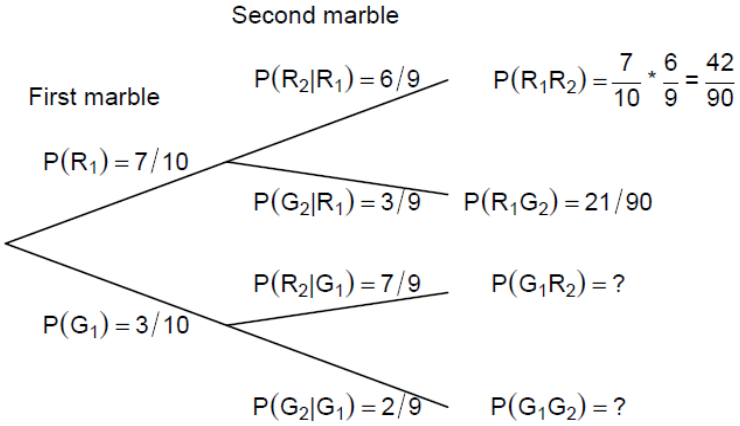
Let A = {a family has two boys} and B = {a family of two has at least one boy} Find $\mathrm{P(A/B)}$

Solution

The event B contains the following outcomes: $\mathrm{B} = \{(\mathrm{B,B}), (\mathrm{B,G}), (\mathrm{G;B})\}$ and. Only one of these is in A. Thus, $\mathrm{P(A/B)} = \frac{1}{3}$. However, if I know that the family has two children, and I see one of the children and it's a boy, then the probability suddenly changes to 1/2. There is a subtle difference in the language and this changes the conditional probability

5.7.1 Tree Diagrams in conditional probability

Suppose we are drawing marbles from a bag that initially contains 7 red and 3 green marbles. The drawing is without replacement that is after we draw the first marble, we do not put it back. Let's denote the events $R_1 = \{\text{the first marble is red}\}$ $R_2 = \{\text{the second marble is red}\}$ $G_1 = \{\text{tthe first marble is green}\}$ and so on. Let's fill out the tree representing the consecutive choices.



The conditional probability $P(R_2 / R_1)$ can be obtained directly from reasoning that after we took the first red marble, there remain 6 red and 3 green marbles. On the other hand, we could use the formula to get $P(R_2 / R_1) = \frac{P(R_1 R_2)}{P(R_1)} = \frac{42/90}{7/10} = \frac{2}{3}$ where the probability $P(R_2\ R_1)$ {same as $P(R_1 R_2)$} can be obtained from counting the outcomes $P(R_1 R_2) = \frac{{}_7C_2}{{}_{10}C_2} = \frac{7}{15}$

Question: Find $P(R_2)$ and $P(R_1 / R_2)$.

Example 2

Suppose that of all individuals buying a certain digital camera, 60% include an optional memory card in their purchase, 40% include a set of batteries, and 30% include both a card and batteries. Consider randomly selecting a buyer and let $A = \{\text{memory card purchased}\}$ and $B = \{\text{battery purchased}\}$. Then find P(A/B) and P(B/A).

Solution

From given information, we have $P(A) = 0.60$, $P(B) = 0.40$ and

$P(\text{both purchased}) = P(A/B) = 0.30$

Given that the selected individual purchased an extra battery, the probability that an optional card was also purchased is $P(A/B) = \frac{P(AB)}{P(B)} = \frac{0.30}{0.40} = 0.75$

That is, of all those purchasing an extra battery, 75% purchased an optional memory card.

Similarly $P(\text{battery j memory card}) = P(B/A) = \frac{P(AB)}{P(A)} = \frac{0.30}{0.60} = 0.5$

Notice that $P(A/B) \neq P(A)$ and $P(B/A) \neq P(B)$, that is, the events A and B are dependent.

Remark: The tree diagram may become tedious especially when the tree grows beyond 4 stages. In such a case we can make use of binomial formular which is applicable when:

i) The experiment's outcome can be classified into 2 categories success and failure with probabilities p and 1-p respectively

ii) The experiment is to be repeated n independent times

iii) Our interest is the number of successes

The probability of observing x successes out of n trials is given by:-

$$P(x) = {}_{n}C_{x} \times p^{x}(1-p)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n$$

Example

A fair coin is tossed 10 times.Whai is the probability of observing exactly 8 heads?

Solution

$n = 10 \quad p = 0.5$ 5 and x = 8 successes Therefore

$$P(X = 8) = {}_{10}C_{8} \times 0.5^{8}(1-0.5)^{10-8} = {}_{10}C_{8} \times 0.5^{10} \approx 0.044$$

Exercises

1) A pair of fair dice is rolled once. If the sum of the scores showing up is 6 , find the probability that one of the dice shows a 2.
2) A consumer research organisation has studied the services and warranty provided by 50 new car dealers in a certain city. It's findings are as follows

In Business for	**Good services and a warranty**	**Poor services and a warranty**
At least 10 yeras	16	4
Less than 10 years	10	20

If a person randomly selects one of these new car dealers ;

a) What is the probability that he gets one who provides good services and a warranty

b) Who has been in business for at least 10 years, what is the probability that hehe provides good services and a warranty

c) What is the probability that one of these new car dealers who has been in business for less than 10 years will provide good services and a warranty?

3) Three machines A, B and C produces 50%, 30% and 20% respectively of the total number of items in a factory. The percentage of defective outputs of these machines are 3%, 4% and 5% respectively. If an item is selected at random:-

a) Find the probability that it is defective

b) And found to be defective, what is the probability that it was produced by machine A?

4) A year has 53 Sundays. What is the conditional probability that it is a leap year?
5) The probability that a majority of the stockholders of a company will attend a special meeting is 0.5. If the majority attends, then the probability that an important merger will be approved is 0.9. What is the probability that a majority will attend and the merger will be approved?
6) Let events A, B have positive probabilities. Show that, if P(A/B) = P(A) then also P(B/A) = P(B).
7) The cards numbered 1 through 10 are placed in a hat, mixed up, and then one of the cards is drawn. If we are told that the number on the drawn card is at least five, what is the probability that it is ten?
8) In the roll of a fair die, consider the events A = {2, 4, 6} = “even numbers" and B = {4, 5, 6} =”high scores". Find the probability that die showing an even number given that it is a high score.
9) There are two urns. In the first urn there are 3 white and 2 black balls and in the second urn there 1 white and 4 black balls. From a randomly chosen urn, one ball is drawn. What is the probability that the ball is white?
10) The level of college attainment of US population by racial and ethnic group in 1998 is given in the following table

Racial or Ethnic Group	No of Adults (Millions)	%age with Associate's Degree	%age With Bachelor's Degree	%age with Graduate or Professional Degree
Native Americans	1.1	6.4	6.1	3.3

Blacks	16.8	5.3	7.5	3.8
Asians	4.3	7.7	22.7	13.9
Hispanics	11.2	4.8	5.9	3.3
Whites	132.0	6.3	13.9	7.7

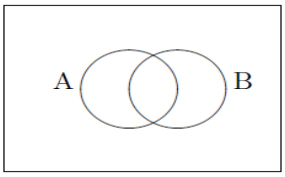
The percentages given in the right three columns are conditional percentages.

a) How many Asians have had a graduate or professional degree in 1998?
b) What percent of all adult Americans has had a Bachelor's degree?
c) Given that the person had an Associate's degree, what is the probability that the person was Hispanic?

11) The dealer's lot contains 40 cars arranged in 5 rows and 8 columns. We pick one car at random. Are the events A = {the car comes from an odd-numbered row} and B = {the car comes from one of the last 4 columns} independent? Prove your point of view.

12) You have sent applications to two colleges. If you are considering your chances to be accepted to either college as 60%, and believe the results are statistically independent, what is the probability that you'll be accepted to at least one? How will your answer change if you applied to 5 colleges?

13) In a high school class, 50% of the students took Spanish, 25% took French and 30% of the students took neither. Let A be the event that a randomly chosen student took Spanish, and B be the event that a student took French. Fill in either the Venn diagram or a 2-way table and answer the questions:



	B	B'	
A			
A'			

a) Describe in words the meaning of the event AB' . Find the probability of this event.
b) Are the events A, B independent? Explain with numbers why or why not.
c) If it is known that the student took Spanish, what are the chances that she also took French?

14) One half of all female physicists are married. Among those married, 50% are married to other physicists, 29% to scientists other than physicists and 21% to non-scientists’. Among male physicists, 74% are married. Among them, 7% are married to other physicists, 11% to scientists other than physicists and 82% to non-scientists What percent of all physicists are female? [Hint: This problem can be solved as is, but if you want to, assume that physicists comprise 1% of all population.]

15) Error-correcting codes are designed to withstand errors in data being sent over communication lines. Suppose we are sending a binary signal (consisting of a sequence of 0's and 1's), and during transmission, any bit may get flipped with probability p, independently of any other bit. However, we might choose to repeat each bit 3 times. For example, if we want to send a sequence 010, we will code it as 000111000. If one of the three bits flips, say, the receiver gets the sequence 001111000, he will still be able to decode it as 010 by majority voting. That is, reading the first three bits, 001, he will interpret it as an attempt to send 000. However, if two of the three bits are flipped, for example 011, this will be interpreted as an attempt to send 111, and thus decoded incorrectly. What is the probability of a bit being decoded incorrectly under this scheme?

5.8 Bayes’ Rule

Events $B_1, B_2, \ldots, B_K$ are said to be a partition of the sample space S if the following two conditions are satisfied. i) $B_i B_J = \phi$ for each pair i, j and ii) $B_1 \cup B_2 \cup \ldots \cup B_K = S$

This situation often arises when the statistics are available in subgroups of a population. For example, an insurance company might know accident rates for each age group B_i. This will give the company conditional probabilities $P(A/B_i)$ (if we denote A = {event of accident}.

Question: if we know all the conditional probabilities $P(A/B_i)$, how do we find the unconditional P(A)?

Consider a case when k = 2:

The event A can be written as the union of mutually exclusive events AB_1 and AB_2, that is $A = AB_1 \cup AB_2$ it follows that $P(A) = P(AB_1) + P(AB_2)$

If the conditional probabilities $P(A/B_1))$ and $P(A/B_2))$ are known, that is

$$P(A/B_1) = \frac{P(AB_1)}{P(B_1)} \text{ and } P(A/B_2) = \frac{P(AB_2)}{P(B_2)} \quad \text{then} \quad P(A) = P(B_1) \times P(A/B_1) + P(B_2) \times P(A/B_2)$$

Suppose we want to find probability of the form $P(B_i/A)$, which can be written as

$$P(B_i/A) = \frac{P(AB_i)}{P(A)} = \frac{P(B_i) \times P(A/B_i)}{P(A)} = \frac{P(B_i) \times P(A/B_i)}{P(B_1) \times P(A/B_1) + P(B_2) \times P(A/B_2)}$$

This calculation generalizes to k > 2 events as follows.

Theorem

If $B_1, B_2, \ldots, B_K$ form a partition of the sample space S such that $P(B_i) \neq 0$ for i = 1, 2, . . . k; then for any event $A \subseteq S$,

$$P(A) = \sum_{i=1}^{k} P(AB_i) = \sum_{i=1}^{k} P(B_i) \times P(A/B_i) \quad \text{Subsequently,} \quad P(B_i/A) = \frac{P(AB_i)}{P(A)} = \frac{P(B_i) \times P(A/B_i)}{\sum_{i=1}^{k} P(B_i) \times P(A/B_i)}$$

This last equation is often called Law of Total Probability.

Example 1

A rare genetic disease (occuring in 1 out of 1000 people) is diagnosed using a DNA screening test. The test has false positive rate of 0.5%, meaning that $P(\text{test positive } / \text{ no disease}) = 0.005$. Given that a person has tested positive, what is the probability that this person actually has the disease? First, guess the answer, then read on.

Solution

Let's reason in terms of actual numbers of people, for a change.

Imagine 1000 people, 1 of them having the disease. How many out of 1000 will test positive? One that actually has the disease, and about 5 disease-free people who would test false positive. Thus, $P(\text{disease/test positive}) \approx \frac{1}{6}$.

It is left as an exercise for the reader to write down the formal probability calculation.

Example 2

At a certain assembly plant, three machines make 30%, 45%, and 25%, respectively, of the products. It is known from the past experience that 2%, 3% and 2% of the products made by each machine, respectively, are defective. Now, suppose that a finished product is randomly selected.

a) What is the probability that it is defective?
b) If a product were chosen randomly and found to be defective, what is the probability that it was made by machine 3?

Solution

Consider the following events:

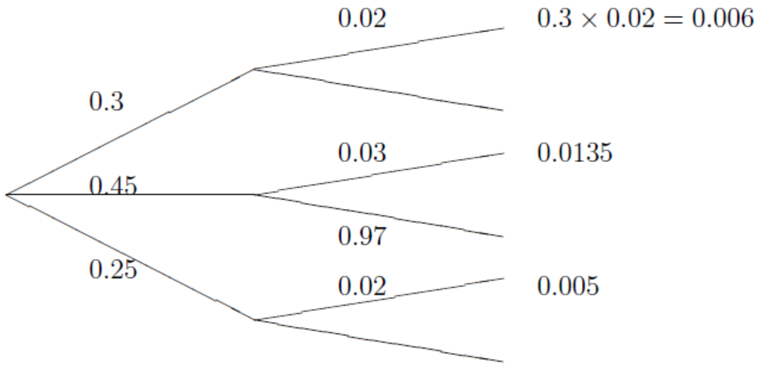
A: the product is defective and B_i : the product is made by machine i=1, 2, 3,
Applying additive and multiplicative rules, we can write

(a) $P(A) = P(B_1) \times P(A/B_1) + P(B_2) \times P(A/B_2) + P(B_3) \times P(A/B_3)$

$$= (0.3)(0.02) + (0.45)(0.03) + (0.25)(0.02) = 0.006 + 0.0135 + 0.005 = 0.0245$$

(b) Using Bayes' rule $P(B_3/A) = \dfrac{P(B_3) \times P(A/B_3)}{P(A)} = \dfrac{0.005}{0.0245} = 0.2041$

This calculation can also be represented using a tree diagram as follows



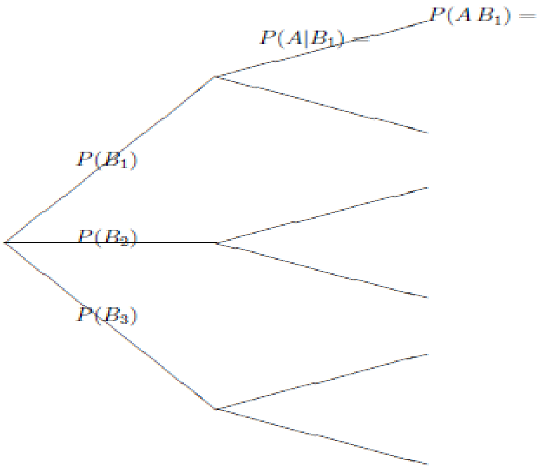
Here, the first branching represents probabilities of the events B_1 and the second branching represents conditional probabilities $P(A/B_1)$. The probabilities of intersections, given by the products, are on the right. P(A) is their sum.

Exercises

1) Lucy is undecided as to whether to take a Math course or a Chemistry course. She estimates that her probability of receiving an A grade would be 0.5 in a math course, and $\frac{2}{3}$ in a chemistry course. If Lucy decides to base her decision on the flip of a fair coin, what is the probability that she gets an A?
2) Of the customers at a gas station, 70% use regular gas, and 30% use diesel. Of the customers who use regular gas, 60% will fill the tank completely, and of those who use diesel, 80% will fill the tank completely.
   a) What percent of all customers will fill the tank completely?
   b) If a customer has filled up completely, what is the probability it was a customer buying diesel?
3) In 2004, 57% of White households directly and/or indirectly owned stocks, compared to 26% of Black households and 19% of Hispanic households. The data for Asian households is not given, but let's assume the same rate as for Whites. Additionally, 77% of households are classifieds as either White or Asian, 12% as African American, and 11% as Hispanic.
   a) What proportion of all families owned stocks?
   b) If a family owned stock, what is the probability it was White/Asian?
4) Drawer one has five pairs of white and three pairs of red socks, while drawer two has three pairs of white and seven pairs of red socks. One drawer is selected at random a pair of socks is selected at random from that drawer.
   a) What is the probability that it is a white pair of socks?
   b) Suppose a white pair of socks is obtained. What is the probability that it came from drawer two?
5) For an on-line electronics retailer, 5% of customers who buy Zony digital cameras will return them, 3% of customers who buy Lucky Star digital cameras will return them, and 8% of customers who buy any other brand will return them. Also, among all digital

cameras bought, there are 20% Zony's and 30% Lucky Stars. Fill in the tree diagram and answer the questions.

a) What percent of all cameras are returned?
b) If the camera was just returned, what is the probability it is a Lucky Star?
c) What percent of all cameras sold were Zony and were not returned?



6) Three newspapers, A, B, and C are published in a certain city. It is estimated from a survey that that of the adult population: 20% read A, 16% read B, 14% read C, 8% read both A and B, 5% read both A and C, 4% read both B and C, 2% read all three. What percentage reads at least one of the papers? Of those that read at least one, what percentage reads both A and B?

7) Suppose $P(A/B) = 0.3$, $P(B) = 0.4$ and $P(B/A) = 0.6$. Find $P(A)$ and $P(A \cup B)$

8) This is the famous Monty Hall problem f A contestant on a game show is asked to choose among 3 doors. There is a prize behind one door and nothing behind the other two. You (the contestant) have chosen one door. Then, the host is flinging one other door open, and there's nothing behind it. What is the best strategy? Should you switch to the remaining door, or just stay with the door you have chosen? What is your probability of success (getting the prize) for either strategy?

9) There are two children in a family. We overheard about one of them referred to as a boy.
   a) Find the probability that there are 2 boys in the family.
   b) Suppose that the oldest child is a boy. Again, find the probability that there are 2 boys in the family.g [Why is it different from part (a)?]

10) At a university, two students were doing well for the entire semester but failed to show up for a final exam. Their excuse was that they travelled out of state and had a at tire. The professor gave them the exam in separate rooms, with one question worth 95 points: \which tire was it?". Find the probability that both students mentioned the same tire.

11) In firing the company's CEO, the argument was that during the six years of her tenure, for the last three years the company's market share was lower than for the first three years. The CEO claims bad luck. Find the probability that, given six random numbers, the last three are the lowest among six.

6 RNNDOM VARIABLES

In this section, we will consider random quantities that are usually called random variables.

Introduction

In application of probability, we are often interested in a number associated with the outcome of a random experiment. Such a quantity whose value is determined by the outcome of a random experiment is called a **random variable**. It can also be defined as any quantity or attribute whose value varies from one unit of the population to another.
A **discrete** random variable is function whose range is finite and/or countable, Ie it can only assume values in a finite or count ably infinite set of values. A **continuous** random variable is one that can take any value in an interval of real numbers. (There are *unaccountably* many real numbers in an interval of positive length.)

6.1 Discrete Random Variables and Probability Mass Function

A random variable X is said to be discrete if it can take on only a finite or countable number of possible values x. Consider the experiment of flipping a fair coin three times. The number of tails that appear is noted as a discrete random variable. *X*= *"*number of tails that appear in 3 flips of a fair coin". There are 8 possible outcomes of the experiment: namely the sample space consists of

$$S=\{\text{HHH}, \text{HHT}, \text{HTH}, \text{HTT}, \text{THH}, \text{THT}, \text{TTH}, \text{TTT}\}$$

$$\text{X}=\{\ 0 \quad 1, \quad 1, \quad 2\ , \quad 1, \quad 2, \quad 2, \quad 3\}$$

are the corresponding values taken by the random variable *X.*

Now, what are the possible values that *X* takes on and what are the probabilities of *X* taking a particular value?

From the above we see that the possible values of *X* are the 4 values

$$\text{X}=\{0,1,2,3\}$$

Ie the sample space is a disjoint union of the 4 events { *X* = *j* } for *j*=0,1,2,3
Specifically in our example :

$\{\text{X}=0\}=\{\text{HHH}\}$ $\{\text{X}=1\}=\{\text{HHT, HTH, THH}\}$

$\{\text{X}=2\}=\{\text{TTH, HTT, THT}\}$ $\{\text{X}=3\}=\{\text{TTT}\}$

Since for a fair coin we assume that each element of the sample space is equally likely (with probability $\frac{1}{8}$, we find that the probabilities for the various values of *X*, called the *probability distribution* of *X* or the *probability mass function (pmf).* can be summarized in the following table listing the possible values beside the probability of that value

x	0	1	2	3
P(X=x)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Note: The probability that *X* takes on the value *x*, ie $\text{p}(\text{X}=\text{x})$, is defined as the sum of the probabilities of all points in *S* that are assigned the value *x*.

We can say that this pmf places mass $\frac{3}{8}$ on the value $\text{X}=2$.

The "masses" (or probabilities) for a pmf should be between 0 and 1.
The total mass (i.e. total probability) must add up to 1.

Definition: The **probability mass function** of a discrete variable is a, table, formula or graph that specifies the proportion (or probabilities) associated with each possible value the random variable can take. The mass function $\text{P}(\text{X}=\text{x})$ (or just p(x) has the following properties:

$$0\le \text{p}(\text{x})\le 1 \text{ and} \sum_{\text{all x}} \text{p}(\text{x})=1$$

More generally, let X have the following properties

i) It is a discrete variable that can only assume values $x_1, x_2, ... x_n$

ii) The probabilities associated with these values are $P(X=x_1)=p_1$, $P(X=x_2)=p_2$ …… $P(X=x_n)=p_n$

Then X is a discrete random variable if $0 \le p_i \le 1$ and $\sum_{i=1}^{n} p_i = 1$

Remark: We denote random variables with capital letters while realized or particular values are denoted by lower case letters.

Example 1

Two tetrahedral dice are rolled together once and the sum of the scores facing down was noted. Find the pmf of the random variable 'the sum of the scores facing down.'

Solution

+	1	2	3	4
1	2	3	4	5
2	3	4	5	6
3	4	5	6	7
4	5	6	7	8

$X = \{1, 2, 3, 4, 5, 6, 7, 8\}$

Therefore the pmf is given by the table below

x	2	3	4	5	6	7	8
P(X=x)	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{3}{16}$	$\frac{1}{4}$	$\frac{3}{16}$	$\frac{1}{8}$	$\frac{1}{16}$

This can also be written as a function

$$P(X=x)=\begin{cases} \frac{x-1}{16} \text{ for } x = 2, 3, 4, 5 \\ \frac{9-x}{16} \text{ for } x = 6, 7, 8 \end{cases}$$

Example 2

The pmf of a discrete random variable W is given by the table below

w	-3	-2	-1	0	1
P(W=w)	0.1	0.25	0.3	0.15	d

Find the value of the constant d, $P(-3 \le \text{w} < 0)$, $P(\text{w} > \text{-}1)$ and $P(-1 < \text{w} < 1)$

Solution

$$\sum_{\text{all w}} \text{p(W = w)} = 1 \;\Rightarrow\; 0.1 + 0.25 + 0.3 + 0.15 + d = 1 \;\Rightarrow\; d = 0.2$$

$$P(-3 \le \text{w} < 0) = P(W=-3) + P(W=-2) + P(W=-1) = 0.65$$

$$P(\text{w} > \text{-}1) = P(\text{w} = 0) + P(\text{w} = 1) = 0.15 + 0.2 = 0.35$$

$$P(-1 < \text{w} < 1) = P(W=0) = 0.15$$

Example 3

A discrete random variable Y has a pmf given by the table below

y	0	1	2	3	4
P(Y=y)	c	2c	5c	10c	17c

Find the value of the constant c hence computes $P(1 \le \text{Y} < 3)$

Solution

$$\sum_{\text{ally}} \text{p(Y = y)} = 1 \;\Rightarrow\; c(1+2+5+10+17) = 1 \;\Rightarrow\; c = \tfrac{1}{35}$$

$$P(1 \le \text{Y} < 3) = P(Y=1) + P(Y=2) = \tfrac{2}{35} + \tfrac{5}{35} = \tfrac{1}{5}$$

Exercise

1. A die is loaded such that the probability of a face showing up is proportional to the face number. Determine the probability of each sample point.
2. Roll a fair die and let X be the square of the score that show up. Write down the probability distribution of X hence compute $P(X < 15)$ and $P(3 \le \text{X} < 30)$
3. Let X be the random variable the number of fours observed when two dice are rolled together once. Show that X is a discrete random variable.

4. The pmf of a discrete random variable X is given by $P(X = x) = kx$ for $x = 1, 2, 3, 4, 5, 6$ Find the value of the constant k, $P(X < 4)$ and $P(3 \leq \mathrm{X} < 6)$
5. A fair coin is flip until a head appears. Let N represent the number of tosses required to realize a head. Find the pmf of N c , $P(\mathrm{N} < 2)$ and $P(\mathrm{N} \geq 2)$
6. A discrete random variable Y has a pmf given by $P(Y = y) = c\left(\frac{3}{4}\right)^x$ for $y = 0, 1, 2, ...$ Find the value of the constant c , $P(\mathrm{X} < 3)$ and $P(\mathrm{X} \geq 3)$
7. Verify that $\mathrm{f(x)} = \dfrac{2x}{k(k+1)}$ for $y = 0, 1, 2, ...k$ can serve as a pmf of a random variable X.
8. For each of the following determine c so that the function can serve as a pmf of a random variable X.
   a. $\mathrm{f(x)} = c$ for $x = 1, 2, 3, 4, 5$
   b. $\mathrm{f(x)} = cx$ for $x = 1, 2, 3, 4, 5$
   c. $\mathrm{f(x)} = cx^2$ for $x = 0, 1, 2, ...k$
   d. $\mathrm{f(x)} = \frac{c}{2}$ for $x = -1, 0, 1, 2$
   e. $\mathrm{f(x)} = \frac{(x-2)}{c}$ for $x = 1, 2, 3, 4, 5$
   f. $\mathrm{f(x)} = \frac{(x^2-x+1)}{c}$ for $x = 1, 2, 3, 4, 5$
   g. $\mathrm{f(x)} = c\left(x^2 + 1\right)$ for $x = 0, 1, 2, 3$
   h. g) $\mathrm{f(x)} = cx\left({}_3C_x\right)$ for $x = 1, 2, 3$
   i. $\mathrm{f(x)} = c\left(\frac{1}{6}\right)^x$ for $x = 0, 1, 2, 3.....$
   j. $\mathrm{f(x)} = c2^{-x}$ for $x =$ for $x = 0, 1, 2, ...$
9. A coin is loaded so that heads is three times as likely as the tails.
   a. For 3 independent tosses of the coin find the pmf of the total number of heads realized and the probability of realizing at most 2 heads.
   b. A game is played such that you earn 2 points for a head and loss 5 points for a tail. Write down the probability distribution of the total scores after 4 independent tosses of the coin
10. For an on-line electronics retailer, X = “the number of Zony digital cameras returned per day” follows the distribution given by

x	0	1	2	3	4	5
P(X=x)	0.05	0.1	t	0.2	0.25	0.1

    Find the value of t and P(X > 3)
11. Out of 5 components, 3 are domestic and 2 are imported. 3 components are selected at random (without replacement). Obtain the PMF of X =”number of domestic components picked” (make a table).

6.2 Continuous Random Variables and Probability Density Function

A **continuous** random variable can assume any value in an interval on the real line or in a collection of intervals. The sample space is uncountable. For instance, suppose an experiment involves observing the arrival of cars at a certain period of time along a highway on a particular day. Let T denote the time that lapses before the 1st arrival, the T is a continuous random variable that assumes values in the interval $[0, \infty)$

Definition: A random variable *X* is *continuous* if there exists a nonnegative function f so that, for every interval *B*, $P(X \in B) = \int_B \mathrm{f(x)\,dx}$. The function $\mathrm{f = f(x)}$ is called the *probability density function* of *X*.

Definition: Let X be a continuous random variable that assumes values in the interval $(-\infty, \infty)$, The f(x) is said to be a probability density function (pdf) of X if it satisfies the following conditions

i) $\mathrm{f(x)} \geq 0$ for all x , ii) $p(a \leq x \leq b) = \int_{a}^{b} \mathrm{f(x)\,dx} = 1$ and iii) $\int_{-\infty}^{\infty} \mathrm{f(x)\,dx} = 1$

The support of a continuous random variable is the smallest interval containing all values of x where f(x) >= 0.

Remark A crucial property is that, for any real number x, we have $\mathrm{P(X = x) = 0}$ (implying there is no difference between $\mathrm{P(X \leq x)}$ and $\mathrm{P(X < x)}$); that is it is not possible to talk about the probability of the random variable assuming a particular value. Instead, we talk about the probability of the random variable assuming a value within a given interval. The probability of the random variable assuming a value within some given interval from $x = a$ to $x = b$ is defined to be the area under the graph of the probability density function between $x = a$ and $x = b$.

Example

Let X be a continuous random variable. Show that the function

$\mathrm{f(x)} = \begin{cases} \frac{1}{2}x, & 0 \leq x \leq 2 \\ 0, & elsewhere \end{cases}$ is a pdf of X hence compute $P(0 \leq \mathrm{X} < 1)$ and $P(-1 < \mathrm{X} < 1)$

Solution

$\mathrm{f(x)} \geq 0$ for all x in the interval $0 \leq x \leq 2$ and $\int_{0}^{2} \frac{1}{2}\mathrm{xdx} = \left[\frac{x^2}{4}\right]_0^2 = 1$. Therefore f(x)is indeed a pdf of X.

Now $P(0 \leq \mathrm{X} < 1) = \int_{0}^{1} \frac{1}{2}\mathrm{xdx} = \left[\frac{x^2}{4}\right]_0^1 = \frac{1}{4}$ and

$P(-1 < \mathrm{X} < 1) = P(-1 < \mathrm{X} < 0) + P(0 < \mathrm{X} < 1) = 0 + \frac{1}{4} = \frac{1}{4}$

Exercise

1) Suppose that the random variable X has p.d.f. given by $\mathrm{f(x)} = \begin{cases} cx, & 0 \leq x \leq 1 \\ 0, & elsewhere \end{cases}$ Find the value of the constant c hence determine m so that $P(\mathrm{X} \leq \mathrm{m}) = \frac{1}{2}$

2) Let X be a continuous random variable with pdf $\mathrm{f(x)} = \begin{cases} \frac{x}{5} + k, & 0 \leq x \leq 3 \\ 0, & elsewhere \end{cases}$ Find the value of the constant k hence compute $P(1 < \mathrm{X} < 3)$

3) 2. A continuous random variable Y has the pdf given by $\mathrm{f(y)} = \begin{cases} k(1+y), & 4 \leq x \leq 7 \\ 0, & elsewhere \end{cases}$ Find the value of the constant k hence compute $P(Y < 5)$ and $P(5 < \mathrm{Y} < 6)$

4) 3. A continuous random variable X has the pdf given by $\mathrm{f(x)} = \begin{cases} k(1+x)^2, & -2 \leq x \leq 0 \\ 4k & 0 < x \leq \frac{4}{3} \\ 0, & elsewhere \end{cases}$

Find the value of the constant k hence compute $P(X > 1)$ and $P(-1 < \mathrm{X} < 1)$

5) 4. A continuous random variable X has the pdf given by $\mathrm{f(x)} = \begin{cases} kx, & 0 \le x \le 2 \\ k(4-x) & 2 < x \le 4 \\ 0, & elsewhere \end{cases}$

Find the value of the constant k hence compute $P(X > 3)$ and $P(1 < \mathrm{X} < 3)$

6.3 Distribution Function of a Random Variables

Definition: For any random variable X, we define the **cumulative distribution function (CDF)**, *F(x)* as $\mathrm{F(x)} = P(X \le x)$ for every x.

If X is a discrete random variable with pmf f(x), then $\mathrm{F(x)} = \sum_{t=-\infty}^{x} \mathrm{f(t)}$ \\ t is introduced to facilitate summation// However, if X is a continuous random variable with pdf f(x), then $\mathrm{F(x)} = \int_{-\infty}^{x} \mathrm{f(t)dt}$ \\ Again here t is introduced to facilitate integration//

Properties of any cumulative distribution function

- $\lim_{x\to\infty} \mathrm{F(x)} = 1$ and $\lim_{x\to-\infty} \mathrm{F(x)} = 0$
- F(x) is a non-decreasing function.
- F (x) is a right continuous function of *x*. In other words $\lim_{t\to x} \mathrm{F(t)} = \mathrm{F(x)}$

Reminder If the c.d.f. of X is F(x) and the p.d.f. is f(x), then differentiate F(x) to get f(x), and integrate f(x) to get F(x);

Theorem: For any random variable X and real values a < b, $P(a \le X \le b) = \mathrm{F}(b) - \mathrm{F}(a)$

Example 1

Let X be a discrete random variable with pmf given by $\mathrm{f(x)} = \begin{cases} \frac{1}{20}(1+x) & \text{for } x = 1,2,3,4,5 \\ 0, & elsewhere \end{cases}$.

Determine the cdf of X hence compute $P(X > 3)$

Solution

$$\mathrm{F(x)} = \sum_{t=-\infty}^{x} \mathrm{f(t)} = \tfrac{1}{20}\sum_{t=1}^{x}(\mathrm{x}+1) = \tfrac{1}{20}(2+3+....+x) = \tfrac{1}{20}\left\{\tfrac{x}{2}[4+(x-1)]\right\} = \tfrac{x(x+3)}{40}$$

$$\mathrm{F(x)} = \begin{cases} 0 & \text{for } x < 1 \\ \frac{x(x+3)}{40} & \text{for } x = 1,2,3,4,5 \\ 1 & \text{for } x > 5 \end{cases} \quad \text{Recall for an AP } S_n = \tfrac{n}{2}[2a+(n-1)d]$$

$$P(X > 3) = 1 - P(X \le 3) = 1 - \frac{3(6)}{40} = \frac{11}{20}$$

***Example** 2*

Suppose is a continuous random variable whose pdf f(x) is given by $\mathrm{f(x)} = \begin{cases} \frac{1}{2}x, & 0 \le x \le 2 \\ 0, & elsewhere \end{cases}$.

Obtain the cdf of X hence compute $P(X > \frac{2}{3})$

Solution

$$\mathrm{F(x)} = \int_{-\infty}^{x} \mathrm{f(t)dt} \ = \int_{0}^{x} \tfrac{1}{2}\,\mathrm{tdt} = \left[\frac{t^2}{4}\right]_0^x = \frac{x^2}{4} \quad \text{thus } \mathrm{F(x)} = \begin{cases} 0, & x < 0 \\ \frac{x^2}{4}, & 0 \le x \le 2 \\ 1, & x > 2 \end{cases}$$

$$P\left(X > \tfrac{2}{3}\right) = 1 - P\left(X \le \tfrac{2}{3}\right) = 1 - \tfrac{1}{4}\left(\tfrac{2}{3}\right)^2 = \tfrac{8}{9}$$

Exercise

1. The pdf of a continuous random variable X is given by $\mathrm{f(x)} = \begin{cases} C/\sqrt{x}, & 0 \le x \le 4 \\ 0, & elsewhere \end{cases}$ Find the value of the constant C, the cdf of X and $P(X \ge 1)$
2. The pdf of a random variable X is given by $\mathrm{g(x)} = \begin{cases} kx(1-x), & 0 \le x \le 1 \\ 0, & elsewhere \end{cases}$ Find the value of the constant k, the cdf of X and the value of m such that $\mathrm{G(x)} = \frac{1}{2}$
3. Find the cdf of a random variable Y whose pdf is given by;

   a) $\mathrm{f(x)} = \begin{cases} 1/3, & 0 \le x \le 1 \\ 1/3, & 2 \le x \le 4 \\ 0, & elsewhere \end{cases}$ b) $\mathrm{f(x)} = \begin{cases} x/2, & 0 \le x \le 1 \\ 1/2, & 1 \le x \le 2 \\ (3-x)/2, & 2 \le x \le 3 \\ 0, & elsewhere \end{cases}$
4. If the cdf of a random variable Y is given by $\mathrm{F(x)} = 1 - \dfrac{9}{y^2}$ for $Y \ge 3$ and $\mathrm{F(x)} = 0$ for $Y < 3$, find $P(X \le 5)$, $P(X > 8)$ and the pdf of X.

6.3 Expectation and Variance of a Random Variable

6.2.2 Expected Values

One of the most important things we'd like to know about a random variable is: what value does it take on average? What is the average price of a computer? What is the average value of a number that rolls on a die? The value is found as the average of all possible values, weighted by how often they occur (i.e. probability)

Definition: Let *X* be a discrete r.v. with probability function *p*(x). Then the **expected value** of X, denoted $E(X)$ or μ, is given by $\mathrm{E(x)} = \mu = \sum_{x=-\infty}^{\infty} xp(X = x).$

Similarly for a continuous random variable X with pdf f(x), $\mathrm{E(x)} = \mu = \int_{-\infty}^{\infty} x\mathrm{f}(x)dx.$

Theorem: Let X be a discrete r.v. with probability function *p*(*X*=*x*) and let *g*(*x*) be a real-valued function of *X*. ie $g: \mathbb{R} \to \mathbb{R}$, then the expected value of *g*(*x*) is given by

$$\mathrm{E}[\mathrm{g(x)}] = \sum_{x=-\infty}^{\infty} g(x)p(X = x).$$

Similarly for a continuous random variable X with pdf f(x), $\mathrm{E}[\mathrm{g(x)}] = \int_{-\infty}^{\infty} \mathrm{g(x)} \times \mathrm{f}(x)dx.$

Theorem: Let *X* be a discrete r.v. with probability function *p*(x). Then

(i) *E*(*c*) = *c*, where *c* is any real constant;

(ii) $\mathrm{E}[\mathrm{ax}+\mathrm{b}] == a\mu+b$ where a and b are constants

(iii) $\mathrm{E}[\mathrm{kg(x)}]=k\mathrm{E}[\mathrm{g(x)}]$ where $g(x)$ is a real-valued function of X

(iv) $\mathrm{E}[\mathrm{ag}_1(\mathrm{x}) \pm b\mathrm{g}_2(\mathrm{x})]=a\mathrm{E}[\mathrm{g}_1(\mathrm{x})] \pm b\mathrm{E}[\mathrm{g}_2(\mathrm{x})]$ and in general $\mathrm{E}\left[\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{c}_{\mathrm{i}}\mathrm{g}_{\mathrm{i}}(\mathrm{x})\right]=\sum_{\mathrm{i}=1}^{n} c_i\mathrm{E}[\mathrm{g}_{\mathrm{i}}(\mathrm{x})]$

where $\mathrm{g}_{\mathrm{i's}}(\mathrm{x})$ are real-valued functions of *X*.

This property of expectation is called *linearity property*

Remark: This theorem will also hold for a continuous random variable but we need to replace all the summation signs with integral signs.

Proof

(i) $\mathrm{E}[\mathrm{c}]=\sum_{all\ x} cP(X=x)=c\sum_{all\ x} P(X=x)=c(1)=c$

(ii) $\mathrm{E}[\mathrm{ax}+\mathrm{b}]=\sum_{all\ x}(\mathrm{ax}+\mathrm{b})P(x)=\sum_{all\ x}\mathrm{ax}P(x)+\sum_{all\ x}\mathrm{b}P(x)=a\sum_{all\ x} xP(x)+b\sum_{all\ x} P(x)=a\mu+b$

(iii) $\mathrm{E}[\mathrm{kg(x)}]=\sum_{all\ x} kg(x)P(X=x)=k\sum_{all\ x} g(x)P(X=x)=k\mathrm{E}[\mathrm{g(x)}]$

(iv) $\mathrm{E}[\mathrm{ag}_1(\mathrm{x}) \pm b\mathrm{g}_2(\mathrm{x})]=\mathrm{E}[\mathrm{ag}_1(\mathrm{x})] \pm E[b\mathrm{g}_2(\mathrm{x})]=a\mathrm{E}[\mathrm{g}_1(\mathrm{x})] \pm b\mathrm{E}[\mathrm{g}_2(\mathrm{x})]$ from part iii

6.2.3 Variance and Standard Deviation

Definition: Let *X* be a r.v with mean $E(X)=\mu$, the **variance** of X, denoted σ^2 or Var(X), is given by $Var(X)=\sigma^2=E(X-\mu)^2$. The units for variance are square units. The quantity that has the correct units is **standard deviation**, denoted σ. It's actually the positive square root of Var(X).

$$\sigma=\sqrt{Var(X)}=\sqrt{E(X-\mu)^2}.$$

Theorem: $Var(X)=E(X-\mu)^2=E(X)^2-\mu^2$

Proof:

$Var(X)=E(X-\mu)^2=E\left(X^2-2X\mu+\mu^2\right)=E(X)^2-2\mu E(X)+\mu^2=E(X)^2-\mu^2$ Since $E(X)=\mu$

Theorem: $Var(aX+b)=a^2\,\mathrm{var}(X)$

Proof:

Recall that $E[aX+b]=a\mu+b$ therefore

$Var(aX+b)=E[(aX+b)-(a\mu+b)]^2=E[a(X-\mu)]^2==E\left[a^2(X-\mu)^2\right]=a^2E\left[(X-\mu)^2\right]=a^2\,\mathrm{var}(X)$

Remark

(i) The expected value of X always lies between the smallest and largest values of X.

(ii) In computations, bear in mind that variance cannot be negative!

Example 1

Given a probability distribution of X as below, find the mean and standard deviation of X.

x	0	1	2	3
P(X=x)	1/8	1/4	3/8	1/4

Solution

x	0	1	2	3	total
$p(X=x)$	⅛	¼	⅜	¼	1
$xp(X=x)$	0	¼	¾	¾	7⁄4
$x^2p(X=x)$	0	¼	3⁄2	9⁄4	4

$E(X)=\mu=\sum_{x=0}^{3} xp(X=x)=1.75$ and

standard deviation

$\sigma=\sqrt{E(X^2)-\mu^2} \quad =\sqrt{4-1.75^2}\approx 0.968246$

Example 2

The probability distribution of a r.v X is as shown below, find the mean and standard deviation of; a) X b) $Y = 12X + 6$.

x	0	1	2
P(X=x)	1/6	1/2	1/3

Solution

x	0	1	2	total
$p(X=x)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$	1
$xp(X=x)$	0	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{7}{6}$
$x^2p(X=x)$	0	$\frac{1}{2}$	$\frac{4}{3}$	$\frac{11}{6}$

$$E(X) = \mu = \sum_{x=0}^{2} xp(X=x) = \tfrac{7}{6} \text{ and}$$

$$E(X^2) = \sum_{x=0}^{2} x^2 p(X=x) = \tfrac{11}{6}$$

Standard deviation $\sigma = \sqrt{E(X^2) - \mu^2} = \sqrt{\tfrac{11}{6} - (\tfrac{7}{6})^2} = \sqrt{\tfrac{17}{6}} \approx 1.6833$

Now $E(Y) = 12E(X) + 6 = 12(\tfrac{7}{6}) + 6 = 20$

$Var(Y) = Var(12X + 6) = 12^2 \times Var(X) = 144 \times \sqrt{\tfrac{17}{6}} \approx 242.38812$

***Example** 3*

A continuous random variable X has a pdf given by $\text{f(x)} = \begin{cases} \frac{1}{2}x, & 0 \le x \le 2 \\ 0, & elsewhere \end{cases}$, find the mean and standard of X

Solution

$$\text{E(x)} = \int_{-\infty}^{\infty} x\text{f}(x)dx = \int_0^2 \tfrac{1}{2}x^2\text{dx} = \left[\frac{x^3}{6}\right]_0^2 = \frac{4}{3} \text{ and } \text{E(x}^2) = \int_{-\infty}^{\infty} x^2\text{f}(x)dx = \int_0^2 \tfrac{1}{2}x^3\text{dx} = \left[\frac{x^4}{8}\right]_0^2 = 2$$

Standard deviation $\sigma = \sqrt{E(X^2) - \mu^2} = \sqrt{2 - (\tfrac{4}{3})^2} = \frac{\sqrt{2}}{3}$

Exercise

1. Suppose X has a probability mass function given by the table below

x	2	3	4	5	6
P(X=x)	0.01	0.25	0.4	0.3	0.04

Find the mean and variance of; X

2. Suppose X has a probability mass function given by the table below

x	11	12	13	14	15
P(X=x)	0.4	0.2	0.2	0.1	0.1

Find the mean and variance of; X

3. Let *X* be a random variable with *P*(*X* = 1) = 0.2, *P*(*X* = 2) = 0.3, and *P*(*X* = 3) = 0.5. What is the expected value and standard deviation of; a)*X* *b)* $Y = 5X - 10$?

4. A random variable W has the probability distribution shown below,

w	0	1	2	3
P(W=w)	2d	0.3	d	0.1

Find the values of the constant d hence determine the mean and variance of W. Also find the mean and variance of $Y = 10X + 25$

5. A random variable X has the probability distribution shown below,

x	1	2	3	4	5
P(X=x)	7c	5c	4c	3c	c

Find the values of the constant c hence determine the mean and variance of X.

6. The random variable Z has the probability distribution shown below,

z	2	3	5	7	11
P(Z=z)	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{4}$	x	y

If $E(Z)=4\frac{2}{3}$, find the values of x and y hence determine the variance of Z

7. A discrete random variable M has the probability distribution $\text{f(m)}=\begin{cases}\frac{m}{36},\ m=1,2,3,...,8\\ 0,\ elsewhere\end{cases}$, find the mean and variance of M

8. For a discrete random variable Y the probability distribution is $\text{f(y)}=\begin{cases}\frac{5-y}{10},\ y=1,2,3,4\\ 0,\ elsewhere\end{cases}$, calculate $E(Y)$ and var(Y)

9. Suppose X has a pmf given by $\text{f(x)}=\begin{cases}kx\ \text{ for } x=1,2,3,4\\ 0,\ elsewhere\end{cases}$, find the value of the constant k hence obtain the mean and variance of X

10. A team of 3 is to be chosen from 4 girl and 6 boys. If X is the number of girls in the team, find the probability distribution of X hence determine the mean and variance of X

11. A fair six sided die has; ‘1’ on one face, ‘2’ on two of it’s faces and ‘3’ on the remaining three faces. The die is rolled twice. If T is the total score write down the probability distribution of T hence determine;
    a) the probability that T is more than 4 b) the mean and variance of T

12. The pdf of a continuous r.v R is given by $\text{f(r)}=\begin{cases}kr\ \text{ for } 0\le r\le 4\\ 0,\ elsewhere\end{cases}$, (a) Determine *c*. hence Compute $\text{P}(10\le r\le 2)$, E(X) and Var(X).

13. A continuous r.v M has the pdf given by $\text{f(m)}=\begin{cases}k(1-\frac{m}{10})\ \text{ for } M\le 10\\ 0,\ elsewhere\end{cases}$, find the value of the constant k, the mean and the variance of X

14. A continuous r.v X has the pdf given by $\text{f(x)}=\begin{cases}k(1-x)\ \text{ for } 0\le x\le 1\\ 0,\ elsewhere\end{cases}$, findt the value of the constant k. Also find the mean and the variance of X

15. The lifetime of new bus engines, T years, has continuous pdf $\text{f(t)}=\begin{cases}{}^{d}\!/_{t^2}\ if\ x\ge 1\\ 0,\ if\ x<1\end{cases}$ find the value of the constant d hence determine the mean and standard deviation of T

16. An archer shoots an arrow at a target. The distance of the arrow from the centre of the target is a random variable X whose p.d.f. is given by $\text{f(x)}=\begin{cases}k(3+2\text{x - x}^2)\ if\ x\le 3\\ 0,\ \ if\ x>3\end{cases}$ find the value of the constant k. Also find the mean and standard deviation of X

17. A continuous r.v X has the pdf given by $\text{f(x)}=\begin{cases}k(1+x),\ -1\le x<0\\ 2k(1-x),\ 0\le x\le 1\\ 0,\ elsewhere\end{cases}$, find the value of the constant k. Also find the mean and the variance of X

18. A continuous r.v X has the pdf given by $\text{f(x)}=\begin{cases}e^{-x}\ \text{for } x>0\\ 0,\ elsewhere\end{cases}$, find the mean and standard deviation of; a) X b) $\text{Y}=e^{\frac{3}{4}x}$

Probability Distribution

Discrete Distribution

Among the discrete distributions that we will look at includes the Bernoulli, binomial, Poisson, geometric and hyper geometric

a) ..Bernoulli distribution

Definition: A Bernoulli trial is a random experiment in which there are only two possible outcomes - success and failure. Eg

- Tossing a coin and considering heads as success and tails as failure.
- Checking items from a production line: success = not defective, failure = defective.
- Phoning a call centre: success = operator free; failure = no operator free.

A Bernoulli random variable X takes the values 0 and 1 and $P(X=1)=p$ and $P(X=0)=1-p$

Definition: A random variable X is said to be a real Bernoulli distribution if it's pmf is given by;

$$P(X=x)=\begin{cases} p^x(1-p)^{1-x} & for\ x=0,1 \\ 0 & otherwise \end{cases}$$

We abbreviate this as $\text{X} \sim \text{B(p)}$ ie p is the only parameter here. It can be easily checked that the mean and variance of a Bernoulli random variable are $\mu = p$ and $\sigma^2 = p(1-p1)$

b)..Binomial Distribution

Consider a sequence of *n* independent, Bernoulli trials, with each trial having two possible outcomes, *success* or *failure*. Let *p* be the probability of a success for any single trial. Let*X* denote the number of successes on *n* trials. The random variable *X* is said to have a **binomial distribution** and has probability mass function

$$P(X=x)={}_nC_x \times p^x(1-p)^{n-x} \ \ for\ x=0,1,2.....n$$

We abbreviate this as $\text{X} \sim \text{Bin}\ (n,\text{p})$ read as "X follows a binomial distribution with parameters n and p". ${}_nC_x$ counts the number of outcomes that include exactly x successes and $n-x$ failures.The mean and variance of a Binomial random variable are $\mu = np$ and $\sigma^2 = np(1-p1)$ respectively.

Let's check to make sure that if *X* has a binomial distribution, then $\sum_{x=0}^{n} P(X=x)=1$. We will need the binomial expansion for any polynomial:

$$(p+q)^n = \sum_{x=0}^{n} {}_nC_x\left(p^x q^{n-x}\right) \text{so } \sum_{x=0}^{n} {}_nC_x\left(p^x(1-p)^{n-x}\right)=\left[p+(1-p)\right]^n = 1^n = 1 \text{ as required.}$$

Example 1

A biased coin is tossed 6 times. The probability of heads on any toss is 0:3. Let X denote the number of heads that come up. Calculate: (i) $P(X=2)$ (ii) $P(X=3)$ (iii) $P(1<X<5)$

Solution

If we call heads a success then X has a binomial distribution with parameters n=6 and p=0:3.

(i) $P(X=2)={}_6C_2(0.3)^2(0.7)^4 = 0.324135$

(ii) $P(X=3)={}_6C_3(0.3)^3(0.7)^3 = 0.18522$

(iii) $$\begin{aligned} P(1<X\le 5) &= P(X=2)+P(X=3)+P(X=4)+P(X=5) \\ &= 0.324+0.185+0.059+0.01=0.578 \end{aligned}$$

Example 2

A quality control engineer is in charge of testing whether or not 90% of the DVD players produced by his company conform to specifications. To do this, the engineer randomly selects a batch of 12 DVD players from each day's production. The day's production is acceptable provided no more than 1 DVD player fails to meet specifications'. Otherwise, the entire day's production has to be tested.

a) What is the probability that the engineer incorrectly passes a day's production as acceptable if only 80% of the day's DVD players actually conform to speciication?

b) What is the probability that the engineer unnecessarily requires the entire day's production to be tested if in fact 90% of the DVD players conform to speciffications?

Solution

a) Let X denote the number of DVD players in the sample that fail to meet speciffications. In part (i) we want $P(X \le 1)$ with binomial parameters $n = 12$ and $\mathrm{p} = 0.2$

$$P(X \le 1) = P(X = 0) + P(X = 1) = {}_{12}C_0(0.2)^0(0.8)^{12} + {}_{12}C_1(0.2)^1(0.8)^{11} = 0.069 + 0.206 = 0.275$$

b) We now want $P(X > 1)$ with parameters $n = 12$ and $\mathrm{p} = 0.1$.

$$\mathrm{P}(X \le 1) = \mathrm{P}(X = 0) + \mathrm{P}(X = 1) = {}_{12}C_0(0.1)^0(0.9)^{12} + {}_{12}C_1(0.1)^1(0.9)^{11} = 0.659$$

So $\mathrm{P(X>1)} = 0.34$

Example 3

Bits are sent over a communications channel in packets of 12. If the probability of a bit being corrupted over this channel is 0:1 and such errors are independent, what is the probability that no more than 2 bits in a packet are corrupted?

If 6 packets are sent over the channel, what is the probability that at least one packet will contain 3 or more corrupted bits?

Let X denote the number of packets containing 3 or more corrupted bits. What is the probability that X will exceed its mean by more than 2 standard deviations?

Solution

Let C denote the number of corrupted bits in a packet. Then in the first question, we want

$$P(C \le 2) = P(C = 0) + P(C = 1) + P(C = 2)$$

$$= {}_{12}C_0(0.1)^0(0.9)^{12} + {}_{12}C_1(0.1)^1(0.9)^{11} + {}_{12}C_2(0.1)^2(0.9)^{10}$$

$$= 0.282 + 0.377 + 0.23 = 0.889.$$

Therefore the probability of a packet containing 3 or more corrupted bits is

$P(C \ge 3) = 1 - P(C \le 2) = 1 - 0.889 = 0.111.$

Let X be the number of packets containing 3 or more corrupted bits. X can be modelled with a binomial distribution with parameters $n = 6$ and $\mathrm{p} = 0.111$. The probability that at least one packet will contain 3 or more corrupted bits is:

$P(X \ge 1) = 1 - P(X = 0) = 1 - {}_6C_0(0.111)^0(0.889)^6 = 0.494.$

The mean of X is $\mathrm{E(X)} = 6(0.111) = 0.666$ and its standard deviation is

$= \sqrt{6(0.111)(0.889)} = 0.77$

So the probability that X exceeds its mean by more than 2 standard deviations is

$P(X > \mu + 2\sigma) = P(X > 2.2) = P(X \ge 3)$ since X is discrete.

Now $P(X \ge 3) = 1 - P(X \le 2) = 1 - \{P(X = 0) + P(X = 1) + P(X = 2)\}$

$$= 1 - \left[{}_6C_0(0.111)^0(0.889)^6 + {}_6C_1(0.111)^1(0.889)^5 + {}_6C_2(0.111)^2(0.889)^4\right]$$

$$= 1 - (0.4936 + 0.3698 + 0.1026) = 0.032$$

Exercise

1. A fair coin is tossed 10 times. What is the probability that exactly 6 heads will occur.

2. If 3% of the electric bulbs manufactured by a company are defective find the probability that in a sample of 100 bulbs exactly 5 bulbs are defective.
3. An oil exploration firm is formed with enough capital to finance 10 explorations. The probability of a particular exploration being successful is 0.1. Find mean and variance of the number of successful explorations.
4. Emily hits 60% of her free throws in basketball games. She had 25 free throws in last week's game.
   a) What is the expected number and the standard deviation of Emily's hit ?
   b) Suppose Emily had 7 free throws in yesterday's game.What is the probability that she made at least 5 hits?
5. A coin is loaded so that heads has 60% chance of showing up. This coin is tossed 3 times.
   a) What are the mean and the standard deviation of the number of heads that turned out?
   b) What is the probability that the head turns out at least twice?
   c) What is the probability that an odd number of heads turn out in 3 flips?
6. According to the 2009 current Population Survey conducted by the U.S. Census Bureau, 40% of the U.S. population 25 years old and above have completed a bachelor's degree or more. Given a random sample of 50 people 25 years old or above, What is expected number of people and the standard deviation of the number of people who have completed a bachelor's degree.
7. Joe throws a fair die six times and face nimber 3 appeared twice. It he incredibly lucky or unusual?
8. If the probability of being a smoker among a group of cases with lung cancer is .6, what's the probability that in a group of 8 cases you have; (a) less than 2 smokers? (b0 More than 5? (c) What are the expected value and variance of the number of smokers?
9. Suppose 90% of the cars on Thika super highways does over 17 km per litre.
   a) What is the expected number and the standard deviation of cars on Thika super highways that will do over 17 km per litre. in a sample of 15 cars ?
   b) What is the probability that in a sample of 15 cars exactly 10 of these will do over 17 km per litre?

c). ..Poisson distribution

Named after the French mathematician Simeon Poisson, the distribution is used to model the number of events, (such as the number of telephone calls at a business, number of customers in waiting lines, number of defects in a given surface area, airplane arrivals, or the number of accidents at an intersection), occurring within a given time interval. Other such random events where Poisson distribution can apply includes;

- the number of hits to your web site in a day
- the number of calls that arrive in each day on your mobile phone
- the rate of job submissions in a busy computer centre per minute.
- the number of messages arriving to a computer server in any one hour.

Poisson probabilities are useful when there are a large number of independent trials with a small probability of success on a single trial and the variables occur over a period of time. It can also be used when a density of items is distributed over a given area or volume. The formula for the Poisson probability mass function is $P(X = x) = \dfrac{\lambda^x e^{-\lambda}}{x!},\ x = 0,1,2,....$ This is abbreviated as $\mathrm{X} \sim \mathrm{Po}\,(\lambda)$. λ is the shape parameter which indicates the average number of events in the given time interval. The mean and variance of this distribution are equal ie $\mu = \sigma^2 = \lambda$

Let's check to make sure that if X has a poisson distribution, then $\sum_{x=0}^{\infty} P(X=x)=1$. We will need to recall that $e^{\lambda}=1+\frac{\lambda}{1!}+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}+\frac{\lambda^4}{4!}+.....$ Now $\sum_{x=0}^{\infty}\frac{\lambda^x e^{-\lambda}}{x!}=e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!}=e^{-\lambda}e^{\lambda}=e^{0}=1$

Remark

The major difference between Poisson and Binomial distributions is that the Poisson does not have a fixed number of trials. Instead, it uses the fixed interval of time or space in which the number of successes is recorded.

Example: 1

Consider a computer system with Poisson job-arrival stream at an average of 2 per minute. Determine the probability that in any one-minute interval there will be

(i) 0 jobs;

(ii) exactly 2 jobs;

(iv) at most 3 arrivals.

(v) more than 3 arrivals

Solution

Job Arrivals with λ=2

(i) No job arrivals: $\text{P(X}=0)=e^{-2}=0.1353353$

(ii) Exactly 3 job arrivals: $P(X=3)=\frac{2^3e^{-2}}{3!}=0.1804470$

(iii) At most 3 arrivals

$$\text{P}(X\le 3)=\text{P}(X=0)+\text{P}(X=1)+\text{P}(X=2)+\text{P}(X=3)=\left(1+\frac{2}{1}+\frac{2^2}{2}+\frac{2^3}{3!}\right)e^{-2}=0.8571$$

(iv) more than 3 arrivals

$$P(X>3)=1-P(X\le 3)==1\text{-}0.8571=0.1429$$

Example: 2

If there are 500 customers per eight-hour day in a check-out lane, what is the probability that there will be exactly 3 in line during any five-minute period?

Solution

The expected value during any one five minute period would be 500 / 96 = 5.2083333. The 96 is because there are 96 five-minute periods in eight hours. So, you expect about 5.2 customers in 5 minutes and want to know the probability of getting exactly 3.

$$P(X=3)=\frac{(\text{-500/96})\,^{3}e^{--500/96}}{3!}=0.1288\,\text{(approx)}$$

Example: 3

If new cases of West Nile in New England are occurring at a rate of about 2 per month, then what's the probability that exactly 4 cases will occur in the next 3 months?

Solution

X ~ Poisson (λ=2/month)

$$\text{P(X}=4\,\text{in 3 months)}=\frac{(2*3)^4e^{-(2*3)}}{4!}=\frac{6^4e^{-(6)}}{4!}=13.4\%$$

Exactly 6 cases

$$\text{P(X}=6\,\text{in 3 months)}=\frac{(2*3)^6e^{-(2*3)}}{6!}=\frac{6^6e^{-(6)}}{6!}=16\%$$

Exercise

1. Calculate the Poisson Distribution whose λ (Average Rate of Success)) is 3 & X (Poisson Random Variable) is 6.
2. Customers arrive at a checkout counter according to a Poisson distribution at an average of 7 per hour. During a given hour, what are the probabilities that
   a) No more than 3 customers arrive?
   b) At least 2 customers arrive?
   c) Exactly 5 customers arrive?
3. Manufacturer of television set knows that on an average 5% of their product is defective. They sells television sets in consignment of 100 and guarantees that not more than 2 set will be defective. What is the probability that the TV set will fail to meet the guaranteed quality?
4. It is known from the past experience that in a certain plant there are on the average of 4 industrial accidents per month. Find the probability that in a given year will be less that 3 accidents.
5. Suppose that the change of an individual coal miner being killed in a mining accident during a year is 1.1499. Use the Poisson Distribution to calculate the probability that in the mine employing 350 miners- there will be at least one accident in a year.
6. The number of road construction projects that take place at any one time in a certain city follows a Poisson distribution with a mean of 3. Find the probability that exactly five road construction projects are currently taking place in this city. (0.100819)
7. The number of road construction projects that take place at any one time in a certain city follows a Poisson distribution with a mean of 7. Find the probability that more than four road construction projects are currently taking place in the city. (0.827008)
8. The number of traffic accidents that occur on a particular stretch of road during a month follows a Poisson distribution with a mean of 7.6. Find the probability that less than three accidents will occur next month on this stretch of road. (0.018757)
9. The number of traffic accidents that occur on a particular stretch of road during a month follows a Poisson distribution with a mean of 7. Find the probability of observing exactly three accidents on this stretch of road next month. (0.052129)
10. The number of traffic accidents that occur on a particular stretch of road during a month follows a Poisson distribution with a mean of 6.8. Find the probability that the next two months will both result in four accidents each occurring on this stretch of road. (0.00985)
11. Suppose the number of babies born during an 8-hour shift at a hospital's maternity wing follows a Poisson distribution with a mean of 6 an hour. Find the probability that five babies are born during a particular 1-hour period in this maternity wing. (0.160623)
12. The university policy department must write, on average, five tickets per day to keep department revenues at budgeted levels. Suppose the number of tickets written per day follows a Poisson distribution with a mean of 8.8 tickets per day. Find the probability that less than six tickets are written on a randomly selected day from this distribution. (0.128387)
13. A taxi firm has two cars which it hires out day by day. The number of demands for a car on each day is distributed as Poisson distribution with mean 1.5. Calculate the proportion of days on which neither car is used and the proportion of days on which some demands is refused
14. If calls to your cell phone are a Poisson process with a constant rate λ=0.5 calls per hour, what's the probability that, if you forget to turn your phone off in a 3 hour lecture, your phone rings during that time? How many phone calls do you expect to get during this lecture?
15. The manufacturer of the disk drives in one of the well-known brands of microcomputers expects 2% of the disk drives to malfunction during the microcomputer's warranty period. Calculate the probability that in a sample of 100 disk drives, that not more than three will malfunction
16. The average number of defects per wafer (defect density) is 3. The redundancy built into the design allows for up to 4 defects per wafer. What is the probability that the redundancy will not be sufficient if the defects follow a Poisson distribution?

17. The mean number of errors due to a particular bug occurring in a minute is 0.0001
    a) What is the probability that no error will occur in 20 minutes?
    b) How long would the program need to run to ensure that there will be a 99.95% chance that an error wills showup to highlight this bug?

Properties of Poisson

- The mean and variance are both equal to λ.
- The sum of independent Poisson variables is a further Poisson variable with mean equal to the sum of the individual means.
- As well as cropping up in the situations already mentioned, the Poisson distribution provides an approximation for the Binomial distribution.

d)..Geometric Distribution

Suppose a Bernoulli trial with success probability *p* is performed repeatedly until the first success appears we want to find the probability that the first success occurs on the yth trial. ie let *Y* denote the number of trials needed to obtain the first success. The sample space S={*s;fs;ffs, fffs ffffs ...}.* This is an *infinite* sample space (though it is still discrete). What is the probability of a sample point, say $\mathrm{P}(fffs)=\mathrm{P}(Y=4)$ *)*? Since successive trials are independent (this is implicit in the statement of the problem), we have $\mathrm{P}(fffs)=\mathrm{P}(Y=4)=\mathrm{q}^3\mathrm{p}$ where $q=1-p$ and $0\le p\le 1$

Definition: A r.v. *Y* is said to have a **geometric probability distribution** if and only if

$$\mathrm{P}(Y=y)=\begin{cases} pq^{y-1}\text{for } y=1,2,3..... \text{ where } q=1-p \\ 0 \quad\quad \textit{otherwise}\end{cases}.$$

This is abbreviated as *X ~ Geom*(*p*).

The only parameter for this geometric distribution is p (ie the probability of success in each trial).

To be sure everything is consistent; we should check that the probabilities of all the sample points add up to 1. Now

$$\sum_{y=1}^{\infty}P(Y=y)=\sum_{y=1}^{\infty}pq^{y-1}=\frac{p}{1-q}=1$$

Recall sum to infinity of a convergent G.P is $s=\dfrac{a}{1-r}$

The cdf of a geometric distributions given by

$$\mathrm{F(y)=P(Y\le y)=P(Y=1)+P(Y=2)+P(Y=3)+...+P(Y=y)}$$

$$=p+pq+pq^2+.....pq^{y-1}=\frac{p(1-q^y)}{1-q}=1-q^y$$

Let *Y*~ Geo(*p*), then $\mu=E(Y)=\dfrac{1}{p}$ and $Var(X)=\sigma^2=\dfrac{q}{p^2}$ Show?

Example:1

A sharpshooter normally hits the target 70% of the time.

a) Find the probability that her first hit is on the second shot
b) Find the mean and standard deviation of the number of shots required to realize the 1st hit

Solution

Let X be the random variable ‘the number of shoots required to realize the 1st hit’

$x\sim Geo(0.7)$ and $P(X=x)=0.7(1-0.7)^{x-1},\ x=1,2,3,....$

a) $P(X=2)=p(1-\rho)=0.7(0.3)=0.21$

b) $\mu = \frac{1}{\rho} = \frac{1}{0.7} = 1.428571$ and $\sigma = \frac{\sqrt{1\text{-}\mathrm{p}}}{\mathrm{p}} = \frac{\sqrt{1-0.7}}{0.7} \approx 0.78$

Example:2

The State Department is trying to identify an individual who speaks Farsi to fill a foreign embassy position. They have determined that 4% of the applicant pool are fluent in Farsi.

a) If applicants are contacted randomly, how many individuals can they expect to interview in order to find one who is fluent in Farsi?

b) What is the probability that they will have to interview more than 25 until they find one who speaks Farsi?

Solution

a) $\mu = \frac{1}{\rho} = \frac{1}{0.04} = 25$

b) $P(X \le 25) = (1-\rho)^n = 1-(0.04)^{25} = 1 \;\Rightarrow\; P(X > 25) = 1 - P(X \le 25) = 0$

Example:3

From past experience it is known that 3% of accounts in a large accounting population are in error. What is the probability that 5 accounts are audited before an account in error is found? What is the probability that the first account in error occurs in the first five accounts audited?

Solution

$\mathrm{P(Y=5)} = 0.03(0.97)^4 = 0.02655878$ $\qquad$ $\mathrm{P(Y \le 5)} = 1 - 0.97^5 = 0.14126597$

Exercise

1. Over a very long period of time, it has been noted that on Friday’s 25% of the customers at the drive-in window at the bank make deposits. What is the probability that it takes 4 customers at the drive-in window before the first one makes a deposit.
2. It is estimated that 45% of people in Fast-Food restaurants order a diet drink with their lunch. Find the probability that the fourth person orders a diet drink. Also find the probability that the first diet drinker of th e day occurs before the 5th person.
3. What is the probability of rolling a sum of seven in fewer than three rolls of a pair of dice? Hint (The random variable, X, is the number of rolls before a sum of 7.)
4. In New York City at rush hour, the chance that a taxicab passes someone and is available is 15%. a) How many cabs can you expect to pass you for you to find one that is free and b) what is the probability that more than 10 cabs pass you before you find one that is free.
5. An urn contains N white and M black balls. Balls are randomly selected, one at a time, until a black ball is obtained. If we assume that each selected ball is replaced before the next one is drawn, what is;
   a) the probability that exactly n draws are needed?
   b) the probability that at least k draws are needed?
   c) the expected value and Variance of the number of balls drawn?
6. In a gambling game a player tosses a coin until a head appears. He then receives \$2*n* , where *n* is the number of tosses.
   a) What is the probability that the player receives \$8.00 in one play of the game?
   b) If the player must pay \$5.00 to play, what is the win/loss per game?
7. An oil prospector will drill a succession of holes in a given area to find a productive well. The probability of success is 0.2.
   a) What is the probability that the 3rd hole drilled is the first to yield a productive well?
   b) If the prospector can afford to drill at most 10 well, what is the probability that he will fail to find a productive well?

8. A well-travelled highway has itstraffic lights green for 82% of the time. If a person travelling the road goes through 8 traffic intersections, complete the chart to find a) the probability that the first red light occur on the nth traffic light and b) the cumulative probability that the person will hit the red light on or before the nth traffic light.
9. An oil prospector will drill a succession of holes in a given area to find a productive well. The probability of success is 0.2.
   a) What is the probability that the 3rd hole drilled is the first to yield a productive well?
   b) If the prospector can afford to drill at most 10 well, what is the probability that he will fail to find a productive well?

Continuous Distribution

a)…Uniform (Rectangular) Distribution

A **uniform density function** is a density function that is constant, (*Ie all the values are* equally likely outcomes over the domain). It's often referred as the *Rectangular distribution* because the graph of the pdf has the form of a rectangle, making it the simplest kind of density function. The uniform distribution lies between two values on the x-axis. The total area is equal to 1.0 or 100% within the rectangle

Definition: A random variable X has a uniform distribution over the range $[\mathrm{a,b}]$ If

$$\mathrm{f(x)}=\begin{cases}\frac{1}{b-a}, & a\le x\le b\\ 0, & elsewhere\end{cases}$$ We denote this distribution by $\mathrm{X \sim U[a,b]}$

where: a = smallest value the variable can assume and b = largest value.

The expected Value and the Variance of *X* are given by $\mu=\dfrac{a+b}{2}$ and $\sigma^2=\dfrac{(b-a)^2}{12}$ respectively. The cdf F(x) is given by $\mathrm{F(x)}=\dfrac{1}{b-a}\int_a^x dt=\dfrac{x-a}{b-a}$ $\Rightarrow$ $\mathrm{F(x)}=\begin{cases}0 & x<a\\ \frac{x-a}{b-a}, & a\le x\le b\\ 1 & x>b\end{cases}$

Example

Prof Hinga travels always by plane. From past experience he feels that take off time is uniformly distributed between 80 and 120 minutes after check in. Determine the probability that: a) he waits for more than 15 minutes for take off after check in. b) the waiting time will be between 1.5 standard deviation from the mean,

Solution

$$\mathrm{X \sim U(80,120)} \quad \Rightarrow \quad \mathrm{f(x)}=\begin{cases}\frac{1}{40}, & 80\le x\le 120\\ 0, & elsewhere\end{cases}$$

$$\mathrm{P(X>105)=1\text{-}P(X\le 105)=1\text{-}}\tfrac{105-80}{40}=\tfrac{3}{8}$$

$$\mathrm{P}(\mu-1.5\sigma\le x\le\mu+1.5\sigma)=\int_{\mu-1.5\sigma}^{\mu+1.5\sigma}\tfrac{1}{40}dx=\tfrac{1}{40}\left[x\right]_{\frac{1}{4\mu-1.5\sigma 0}}^{\mu-1.5\sigma}=\frac{3\sigma}{40} \qquad \text{But } \sigma=\frac{b-a}{\sqrt{12}}=\frac{40}{\sqrt{12}}$$

$$\mathrm{P}(\mu-1.5\sigma\le x\le\mu+1.5\sigma)=\frac{3\sigma}{40}=\frac{3}{\sqrt{12}}$$

Exercise

1. Uniform: The amount of time, in minutes, that a person must wait for a bus is uniformly distributed between 0 and 15 minutes, inclusive. What is the probability that a person waits fewer than 12.5 minutes? What is the probability that will be between 0.5 standard deviation from the mean,

2. Slater customers are charged for the amount of salad they take. Sampling suggests that the amount of salad taken is uniformly distributed between 5 ounces and 15 ounces. Let *x* = salad plate filling weight, find the expected Value and the Variance of *x.* What is the probability hat a customer will take between 12 and 15 ounces of salad?
3. The average number of donuts a nine-year old child eats per month is uniformly distributed from 0.5 to 4 donuts, inclusive. Determine the probability that a randomly selected nine-year old child eats an average of;
    a) more than two donuts
    b) more than two donuts given that his or her amount is more than 1.5 donuts.
4. Starting at 5 pm every half hour there is a flight from Nairobi to Mombasa. Suppose that none of these plane tickets are completely sold out and they always have room for passagers. A person who wants to fly to Mombasa arrives at the airport at a random time between 8.45 AM and (.45 AM. Determine the probability that he waits for
    a) At most 10 minutes
    b) At least 15 minutes

b)…Exponential Distribution

The exponential distribution is often concerned with the amount of time until some specific event occurs. For example, the amount of time (beginning now) until an earthquake occurs has an exponential distribution. Other examples include the length, in minutes, of long distance business telephone calls, and the amount of time, in months, a car battery lasts. It can be shown, too, that the amount of change that you have in your pocket or purse follows an exponential distribution. Values for an exponential random variable occur in the following way. There are fewer large values and more small values. For example, the amount of money customers spend in one trip to the supermarket follows an exponential distribution. There are more people that spend less money and fewer people that spend large amounts of money.

The exponential distribution is widely used in the field of reliability. Reliability deals with the amount of time a product lasts

In brief this distribution is commonly used to model waiting times between occurrences of rare events, lifetimes of electrical or mechanical devices

Definition: A RV X is said to have an exponential distribution with parameter $\lambda > 0$ if the pdf of X is: $\text{f(x)} = \begin{cases} \lambda e^{-\lambda x} \text{for x} \geq 0 \text{ and } \lambda > 0 \\ 0 \quad otherwise \end{cases}$ we abbreviate this as $\text{X} \sim \exp(\lambda)$

λ is called the *rate parameter*

The mean and variance of this distribution are $\mu = \frac{1}{\lambda}$ and $\sigma^2 = \frac{1}{\lambda^2}$ respectively

The cumulative distribution function is F(x) is given by $\text{F(x)} = \begin{cases} 1 - e^{-\lambda x} \text{for x} \geq 0 \\ 0 \quad otherwise \end{cases}$

Example

Torch batteries have a lifespan T years with pdf $\text{f(t)} = \begin{cases} 0.01 e^{-0.01t}, \text{T} \geq 0 \\ 0 \quad otherwise \end{cases}$. Determine the probability that the battery; a) Falls before 25 hours. b) life is between 35 and 50 hours. *c)* life exceeds 120 hours. d) life exceeds the mean lifespan.

Solution

a) $\text{P(T} < 25) = \text{F(25)} = \int_0^{25} e^{-0.01t} dt = 1 - e^{-0.01(25)} \approx 0.2212$

b) $\mathrm{P}(35 \le \mathrm{T} \le 50) = \int_{35}^{50} e^{-0.01t}dt = e^{-0.35} - e^{-0.50} \approx 0.0982$

c) $\mathrm{P}(\mathrm{T} > 120) = \int_{120}^{\infty} e^{-0.01t}dt = e^{-1.2} - 0 \approx 0.3012$

d) $\mu = \dfrac{1}{0.01} = 100 \Rightarrow \mathrm{P}(\mathrm{T} > 100) = \int_{100}^{\infty} e^{-0.01t}dt = e^{-1} \approx 0.3679$

Exercise:

1. Jobs are sent to a printer at an average of 3 jobs per hour.
   a) What is the expected time between jobs?
   b) What is the probability that the next job is sent within 5 minutes?
2. The time required to repair a machine is an exponential random variable with rate λ= 0.5 downs/hour
   a) what is the probability that a repair time exceeds 2 hours?
   b) what is the probability that the repair time will take at least 4 hours given that the repair man has been working on the machine for 3 hours?
3. Buses arrive to a bus stop according to an exponential distribution with rate λ= 4 busses/hour. If you arrived at 8:00 am to the bus stop,
   a) what is the expected time of the next bus?
   b) Assume you asked one of the people waiting for the bus about the arrival time of the last bus and he told you that the last bus left at 7:40 am. What is the expected time of the next bus?
4. Break downs occur on an old car with rate λ= 5 break-downs/month. The owner of the car is planning to have a trip on his car for 4 days.
   a) What is the probability that he will return home safely on his car.
   b) If the car broke down the second day of the trip and the car was fixed, what is the probability that he doesn't return home safely on his car.
5. Suppose that the amount of time one spends in a bank is exponentially distributed with mean 10 minutes. What is the probability that a customer will spend more than 15 minutes in the bank? What is the probability that a customer will spend more than 15 minutes in the bank given that he is still in the bank after 10 minutes?
6. Suppose the lifespan in hundreds of hours, T, of a light bulb of a home lamp is exponentially distributed with lambda = 0.2. compute the probability that the light bulb will last more than 700 hours Also, the probability that the light bulb will last more than 900 hours
7. Let X = amount of time (in minutes) a postal clerk spends with his/her customer. The time is known to have an exponential distribution with the average amount of time equal to 4 minutes.
   a) Find the probability that a clerk spends four to five minutes with a randomly selected customer.
   b) Half of all customers are finished within how long? (Find median)
   c) Which is larger, the mean or the median?
8. On the average, a certain computer part lasts 10 years. The length of time the computer part lasts is exponentially distributed.
   a) What is the probability that a computer part lasts more than 7 years?
   b) On the average, how long would 5 computer parts last if they are used one after another?
   c) Eighty percent of computer parts last at most how long?
   d) What is the probability that a computer part lasts between 9 and 11 years?
9. Suppose that the length of a phone call, in minutes, is an exponential random variable with decay parameter = 1/12 . If another person arrives at a public telephone just before you, find the probability that you will have to wait more than 5 minutes. Let X = the length of a phone call, in minutes. What is median mean and standard deviation of X?

c).. The Normal Distribution

The normal, or Gaussian, distribution is one of the most important distributions in probability theory. It is widely used in statistical inference. One reason for this is that sums of random variables often approximately follow a normal distribution.

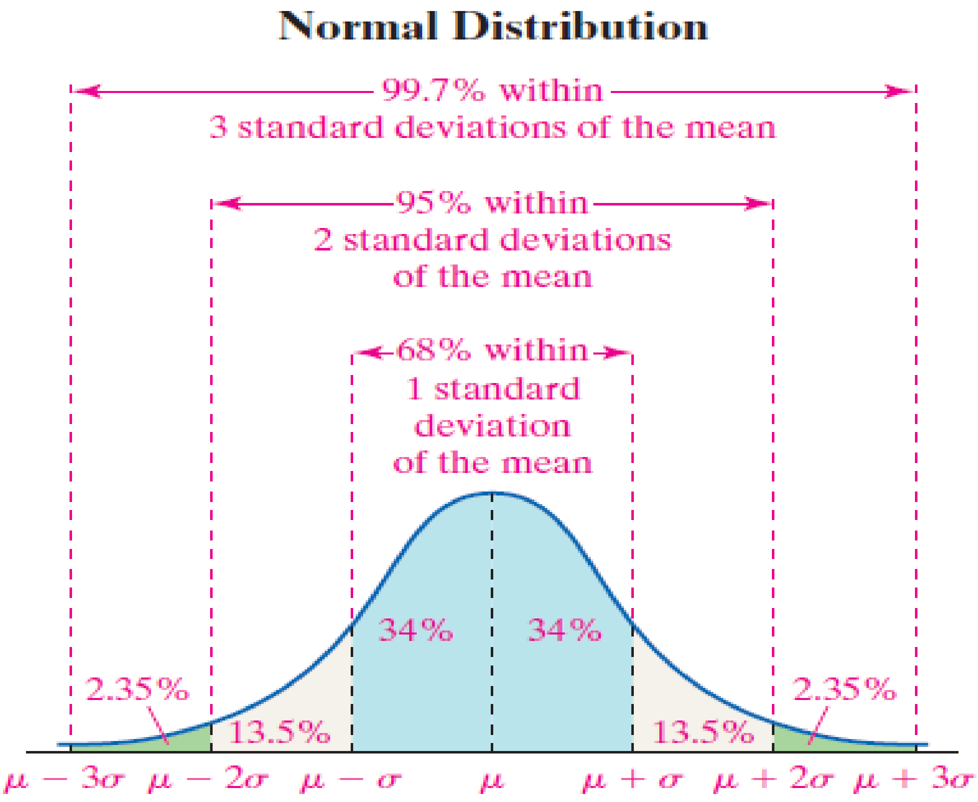
Definition A r.v X has a normal distribution with parameters μ and σ^2, abbreviated $X \sim N(\mu, \sigma^2)$ if it has probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \text{ for } -\infty < x < \infty \text{ and } \sigma > 0$$

Where μ is the mean and σ is the standard deviation.

Properties of normal distribution

1) The normal distribution curve is bell-shaped and symmetric, about the mean
2) The curve is asymptotic to the horizontal axis at the extremes.
3) The highest point on the normal curve is at the mean, which is also the median and mode.
4) The mean can be any numerical value: negative, zero, or positive
5) The standard deviation determines the width of the curve: larger values result in wider, flatter curves
6) Probabilities for the normal random variable are given by areas under the curve. The total area under the curve is 1 (0.5 to the left of the mean and 0.5 to the right).
7) It has inflection points at $\mu-\sigma$ and $\mu+\sigma$.
8) Empirical Rule:
   a) 68.26% of values of a normal random variable are within ± 1 standard deviation of its mean. ie $P(\mu-\sigma \le X \le \mu+\sigma) = 0.6826$
   b) 95.44% of values of a normal random variable are within ± 2 standard deviation of its mean. ie $P(\mu-2\sigma \le X \le \mu+2\sigma) = 0.9544$
   c) 99.72% of values of a normal random variable are within ± 3 standard deviation of its mean. ie $P(\mu-3\sigma \le X \le \mu+3\sigma) = 0.9972$



Standard Normal Probability Distribution

A random variable having a normal distribution with a mean of 0 and a varuance of 1 is said to have a **standard normal** probability distribution

Definition The random variable Z is said to have the standard normal distribution if $\mathrm{Z} \sim \mathrm{N}(0,1)$. Therefore, the density of Z, which is usually denoted $\phi(\mathrm{z})$ is given by;

$$\phi(\mathrm{z}) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\tfrac{1}{2} z^2\right\} \text{ for } -\infty < z < \infty$$

The cumulative distribution function of a standard normal random variable is denoted $\Phi(z)$, and is given by

$$\Phi(z) = \int_{-\infty}^{z} \phi(\mathrm{t})dt$$

Consider $\mathrm{Z} \sim \mathrm{N}(0,1)$ and let $\mathrm{X} = \mu + \sigma Z$ for $\sigma > 0$. Then $\mathrm{X} \sim \mathrm{N}(\mu, \sigma^2)$ But we know that

$f(\mathrm{x}) = \dfrac{1}{\sigma}\phi\left(\dfrac{\mathrm{X}-\mu}{\sigma}\right)$ from which the claim follows. Conversely, if $\mathrm{X} \sim \mathrm{N}(\mu, \sigma^2)$, then $\mathrm{Z} = \dfrac{\mathrm{X}-\mu}{\sigma} \sim \mathrm{N}(0,1)$. It is also easily shown that the cumulative distribution function satisfies

$$\mathrm{F(x)} = \Phi\left(\frac{\mathrm{X}-\mu}{\sigma}\right)$$

and so the cumulative probabilities for any normal random variable can be calculated using the tables for the standard normal distribution..

Definition A variable X is said to be standardized if it has been adjusted (or transformed) such that its mean equals 0 and its standard deviation equals 1. Standardization can be accomplished using the formula for a z-score: $\mathrm{Z} = \dfrac{\mathrm{X}-\mu}{\sigma} \sim \mathrm{N}(0,1)$. The z-score represents the number of standard deviations that a data value is away fromthe m ean.

Computing Normal Probabilities

It is very important to understand how the standardized normal distribution works, so we will spend some time here going over it. There is no simple analytic expression for $\Phi(z)$ in terms of elementary functions. but the values of $\Phi(z)$ has been exhaustively tabulated. This greatly simplifies the task of computing normal probabilities..

Table 1 below reports the cumulative normal probabilities for normally distributed variables in standardized form (i.e. Z-scores). That is, this table reports $\mathrm{P(Z \le z)} = \Phi(z)$). For a given value of Z, the table reports what proportion of the distribution lies below that value. For example, $\mathrm{P(Z \le 0)} = \Phi(0) = 0.5$; half the area of the standardized normal curve lies to the left of $Z = 0$.

Theorem: It may be useful to keep in mind that

i) $\mathrm{P(Z > z)} = 1 - \Phi(z)$ complementary law

ii) $\mathrm{P(Z \le -z)} = \mathrm{P(Z \ge z)} = 1 - \Phi(z)$ ie due to symmetry

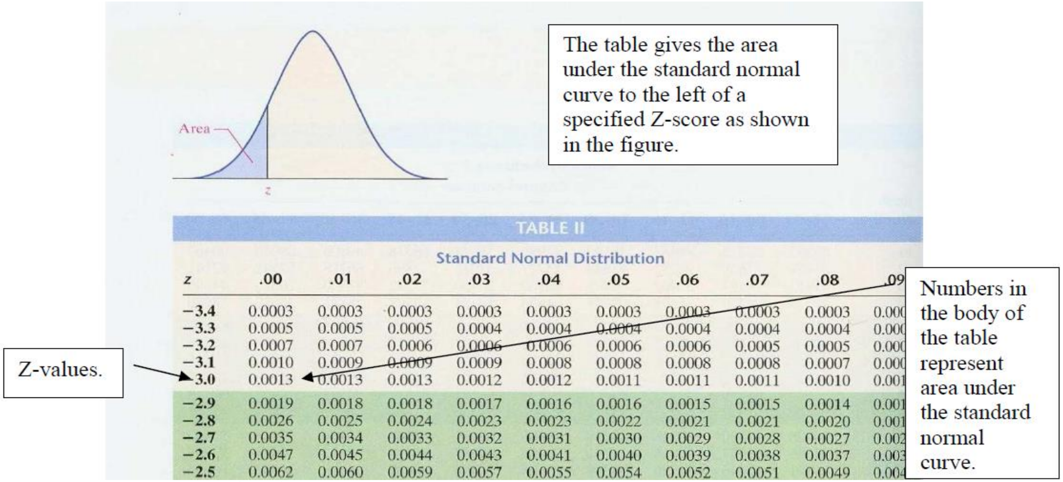
$\Rightarrow \quad \Phi(z) + \Phi(-z) = 1$ Since $\mathrm{P(Z \le z)} + \mathrm{P(Z \ge z)} = 1$

iii) $\mathrm{P(a} \le z \le \mathrm{b)} = \Phi(b) - \Phi(a)$

iv) $\mathrm{P(\text{-}a} \le z \le \mathrm{a)} = 2\Phi(a) - 1$since $\mathrm{P(\text{-}a} \le z \le \mathrm{a)} = \Phi(a) - \Phi(-a) = \Phi(a) - [1 - \Phi(a)] = 2\,\Phi(a) - 1$

v) If we now make $\Phi(a)$ the subject, then $\Phi(a) = \frac{1}{2}[1 + \mathrm{P(\text{-}a} \le z \le \mathrm{a)}]$

Table 1



The table gives the area under the standard normal curve to the left of a specified Z-score as shown in the figure.

TABLE II

Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.000
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.000
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.000
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.000
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.001
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.001
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.001
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.002
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.003
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.004

Z-values.

Numbers in the body of the table represent area under the standard normal curve.

Example 1 Given $Z \sim N(0,1)$, find;

a) $P(Z \leq z)$ if $z = 1.65, -1.65, 1.0, -1.0$

b) $P(Z > z)$ for $z = 1.02, -1.65$

c) $P(0.365 \leq z \leq 1.75)$

d) $P(-0.696 \leq z \leq 1.865)$

e) $P(-2.345 \leq z \leq -1.65)$

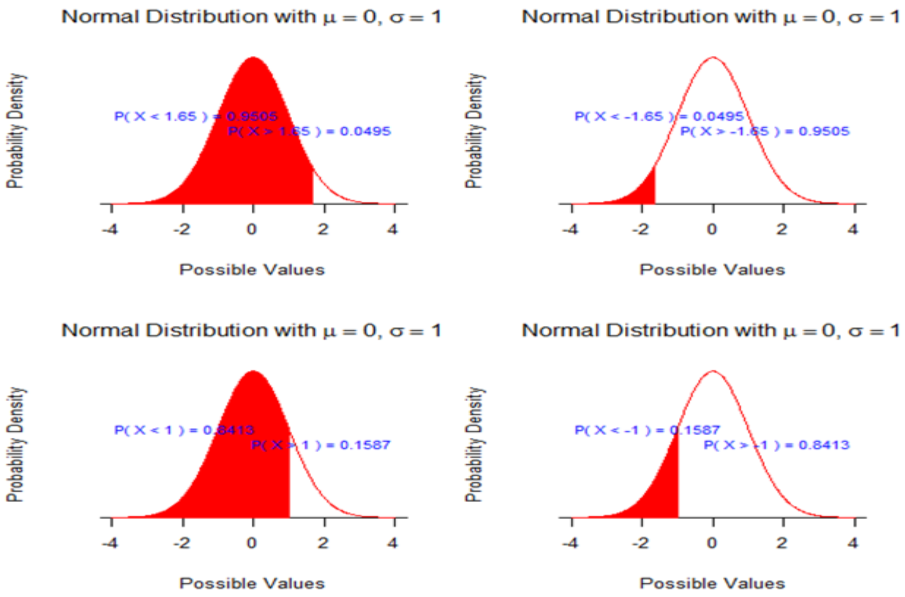
f) $P(|z| \leq 1.43)$

Solution

a) Look up and report the value for $\Phi(z)$ from the standard normal probabilities table

$P(Z \leq 1.65) = \Phi(1.65) = 0.9505 \quad \Phi(-1.65) = 0.0495 \quad \Phi(1.0) = 0.8413$

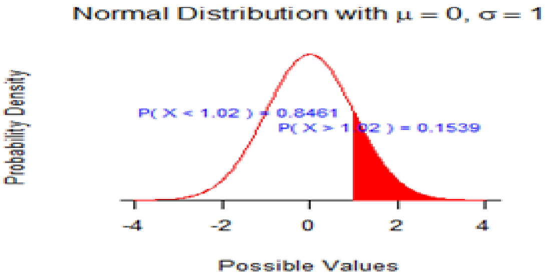
$\Phi(-1.0) = 0.1587$

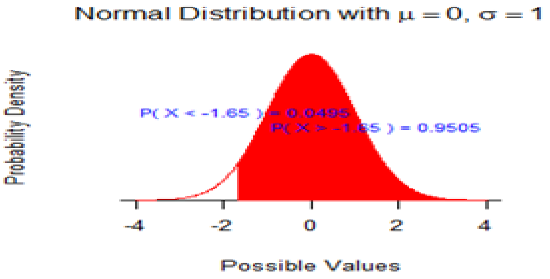


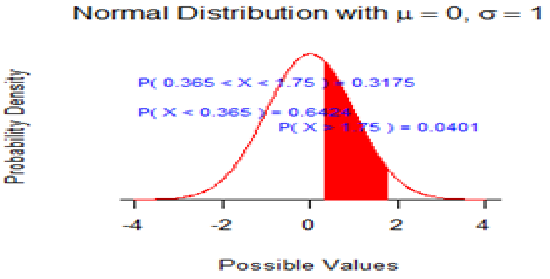
b) $P(Z > z) = \Phi(-z)$ Thus $P(Z > 1.02) = \Phi(-1.02) = 0.1515 \quad P(Z > -1.65) = \Phi(1.65) = 0.9505$

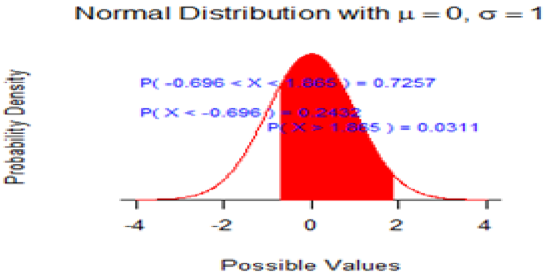
c) $P(0.365 \leq z \leq 1.75) = \Phi(1.75) - \Phi(0.365) = 0.9599 - 0.6350 = 0.3249$

d) $P(-0.696 \leq z \leq 1.865) = \Phi(1.865) - \Phi(-0.696) = 0.9689 - 0.2432 = 0.3249 = 0.7257$









e) $P(-2.345 \le z \le -1.65) = \Phi(-1.65) - \Phi(-2.345) = 0.0505 - 0.0095 = 0.0410$

f) $P(|z| \le 1.43) = P(-1.43 \le z \le 1.43) = 2\Phi(1.43) - 1 = 2(0.9236) - 1 = 0.8472$

Example 2

If $Z \sim N(0,1)$, find the value of t for which;

a) $P(Z \le t) == 0.6026,\ \ 0.9750,\ \ 0.3446$

b) $P(Z > t) == 0.4026,\ 0.7265,\ 0.5446$

c) $P(-0.28 \le z \le t) = 0.2665$

d) $P(-t \le z \le t) = 0.9972,\ 0.9505,\ 0.9750$

Solution

Here we find the probability value in Table I, and report the corresponding value for Z.

a) $\Phi(t) = 0.6026 \ \Rightarrow\ t = 0.26 \quad \Phi(t) = 0.950 \ \Rightarrow\ t = 1.96 \quad \Phi(t) = 0.3446 \ \Rightarrow\ t = -0.40$

b) $P(Z > t) = 0.4026 \Rightarrow \Phi(t) = 0.5974 \Rightarrow t = 0.25$

$P(Z > t) = 0.7265 \Rightarrow \Phi(t) = 0.2735 \Rightarrow t = -0.60$

$P(Z > t) = 0.5446 \Rightarrow \Phi(t) = 0.4554 \Rightarrow t = -0.11$

c) $P(-0.28 \le z \le t) = \Phi(t) - \Phi(-0.28) = 0.2665 \Rightarrow \ \Phi(t) = 0.3897 + 0.2665 \ \Rightarrow t = 0.40$

d) $P(-t \le z \le t) = 2\ \Phi(t) - 1 = 0.9972 \ \Rightarrow\ \Phi(t) = 0.9986 \ \Rightarrow t = 2.99$

$P(-t \le z \le t) = 2\ \Phi(t) - 1 = 0.9505 \ \Rightarrow\ \Phi(t) = 0.9753 \ \Rightarrow t = 1.96$

$P(-t \le z \le t) = 2\ \Phi(t) - 1 = 0.9750 \Rightarrow \ \Phi(t) = 0.9875 \ \Rightarrow t = 2.24$

Exercise

1..Given $Z \sim N(0,1)$, find;

a) $P(Z \le z)$ if $z = 1.95,\ -1.89,\ 1.074,\ \ -1.53$

b) $P(Z > z)$ for $z = 1.72,\ -1.15$

c) $P(0 \le z \le 1.05)$

d) $P(-1.396 \le z \le 1.125)$

e) $P(-1.96 \le z \le -1.65)$

f) $P(|z| \le 2.33)$

2..If $Z \sim N(0,1)$, find the value of z for which;

a) $P(Z \le a) == 0.973,\ \ 0.6693,\ \ 0.4634$

b) $P(Z > a) == 0.3719,\ 0.9545,\ 0.7546$

c) $P(-1.21 \le z \le t) = 0.6965$

d) $P(|z| \le t) = 0.9544,\ 0.9905,\ 0.3750$

Let $X \sim N(\mu, \sigma^2)$ then $P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$where

$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

Example 1

A r.v $\mathrm{X} \sim \mathrm{N}(50, 25)$ compute $\mathrm{P}(45 \le z \le 60)$

Solution

$\mu = 50$ and $\sigma = 5 \;\Rightarrow\; Z = \frac{\mathrm{x}-50}{5} \sim N(0,1)$

$\mathrm{P}(45 \le X \le 60) = \mathrm{P}\left(\frac{45-50}{5} \le Z \le \frac{60-50}{5}\right) = \Phi(2) - \Phi(-1) = 0.9772 - 0.1587 = 0.8185$

Example 2

Suppose $X \sim \mathrm{N}(30, 16)$. Find; a) $P(X < 40)$ b) $P(X > 21)$ c) $P(30 < X < 35)$

Solution

$X \sim \mathrm{N}(30, 16) \Rightarrow Z = \frac{\mathrm{X}-30}{4} \sim \mathrm{N}(0, 1)$

a) $\mathrm{P}(X \le 40) = \mathrm{P}\left(Z \le \frac{40-30}{4}\right) = \Phi(2.5) = 0.9938$

b) $P(X > 21) = \mathrm{P}\left(Z > \frac{21-30}{4}\right) = \mathrm{P}(Z > -2.25) = \mathrm{P}(Z \le 2.25) = \Phi(2.25) = 0.9878$

c) $P(30 < X < 35) = P\left(\frac{30-30}{4} \le Z \le \frac{35-30}{4}\right) = P(0 < Z < 1.25) = 0.8944 - 0.5 = 0.3944$

Example 3

The top 5% of applicants (as measured by GRE scores) will receive scholarships. If $\mathrm{GRE} \sim \mathrm{N}(500, 100^2)$, how high does your GRE score have to be to qualify for a scholarship?

Solution

Let $\mathrm{X} = \mathrm{GRE}$. We want to find x such that $\mathrm{P(X \ge x)} = 0.05$ This is too hard to solve as it stands - so instead, compute $Z = \frac{\mathrm{X}-500}{100} \sim \mathrm{N}(0, 1)$ and find z for the problem,

$$\mathrm{P}(\mathrm{Z} \ge z) = 1 - \Phi(z) = 0.05 \quad \Rightarrow \quad \Phi(z) = 0.95 \quad \Rightarrow \quad z = 1.645$$

To find the equivalent x, compute $\mathrm{X} = \mu + \sigma \mathrm{Z} \;\Rightarrow\; x = 500 + 100(1.645) = 66.5$

Thus, your GRE score needs to be 665 or higher to qualify for a scholarship.

Example 4

Family income is believed to be normally distributed with a mean of \$25000 and a standard deviation on \$10000. If the poverty level is \$10,000, what percentage of the population lives in poverty? A new tax law is expected to benefit "middle income" families, those with incomes between \$20,000 and \$30,000. What percentage of the population will benefit from the law?

Solution

Let X = Family income. We want to find P(X ≤ \$10,000)., so

$X \sim \mathrm{N}(25000, 10000^2) \Rightarrow Z = \frac{\mathrm{X}-25000}{10000} \sim \mathrm{N}(0, 1)$

$\mathrm{P}(X \le 10{,}000) = \mathrm{P}(Z \le -1.5) = \Phi(-1.5) = 0.0668$.

Hence, a slightly below 7% of the population lives in poverty.

$\mathrm{P}(20{,}000 \le X \le 30{,}000) = \mathrm{P}(-0.5 \le Z \le 0.5) = 2\Phi(0.5) - 1 = 2 \times 0.6915 - 1 - 0.383$

Thus, about 38% of the taxpayers will benefit from the new law.

Exercise

1) Suppose $X \sim \mathrm{N}(130, 25)$. Find; a) $P(X < 140)$ b) $P(X > 120)$ c) $P(130 < X < 135)$
2) The random variable X is normally distributed with mean 500 and standard deviation 100. Find; (i) $P(X < 400)$, (ii) $P(X > 620)$ (iii) the 90th percentile (iv) the lower and upper quartiles. Use graphs with labels to illustrate your answers.
3) A radar unit is used to measure speeds of cars on a motorway. The speeds are normally distributed with a mean of 90 km/hr and a standard deviation of 10 km/hr. What is the probability that a car picked at random is travelling at more than 100 km/hr?

4) For a certain type of computers, the length of time bewteen charges of the battery is normally distributed with a mean of 50 hours and a standard deviation of 15 hours. John owns one of these computers and wants to know the probability that the length of time will be between 50 and 70 hours

5) Entry to a certain University is determined by a national test. The scores on this test are normally distributed with a mean of 500 and a standard deviation of 100. Tom wants to be admitted to this university and he knows that he must score better than at least 70% of the students who took the test. Tom takes the test and scores 585. Will he be admitted to this university?

6) A large group of students took a test in Physics and the final grades have a mean of 70 and a standard deviation of 10. If we can approximate the distribution of these grades by a normal distribution, what percent of the student; (a) scored higher than 80? (b) should pass the test (grades≥60)? (c) should fail the test (grades<60)?

7) A machine produces boltswhich are N(4 0.09) where measurements are in cm. Bolts are measured accurately and any bolt smaller than 3.5 cm or larger than 4.4 cm is rejected. Out of 500 bolts how many would be accepted? Ans 430

8) Suppose IQ ~ N(100,22.5).a woman wants to form an Egghead society which only admits people with the top 1% IQ score. What should she have to set the cut-off in the test to allow this to happen? Ans 134.9

9) A manufacturer does not know the mean and standard deviation of ball bearing he is producing. However a sieving system rejects all the bearings larger than 2.4 cm and those under 1.8 cm in diameter. Out of 1,000 ball bearings, 8% are rejected as too small and 5.5% as too big. What is the mean and standard deviation of the ball bearings produced? Ans mean=2.08 sigma=0.2

Normal Approximation to Binomial

Introduction

Suppose a fair coin is tossed 10 times, whar is the probability of observing: a) exactly 4 heads b) at most 4 heads?

Solution

Let X be the r.v the number of heads observed then $X \sim \text{Bin}(10, 0.5)$

$$\Rightarrow P(X = x) = {}_{10}C_x(0.5)^x(0.5)^{10-x} = {}_{10}C_x\left(\tfrac{1}{2}\right)^{10} \text{ for x = 0,1,2,….,10}$$

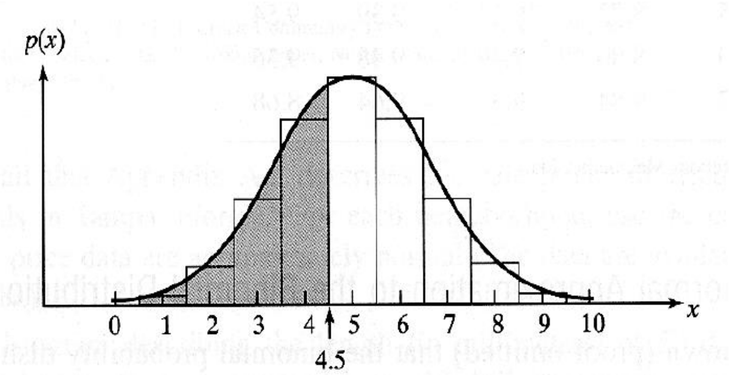
a) $\Rightarrow P(X = 4) = {}_{10}C_4\left(\tfrac{1}{2}\right)^{10} = \dfrac{105}{512} \approx 0.2051$

b) $P(X \le 4) = \left[{}_{10}C_0 + {}_{10}C_1 + {}_{10}C_2 + {}_{10}C_3 + {}_{10}C_4\right](0.5)^{10} = \dfrac{193}{512} \approx 0.3770$

Normal approximation:

Many interesting problems can be addressed via the binomial distribution. However, for large n, it is sometimes difficult to directly compute probabilities for a binomial (*n*, *p*) random variable, X. **Eg**: Compute $P(X \le 12)$ for 25 tosses of a fair coin. Direct calculations can get cumbersome very quickly. . Fortunately, as n becomes large, the binomial distribution becomes more and more symmetric, and begins to converge to a normal distribution. That is, for a large enough n, a binomial variable X is approximately ∼ N(*np, npq*). Hence, the normal distribution can be used to approximate the binomial distribution.

To get a feel for why this might work, let us draw the probability histogram for 10 tosses of a fair coin. The histogram looks bell-shaped, as long as the number of trials is not too small



In general, the distribution of a binomial random variable may be accurately approximated by that of a normal random variable, as long as $np \geq 5$ and $n\text{q} \geq 5$, and assuming that a .continuity correction. is made to account for the fact that we are using a continuous distribution (the normal) to approximate a discrete one (the binomial).

In approximating the distribution of a binomial random variable X, we will use the normal distribution with mean $\mu = np$ and variance $\sigma^2 = npq$ where $q = 1 - p$. Why are these reasonable choices of μ, σ^2?

Continuity Correction

In the binomial, $P(X \leq a) + P(X \geq a+1) = 1$ whenever a is an integer. But if we sum the area under the normal curve corresponding to $P(X \leq a) + P(X \geq a+1)$, this area does not sum to 1 because the area from a to (a + 1) is missing.

The usual way to solve this problem is to associate 1/2 of the interval from a to a + 1 with each adjacent integer. The continuous approximation to the probability $P(X \leq a)$ would thus be $P(X \leq a + \frac{1}{2})$, while the continuous approximation to $P(X \geq a+1)$ would be $P(X \geq a + \frac{1}{2})$. This adjustment is called a continuity correction. More specifically,

$$\underbrace{P(X \leq x) = P(X < x+1)}_{\text{Binomial distribution}} \rightarrow \underbrace{P(X \leq x + 0.5)}_{\text{Normal approximation}} = \underbrace{P\left(z \leq \tfrac{x+0.5-np}{\sqrt{npq}}\right)}_{\text{Standardd Normal approx}}$$

$$\underbrace{P(X \geq x) = P(X > x-1)}_{\text{Binomial distribution}} \rightarrow \underbrace{P(X \geq x - 0.5)}_{\text{Normal approximation}} = \underbrace{P\left(z \leq \tfrac{x-0.5-np}{\sqrt{npq}}\right)}_{\text{Standardd Normal approx}}$$

$$\underbrace{P(a \leq X \leq b) = P(a-1 < X < b+1)}_{\text{Binomial distribution}} \rightarrow = \underbrace{P(a-0.5 \leq X \leq b+0.5)}_{\text{Normal approximation}} = P\left(\tfrac{a-0.5-np}{\sqrt{npq}} \leq z \leq \tfrac{b+0.5-np}{\sqrt{npq}}\right)$$

$$\underbrace{P(X = x) = P(x-1 < X < x+1)}_{\text{Binomial distribution}} \rightarrow \underbrace{P(x-0.5 \leq X \leq x+0.5)}_{\text{Normal approximation}} = \underbrace{P\left(\tfrac{x-0.5-np}{\sqrt{npq}} \leq z \leq \tfrac{x+0.5-np}{\sqrt{npq}}\right)}_{\text{Standardd Normal approx}}$$

NOTE: For the binomial distribution, the values to the right of each = sign are primarily included for illustrative purposes. The equalities which hold in the binomial distribution do not hold in the normal distribution, because there is a gap between consecutive values of a. The normal approximation deals with this by "splitting" the difference.

For example, in the binomial, $P(X \leq 6) = P(X < 7)$ since 6 is the next possible value of X that is less than 7. In the normal, we approximate this by finding $P(X \leq 6.5)$. And, in the binomial, $P(X \geq 6) = P(X > 5)$, because 6 is the next value of X that is greater than 5. In the normal, we approximate this by finding $P(X \geq 5.5)$

Returning to the case of coin tossing Suppose we wish to find $P(X \le 4)$, the probability that the binomial r.v is less than or equal to 4. In the diagram above, the bars represent the binomial distribution with n = 10, p = 0.5. The superimposed curve is a normal density f(x). The mean of the normal is $\mu = np = 5$ and the standard deviation is $\sigma = \sqrt{10(0.5)(0.5)} \approx 1.58$ Using the normal approximation, we need to calculate the probability that our normal r.v is less than or equals to 4.5. ie

$\underbrace{P(X \le 4)}_{Binomial} = \underbrace{P(X \le 4.5)}_{Normal} = \underbrace{P\left(Z \le \tfrac{4.5-5}{1.58}\right)}_{std\ normal} = \Phi(-0.3162) = 0.3759$ which is very close to the actual answer of 0.377

Example 1

Suppose 50% of the population approves of the job the governor is doing, and that 20 individuals are drawn at random from the population. Solve the following, using the normal approximation to the binomial. What is the probability that;.

a) exactly 7 people will support the governor?
b) at least 7 people will support the governor?
c) more than 11 people will support the governor?
d) 11 or fewer will support the governor?

Solution

Note that $n = 20,\ p = 0.5 \Rightarrow \mu = np = 10$ and $\sigma = \sqrt{npq} = \sqrt{5}$ Since $np \ge 5$ and $nq \ge 5$, it is probably safe to assume that $X \sim N(10,5)$

a) $\underbrace{P(X = 7)}_{Binomial} = \underbrace{P(6.5 \le X \le 7.5)}_{Normal} = \underbrace{P\left(\tfrac{6.5-10}{\sqrt{5}} \le Z \le \tfrac{7.5-10}{\sqrt{5}}\right)}_{std\ normal} = \underbrace{P(-1.565 \le Z \le -1.118)}_{std\ normal}$

$= \Phi(-1.118) - \Phi(-1.565) = 0.1318 \text{ - } 0.0588 = 0.0730$

b) $\underbrace{P(X \le 7)}_{Binomial} = \underbrace{P(X \le 7.5)}_{Normal} = \underbrace{P(Z \le -1.118)}_{std\ normal} = 0.1318$

c) $\underbrace{P(X > 11)}_{Binomial} = \underbrace{P(X \ge 11.5)}_{Normal} = \underbrace{P(Z \ge 0.6708)}_{std\ normal} = 1 - \Phi(0.6708) = 1 - 0.7488 = 0.2512$

d) $\underbrace{P(X \le 11)}_{Binomial} = \underbrace{P(X \le 11.5)}_{Normal} = \underbrace{P(Z \le 0.6708)}_{std\ normal} = \Phi(0.6708) = 0.7488$

Example 2

In each of 25 races, the Democrats have a 60% chance of winning. What are the odds that the Democrats will win 19 or more races? Use the normal approximation to the binomial

Solution

Note that $n = 25,\ p = 0.6 \Rightarrow \mu = np = 15$ and $\sigma = \sqrt{npq} = \sqrt{6}$ Since $np > 5$ and $nq > 5$, it is probably safe to assume that $X \sim N(15,6)$. .

Using the normal approximation to the binomial,

$\underbrace{P(X \ge 19)}_{Binomial} = \underbrace{P(X \ge 18.5)}_{Normal} = \underbrace{P(Z \ge 1.4289)}_{std\ normal} = 1 - \Phi(1.4289) = 1 - 0.9235 = 0.0765$

Hence, Democrats have a little less than an 8% chance of winning 19 or more races.

Example 3

Tomorrow morning Iberia flight to Madrid can seat 370 passengers. From past experience, Iberia knows that the probability is 0.90 that a given ticket-holder will show up for the flight. They have sold 400 tickets, deliberately overbooking the flight. How confident can Iberia be that no passenger will need to be .bumped. (denied boarding)?

Solution:

We will assume that the number (X) of passengers showing up for the flight has a binomial distribution with mean $\mu = 400 \times 0.9 = 360$ and standard deviation $\sigma = \sqrt{400 \times 0.9 \times 0.1} = 6$

We want $\underbrace{P(X \le 370)}_{Binomial} = \underbrace{P(X \le 370.5)}_{Normal} = \underbrace{P(Z \le 1.75)}_{std\ normal} = 0.9599$ So the probability that nobody gets bumped is approximately 0.9599. (Almost 96%).

Exercise (Use the normal approximation to the binomial in the entire exercise)

1. A coin is loaded such that heads is thrice as likely as the tails. Find the probability of observing between 4 and 7 heads inclusive with 12 tosses of the coin.
2. Based upon past experience, 40% of all customers at Miller’s Automotive Service Station pay for their purchases with a credit card. If a random sample of 200 customers is selected, what is the *approximate* probability that;
   a) at least 75 pay with a credit card? b) not more than 70 pay with a credit card?
   b) between 70 and 75 customers, inclusive, pay with a credit card?
3. The probability that Ronado scores a goal in any game against a tough opponent in soccer is 0.3. What is the probability that he scores 30 goals in the next 100 games in which he plays?
4. Crafty Computers limited produces PCs. The probability that one of their computers has a virus is 0.25. JKUAT ICSIT buys 300 computers from the company. What is the probability that between 70 and 80 PCs inclusive have a virus? Would you advice the director JKUAT ICSIT to buy Computers from this company in future?
5. For overseas flights, an airline has three different choices on its dessert menu—ice cream, apple pie, and chocolate cake. Based on past experience the airline feels that each dessert is equally likely to be chosen.
   a) If a random sample of four passengers is selected, what is the probability that at least two will choose ice cream for dessert
   b) If a random sample of 21 passengers is selected, what is the *approximate* probability that at least eight will choose ice cream for dessert?
6. In a family of 11 children, what is the probability that there will be more boys than girls?.
7. A baseball player has a long term batting average of 0.300. What is the chance he gets an average of 0.330 or higher in his next 100 bats?
8. Suppose we draw a Simple Random Sample of 1,500 Americans and want to assess whether the representation of blacks in the sample is accurate. We know that about 12% of Americans are black, what is the probability that the sample contains 170 or fewer blacks?
9. Let T be the lifetime in years of new bus engines. Suppose that T is continuous with probability density function $\mathrm{f(t)} = \begin{cases} 0 & \mathrm{for\ T} < 1 \\ {}^{\mathrm{d}}\!/_{\mathrm{t}^3} & \mathrm{for\ T} \ge 1 \end{cases}$ for some constant d.
   a) Find the value of d and the mean and median of T.
   b) Suppose that 240 new bus engines are installed at the same time, and that their lifetimes are independent. By using a normal approximation to the binomial, find the probability that at most 10 of the engines last for 4 years or more.

Sums of Independent Random Variables

Theorem 1

If $X_1, X_2, ..., X_n$ are n independent continuous r.v each with mean μ and variance $\sigma^2 < \infty$ then r.v $\mathrm{W} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i}$ has mean μ and variance $\frac{\sigma^2}{n}$

Theorem 2

If $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ are 2 independent r.v, then
$X_1 \pm X_2 \sim \mathrm{N}(\mu_1 \pm \mu_2\ ,\ \sigma_1^2 + \sigma_2^2)$

Proof

Let the mgf of X_1 be $M_1(t) = e^{t\mu_1 + \frac{1}{2}t^2\sigma_1^2}$ and for X_2 be $M_2(t) = e^{t\mu_2 + \frac{1}{2}t^2\sigma_2^2}$. Let $Y_1 = X_1 + X_2$ so Y_1 has mgf $M_1(t) = M_1(t) \times M_2(t) = e^{t\mu_1 + \frac{1}{2}t^2\sigma_1^2} \times e^{t\mu_2 + \frac{1}{2}t^2\sigma_2^2} = e^{t(\mu_1+\mu_2) + \frac{1}{2}t^2(\sigma_1^2+\sigma_2^2)}$ which is the mgf of a normal r.v with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$

Note $Y_2 = -X_2 \sim N(-\mu_2, +\sigma_2^2) \Rightarrow X_1 - X_2 \sim \mathrm{N}(\mu_1 - \mu_2\ ,\ \sigma_1^2 + \sigma_2^2)$

Theorem 3

If $X_1, X_2, ...\mathrm{X}_n$ are n independent r.vs and each $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$, then the r.v
$X_1 \pm X_2 \pm ..... \pm X_n \sim \mathrm{N}(\mu_1 \pm \mu_2 \pm ..... \pm \mu_n\ ,\ \sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2)$

Proof is by induction (left as exercise to the learner)

Example

If $X \sim \mathrm{N}(60,16)$ and $Y \sim \mathrm{N}(70, 9)$ are 2 independent r.v, Find (a) $P(X + Y \le 140)$
(b) $P(120 \le X + Y \le 135)$ (c) $P(Y - X > 7)$ (d) $P(2 \le Y - X \le 12)$

Solution

$X + Y \sim \mathrm{N}(130,25)$ and $Y - X \sim \mathrm{N}(10,25)$therefore

a) $P(X + Y \le 140) = P(Z \le \frac{140-130}{5}) = \Phi(2) = 0.9772$
b) $P(120 \le X + Y \le 135) = P(\frac{120-130}{5} \le Z \le \frac{135-130}{5}) = \Phi(1) - \Phi(-2) = 0.8413 - 0.0228 = 0.8185$
c) $P(Y - X > 7) = P(Z > \frac{7-10}{5}) = P(Z > -0.6) = P(Z \le 0.6) = \Phi(0.6) = \ 0.7257$
d) $P(2 \le Y - X \le 12) = P(\frac{2-10}{5} \le Z \le \frac{12-10}{5}) = \Phi(0.4) - \Phi(-1.6) = 0.6554 - 0.0548 = 0.6006$

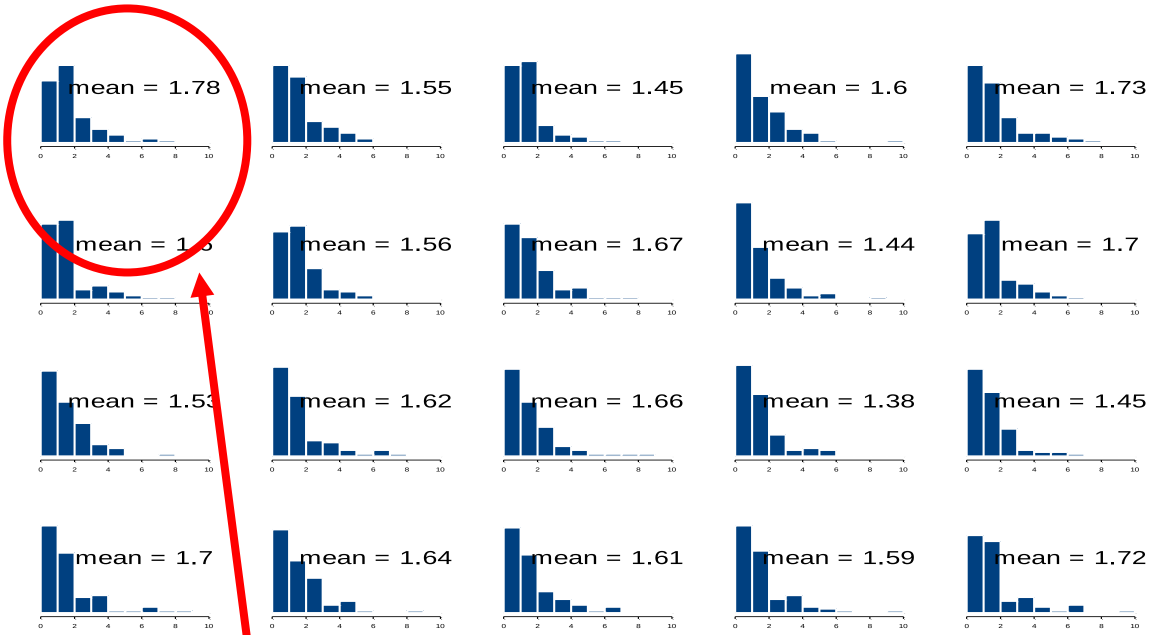
Exercise

1. If $X \sim \mathrm{N}(65,28)$ and $Y \sim \mathrm{N}(85, 36)$ are 2 independent r.v, Find (a) $P(X + Y \le 142)$
(b) $P(134 \le X + Y \le 166)$ (c) $P(Y - X > 4)$ (d) $P(12 \le Y - X \le 24)$
2. Each day Mr. Njoroge walks to the library bto read a newspaper. Total time spent walking is normally distributed with mean 15 minutes and standard deviation 2 minutes. Total time spent in the library is also normally distributed with mean 25 minutes and standard deviation $\sqrt{12}$ minutes. Find the probability that on one day;
   a) he is away from his home for more than 45 minutes.
   b) he spends more time walking than in the library

Sampling Distributions

In many investigations the data of interest can take on many possible values and it is often of interest to estimate the population mean, μ. A common estimator for μ is the sample mean $\overline{x}$. Consider the following set up: We observe a sample of size *n* from some population and compute the mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ .Since the particular individuals included in our sample are *random*, we would observe a different value of $\overline{x}$ if we repeated the procedure. That is, $\overline{x}$ *is also a random quantity.* Its value is determined partly by which people are randomly chosen to be in the sample. If we repeatedly drew samples of size *n* and calculated $\overline{x}$, we could ascertain the sampling distribution of $\overline{x}$.

Many possible samples, many possible $\overline{x}$'s



We only see one!

We will have a better idea of how good our one estimate is if we have good knowledge of how $\bar{x}$ behaves; that is, if we know the probability distribution of $\bar{x}$.

Properties of the Sampling Distribution of the Sample Means (Summary)

When all of the possible sample means are computed, then the following properties are true:

1. The mean of the sample means will be the mean of the population
2. The variance of the sample means will be the variance of the population divided by the sample size.
3. The standard deviation of the sample means (known as the standard error of the mean) will be smaller than the population mean and will be equal to the standard deviation of the population divided by the square root of the sample size.
4. If the population has a normal distribution, then the sample means will have a normal distribution.
5. If the population is not normally distributed, but the sample size is sufficiently large, then the sample means will have an approximately normal distribution. Some books define sufficiently large as at least 30 and others as at leas t25.

Definitions

Central Limit Theorem:- Stats that as the sample size increases, the sampling distribution of the sample means will become approximately normally distributed.

Sampling Distribution of the Sample Means:- Distribution obtained by using the means computed from random samples of a specific size.

Sampling Error :- Difference which occurs between the sample statistic and the population parameter due to the fact that the sample isn't a perfect representation of the population.

Standard Error or the Mean:- The standard deviation of the sampling distribution of the sample means. It is equal to the standard deviation of the population divided by the square root of the sample size.

The Mean and Standard Deviation of $\bar{x}$

What are the mean and standard deviation of $\bar{x}$?

Let's be more specific about what we mean by a sample of size *n*. We consider the sample to be a collection of *n independent and identically distributed* (or *iid) random* variables $X_1, X_2, ..., X_n$ with common mean μ and common standard deviation σ.

Thus, $E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}(n\mu) = \mu$

$$Var(\overline{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(x_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n} \quad \Rightarrow SD(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

The Central Limit Theorem

Now we know that $\overline{x}$ has mean μ and standard deviation $\sigma/\sqrt{n}$, but what is its distribution?

If $X_1, X_2, ..., X_n$ are *normally distributed*, then $\overline{x}$ is also normally distributed. Thus,

$X_i \sim N(\mu, \sigma^2) \;\Rightarrow\; \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \;\Rightarrow\; z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$ **.** If $X_1, X_2, ..., X_n$ are *not* normally distributed, then the Central Limit Theorem tells us that $\overline{x}$ is *approximately N*ormal.

In brief if $X_1, X_2, ..., X_n$ are *iid* random variables with mean μ and finite standard deviation σ. Then for a sufficiently large n, the sampling distribution of $\overline{X}$ is approximately Normal with mean μ and variance $\frac{\sigma^2}{n}$.

Remark: Central limit theorem involves two different distributions: the distribution of the original population and the distribution of the sample means

Example

Intelligence Quotient (IQ) is normally distributed with mean 110 and standard deviation of 10. A moron is a person with IQ less than 80. Find the probability that a randomly chosen person is a moron. Let idiot be defined as one with an IQ less than 90. Find the probability that a randomly chosen person is an idiot. (Hint this random variable is for a single person X)

If a sample of 25 students is available, what is the probability that the **average** IQ exceeds 105? What is the probability that the average IQ exceeds 115 (Hint this random variable is for an average over 25 persons or $\overline{X}$)

Solution

$IQ = X \sim \mathrm{N}(110_1, 10^2)$, and therefore for a sample of 25 people average $IQ = \overline{X} \sim \mathrm{N}(110_1, 4)$

The probability that a randomly chosen person is a moron is given by

$P(X < 80) = P\left(Z < \frac{80-110}{10}\right) = \Phi(-3) = 0.0013$

The probability that a randomly chosen person is an idiot is given by

$P(X < 90) = P\left(Z < \frac{90-110}{10}\right) = \Phi(-2) = 0.0228$

The probability that the **average** IQ exceeds 105 is $P(\overline{X} > 105)$

The random variable under consideration here is the average. Hence, a sampling distribution is relevant when we consider average IQ as the variable of interest, not he IQ of an individual student, but the average over 25 students. Standard deviation of the sampling distribution = Standard Error SE = $\sigma/\sqrt{\mathrm{n}} = 10/5 = 2$. Now

$$P(\overline{X} > 105) = P\left(Z > \frac{105-110}{2}\right) = \underbrace{P(Z \le 2.5)}_{due\,to\,symmetry} = \Phi(2.5) = 0.9938$$

We now find probability that the average IQ exceeds 115 ie

$$P(\overline{X} > 115) = P(Z > \tfrac{115-110}{2}) = \underbrace{P(Z \leq -2.5)}_{due to symmetry} = \Phi(-2.5) = 0.0062$$

Exercise

1) The annual salaries of employees in a large company are approximately normally distributed with a mean of \$50,000 and a standard deviation of \$10,000. If a random sample of 50 employees is taken whar is the probability that their average salary is;
   a) less than \$45,000? B) between \$45,000 and \$65,000? c)more than \$70,000
2) Library usually has 13% of its books checked out. Find the probability that in a sample of 588 books greater than 14% are checked out. ANS= 0.2358
3) The length of similar components produced by a company are approximated by a normal distribution model with a mean of 5 cm and a standard deviation of 0.02 cm.
   a) If a component is chosen at random what is the probability that the length of this component is between 4.98 and 5.02 cm?
   b) what is the probability that the average length of a sample of 25 component is between 4.96 and 5.04 cm?
4) The length of life of an instrument produced by a machine has a normal distribution with a mean of 12 months and standard deviation of 2 months. Find the probability that in a random sample of 4 instrument produced by this machine, the average length of life
   a) less than 10.5 months. b) between 11and 13 months.
5) The time taken to assemble a car in a certain plant is a random variable having a normal distribution of 20 hours and a standard deviation of 2 hours. What is the probability that a car can be assembled at this plant in a period of time
   a) less than 19.5 hours? b) between 20 and 22 hours?

STATISTICAL INFERENCES AND HYPOTHESIS TESTING

In research, one always has some fixed ideas about certain population parameters based on say, prior experiments, surveys or experience. However, these are only ideas. There is therefore a need to ascertain whether these ideas /claims are correct or not.

The ascertaining of claims is done by first collecting information in the form of sample data. We then decide whether our sample observations (statistic) have come from a postulated population or not.

Definitions

A **hypothesis** is a claim (assumption) about a population parameter such as the population mean, the population proportion or the population standard deviation. It's rather a postulated or a stipulated value of a parameter

Example: The mean monthly cell phone bill in this city is μ = \$42

The proportion of adults in this city with cell phones is π = 0.68

On the basis of observation data, one then performs a test to decide whether the postulated hypothesis should be accepted or not. However, we note that the decision aspect is prone to error/risk.

Null Hypothesis (denoted H_0): Statement of zero or no change and is the hypothesis which is to be actually tested for acceptance or rejection. If the original claim includes equality ($\leq$, = or $\geq$), it is the null hypothesis. If the original claim does not include equality ($<$, $\neq$ or $>$) then the null hypothesis is the complement of the original claim. The null hypothesis always includes the equal sign. The decision is based on the null hypothesis.

Eg: The average number of TV sets in U.S. homes is equal to three ($H_0 : \mu = 3$)

It's always about a population parameter, and not about a sample statistic

Ie $H_0 : \mu = 3$ and not $H_0 : \bar{x} = 3$

We begin with the assumption that the null hypothesis is true

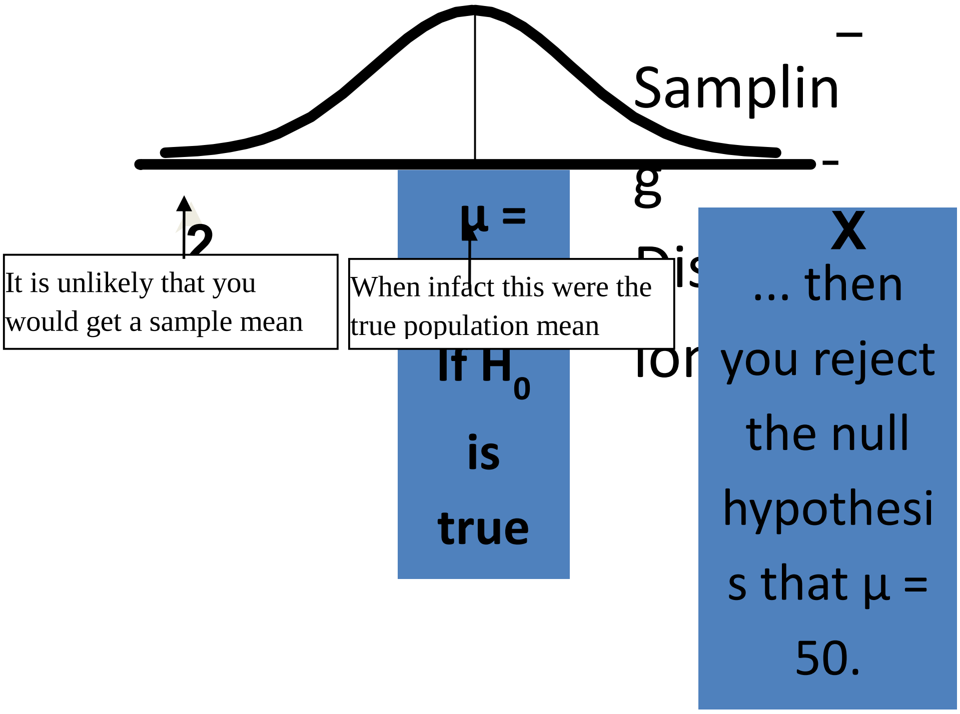
- Similar to the notion of innocent until proven guilty

Alternative Hypothesis (denoted H_1 or H_a): Statement which is true if the null hypothesis is false. it Challenges the status quo. It Is generally the hypothesis that the researcher is trying to prove and it is accepted when H_0 is rejected and vice versa. The type of test (left, right, or two-tail) is based on the alternative hypothesis.

The Hypothesis Testing Process

Claim: The population mean age is 50. Ie Hypothesis H_0: $\mu = 50$, vs H_1: $\mu \neq 50$

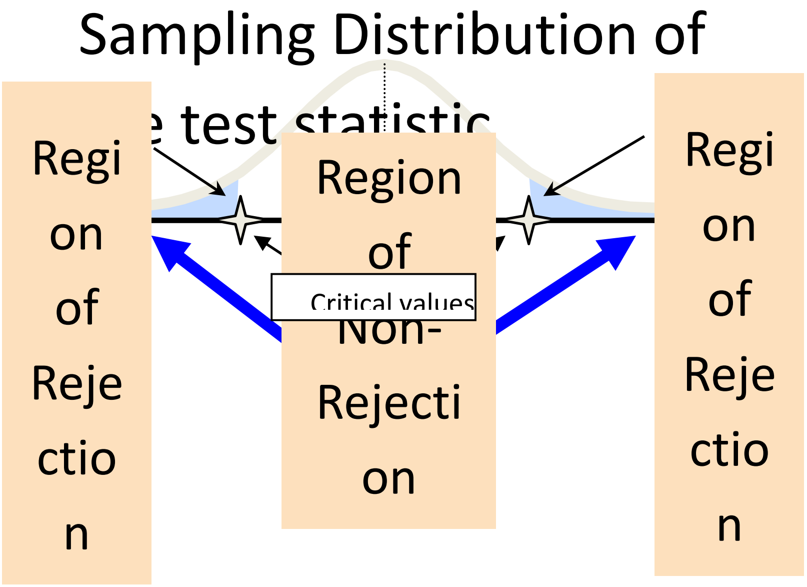
Sample the population and find sample mean. Suppose the sample mean age was $\bar{x} = 20$. This is significantly lower than the claimed mean population age of 50. If the null hypothesis were true, the probability of getting such a different sample mean would be very small, so you reject the null hypothesis .In other words, getting a sample mean of 20 is so unlikely if the population mean was 50, you conclude that the population mean must not be 50.


Sampling Distribution of $\bar{X}$
2
μ = If H_0 is true
It is unlikely that you would get a sample mean
When infact this were the true population mean
... then you reject the null hypothesis that μ = 50.

- If the sample mean is close to the assumed population mean, the null hypothesis is not rejected.

- If the sample mean is far from the assumed population mean, the null hypothesis is rejected.

How far is "far enough" to reject H_0? The critical value of a test statistic creates a "line in the sand" for decision making -- it answers the question of how far is far enough.



"Too Far Away" From Mean of Sampling Distribution

Possible Errors in Hypothesis Test Decision Making

When taking a decision about the acceptance or rejection of a null hypothesis/ alternative hypothesis, there is a risk of committing an error. These errors are of two types:

Type I error; Mistake of rejecting the null hypothesis when it is true (saying false when true). It is usually the more serious error.

The probability of a Type I Error is (denoted α) is Called the level of significance of the test and it is Set by researcher in advance. α= 0.05 and α = 0.01 are common. If no level of significance is given, use α = 0.05. The level of significance is the complement of the level of confidence in estimation.

Type II error: Mistake of failing to reject the null hypothesis when it is false (saying true when false). The probability of a Type II error is denoted by β

Remarks

1) The confidence coefficient (1-α) is the probability of not rejecting H_0 when it is true.
2) The confidence level of a hypothesis test is 100 (1-α)%.
3) The power of a statistical test (1-β) is the probability of rejecting H_0 when it is false

Possible Hypothesis Test Outcomes		
	Actual Situation	
Decision	H_0 True	H_0 False
Do Not Reject H_0	No Error Probability 1 - α	Type II Error Probability β
Reject H_0	Type I Error	No Error

	Probability α	Probability 1 - β

Relationship between Type I & Type II Error

Type I and Type II errors cannot happen at the same time

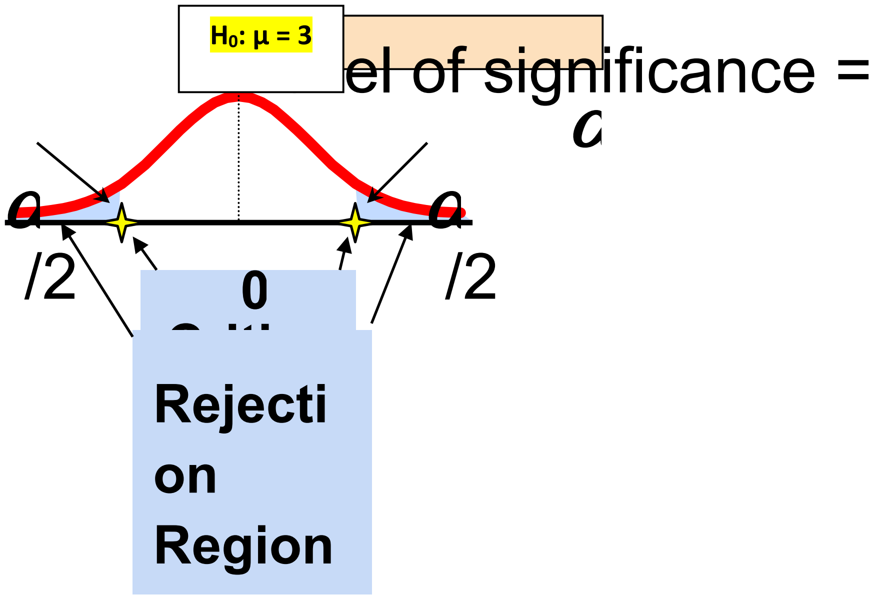
- A Type I error can only occur if H_0 is true
- A Type II error can only occur if H_0 is false

If Type I error probability (α) increases, then Type II error probability (β) decreases

Level of Significance and the Rejection Region

Critical region: Set of all values which would cause us to reject H_0

Critical value(s): The value(s) which separate the critical region from the non-critical region. The critical values are determined independently of the sample statistics.



This is a two-tail test because there is a rejection region in both tails

Test statistic: Sample statistic used to decide whether to reject or fail to reject the null hypothesis

Probability Value (P-value): The probability of getting the results obtained if the null hypothesis is true. If this probability is too small (smaller than the level of significance), then we reject the null hypothesis. If the level of significance is the area beyond the critical values, then the probability value is the area beyond the test statistic.

Decision: A statement based upon the null hypothesis. It is either "reject the null hypothesis" or "fail to reject the null hypothesis". We will never accept the null hypothesis.

Conclusion: A statement which indicates the level of evidence (sufficient or insufficient), at what level of significance, and whether the original claim is rejected (null) or supported (alternative).

Steps in Hypothesis Testing

Any hypothesis testing is done under the assumption that the null hypothesis is true.

Here are the steps to performing hypothesis testing

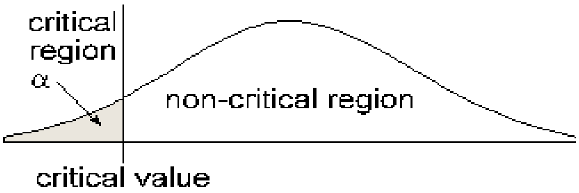
a) Write the null and alternative hypothesis.
b) Use the alternative hypothesis to identify the type of test.

c) specify the level of significance, α and find the critical value using the tables
d) Compute the test statistic
e) Make a decision to reject or fail to reject the null hypothesis.
f) Write the conclusion

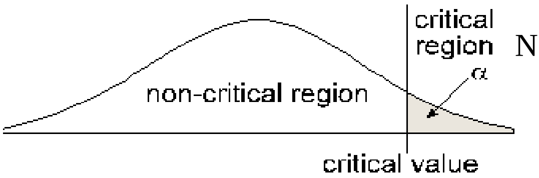
Remarks

The first thing to do when given a claim is to write the claim mathematically (if possible), and decide whether the given claim is the null or alternative hypothesis. If the given claim contains equality, or a statement of no change from the given or accepted condition, then it is the null hypothesis, otherwise, if it represents change, it is the alternative hypothesis.
The type of test is determined by the *Alternative Hypothesis* (H_1)

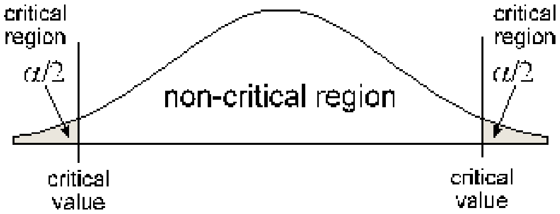
Left Tailed Test
H_1: parameter **<** value
Notice the inequality points to the left
Decision Rule: Reject H_0 if t.s. < c.v.



Right Tailed Test
H_1: parameter **>** value
Notice the inequality points to the right
Decision Rule: Reject H_0 if t.s. > c.v.



Two Tailed Test
H_1: parameter **not equal** to a value
Notice the inequality points to both sides
Decision Rule: Reject H_0 if t.s. < c.v. (left) or t.s. > c.v. (right)



If the test statistic falls into the non rejection region, do not reject the null hypothesis H_0. If the test statistic falls into the rejection region, reject the null hypothesis. Express the managerial conclusion in the context of the problem
Conclusions are sentence answers which include whether there is enough evidence or not (based on the decision) and whether the original claim is supported or rejected. Conclusions are based on the original claim, which may be the null or alternative hypotheses.

Approaches to Hypothesis Testing

There are three approaches to hypothesis testing namely Classical Approach, p vale approach and the confidence interval approach

The Classical Approach

The Classical Approach to hypothesis testing is to compare a test statistic and a critical value. It is best used for distributions which give areas and require you to look up the critical value (like the Student's t distribution) rather than distributions which have you look up a test statistic to find an area (like the normal distribution).

The Classical Approach also has three different decision rules, depending on whether it is a left tail, right tail, or two tail test.

One problem with the Classical Approach is that if a different level of significance is desired, a different critical value must be read from the table.

P-Value Approach

The P-Value, short for Probability Value, Approach to hypothesis testing form a different manner. Instead of comparing z-scores or t-scores as in the classical approach, you're comparing probabilities, or areas.

The level of significance (alpha) is the area in the critical region. That is, the area in the tails to the right or left of the critical values.

The p-value is the area to the right or left of the test statistic. If it is a two tail test, then look up the probability in one tail and double it.

If the test statistic is in the critical region, then the p-value will be less than the level of significance. It does not matter whether it is a left tail, right tail, or two tail test. This rule always holds.

Reject the null hypothesis if the p-value is less than the level of significance.

You will fail to reject the null hypothesis if the p-value is greater than or equal to the level of significance.

The p-value approach is best suited for the normal distribution when doing calculations by hand. However, many statistical packages will give the p-value but not the critical value. This is because it is easier for a computer or calculator to find the probability than it is to find the critical value.

Another benefit of the p-value is that the statistician immediately knows at what level the testing becomes significant. That is, a p-value of 0.06 would be rejected at an 0.10 level of significance, but it would fail to reject at an 0.05 level of significance. Warning: Do not decide on the level of significance after calculating the test statistic and finding the p-value.

Here are a couple of statements to help you keep the level of significance the probability value straight.

The Level of Significance is pre-determined before taking the sample. It **does not** depend on the sample at all. It is the area in the critical region; that is the area beyond the **c**ritical values. It is the probability at which we consider something unusua**l**.

The Probability-Value can only be found after taking the sample. It depends on the sample. It is the area beyond the test statistic**.** It is the probability of getting the results we obtained if the null hypothesis is true.

Confidence Intervals as Hypothesis Tests

Using the confidence interval to perform a hypothesis test only works with a two-tailed test.

a) If the hypothesized value of the parameter lies within the confidence interval with a 1-α level of confidence, then the decision at α level of significance is to fail to reject the null hypothesis.

b) If the hypothesized value of the parameter lies outside the confidence interval with a 1-α level of confidence, then the decision at α level of significance is to reject the null hypothesis.

However, it has a couple of problems.

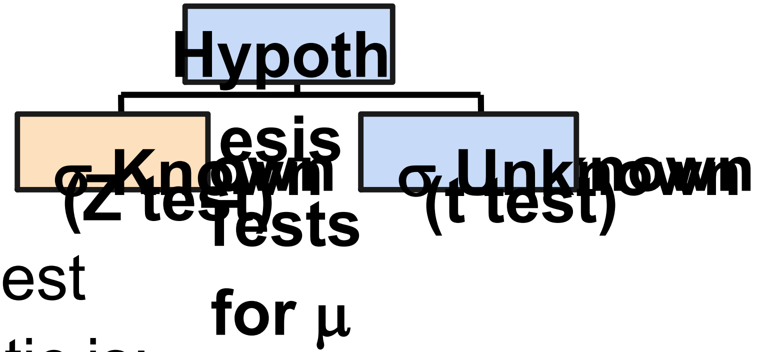
- It only works with two-tail hypothesis tests.
- It requires that you compute the confidence interval first. This involves taking a z-score or t-score and converting it into an x-score, which is more difficult than standardizing an x-score.

Testing a Single Mean

The value for all population parameters in the test statistics come from the null hypothesis. This is true not only for means, but all of the testing we're going to be doing.

The following hypotheses are to be tested: $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ 0r $H_0 : \mu > \mu_0$ 0r $H_0 : \mu < \mu_0$ Where μ_0 is some hypothesised value.

The statistic and the critical values depends on whether **σ**, is known or unknown.



The test statistic is:

a) **Population Standard Deviation Known**

If the population standard deviation **σ**, is known, then the population mean has a normal distribution, and you will be using the z-score formula for sample means. The test statistic is the standard formula you've seen before. $z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

The critical value is obtained from the normal table.

Example 1

Test at 5% level the claim that the true mean # of TV sets in US homes is equal to 3. Suppose the sample results are n = 100, $\bar{x} = 2.84$ (σ = 0.8 is assumed known**)**

Solution

State the appropriate null and alternative hypotheses

- H_0: μ = 3 H_1: μ ≠ 3 (This is a two-tail test)

Determine the appropriate technique

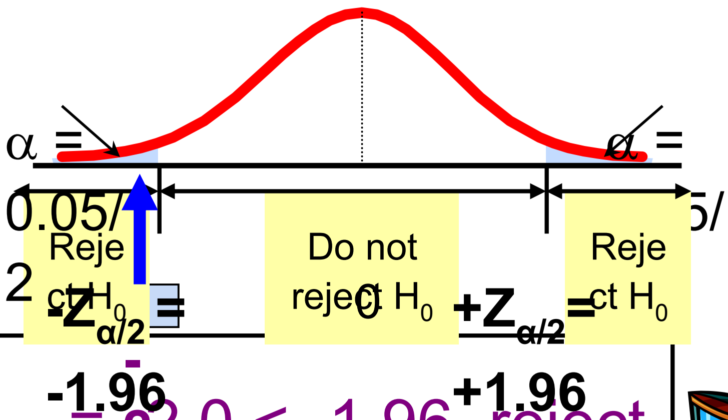
- σ is assumed known so this is a Z test.

Determine the critical values

- For α = 0.05 the critical Z values are ±1.96

Compute the test statistic So the test statistic is:

$$Z_{STAT} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{2.84-3}{0.8/\sqrt{100}} = -2.0$$



Since Z_{STAT} = -2.0 < -1.96, reject the null hypothesis and conclude there is sufficient evidence that the mean number of TVs in US homes is not equal to 3

Exercise

1. A simple random sample of 10 people from a certain population has a mean age of 27. Can we conclude that the mean age of the population is less than 30? The variance is known to be 20. Let α = .05.
2. Bon Air Elementary School has 300 students. The principal of the school thinks that the average IQ of students at Bon Air is at least 110. To prove her point, she administers an IQ test to 20 randomly selected students. Among the sampled students, the average IQ is 108. Assuming variance is known to be 100, should the principal accept or reject her original hypothesis? at 5% level of significance
3. The manager at the Omni Fitness Club in Muskegon, Michigan, believes that the recent remodeling project has greatly improved the club’s appeal for members and that they now stay longer at the club per visit than before the remodeling. Studies show that the previous mean time per visit was 36 minutes, with a standard deviation equal to 11 minutes. A simple random sample of $n = 200$ visits is selected, and the current sample mean is 36.8 minutes. To test the manager’s claim, and partially justify the remodeling project, using $\propto = 0.05$ level, the following steps can be used:
4. The Wilson Glass Company has a contract to supply plate glass for home and commercial windows. The contract specifies that the mean thickness of the glass must be 0.375 inches. The standard deviation, σ, is known to be 0.05 inch. Before sending the first shipment, Wilson managers wish to test whether they are meeting the requirements by selecting a random sample of $n = 100$ thickness measurements.

5. Central bank believes that if consumer confidence is too high, the economy risks over heating. Low confidence is a warning that rcession might be on the way. In either case, the bank may choose to intervene by altering interest rates. The ideal value for the bank's chosen measure is 50. We may assume the measure is normally distributed with standard deviation 10. The bank takes a survey of 25 people. Which returned a sample mean of 54 for the index. What would you advice the bank to do? Use α = .05.
6. A manager will switch to a new technology if the production process exceeds 80 units per hour. The manager asks the company statistician to test the null hypothesis: H_0: μ = 80 against the alternative hypothesis: H_1: μ >80 If there is strong evidence to reject the null hypothesis then the new technology will be adopted. Past experience has shown that the standard deviation is 8. A data set with n = 25 for the new technology has a sample mean of:83 Does this justify adoption of the new technology?

b) Population Standard Deviation Unknown

If the population standard deviation **σ**, is unknown, then the population mean has a student's t distribution, and you will be using the t-score formula for sample means. The test statistic is very similar to that for the z-score, except that sigma has been replaced by s and z has been replaced by t. ie $t = \dfrac{\overline{\text{x}} - \mu}{s/\sqrt{n}}$

The critical value is obtained from the t-table. The degrees of freedom is n-1.

Example 1

A fertilizer mixing machine is set to give 12 kg of nitrate for every 100kg bag of fertilizer. Ten 100kg bags are examined. The percentages of nitrate are as follows: 11, 14, 13, 12, 13, 12, 13, 14, 11, 12. Is there reason to believe that the machine is defective at 5% level of significance?

Solution

Hypothesis H_0: μ = 12 H_1: μ ≠ 12   (This is a two-tail test)

σ is known so this is a t test. use the unbiased estimator ie $s = \sqrt{\dfrac{\sum(x-\overline{x})^2}{n-1}}$

Critical Region based on α = 0.05 and 9 degrees freedom

$t_{9,0.025} = 2.262$ ie reject H_0: μ = 12 if $|t_c| \geq 2.262$

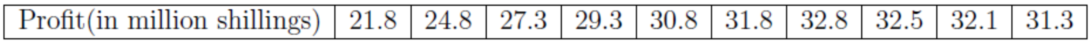
From calculator $\overline{\text{x}} = 12.5$ and $s = 1.0801$

Test statistic $t_c = \dfrac{\overline{\text{x}} - \mu}{s/\sqrt{n}} = \dfrac{12.5 - 12}{1.0801/\sqrt{10}} = 1.4639$

Decision since $|t_c| = 1.2639 < 2.262$, we fail to reject H_0 and conclude that the machine is not defective.

Example 2

The following figures give the end of year profits of ten randomly selected Chemists in Nairobi county.



Profit(in million shillings)	21.8	24.8	27.3	29.3	30.8	31.8	32.8	32.5	32.1	31.3

On the basis of this data, test whether the average profit is greater than 30M KSH at 1% level of significance

Solution

Hypothesis H$_0$: $\mu = 30$ H$_1$: $\mu > 30$ (This is a 1-tail test)

σ is known so this is a t test. use the unbiased estimator ie $s = \sqrt{\dfrac{\sum (x-\bar{x})^2}{n-1}}$

Critical Region based on $\alpha = 0.01$ and 9 degrees freedom

$t_{9,0.01} = 2.82$ ie reject H$_0$: $\mu = 12$ if $|t_c| \geq 2.82$

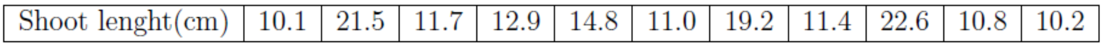
From calculator $\bar{x} = 29.415$ and $s = 3.6601$

Test statistic $t_c = \dfrac{\bar{x}-\mu}{s/\sqrt{n}} = \dfrac{29.415-30}{3.6601/\sqrt{10}} = -0.51$

Decision since $|t_c| = 0.51 < 2.82$, we don't reject H$_0$ and conclude that the average profit is not greater than 30M KSH.

Exercise

1. Identify the critical t value for each of the following tests:
   a. A two-tailed test with α=0.05 and 11 degrees of freedom
   b. A one-tailed test with α=0.01 and n=17
2. Consider a sample with $n = 20$ $\bar{x} = 8.0$ and $s = 2$ Do the following hypothesis tests.
   a) . H$_0$: μ =8.7 H$_1$: $\mu > 8.7$ at α=0.01 b) H$_0$: μ =8.7 H$_1$: $\mu \neq 8.7$ at α=0.05
3. It is widely believed that the average body temperature for healthy adults is 98.6 degrees Fahrenheit. A study was conducted a few years go to examine this belief. The body temperatures of *n* = 130 healthy adults were measured (half male and half female). The average temperature from the sample was found to be $\bar{x} = 98.249$ with a standard deviation *s* = 0.7332. Do these statistics contradict the belief that the average body temperature is 98.6? test at 1% level of significance
4. A study is to be done to determine if the cognitive ability of children living near a lead smelter is negatively impacted by increased exposure to lead. Suppose the average IQ for children in the United States is 100. From a pilot study, the mean and standard deviation were estimated to be $\bar{x} = 89$ and *s* = 14.4 respectively. Test at 5% level whether there is a negative impact.
5. The average cost of a hotel room in New York is said to be $168 per night. To determine if this is true, a random sample of 25 hotels is taken and resulted in $\bar{x} = \$172.5$ and $s = \$15.40$. Test the appropriate hypotheses at α = 0.05.



Shoot lenght(cm)	10.1	21.5	11.7	12.9	14.8	11.0	19.2	11.4	22.6	10.8	10.2

   a) An earlier study reported that the mean shoot length is 15cm.
   b) Test whether the experimental data confirms the old view at 5% level of significance.
6. A sample of eleven plants gave the following shoot lengthsA simple random sample of 14 people from a certain population gives body mass indices as shown in Table 7.2.1. Can we conclude that the BMI is not 35? Let α = .05.

subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14
BM	23	25	21	37	39	21	23	24	32	57	23	26	31	45

7. A company selling licenses for a franchise operation claims that, in the first year, the yield on an initial investment is 10%. Test the company's claim. How should a hypothesis test be stated? If there is strong evidence that the mean return on the investment is below 10% this will give a cautionary warning to a potential investor. Therefore, test the null hypothesis: H$_0$: $\mu = 10$ against

the alternative hypothesis: $H_1 : \mu < 10$ From a sample of n = 10 observations, the sample statistics are: $\bar{x} = 8.82$ and s = 2.40

8. We know the distance that an athlete can jump is normally distributed but we do not know the standard deviation. We record 15 jumps: 7.48 7.34 7.97 5.88 7.48 7.67 7.49 7.48 8.51 5.79 7.13 6.80 6.19 6.95 5.93 Test whether these values are consistent with a mean jump length of 7m. Do you have any reservations about this test?
9. The manufacturing process should give a weight of 20 ounces. Does the data show evidence that the process is operating correctly? Test the null hypothesis: H0 : μ = 20 the process is operating correctly against the alternative: H1 : μ ^{1} 20 the process is not operating correctly From the data set, the sample statistics are: n = 9, $\bar{x} = 20.356$ (ounces) and s = 0.6126

Chi-Square Distribution

The chi-square (χ^2) distribution is obtained from the values of the ratio of the sample variance and population variance multiplied by the degrees of freedom. This occurs when the population is normally distributed with population variance σ^2.

Properties of the Chi-Square

a) Chi-square is non-negative. Is the ratio of two non-negative values, therefore must be non-negative itself.
b) Chi-square is non-symmetric.
c) There are many different chi-square distributions, one for each degree of freedom.
d) The degrees of freedom when working with a single population variance is n-1.

Chi-Square Probabilities

Since the chi-square distribution isn't symmetric, the method for looking up left-tail values is different from the method for looking up right tail values.

a) Area to the right - just use the area given.
b) Area to the left - the table requires the area to the right, so subtract the right critical value given area from one and look this area up in the table.
c) Area in both tails - divide the area by two. Look up this area for the right critical value and one minus this area for the left critical value.

Test on Variance: - Chi-Square Test

The chi-square test is used to test hypothesis of the variance of a normal population, goodness of fit of the theoretical distribution to observed frequency distribution and in testing independence of attributes in a contingency table. In this unit we will focus on testing of hypothesis of the variance of a normal population only.

Assumptions

a) The population has a normal distribution
b) The data is from a random sample
c) The observations must be independent of each other

Testing is done in the same manner as before. Remember, all hypothesis testing is done under the assumption the null hypothesis is true.

Here, we have, for example: : $H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 \neq \sigma_0^2$

Critical region $\chi^2_{tab} = \chi^2_{(n-1),(1-\frac{\alpha}{2})}$ Since H_1 is two sided.

The test statistics $\chi^2_{stat} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{\sum (x-\bar{x})^2}{\sigma_0^2}$

Reject Ho i f $\chi^2_{stat} > \chi^2_{tab}$

Example

An owner of a big firm agrees to purchase the products of a factory if the produced items do not have variance of 0.5mm^2 in their length. To be sure of the specifications, the buyer selects a sample of 18 items from his lot. The length of each item was measured as follows: 18:57 18:10 18:61 18:32 18:33 18:46 18:12 18:34 18:57 18:22 18:63 18:43 18:37 18:64 18:58 18:34 18:43 18:63. On the basis of the sample data, should the buyer purchase the lot at 5% level of significance?

Solution

Hypothesis $H_0 : \sigma^2 = 0.5$ vs $H_1 : \sigma^2 > 0.5$

Critical value

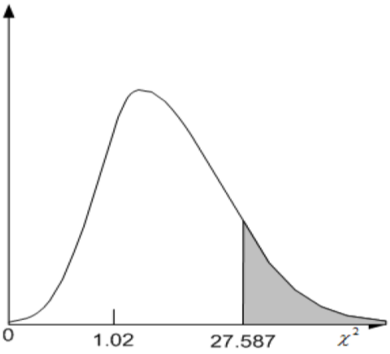
$\chi^2_{tab} = \chi^2_{(n-1),\alpha} = \chi^2_{17,0.05} = 27.587$

Test statistics χ^2_{stat}

$= \frac{(n-1)s^2}{\sigma_0^2} = \frac{17(0.03)}{0.5} = 1.02$

Since $\chi^2_{stat} < \chi^2_{tab}$ we fail to reject H$_o$

Conclusion $\sigma^2 < 0.5$ therefore the buyer should buy the products



Exercise

1. A commercial freezer must hold the selected temperature with little variation. Specifications call for a standard deviation of no more than 4 degrees (or variance of 16 degrees^2). A sample of 16 freezers is tested and yields a sample variance of $s^2 = 24$. Test to see whether the standard deviation specification is exceeded. Use $\alpha = .05$
2. You are an institutional investor evaluating a hedge fund that seeks to deliver a return comparable to the domestic broad market equity index, while keeping the monthly standard deviation in asset value under 5%. According to data in the prospectus, during the last 30 months the monthly standard deviation in asset value was 4.5%. Test this claims statistically, using a significance level of 0.10. Assume that returns are normally distributed and monthly asset values are independent observations.
3. Consider the following information about a fund. The fund has been in existence for 4 years. Over this period it has achieved a mean monthly return of 3% with a sample standard deviation of monthly returns of 5%. Test the claim that the investment disciplines of the fund result in a standard deviation of monthly returns of less than 6%.. (test at the 0.05 level of significance):
4. Response to allergen inhalation in allergic primates. In a study of 12 monkeys, the standard error of the mean for allergen inhalation was found to be .4 for one of the items studied. We wish to know if we may conclude that the population variance is not 4. Use $\alpha = 0.05$

Inferences from Two Samples

Definitions

Dependent Samples: Samples in which the subjects are paired or matched in some way. Dependent samples must have the same sample size, but it is possible to have the same sample size without being dependent.
Independent Samples: Samples which are independent when they are not related. Independent samples may or may not have the same sample size.
Pooled Estimate of the Variance: A weighted average of the two sample variances when the variances are equal. The variances are "close enough" to be considered equal, but not exactly the same, so this pooled estimate brings the two together to find the average variance.

Two-Sample Hypothesis Testing for the Mean

Two-sample hypothesis testing is statistical analysis designed to test if there is a difference between two means from two different populations. For example, a two-sample hypothesis could be used to test if there is a difference in the mean salary between male and female doctors in the New York City area. A two-sample hypothesis test could also be used to test if the mean number of defective parts produced using assembly line A is greater than the mean number of defective parts produced using assembly line B. Similar to one-sample hypothesis tests, a one-tailed or two-tailed test of the null hypothesis can be performed in two-sample hypothesis testing as well. The two-sample hypothesis test of no difference between the mean salaries of male and female doctors in the New York City area is an example of a two-tailed test. The test of whether or not the mean number of defective parts produced on assembly line A is greater than the mean number of defective parts produced on assembly line B is an example of a one-tailed test. The following section provides step-by-step instructions for performing a two-sample test of a hypothesis in Excel.

The hypothesis to be tested may be any of the following

$$\left\{ \begin{matrix} H_0 : \mu_1 \\ = \end{matrix} \right.$$

$$\left\{ \begin{matrix} H_0 : \mu_1 \\ < \end{matrix} \right.$$

$$\left\{ \begin{matrix} H_0 : \mu_1 \\ \geq \end{matrix} \right.$$

Regardless of which hypotheses used, $\mu_1 = \mu_2$ is always assumed to be true.

The sampling distribution involves two means, so it is called the *sampling distribution of the difference between means*. The value that we are interested is the difference between the means, that is: $\bar{x}_1 - \bar{x}_2$

The mean and standard deviation of the sampling distribution of the difference between means are given by: $\mu_1 - \mu_2$ and $S_{\bar{x}_1 - \bar{x}_2}$ respectively. The latter is called the *standard error of the difference between means*. Since the sample standard deviation is again used to estimate the population value, the sampling distribution of the difference between means will also be distributed as *t*. So if H_0 is true the formula becomes: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1 - \bar{x}_2}}$

All we need to do now is determine the formula for the standard error. However, this formula differs depending whether we are dealing with ***independent*** or ***dependent*** groups. With the **independent groups design**, the subjects in each of the two groups are different and unrelated in any way. For a

dependent groups design the most common type is called a *within subjects* or *repeated measures* design, because the same subjects (thus actually only one group) are tested twice.

There are two possible cases when testing two population means, the dependent case and the independent case.

a). Independent Groups

The computational formulas (which will also handle unequal sample sizes) are given by

$S_{\bar{x}_1-\bar{x}_2} = \sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ and $t = \dfrac{\bar{x}_1-\bar{x}_2}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ where $S_p^2 = \dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ is the pooled estimate of the population variance.

Since two variances are used in estimating the *standard error of the difference between means*, the degrees of freedom will equal the sum of the degrees of freedom for each of the variance estimates, that is: $df = (n_1-1)+(n_2-1) = n_1+n_2-2$

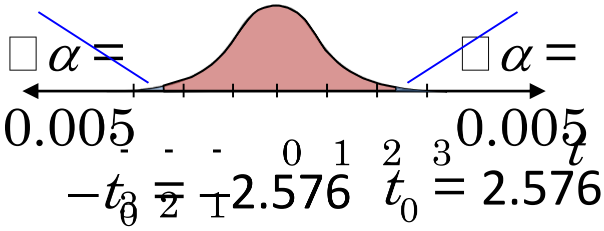
Example 1

A random sample of 17 police officers in Brownsville has a mean annual income of $35,800 and a standard deviation of $7,800. In Greensville, a random sample of 18 police officers has a mean annual income of $35,100 and a standard deviation of $7,375. Test the claim at α = 0.01 that the mean annual incomes in the two cities are not the same. Assume the population variances are equal.

Solution Hypothesis H_0: $\mu_1 = \mu_2$ vs

$H_a: \mu_1 \neq \mu_2$ (Claim)

d.f. = $n_1 + n_2 - 2 = 17 + 18 - 2 = 33$



The pooled estimate of the population variance

$$S_p = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}} = \sqrt{\frac{(17-1)7800^2+(18-1)7375^2}{17+18-2}} = 7584.0355$$

Therefore The test statistic $t = \dfrac{\bar{x}_1-\bar{x}_2}{S_p\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \dfrac{35800-35100}{7584.0355\sqrt{\left(\frac{1}{17}+\frac{1}{18}\right)}} = 0{,}273$

Decision: don’t reject H_0 and conclude that there is no sufficient evidence to support the claim the mean annual income differ.

Example 2

Big Foods Grocery has two grocery stores located in Johnston City. One store is located on First Street and the other on Main Street and each is run by a different manager. Each manager claims that her store's layout maximizes the amounts customers will purchase on impulse. Both managers surveyed a sample of their customers and asked them how much more they spent than they had planned to, in other words, how much did they spend on impulse? The following table shows the sample data collected from the two stores.

First Street	**Main Street**
15.78	15.19
17.73	18.22
10.61	15.38
15.79	15.96
14.22	21.92
13.82	12.87
13.45	12.47
12.86	13.96
10.82	13.79
12.85	13.74
	18.4
	18.57
	17.79
	10.83

Upper-level management at Big Foods Grocery wants to know if there is a difference in the mean amounts purchased on impulse at the two stores. Test at 5% whether there is a difference in the mean amounts purchased on impulse at the two stores.

Solution

Hypothesis $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$ where µ is the mean amount spent on impulse

The critical values here are based on 22 (ie 10+14-2) degrees of freedom. Ie reject H_0 if $|t| \geq 2.074$

The test statistic. $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{S_p^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

Exercise

1. In a study on Serum uric acid levels of individuals with Down's syndrome and normal individuals, the following data was obtained $n_1 = 12, \bar{x}_1 = 4.5$ *and* $S_1^2 = 1$ and $n_2 = 15, \bar{x}_2 = 3.4$ *and* $S_2^2 = 2.25$ Is there a difference between the means of individuals with Down's syndrome and normal individuals?
2. We wish to know if we may conclude, at the 95% confidence level, that smokers, in general, have greater lung damage than do non-smokers.
   Lung destructive index

	n	$\bar{x}$	S
smokers	16	17.5	4.4711
Non smokers	9	12.4	4.8492

3. These data were obtained in a study comparing persons with disabilities with persons without disabilities. A scale known as the Barriers to Health Promotion Activities for Disabled Persons

(BHADP) Scale gave the data. We wish to know if we may conclude, at the 99% confidence level, that persons with disabilities score higher than persons without disabilities.

	n	$\bar{x}$	S
Disabled	132	31.83	7.93
Non Disabled	137	25.07	4.80

4. The values below are the average July temperatures (C) in central England for 1987 to1996 and from 1997 to 2006 (from HadCETdata set).
1987 to1996 data: 15.9 14.7 18.2 16.9 17.3 16.2 15.2 18.0 18.6 16.5
1997 to 2006 data: 16.7 15.5 17.7 15.5 17.2 16.0 17.6 15.8 16.9 19.7
What conclusions do you draw from these data?
5. A stock market data base contains daily closing prices for the company Johnson & Johnson for the year 1999. For the first six months (January to June) observations are recorded for 124 trading days. For the period July to December observations are available for 128 trading days. An exercise of interest is to find a 95% confidence interval estimate for the difference between the population mean closing price in the two sample periods. Stock market prices adjust rapidly to the arrival of new information. Therefore, it is reasonable to consider that the two samples are independent. Summary statistics for the closing prices for the two sample periods are:

	n	$\bar{x}$	s^2
Jan-June	124	$89.96	32.54
July-Dec	128	$97.98	21.45

6. Within a school district, students were randomly assigned to one of two Math teachers - Mrs. Smith and Mrs. Jones. After the assignment, Mrs. Smith had 30 students, and Mrs. Jones had 25 students. At the end of the year, each class took the same standardized test. Mrs. Smith's students had an average test score of 78, with a standard deviation of 10; and Mrs. Jones' students had an average test score of 85, with a standard deviation of 15. Test the hypothesis that Mrs. Smith and Mrs. Jones are equally effective teachers. Use a 0.05 level of significance. (Assume that student performance is approximately normal.)
7. The Acme Company has developed a new battery. The engineer in charge claims that the new battery will operate continuously for *at least* 7 minutes longer than the old battery. To test the claim, the company selects a simple random sample of 100 new batteries and 100 old batteries. The old batteries run continuously for 190 minutes with a standard deviation of 20 minutes; the new batteries, 200 minutes with a standard deviation of 40 minutes. Test the engineer's claim that the new batteries run at least 7 minutes longer than the old. Use a 0.02 level of significance. (Assume that there are no outliers in either sample.)
8. From hospital records, we obtain the following values for these components:

	Treatment	Control
Average Weight	3100 g	2750 g
SD	420	425
n	75	75

With these pieces of information, test at 5% level whether the average weight of the treated sample is higher than for the control,

9. Suppose you are a researcher interested in the factors influencing paper grading by professors. You have a hunch (and/or previous research) might lead you to predict that papers that are typed

are rated higher than papers that are handwritten. Research to date though, has only been correlational and thus little can be said in terms of a cause and effect relationship.

10. Suppose you have 10 freshman students currently taking English as well as an introductory psychology course each write one paper. They should each provide two copies of their paper (one typed and one handwritten). Next, we enlist the aid of 20 English instructors. We randomly assign 10 instructors to each of two groups. Each instructor in one group (the control group) will grade each of the 10 papers that are hand written, while the second group (the experimental group) will grade the same papers that are typed. Based on the following results, does typing a paper influence the grade it receives?

	Written (1)	**W^2**	**Typed (2)**	**T^2**
Sum	**738**	**60,646**	**856**	**73,356**
N	**9**		**10**	
Mean	**82.0**		**85.6**	

b) **Dependent Groups**

The most common type of **Dependent Groups Design** is also called a *Within Subjects* or *Repeated Measures* Design, because the same subjects (thus, actually only one group) are tested twice. There is another situation, though, in which this analysis is sometimes used. It is called the *Matched Groups Design*. In this case, there are two groups, but they are matched on some variable that is highly and positively correlated with the DV.

The idea with the dependent case is to create a new variable, D, which is the difference between the paired values. You will then be testing the mean of this new variable.

Steps in paired sample Hypothesis Testing

a) State the claim mathematically ie State H_0 and H_a.

b) Specify the level of significance and Identify the degrees of freedom.

c) Determine the critical value(s) and the rejection region(s). d.f. = $n - 1$

d) Calculate $\bar{d}$ and S_d where $\bar{d} = \frac{\sum d}{n}$ and $S_d^2 = \frac{\sum d^2 - n\bar{d}^2}{n-1}$

e) Find the standardized test statistic. $t = \frac{\bar{d} - \mu_d}{S_d / \sqrt{n}}$

f) Make a decision to reject or fail to reject the null hypothesis.

g) Interpret the decision in the context of the original claim.

Example:

A reading center claims that students will perform better on a standardized reading test after going through the reading course offered by their center. The table shows the reading scores of 6 students before and after the course. At $\alpha = 0.05$, is there enough evidence to conclude that the students' scores after the course are better than the scores before the course?

Student	1	2	3	4	5	6
Score (before)	85	96	70	76	81	78
Score (after)	88	85	89	86	92	89

Solution

Hypothesis $H_{0:} : \mu_1 = \mu_2$ vs $H_{0:} : \mu_1 > \mu_2$ $\alpha = 0.05$ and d f=6-1=5

Critical value $t_{5,0.05} = 2.015$

Test statistics

d = (score before) – (score after)

Student	1	2	3	4	5	6	
Score (before)	85	96	70	76	81	78	
Score (after)	88	85	89	86	92	89	total
d	-3	11	-19	-10	-11	-11	-43

From the calculator $\overline{d} = -7.167$ and $S_d = 10.245$ implying $t_c = \dfrac{7.167 - 0}{10.245/\sqrt{6}} = -1.714$

Decision: Since $t_c = 1.714 < 2.015$ we fail to reject $H_{0:}$ and conclude that there is no sufficient evidence at the 5% level to support the claim that the students' scores after the course are better than the scores before the course.

Exercise

1. Table gives B (before) and A (after) treatment data for obese female patients in a weight-loss program.

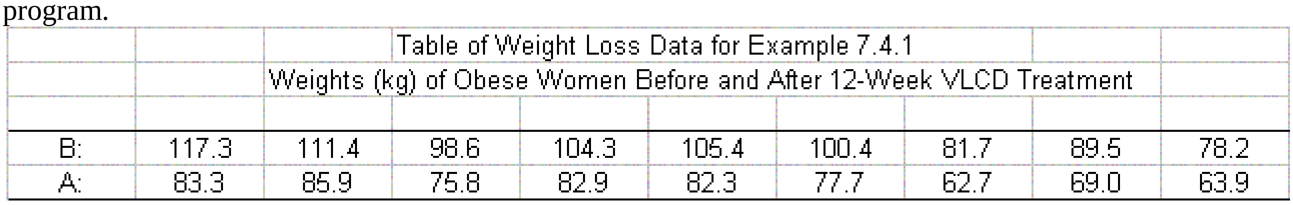


			Table of Weight Loss Data for Example 7.4.1						
		Weights (kg) of Obese Women Before and After 12-Week VLCD Treatment							
B:	117.3	111.4	98.6	104.3	105.4	100.4	81.7	89.5	78.2
A:	83.3	85.9	75.8	82.9	82.3	77.7	62.7	69.0	63.9

Calculate d= A-B for each pair of data. Is the treatment effective in causing weight reduction in these people. Test at 95% confidence level.

2. Suppose you are interested in reactions times to different coloured lights (especially green and red). If a random sample of 10 subjects was tested and gave the information below Test at 5% level whether there is a difference in reaction time for X and Y.

subject	1	2	3	4	5	6	7	8	9	10
X(Red)	18	16	23	30	32	30	31	25	27	21
Y(Green)	22	20	29	35	27	29	33	29	31	24

3 After implementing a series of flawed policies, a senior politician's approval rating is at an all time low. He commissions a market research company to carry out a survey of 120 people to determine their attitude toward him. He then goes on a charm offensive (being nice to old people, conspicuously enjoying the national sport, etc.). The survey is then repeated on the same group of 120 people. We calculate sample mean and variance for the differences. $\overline{d} = 1.24$ and $S_d^2 = 110$. What conclusions do you draw from these results?