A person posing for the camera

Description automatically generated

**Statistical Methods &**

**Data Analysis Summary**

**Ott & Longnecker**

# Contents

[Chapter 1 – Statistics and the Scientific Method 3](#_Toc37415191)

[Chapter 2- Using Surveys and Experimental Studies to Gather data 4](#_Toc37415192)

[Chapter 3 – Data Description 7](#_Toc37415193)

[Purpose Graphic designs 7](#_Toc37415194)

[Histograms: 7](#_Toc37415195)

[Measures of central tendency: 8](#_Toc37415196)

[Measures of variability: 8](#_Toc37415197)

[Chapter 4 – Probability and Probability Distributions 11](#_Toc37415198)

[Conditional probability and independence 11](#_Toc37415199)

[Bayes’ Formula: 12](#_Toc37415200)

[Variables: Discrete and Continuous 12](#_Toc37415201)

[Probability distribution - Properties of Discrete Random Variables: 13](#_Toc37415202)

[Binominal experiment - characteristics: 13](#_Toc37415203)

[Poisson distribution: 13](#_Toc37415204)

[Discrete versus continuous (normal distribution): 13](#_Toc37415205)

[Random sample 14](#_Toc37415206)

[Central Limit Theorem: 14](#_Toc37415207)

# Chapter 1 – Statistics and the Scientific Method

**Statistics** is the science of learning from Data:

1. Defining the problem
2. Collecting the data
3. Summarizing the data
4. Analyzing the data, interpreting analyses and communicating results

**Population** = set of all measurements of interest to the sample collector

**Sample** = Any subset of measurements selected from the population

**Data Mining** => a process by which useful information is obtained from large datasets

**Data Collection**:

* Surveys (passive data collection)
* Experiments (active data collection, as the conditions vary)
* Existing data
* Censuses
* Government records
* Other studies
* External data

# Chapter 2- Using Surveys and Experimental Studies to Gather data

**Observational study** = information recorded by the researcher concerning the subjects without any interference with the process.

**Experimental study** = The researcher actively manipulates certain variables associated with the study, called the **explanatory variables**, and then records their effect on the **response variable** associated with the experimental subjects.

**Confounding variables** = variables not under the control of the researcher that affect the response variables and are not the explanatory variables.

**Comparative study** = two or more methods of achieving delivery methods are compared based on cost effectiveness. Alternatively, several groups are compared based on some common attribute.

**Descriptive study** = the major purpose is to characterize a population or process based on certain attributes.

### Problems to address in a study:

* **Cause-and-effect relationships** = assigning this to get associations between factors

### Types of observational studies:

* **Sample survey**: a study that provides information about a population at a particular point in time
* **Prospective study**: a study that observes a population in the present using a sample survey and proceeds to follow the subjects in the sample forward in time in order to record the occurrence of specific outcomes. (***Cohort studies***)
* **Retrospective study**: a study that observes a population in the present using a sample survey and also collects information about the subjects regarding the occurrence of specific outcomes that have already taken place. (***Case-control studies***)

### Sampling:

* **Target population** = the complete collection of objects whose description is the major goal of the study.
* **Sample** = a subset of the target population
* **Sampled population** = the complete collection of objects that have the potential of being selected in the sample = the population from which the sample is actually selected
* **Observation unit** = the object about which data are collected (f.e. a human / a sample of water from a stream).
* **Sampling unit** = the object that is actually sampled (f.e. the household where the observation unit is the human).
* **Sampling frame** = the list of sampling units (f.e. a list of household addresses in a city)

### How:

* **Simple random sampling** = consists of a selected group of *n* units in such way that each sample of *n* has the same chance of being selected.
* **Stratified random sample** = f.e. divide the population into two groups and than take a simple random sample from each group
* **Ratio estimation** = they not only use measurements on the response of interest but also incorporate measurements on an auxiliary variable.
* **Cluster sampling**  = groups of individuals (f.e. urban areas / city blocks).
* **Systematic sampling** = taking f.e. the 1st, 10th, 20th, etc from a list.

### Problems:

* **Survey nonresponse** = when a portion of the individuals samplesd cannot or will not participate in the survey
  + **Solutions:**
    - Offerering an inducement / incentive
    - Sending reminders / follow-up calls
    - Using statistical techniques to adjust the findings
* **Measurement problems** = when the respondents do not provide the information that the survey seeks.
  + **Solutions:**
    - Make sure the definitions and words in the questions are clear
    - Do not ask leading questions
    - Make sure the interviewee can recall an answer / answer via written records / make the period to recall smaller (f.e. do not ask how often they went to the park last year but for example ask how often they go in a week / shorter period).

### Experimental study:

* **Factors** = Controlled variables
* **Measurements / observations** = Response variables
* **Treatments** = the conditions constructed from the factors
* **Factorial treatment design**  = the treatments are formed by combining levels of the factors.
* **Fractional factorial treatment structure** = only a fraction of the possible treatments are actually used in the experiment.
* **Control treatment** = the benchmark to which effectiveness of each remaining treatment is compared. The benchmark can be:
  + No treatment (f.e. an area without treatment)
  + Placebo control (f.e. treatment with a placebo medicin)
  + Standard method treatment (f.e. the existing method as a benchmark)
* **Experimental unit**  = the physical entity to which the treatment is randomly assigned or the subject that is randomly selected from one of the treatment populations.
* **Replications** = the number of times the treatment is assigned to an experimental unit in a test.
* **Measurement unit** = the physical entity upon which a measurement is taken (f.e. individuals in a group if it was a cluster sample).
* **Experimental error** = used to describe the variations in the responses among the experimental units, reason it is not 0:
  + The natural differences in the experimental units prior to the treatment
  + The variation in the devices that record the measurements
  + The variation in setting the treatment conditions
  + The effect on the response variables of all extraneous factors (other that the treatment factors).
* **Controlling experimental error - sources:**
  + The procedures under which the experiment is conducted (if not followed in a precise manner, the result is in increase in the variance of the response variable)
  + The choice of experimental units and measurement units
  + The procedure by which measurements are taken and recorded
  + The blocking of the experimental units
  + The type of experimental design
  + The use of ancillary variables (**covariates**)

### Designs (experiments)

* Completely randomized design (all variables randomly assigned)
  + Randomized block design (the treatments assigned to all subjects equally).
  + Latin square design (a design having two blocking variables)

# Chapter 3 – Data Description

### Difference between descriptive and inferential statistics:

* Descriptive is mostly for the entire population, inferential is for a sample of the population.
* Good descriptive statistics enable us to make sense of data by reducing a large set of measurements to a few summary measures that provide a good, rough picture of the original measurements.

### Methods for describing:

* Graphical techniques (pie chart, bar chart, histogram, table, etc)
* Numerical descriptive techniques

### Error and probability:

* **Sampling error**: error due to the sample taken from the population, as the actual outcome may be different than from the sample.
* If the data are organized into categories, it’s important to define the categories so that a measurement can be placed into only one category! (f.e. 2.000-4.000 + 4.000-6.000 is less clear than 2.000-3.999 + 4.000-6.000)
* **Probability**: the chance that te selected measurement lies in a particular interval is equal to the fraction of the total number of sampling measurements falling in that interval.

## Purpose Graphic designs

* Bar charts and pie charts are used to display frequency data from qualitative variables
* Histograms are appropriate for displaying frequency data for quantitative variables. Be aware that the fraction of the total number of measurements in an interval is equal to the fraction of the total area under the histogram over the interval.
* Histograms are most useful for describing data sets when the number of data points is fairly large (50+)
* Time series are useful for examining general trends and seasonal or cyclic patterns.

## Histograms

* **Unimodal =** a histogram with one major peak
* **Bimodal =** a histogram with two major peaks (mostly indicates that the sampled data are from two distinct populations)
* **Uniform =** when every interval has essentially the same number of observations
* **Symmetric:** if the shape of the right and left sides are essentially the same
* **Right Skewed:** when the right side of the histogram, containing the larger half of the observations in the data extends a greater distance that the left side
* **Left Skewed:** when the left side externs a much larger distance than the right side

### Specific designs:

* **Stem-and-leaf plot**: for constructing a histogramlike picture of a frequency distribution, show the range of scores, where the scores are concentrated, the shape of the distribution, whether there are any specific values or scores not represented and whether there are any stray or extreme scores. It splits each score /value into two sets of digits, the first is the stem, the second the leaf.
* **Time series**: a pictorial method of presenting changes in a variable over time.

## Measures of central tendency

* **Mode:** the mode is the measurement that occurs most often (with the highest frequency) 🡪 used as a measure of popularity
* **Median:** the median is the middle value when the measurements are arranged from lowest to highest. The median for an even number of measurements is the average of the two middle values. 🡪 used to measure the midpoint of a large set of measurements.
* **The median for grouped data:** this is more difficult to compute, as the actual values of the measurements are unknown:

*L* = lower class limit of the interval that contains the median

*n* = total frequency

*cfb*= the sum of frequencies (cumulative frequency) for all classes before the median class

*fm =* frequency of the class interval containing the median

*w =* interval width

**so the median =** *L + w/fm (.5n – cfb)*

* **Mean:** the sum of the measurements divided by the total number of measurements (average). The population mean is ‘mu’ (**µ**) and the sample mean is ‘y-bar’ (**ȳ**).
* **Trimmed mean:** drops the highest and lowest extreme values and averages the rest. F.e. a 5% trimmed mean drops the lowest and highest 5% measurements, a 10% trimmed mean drops the highest and lowest 10% of the measurements. Used to reduce the impact of very extreme (large or small) values on the mean and thus get a more reliable measure of the central value of the set.
* **Central tendency**: to describe the center of the distribution of measurements (is the median)

Parameters vs. statistics: parameters are numerical descriptive measurements for a **population** whereas statistics are numerical descriptive measurements for a **sample.**

## Measures of variability

* **Variability:** How the measurements vary about the center of the distribution
* **Range:** the difference between the largest and smallest measurements of the set. For grouped data, the range is the difference between the upper limit of the last interval and the lower limit of the first interval. (sensitive to outliers)
* **Percentiles:** The *p*th percentile of a set of *n* measurements arranged in order of the magnitude is that value that has at most *p*% of the measurements below it and at most (100- *p*)% above it.
  + Specific percentiles of interest are the 25th, 50th and 75th percentiles, often called the lower quartile, middle quartile (median) and the upper quartile.
  + The *i*th ordered observation, *y(i)* corresponds to the 100(*i* - .5)/*n* percentile
  + **Quantile** Q(u): a number that divides the sample of *n* data values into two groups so that the specified fraction *u* of the data values is less than or equal to the value of the quantile, Q(u).
  + **Grouped Data formula:** this formula can be used to approximate the percentiles for the original data, where:
    - *P =* percentile of interest
    - *L =* lower limit of the class interval that includes the percentile of interest
    - *n =* total frequency
    - *cfb =* cumulative frequency for all class intervals before the percentile class
    - *fp =* frequency of the class interval that includes the percentile of interest
    - *w =* interval width
    - *formula = P = L + w/fp (Pn – cfb)*
* **Interquartile range (IQR):** is the difference between the upper and lower quartiles
  + IQR = 75th percentile – 25th percentile.
  + Ignores the extremes in a dataset completely, measures only the distance to cover the middle 50% of the data. Used in boxplots.
* **Deviation:** *yi - ȳ* of a measurement *yi* from the mean *ȳ* of the set of measurements.
* **Variance:** the variance of a set of *n* measurements *y1, y2, …, yn* with mean *ȳ* is the sum of the squared deviations divided by *n-1:*



* + The symbol s2 represents the sample variance and the symbol σ2 represents the population variance
* **Standard deviation:** defined to be the positive square root of the variance
  + The symbol s represents the sample standard deviation and the symbol σ represents the population standard deviation
* **Median absolute deviation**: the MAD of a set of *n* measurements *y1, y2, …, yn*  with median *ỹ* is the median of the absolute deviations of the *n* measurements about the median:
  + **MAD =** *median {| y1 - ỹ |, | y2 - ỹ |, …., | yn- ỹ |}/.6745*

### Summary values for minimal description

In most datasets we need a **minimum of five summary values** to provide a minimal description of the dataset:

* Smallest value = *y(1)*
* Lower quartile = *Q(.25)*
* Median = middle value
* Upper quartile = *Q(.75)*
* Largest value = *y(n)*

### Empirical rule

Given a set of *n* measurements possessing a mound-shaped histogram, then

* The interval *ȳ s* contains approximately 68% of the measurements
* The interval *ȳ 2s* contains approximately 95% of the measurements
* The interval *ȳ 3s* contains approximately 99.7% of the measurements

### Exploratory data analysis (EDA)

* **Box-and-Whiskers plot** (boxplot), where:
  + ***Lower inner fence***: *Q1 –1,5\*(IQR) 🡪 Q1* = first Quartile and IQR isInterquartile range
  + ***Upper inner fence***: *Q3 +1,5\*(IQR) 🡪 Q3* = third Quartile
  + ***Lower outer fence***: *Q1 –3\*(IQR)*
  + ***Upper outer fence***: *Q3 +3\*(IQR)*
    - ***Mild outlier*** = any data value beyond the inner fence on either side - marked
    - ***Extreme outlier*** = any data value beyond an outer fence on either side – “
    - ***Lower adjacent value*** = the smallest data value that is not an outlier
    - ***Upper adjacent value*** = the largest data value that is not an outlier

### Mutiple variables – graphs and correlation:

* **Contingency table**: the rows of the table identify the categories of one variable, the columns identify the categories of the other variable.
* Stacked bar graph
* Clustered bar graph
* Scatterplot
* **Correlation coefficient:** measures the strength of the linear relationship between two quantitative variables. The correlation coefficient is usually denoted as *r*.
  + A positive value for *r* indicates a positive association between the two variables, and a negative value for *r* indicates a negative association between the two variables.
  + The value of *r* is a number between -1 and +1. When the value of *r* is very close to 1, the points in the scatterplot will lie close to a straight line
  + Because the two variables are standardized in the calculation of *r*, the value of *r* does not change if we alter the units of *x* or y*.* The same value of *r* will be obtained no matter what units are used for *x* and *y.* Correlation is a unit-free measure of association.
  + Correlation measures the degree of the straight-line relationship between two variables. The correlation coefficient does *not* describe the closeness of the points (*x,*y) to a curved relationship, no matter how strong the relationship. Unless *r* I greater than .6, there is very little trend in the scatterplot.

# Chapter 4 – Probability and Probability Distributions

**Probability** = the language of uncertainty, the mechanism for making inferences from sample results that are used to draw conclusions about populations.

Three definitions for the probability of an outcome:

* **Classical interpretation of probability:** Arose from games of chance (50% chance a coin flips ‘head’). Each possible distinct result is called an ***outcome***. An ***event*** is identified as a collection of outcomes.
  + The probability of an event (*E*) under the classical interpretation of probability is computed by taking the ratio of the number of outcomes (*Ne*) favorable to event *E* to the total number of possible outcomes, *N*
  + 
  + Assumption: All outcomes are equally likely
* **Relative frequency concept of probability:** This is an empirical approach to probability. IF an experiment is repeated a large number of times (*n)* and event *E* occurs on *ne (..%)* of these trials, then the probability of event *E* is approximately:
  + 
  + This is the most reasonable
* **Personal or subjective probability**: used for a one-shot statement. (there are no repetitions of an experiment possible). Problem: they vary from person to person and can not be checked / verified.

**Simulation:** a technique that produces outcomes having the same probability of occurrence as the real situation events.

### Probability laws and event relations:

* The probability of an event (*A)* lies anywhere in the interval from 0 (the occurrence of event *A* is impossible) to 1 (the occurrence of the event *A* is a ‘sure thing’). The formula:
  + 0
* **Mutually exclusive:** Two events (*A* and *B*) are said to be mutually exclusive if (when the experiment is performed a single time) the occurrence of one of the events excludes the possibility of the occurrence of the other event.
  + **Probability**: *P(*either *A* or *B*) = *P(A) + P(B)*
* **Complement:** the complement of an event *A* is the event that *A* **DOES NOT** occur.
  + The complement of *A* is denoted by the symbol *Ā*
  + It occurs when *A* does not
  + *P(A) + P(Ā) = 1* (the probability that *A* occurs+the probability that *A* doesn’t occur =1)
* **Union:** the union of two events (*A & B)* is the set of all outcomes that are included in either *A* or *B*.
  + The union is denoted as
  + It occurs when either *A* or *B* occurs
  + = probability of union
* **Intersection:** the intersection of two events (*A & B)* is the set of all outcomes that are included in both *A* or *B* (or both).
  + The intersection is denoted as
  + It occurs when *A* and *B* occur.

## Conditional probability and independence

* **Conditional probability:** Consider two events *A* and *B* with nonzero probabilities *P(A)* and *P(B).*
  + The conditional probability of event *A,* given event *B*, is:
  + The conditional probability of event *B,* given event *A,* is:
* **Unconditional or marginal probability:** this is the proportion of the times event *A* occurs in very large(infinitely large) number of repetitions of the experiment.
* **The probability of the intersection:** of two events *A* and *B* is:
* **Independent events:** two events *A* and *B* are independent events if the occurrence of event *A* is not dependent of the occurrence of event *B* (most samples from the population will be independent samples):
  + or
* **Dependent events:** when the occurrence of *A* depends on the occurrence of *B*.

## Bayes’ Formula

Bayes’ Formula describes the probability of an event, based on prior knowledge of conditions that might be related to the event. The probability of *E* being present in the samples when the test yields a positive result *P(E*|Pos).

* **Bayes’ Formula**: if *A* and *B* are any events whose probabilities are not 0 or 1, then:
* **False positive**: a positive result in a test where the presence is incorrect
* **False negative**:a negative result in a test where the absence is incorrect
* **Sensitivity**:The True positive rate *P(test is positive|presence is correct)*
* **Specificity:** The True negative rate *P(test is negative|absence is correct)*

## Variables: Discrete and Continuous

* **Qualitative random variable**: There are a finite (an typically quite small) number of possible outcomes associated with this variable. For example: no accident, minor accident or major accident. These are categories and they are not different in any measurable numerical amount.
* **Quantitative random variable:** the responses vary in numerical magnitude = **random variable** (we can do so much with this, f.e. we can average the resulting quantities, find standard deviations and asses probable errors). This is denoted by the symbol and we are interested in the values that can assume (*numerical outcomes)*.
* **Discrete random variable:** when observations on a quantitative random variable can assume only a countable number of values. Examples:
  + Change in the number of accidents per month at an intersection after a new signaling device has been installed
  + Number of ‘dead persons’ voting in the last mayoral election in a major midwestern city.
* **Continuous random variable:** when observations on a quantitative random variable can assume any one of the uncountable number of values in a line interval. Examples:
  + Temperature: f.e. the daily maximum temperature in a place.
  + Pressure
  + Height
  + Weight
  + Distance
* The distinction between discrete and continuous random variables is pertinent when we are seeking the probabilities associated with specific values of a random variable.

## Probability distribution - Properties of Discrete Random Variables

* The probability associated with every value of y lies between 0 and 1.
* The sum of the probabilities for all values of y is equal to 1.
* The probabilities for a discrete random variable are additive. Hence, the probability that y = 1 or 2 is equal to P(1) + P(2)

## Binominal experiment - characteristics

* The experiment consists of *n* identical trials
* Each trial results in one of two outcomes. We will label one outcome a success and the other a failure.
* The probability of success on a single trial is equal to and remains the same from trial to trial.
* The trials are independent: that is, the outcome of one trial does not influence the outcome of another trial
* The random variable *y* is the number of successes observed during the *n* trials.
* Formula:
  + *n* = number of trials
  + = probability of success on a single trial
  + = probability of failure on a single trial
  + = number of successes in *n* trials
* = *n* factorial, if *n* = 4, than 4! = (4)(3)(2)(1) = 24. Note that 0! is equal to 1.
* Mean of a Binominal Probability Distribution: (n is number of trials, π is the probability of success in a given trial.)
* Standard deviation of a Binominal Probability Distribution: (n is number of trials, π is the probability of success in a given trial.)

## Poisson distribution

* y = the number of events occurring during a fixed time interval or length *t* or a fixed region *R* or area or volume *m(R)*.
* Events occur one at a time: Two or more events do not occur precisely at the same time or in the same space
* The occurrence of an event in a given period of time or region of space is independent of the occurrence of the event in a nonoverlapping time period or region of space. The (non)occurrence doesn’t affect the probability of an other event.
* The expected number of events during one period or in one region, µ, is the same as the expected number of events in any other period or region.
* Formula: , where *e* is a naturally occurring constant approximately equal to 2.71828.

## Discrete versus continuous (normal distribution)

* Discrete random variables (such as binomial) have possible values that are distinct and separate, such as 0, 1, or 2.
* Continuous random variables form a whole interval (or range, or continuum) (example: the 1-year return per dollar invested in a common stock could range from 0 to some quite large value). It is one that can assume values associated with infinitely many points in a line interval.
  + The possibility that a continuous variable falls in an interval, between points *a* and *b* is:
  + **Normal curve**: the relative frequency histogram for the normal random variable is a smooth, bell-shaped curve. If *y* represents the normal random variable, than *f(y)* represents the height of the probability distribution:
    - Normal distribution: , where µ and σ are the mean and standard deviation of the population of y-values.
  + **Area under a normal curve / z-score:** to determine the probability that a measurement will be less than some value *y*, we first calculate the number of standard deviations that *y* lies away from the mean by using the formula:
    - and the probability is calculated in R using ‘pnorm(z)’
  + An important aspect of the normal distribution is that we can easily find the percentiles of the distribution:
    - **The 100*p*th percentile:** of a distribution is that value, , such that 100*p*% of the populations fall below and 100(1-*p*)% are above .
    - Formula: , calculated in R using ‘qnorm(z,µ,σ)’

## Random sample

* A sample of *n* measurements selected from a population is said to be a random sample if every different sample of size *n* from the population has an equal probability of being selected.
* Random number tables are constructed in such a way that no matter where you start in the table and no matter in which direction you move, the digits occur randomly and with equal probability.
* Code in R: sample(seq(I:N), n, replace = False)
* Code in Python: numpy.random.choice(a, size=None, replace=True, p=None). The p argument corresponds to the prob argument in the sample()function. Or use pandas: df['a'].sample(4, replace=True, weights=df['b']), where the prob argument in R's sample() corresponds to weights in the pandas methods.
* **Sampling distribution** = the probability distribution of a sample statistic.
* **Standard error of ȳ** = This is

## Central Limit Theorem

Provide theoretical justification for approximating the true sampling distribution of many sample statistics with the normal distribution. This is very important. The minimum sample size for the central limit theorem to hold is *n* > 30, unless heavily skewed.

We subset two different approaches:

#### Central Limit Theorem for ȳ

Let ȳ denote the sample mean computed from a random sample of *n* measurements from a population having a mean and a finite standard deviation . Let denote the mean and standard deviation of the sampling distribution of ȳ, respectively. Based on repeated random samples of size *n* from the population, we can conclude the following:

1. =
2. =
3. When *n* is large, the sampling distribution of ȳ will be approximately normal (with the approximation becoming more precise as *n* increases)
4. When the population distribution is normal, the sampling distribution of ȳ is exactly normal for any sample size *n.*

#### Central Limit Theorem for Σy

Let Σy denote the sum of a random sample of *n* measurements from a population having a mean and finite standard deviation . Let denote the mean and standard deviation of the sampling distribution of Σy, respectively. Based on repeated random samples of size *n* from the population, we can conclude the following:

1. =
2. =
3. When *n* is large, the sampling distribution of Σy will be approximately normal (with the approximation becoming more precise as *n* increases)
4. When the population distribution is normal, the sampling distribution of Σy is exactly normal for any sample size *n.*

Sampling distributions can be **interpreted** in at least 2 ways:

1. Using the long-run relative frequency approach