Class notes 1

Data analysis

Population vs. sample

The foundation of statistical analysis is the distinction between the underlying population and the sample.

Definitions

Population: The complete set of objects under study.

Sample: Any subset of the population.

When performing statistical analysis, one must make decisions about what statistical procedures that are appropriate for the research project. The choice of modeling framework depends on the goal of the analysis and the formulation of the hypotheses to be evaluated. Obviously, to make such decisions, we must think through the type of data needed, and how the data can be obtained. Also, keep in mind that the task of data collection (sampling) is often cumbersome and costly. Thus, the amount of data available to the analyst is often limited.

Sampling

Statistical inference makes use of information from a sample to draw conclusions (inferences) about the population from which the sample was taken. The main objective when collecting the data is to obtain a sample that reflects the population. It is additionally required that objects (or individuals) are selected independently. That is, the selection of one object from the population does not influence the selection of another object. The way to achieve this is by randomly select the objects entering the sample. Throughout the sampling process, each (remaining) object must have an equal chance of being selected.

One distinguish between sampling *with replacement* and sampling *without replacement.* When sampling with replacement is performed, the same object may appear several times in the final sample. This type of sampling is not commonplace in practice. For instance, when performing a survey, we do not distribute the same questionnaire to the same individual more than one time. One can show mathematically that sampling without replacement violates the independence requirement. Fortunately, the violation may be inconsequential if the population is reasonably large. The sampling scheme just described is typically referred to as *Simple random sampling*.

In what follows, whenever the term *random sample* is used, we can assume independence among the objects and that the objects in the population have equal chance to be chosen when the sample is obtained.

Variables and data types

Next, we discuss different types of variables and their measurement scales. We consider four levels of measurement scales

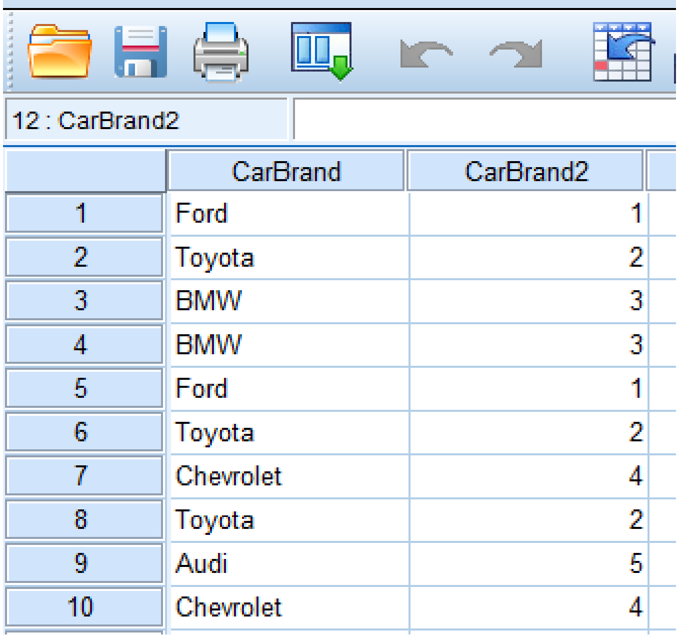
1. Nominal scale (lowest level)
2. Ordinal scale

3. Interval scale
4. Ratio scale (highest level)

The first and second level are referred to as *non-parametric*, whereas the third and fourth level are referred to as *parametric*. The distinction between non-parametric and parametric measurement levels follows from the presentation below.

Nominal scale

The purpose of nominal data is to classify (or categorize) objects. The classification is done by labeling the objects in the data. Note that data on a nominal scale does not have a natural ordering. Common examples of nominal data are *Gender*, *Nationality*, *Company*, *Sector*, *Brands names* etc. Nominal data are often referred to as qualitative data, since this type of data can be used to describe various qualities or traits associated with the objects. In the following example, we have collected some data on car brands.



12 : CarBrand2

	CarBrand	CarBrand2
1	Ford	1
2	Toyota	2
3	BMW	3
4	BMW	3
5	Ford	1
6	Toyota	2
7	Chevrolet	4
8	Toyota	2
9	Audi	5
10	Chevrolet	4

Note that *CarBrand* does not take any numerical values. We may refer to *CarBrand* as a *string* variable, since its "values" are just strings of text. Quite often, one assigns a specific value to each category in the data, as is done for *CarBrand2*. However, this does not really change anything. For instance, it would not make sense to perform any form of computation using the numbers of *Carbrand2*. We can only use *CarBrand2* for classification purposes.

Ordinal scale

In contrast to nominal data, ordinal data have a natural ordering. This allows us to rank the objects. However, there is no guarantee that the interval between the points on the scale remains the same for all points. To make this clearer, consider the case of a 5-point scale where each point on the scale is coded using the numbers 1 – 5. Then, the difference between points 2 and 3 may not be the same as the difference between points 4 and 5. Although, commonly done, it is incorrect to apply standard parametric statistical techniques (such as computing sample means, variances and correlations) to this type of data. We discuss this in more detail below.

Interval scale

As in the previous case, data measured on an interval scale have a natural order. The distinction is that the interval between the points remains constant for all points. For instance, the difference between the points 2 and 3 on the scale is the same as the difference between points 6 and 7. Data measured on this level allows the researcher to apply parametric statistical techniques.

Note that an interval scale does not have an absolute zero point. If the scale in question contains a zero, its location may be somewhat arbitrary. For instance, consider the centigrade scale. We know that the zero point represent the temperature at which water freezes. However, the zero point on the scale does not imply the absence of temperature.

Ratio scale

Data measured on a ratio scale have a well-defined zero point. Variables such as *Income*, *market share*, *Height*, *Weight*, *Length* etc. belong to this type of data. As opposed to the points on an interval scale, it makes sense to interpret the ratio of two points on the scale. For instance, if we consider the length of a rod measured in centimeters. Then, we know that a 30 cm rod is three times the length of a 10 cm rod, since 30cm/10cm = 3. Again, data measured on this level allows the researcher to apply parametric statistical techniques.

Continuous versus discrete data

Not only the measurement scale is important, we also have to consider whether the data is continuous (also called metric) or discrete. A continuous random variable can take any value (with any number of decimals) within its range. A discrete random variable can only take specific values (for instance the number of dots obtained from rolling a dice). Discrete variables often come in the form of counts. For instance, consider an insurance company. Before a quote for car insurance can be made, the company would like to know how many (or the number of) previous car accidents the car owner has previously been involved in. The number of car accidents would then take the values 0, 1, 2, …

We note that variables measured on a nominal or ordinal scale are categorical and can therefore not be continuous.

What is important?

The important distinction is between nominal/ordinal data and interval/ratio data. The distinction is important since it dictates which statistical procedures that are available to the researcher. As indicated, data on an interval/ratio scale are more informative, and hence offers the researcher better possibility to investigate the data. It is not surprising that applied researchers sometimes treat ordinal data as being measured on an interval scale. Doing so, allows the researcher (conveniently) to apply parametric statistical techniques.

An interesting case is the *Likert* scale, which is a scale that is popular in the social sciences. The use of Likert scale variables is commonplace when designing a questionnaire. A respondent taken a questionnaire may be asked to evaluate a series of questions on a scale from $1-5$, where

1 : Strongly disagree

2 : Disagree

3 : Neither agree nor disagree

4 : Agree

5 : Strongly agree

In the strictest sense, Likert scale variables are ordinal. To better accommodate an interval scale, it is common to allow for a wider range of responses in the questionnaire. A wider range of responses could for instance be a scale from 1 – 9. Even with the wider range scale, the practice of treating the data as if it were measured on an interval scale remains questionable.

There are specialized statistical techniques that allows the researcher to handle ordinal data. These techniques include non-parametric statistics and various types of regression designed specifically for the purpose.

It is often useful to combine data of different scales. For instance, consider the case when interval/ratio data are combined with nominal data. Such combination of data allows the researcher to perform statistical comparisons across groups of objects. ANOVA analysis is a statistical technique developed for such a purpose.

Summarizing the data

A statistical analysis often starts by computing various sample statistics to summarize the information in the data. Sample statistics, also referred to as descriptive statistics, are powerful tools since computing such statistics allows the analyst to get an overview of the data quickly.

Measures of center

Common measures of center are the *sample mean* (or the average) and the *median*. When summarizing the data, we are interested in knowing what is a "typical" value of the variable being studied. The answer is found by computing the sample mean or the median.

Sample mean: let $x_1, x_2, \dots, x_n$ be a data sequence consisting of n element (sample size is n). The sample mean is then given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median: the median, typically denoted $\tilde{x}$, is found by first writing the data sequence in ascending order (from smallest to largest). The median is then the middle number in the ordered sequence. This definition applies if the number of elements n is an odd number. If n is an even number, the median is the average of the two middle numbers. Formally, this is stated in the following way:

Let $x_1, x_2, \dots, x_n$ be an ordered data sequence. Then,

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \dfrac{x_{n/2} + x_{(n/2)+1}}{2} & \text{if } n \text{ is even} \end{cases}$$

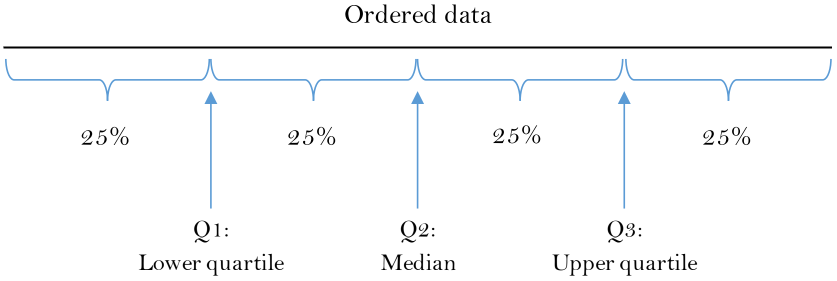
where the subscript refers to the position in the ordered sequence.

Other location measures

While the sample mean and the median are referred to as measures of center, they are also location measures. Other measures of location are the *mode* and *quartiles.*

Mode: the mode is simply the most frequent value in the data sequence. By its definition, the mode is only (or almost only) relevant when the data is discrete.

Quartiles: like the median, quartiles are measures based on an ordered sequence. In fact, the second quartile is the median. The following diagram indicates the position of the quartiles in an ordered sequence:



There are quite many procedures for how to obtain the lower and upper quartiles (Q1 and Q3). Here, we will use the so-called Tukey procedure, which works in two steps

1. Split the data into two sub-sequences at the median. If n is odd, include the median in each sub-sequence.
2. Find the medians of the two sub-sequences. The two medians are then Q1 and Q3.

Measures of spread

The most important measures of spread are the sample variance and the standard deviation. As before, let $x_1, x_2, \dots, x_n$ be a data sequence of length n.

The sample variance is then given by the expression

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

and the standard deviation is simply the square root of the variance

$$s_X = \sqrt{s_X^2}$$

A third measure of spread is the *Interquartile range* (or simply IQR), which is defined as

$$IQR = Q_3 - Q_1$$

Measures of co-variation (two-variable analysis)

Working with two variables, we might be interested in their relationship. Important measures of co-variation are the covariance and the correlation. Note that covariance and correlation are restricted to measure only *linear dependence* between two variables. Thus, nonlinear dependence between the two variables is not be captured by these two measures.

let $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ be a sample of n pairs. The sample covariance is then found by

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

A negative (positive) value of s_{XY} indicates a negative (positive) (linear) relationship between X and Y. A value of zero indicates no (linear) relationship.

The problem with the sample covariance as a descriptive measure is that it is a scale dependent. Thus, it becomes difficult to interpret the degree covariation purely based on the value of s_{XY}. To avoid this problem, one can compute the correlation, which is a scale free measure of covariation. The correlation between X and Y is given by the expression

$$r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y}$$

Note that

$$-1 \leq r_{XY} \leq 1$$

where r_{XY} close to -1 indicates a near perfect negative relationship, r_{XY} close to 1 indicates a near perfect positive relationship and $r_{XY} = 0$ indicates no relationship.

Sample statistics versus population parameters

For any sample statistic there exist a population counterpart. In general, values obtained from a sample are known as *sample statistics* (or just statistics) and values associated with the population are known as *parameters*. For instance, consider $\bar{x}$ and its population counterpart μ. The sample mean $\bar{x}$ is obtained from the realized sample whereas the parameter μ is the population mean.

Estimation

In statistics, a distinction between an *estimator* and an *estimate* is made. An estimator is a function (a mathematical expression) of the sample. An estimate is the result from the actual application of the function to a particular sample. The sequence of random variables $X_1, X_2, \dots, X_n$ (the upper case Xs) is representative of the sample, whereas the sequence $x_1, x_2, \dots, x_n$ (the lower case xs) is a particular realization of the sample (simply a sequence of numbers).

Next, a few examples of estimators are provided. To this aim, let $X_1, X_2, \dots, X_n$ be a random sample of size n. We then have

Description:	**Parameter:**	**Estimator:**
Population mean	μ	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Population variance	σ^2	$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
Population std. dev.	σ	$S_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$
Population covariance	σ_{XY}	$S_{YX} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$
Population correlation	ρ_{XY}	$R_{YX} = \frac{S_{XY}}{S_X * S_Y}$

Obviously, there are many more estimators that could be added to the list.

A word on notation. As we will see later, some statistical procedures involve more than one random variable. In that case, we will expand the notation by adding subscripts to the parameters at appropriate places. For instance, if the aim is to study the difference between two means, we may use μ_1 and μ_2 to denote the population means of X_1 and X_2, respectively.

Properties of an estimator

When an estimator is applied to a random sample, the estimator itself becomes a random variable. We will be concerned about the quality of the computed estimates. It is therefore of interest to discuss the properties of an estimator. Three important properties are:

Property 1: *Consistency*: The definition of consistency is rather technical. Thus, we shall satisfy our self with a somewhat rough idea. We may say (slightly incorrect) that an estimator is consistent if the estimates obtained from it *converge* to the true value of the parameter being estimated as n tend to infinity. It follows that consistency is a *large sample* property.

Property 2: *Unbiasedness*: Refers to the property that the estimator on average produce estimates that are equal to the true population values. If we were to redo the analysis, by collecting many samples of size n from the same population, the average estimates will on average tend to the true value of the parameter being estimated. In contrast to consistency, unbiasedness is a property that does not depend on sample size.

Property 3: *Efficiency*: Relates to the variation in the estimator. The goal is to obtain estimates that are as precise as possible. It is therefore desirable to have an estimator with as low variance as possible. As an example, consider a random sample from a normal population. It can be shown than the sample mean, given by $\bar{X}$, and the sample median, given by $\tilde{X}$, are both unbiased estimators of the parameter μ. It can further be shown that the variance of $\bar{X}$ is less than the variance of $\tilde{X}$. It follows that $\bar{X}$ is a more efficient estimator of μ as compared to $\tilde{X}$.

Interval estimation

Thus far, we have discussed point estimation. For instance, when applying or $\bar{x}$, we are computing a point estimate (a single number). Another type of estimators is known as interval estimators or *confidence intervals.* In general, a confidence interval takes the general form

$$PE \ \pm ME = [PE - ME, PE + ME]$$

where PE is a point estimate and ME is the so-called *margin error.* A confidence interval is typically stated in terms of a percentage value called the *confidence level.* The confidence level expresses the probability of the true (population) parameter being contained within the bounds of the interval. Obviously, we want that probability to be large. Thus, confidence levels of 90 %, 95 % or 99 % are common. To aid the computation, it is useful to introduce the quantity α. This quantity allows us to express the confidence level as $100 \cdot (1 - \alpha)\%$. It follows that values of α equal to 0.1, 0.05 and 0.01 correspond to confidence levels of 90 %, 95 % and 99 %, respectively.

When computing a confidence interval, one is applying a procedure for which the true parameter is contained in the interval with probability $100 \cdot (1 - \alpha)\%$. That is, if we were to compute a large number of confidence intervals, based on a large number of samples (drawn from the same population), we would expect the true parameter to be contained in the intervals in $100 \cdot (1 - \alpha)\%$ of the cases. *IMPORTANT: It would be wrong to claim that there is a* $100 \cdot (1 - \alpha)\%$ *chance that the true parameter falls within the bounds of an already computed confidence interval.*

Confidence interval for the population mean

A two-sided $100 \cdot (1-\alpha)\%$ confidence interval for the population parameter μ is given by

$$\bar{x} \pm ME = [\bar{x} - ME,\ \ \bar{x} + ME]$$

where

$$\begin{aligned} ME &= t_{\alpha/2} \cdot SE(\bar{x}) \\ &= t_{\alpha/2} \cdot \frac{s_X}{\sqrt{n}} \end{aligned}$$

In this expression, $t_{\alpha/2}$ is the critical value and $s_X/\sqrt{n}$ is the realized standard error associated with $\bar{X}$. The confidence interval is now stated as

$$\bar{x} \pm t_{\alpha/2} \cdot \frac{s_X}{\sqrt{n}}$$

The quantity $t_{\alpha/2}$ is a value satisfying the expression

$$P\left(T \geq t_{\alpha/2}\right) = \frac{\alpha}{2}$$

where the degrees of freedom is $df = n - 1$.

Assumptions:

In this case, the assumptions are:

1. The sample is random.
2. Distributional assumption:
   a. *Small sample size*: if n is small, it must be assumed that the sample is drawn from a normal population (or at least approximately normal).
   b. *Large sample size*: if n is large, the CLT applies and no distributional assumption is needed.

Example (taken from http://www.stat.columbia.edu/~martin/W2024/R2.pdf)

An outbreak of Salmonella related illness was attributed to ice cream produced at a certain factory. Scientists measured the level of Salmonella in 9 randomly sampled batches of ice cream. The Salmonella levels, as measured in MPN/g, were:

0.593 0.142 0.329 0.691 0.231 0.793 0.519 0.392 0.418

For simplicity, let X denote MPN/g. Compute a 95 % confidence interval for the mean level of salmonella in the ice cream

Confidence interval for the difference in two population means (independent samples)

It may be of interest to compare two population means. Let μ_1 denote the mean of the first population and let μ_2 denote the mean of the second population. Specifically, our aim is to obtain a two-sided confidence interval for the difference in the means, here given by $\mu_1 - \mu_2$.

Let $X_{1,1}, X_{1,2}, \dots, X_{1,n_1}$ and $X_{2,1}, X_{2,2}, \dots, X_{2,n_1}$ be two independent samples. The estimator for the difference in means is $\bar{X}_1 - \bar{X}_2$. A two-sided $100 \cdot (1 - \alpha)\%$ confidence interval for the difference in means is obtained by

$$(\bar{x}_1 - \bar{x}_2) \pm ME = [(\bar{x}_1 - \bar{x}_2) - ME, (\bar{x}_1 - \bar{x}_2) + ME]$$

Now, the margin error takes the form

$$\begin{aligned} ME &= t_{\alpha/2} \cdot SE(\bar{x}_1 - \bar{x}_2) \\ &= t_{\alpha/2} \cdot \sqrt{\frac{s_{X_1}^2}{n_1} + \frac{s_{X_2}^2}{n_2}} \end{aligned}$$

where $s_{X_1}^2$ and $s_{X_2}^2$ are sample variances based on the data, and $t_{\alpha/2}$ is the critical value. We can write the confidence interval as

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \cdot \sqrt{\frac{s_{X_1}^2}{n_1} + \frac{s_{X_2}^2}{n_2}}$$

The critical value is found the same way as before, but the degrees of freedom is a more involved calculation. The following expression can be used

$$df = \frac{\left(\frac{s_{X_1}^2}{n_1} + \frac{s_{X_2}^2}{n_2}\right)^2}{\frac{\left(s_{X_1}^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_{X_2}^2/n_2\right)^2}{n_2 - 1}}$$

The result of the formula is typically a decimal number. Thus, it is necessary to round down the number to the nearest integer value.

Assumptions:

For the independent samples confidence interval for the difference in means, the assumptions are:

1. The samples are random.
2. The two groups, from which the two samples are taken, are independent.
3. Distributional assumption:
   a. *Small sample size*: if n is small, it must be assumed that the two samples are drawn from normal populations (or at least approximately normal).
   b. *Large sample size*: if n is large, the CLT applies and no distributional assumption is needed.

Example (taken from Triola)

People spend around 5 billion USD annually for the purchase of magnets used to treat a wide variety of pains. Researchers conducted a study to determine whether magnets are effective in treating back pain.

Pain was measured using the so-called visual analog scale. The results from the study are given in the following summary

Reduction in pain level after magnet treatment (treatment group):

$n_1 = 20, \;\; \bar{x}_1 = 0.49, \;\; s_{X_1} = 0.96$

Reduction in pain level after placebo treatment (control group):

$n_2 = 20, \;\; \bar{x}_2 = 0.44, \;\; s_{X_2} = 1.40$

Compute a 95 % confidence interval for the difference in mean pain reduction across the groups.