Descriptive Statistics, Sampling Distributions and Methods of Point Estimation

Polychronis Economou

BIOSTATISTICS – DATA ANALYTICS

Biomedical Engineering

Outline

Outline

Probability and Statistics

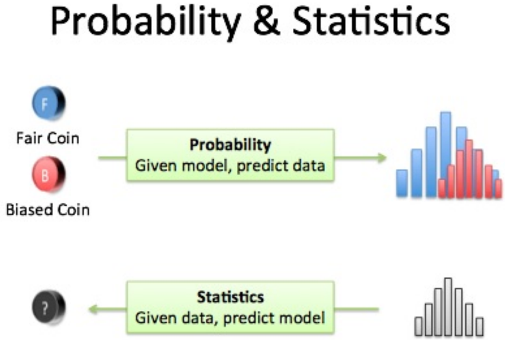


Chart taken from: `https://betterexplained.com/articles/a-brief-introduction-to-probability-statistics/`

Probability and Statistics

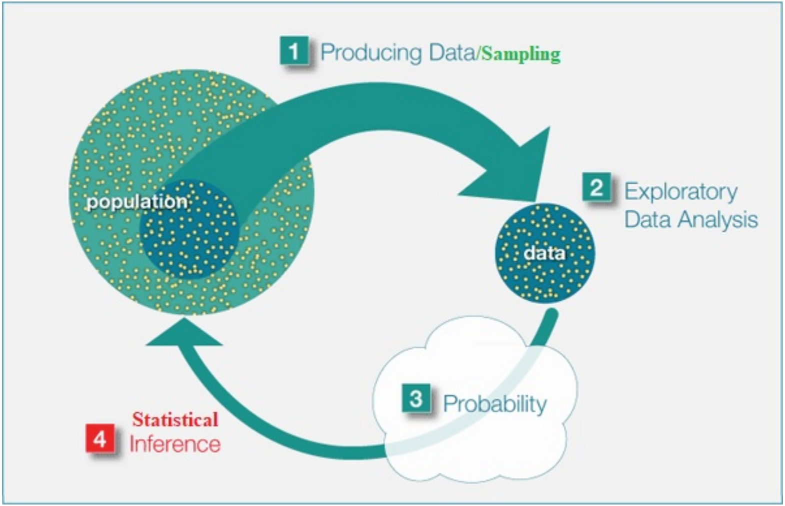


Chart taken from:
https://bolt.mph.ufl.edu/6050-6052/unit-4/

Population and sample

A **population** is a collection of people, items or events about which you want to make inferences.

Typically, the population is very large, making a census of all the units in the population impractical or impossible.

However, it is possible, to study a subset of the population. Such a subset is called "**sample**".

The process of designing and obtaining a **representative sample** of a desired population requires care.

In what follows we will assume that our sample is a **random** (representative) **sample**.

Outline

Descriptive Statistics

Descriptive Statistics are the basis of the initial description of the data and the first step of a more extensive statistical analysis.

Descriptive statistics is the term given to the analysis of data that helps describe and summarize data in a meaningful way such that, for example, patterns might emerge from the data.

Univariate analysis

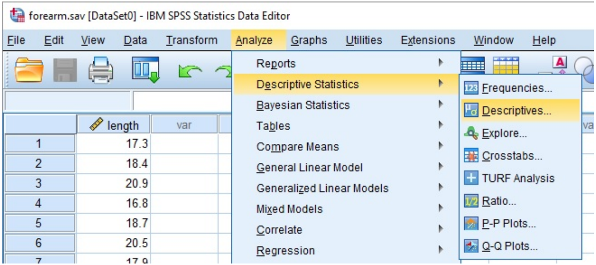
- **Central tendency**
  - mean
  - median
  - mode
  - quantiles
  - ...
- **Dispersion**
  - range
  - interquantile range
  - variance
  - standard deviation
  - ...
- **Distribution shape**
  - skewness
  - kurtosis
  - ...
- **Graphical representation**
  - bar charts
  - histograms
  - box plots
  - ...

Bivariate or Multivariate analysis

- Cross-tabulations and contingency tables
- Graphical representation (for example scatterplots)
- Quantitative measures of dependence
- Descriptions of conditional distributions

Example – Univariate case

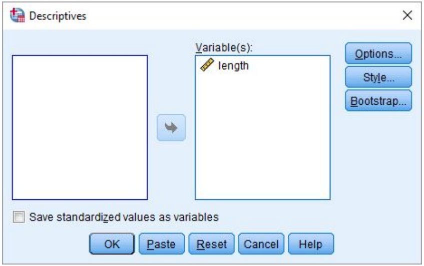
Measurements of the length of the forearm (in inches) made on 140 adult males.

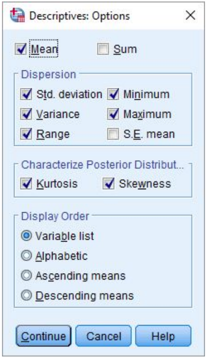


Reference: Pearson, K. and Lee, A. (1903), On the laws of inheritance in man. I. Inheritance of physical characters. Biometrika, 2, 357-462.

Example – Univariate case

Measurements of the length of the forearm (in inches) made on 140 adult males.





Example – Univariate case

Measurements of the length of the forearm (in inches) made on 140 adult males.

```
SAVE OUTFILE='F:\mathimata\Mathimata 2019-2020\Biomed\pointestimation\forearm.sav'
  /COMPRESSED.
DESCRIPTIVES VARIABLES=length
  /STATISTICS=MEAN STDDEV VARIANCE RANGE MIN MAX KURTOSIS SKEWNESS.
```

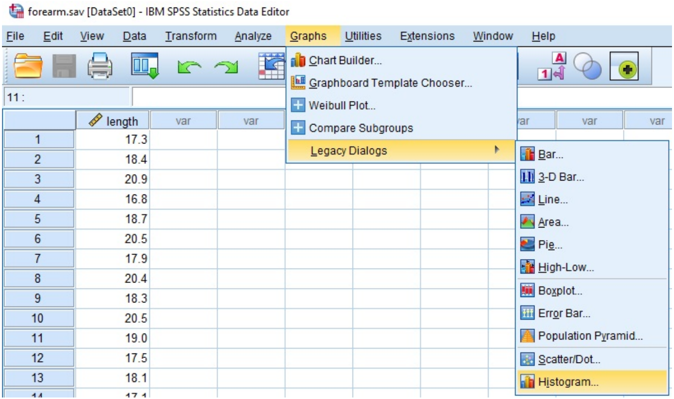
Descriptives

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
length	140	5.3	16.1	21.4	18.802	1.1205	1.255	-.110	.205	-.400	.407
Valid N (listwise)	140										

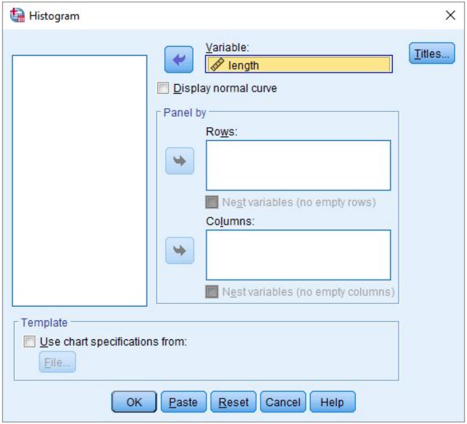
Example – Univariate case – Plots

Measurements of the length of the forearm (in inches) made on 140 adult males.



Example – Univariate case – Plots

Measurements of the length of the forearm (in inches) made on 140 adult males.

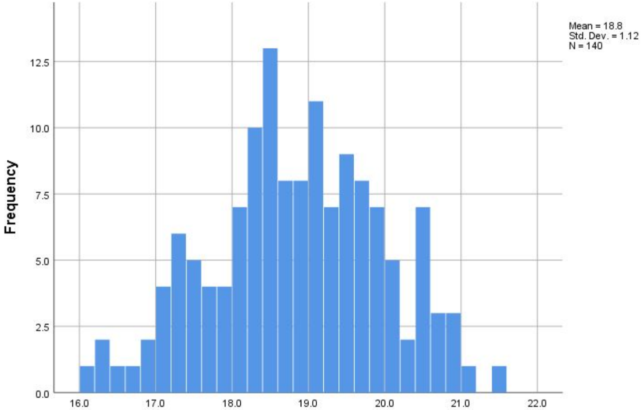


Example – Univariate case – Plots

Measurements of the length of the forearm (in inches) made on 140 adult males.

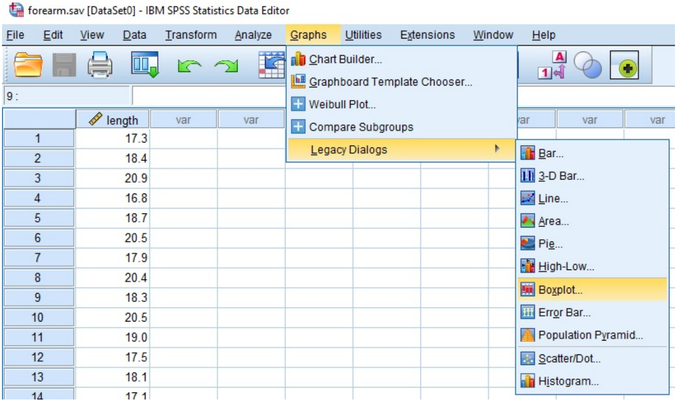
```
GRAPH
  /HISTOGRAM=length.
```

Graph



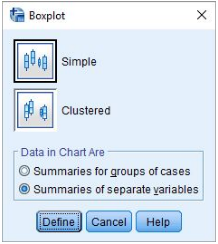
Example – Univariate case – Plots

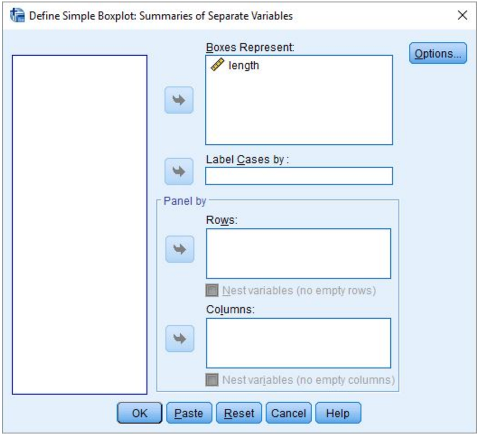
Measurements of the length of the forearm (in inches) made on 140 adult males.



Example – Univariate case – Plots

Measurements of the length of the forearm (in inches) made on 140 adult males.





Example – Univariate case – Plots

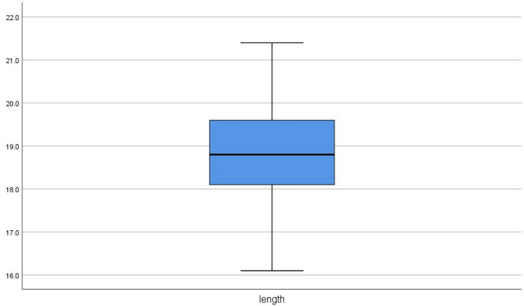
Measurements of the length of the forearm (in inches) made on 140 adult males.

```
EXAMINE VARIABLES=length
  /COMPARE VARIABLE
  /PLOT=BOXPLOT
  /STATISTICS=NONE
  /NOTOTAL
  /MISSING=LISTWISE.
```

Explore

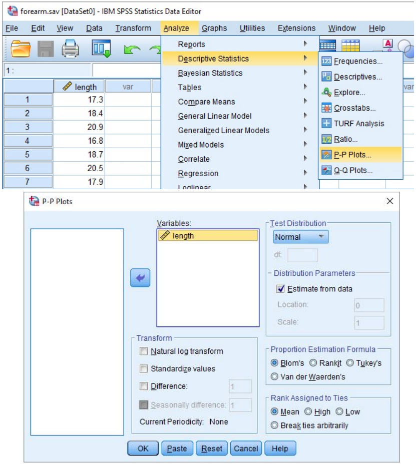
Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
length	140	100.0%	0	0.0%	140	100.0%



Example – Univariate case – Plots

Measurements of the length of the forearm (in inches) made on 140 adult males.



Example – Univariate case – Plots

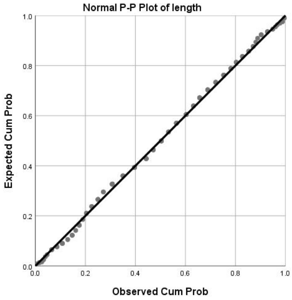
Measurements of the length of the forearm (in inches) made on 140 adult males.

Estimated Distribution Parameters

		length
Normal Distribution	Location	18.802
	Scale	1.1205

The cases are unweighted.

length



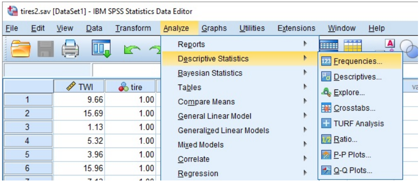
Example – Univariate case – Frequencies

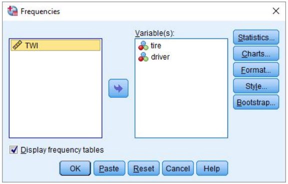
T.W.I. (Tread Wear Indicator)

75 tyres

3 types of tyres

2 drivers





Example – Univariate case – Frequencies

T.W.I. (Tread Wear Indicator)

```
FREQUENCIES VARIABLES=tire driver
  /ORDER=ANALYSIS.
```

Frequencies

Statistics

		tire	driver
N	Valid	75	75
	Missing	0	0

Frequency Table

tire

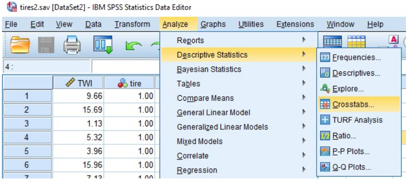
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	25	33.3	33.3	33.3
	2.00	27	36.0	36.0	69.3
	3.00	23	30.7	30.7	100.0
	Total	75	100.0	100.0	

driver

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	35	46.7	46.7	46.7
	1.00	40	53.3	53.3	100.0
	Total	75	100.0	100.0	

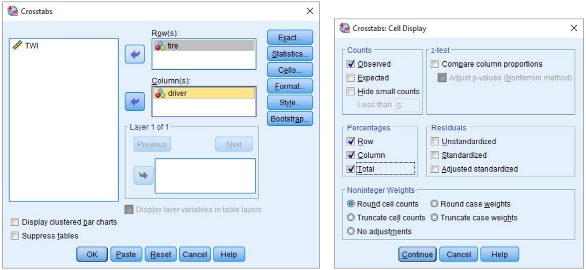
Example – Bivariate case – Tabulate

T.W.I. (Tread Wear Indicator)



Example – Bivariate case – Tabulate

T.W.I. (Tread Wear Indicator)



Example – Bivariate case – Tabulate

T.W.I. (Tread Wear Indicator)

```
CROSSTABS
  /TABLES=tire BY driver
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT ROW COLUMN TOTAL
  /COUNT ROUND CELL.
```

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
tire * driver	75	100.0%	0	0.0%	75	100.0%

tire * driver Crosstabulation

			driver		
			.00	1.00	Total
tire	1.00	Count	12	13	25
		% within tire	48.0%	52.0%	100.0%
		% within driver	34.3%	32.5%	33.3%
		% of Total	16.0%	17.3%	33.3%
	2.00	Count	11	16	27
		% within tire	40.7%	59.3%	100.0%
		% within driver	31.4%	40.0%	36.0%
		% of Total	14.7%	21.3%	36.0%
	3.00	Count	12	11	23
		% within tire	52.2%	47.8%	100.0%
		% within driver	34.3%	27.5%	30.7%
		% of Total	16.0%	14.7%	30.7%
Total		Count	35	40	75
		% within tire	46.7%	53.3%	100.0%
		% within driver	100.0%	100.0%	100.0%
		% of Total	46.7%	53.3%	100.0%

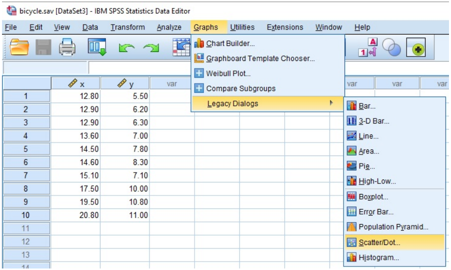
Example – Bivariate case

x: distance between bicycles and the center of the road

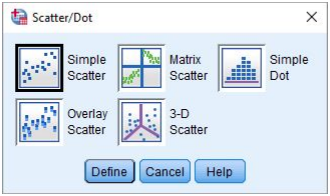
y: distance between bicycles and overtaking cars

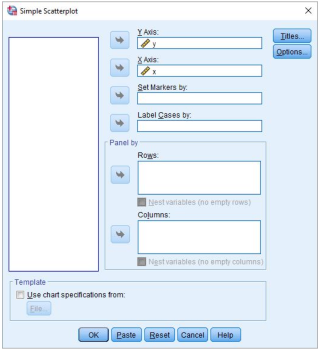
x	12.8	12.9	12.9	13.6	14.5	14.6	15.1	17.5	19.5	20.8
y	5.5	6.2	6.3	7.0	7.8	8.3	7.1	10.0	10.8	11

Reference: Kroll, B.J. & Ramey, M.R. (1977). Effects of Bike Lanes on Driver and Bicyclist Behavior. *Journal of Transportation Engineering*, ASCE, 103 (2), pp. 243-256



Example – Bivariate case



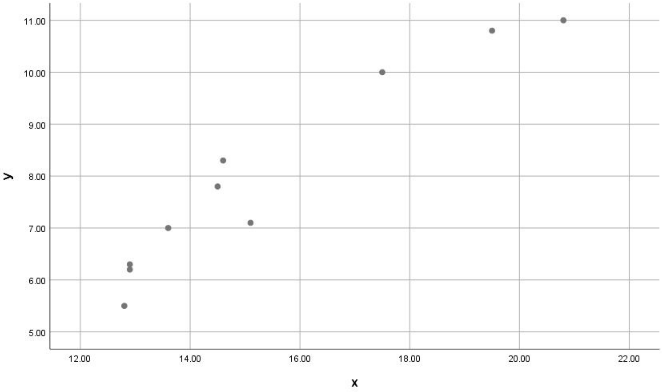


Example – Bivariate case

```
SAVE OUTFILE='F:\mathimata\Mathimata 2019-2020\Biomed\pointestimation\bicycle.sav'
  /COMPRESSED.
GRAPH
  /SCATTERPLOT(BIVAR)=x WITH y
  /MISSING=LISTWISE.
```

Graph

[DataSet3] F:\mathimata\Mathimata 2019-2020\Biomed\pointestimation\bicycle.sav



Outline

Introduction

Before the data is collected, the observations $X_1, X_2, \ldots, X_n$ are considered to be random variables.

So, any function (or **statistic**) of the observations like

- the mean $\bar{X}$
- the variance S^2
- the $X_{(n)}$
- $\ldots$

is also a random variable.

$\hookrightarrow$ Since every statistic is a random variable, it follows a distribution.

$\hookrightarrow$ These distributions are called **sampling distributions**.

Notation

Population parameters
Greek letters: μ,σ^2,δ....

Statistics
$T = T(X_1, X_2, \ldots, X_n)$, $S = S(X_1, X_2, \ldots, X_n)$...

Realization of statistics
Lowercase Latin letters:. $x_1, x_2, \ldots, x_n$, t, $\bar{x}$, s...

- **The corresponding random variables**
  Uppercase Latin letters $(X_1, X_2, \ldots, X_n,\ T,\ \bar{X},\ S\ldots)$

Sampling distribution of $\bar{X}$

The value we get for $\bar{X}$ (the sample mean) depends on the sample chosen.

Since the sample is random, $\bar{X}$ is a random variable!

As a random variable it follows a distribution.

$\hookrightarrow$ It is natural to assume that $\bar{X}$ would be close to μ (the population mean) but there will be some variability in $\bar{X}$ before it is observed because we use random sampling to choose our sample of size n.

$\hookrightarrow$ **The sampling distribution** of $\bar{X}$

- tells us what kind of values are likely to occur for $\bar{X}$.
- puts a probability distribution on the possible values for $\bar{X}$.

$\bar{X}$ and S^2

Theorem

Let $X_1, \ldots, X_n$ be a random sample from a population with mean $\mu < +\infty$ and variance $\sigma^2 < +\infty$, then

- $E(\bar{X}) = \mu$

$$E(\bar{X}) = E\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \sum_{i=1}^{n} \frac{1}{n} E(X_i) = \sum_{i=1}^{n} \frac{1}{n} \mu = \frac{n\mu}{n} = \mu$$

- $Var(\bar{X}) = \frac{\sigma^2}{n}$

$$V(\bar{X}) = V\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \sum_{i=1}^{n} \frac{1}{n^2} V(X_i) = \sum_{i=1}^{n} \frac{1}{n^2} \sigma^2 = \frac{n\sigma^2}{n^2} = \sigma^2/n$$

- $E(S^2) = \sigma^2$

$*$ A statistic $T = T(X_1, \ldots, X_n)$ is called **unbiased estimator** of the parameter θ, if $E(T) = \theta$.

Expected value of S^2

$$
\begin{aligned}
E(S^2) &= E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right) = \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\mu-\bar{X}+\mu)^2\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}\left((X_i-\mu)^2-2(X_i-\mu)(\bar{X}-\mu)+(\bar{X}-\mu)^2\right)\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\mu)^2-2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)+\sum_{i=1}^{n}(\bar{X}-\mu)^2\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\mu)^2-2(\bar{X}-\mu)(n\bar{X}-n\mu)+n(\bar{X}-\mu)^2\right) = \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\mu)^2-n(\bar{X}-\mu)^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}E\left[(X_i-\mu)^2\right]-nE[(\bar{X}-\mu)^2]\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n}\sigma^2-n\frac{\sigma^2}{n}\right) = \frac{1}{n-1}\left(n\sigma^2-\sigma^2\right) \\
&= \sigma^2
\end{aligned}
$$

Central Limit Theorem

Suppose that $X_1, X_2, \ldots, X_n$ are independent and identically distributed with finite mean μ and variance σ^2. If n is large enough, then

$$\bar{X} \sim N(\mu, \sigma^2/n) \quad \text{approximately}$$

See MATLAB file: *illustration_CLT.m*

Other sampling distributions

The Central Limit Theorem provides the sampling distribution of $\bar{X}$ for any population for large enough samples.

There are cases although that we are not (only) interesting in $\bar{X}$.

In such cases the

- Student t-distribution
- Chi-squared distribution
- F-Ratio distribution

are useful choices to describe sampling distributions for interesting statistics like the

- $\bar{X}$ when the population variance is unknown (Student t-distribution)
- S^2 (Chi-squared distribution)
- S_1^2/S_2^2 (F-Ratio distribution)

Beyond Central Limit Theorem

In many cases we want to make inference for a population parameter that is neither the mean nor the variance.

Example: The parameter λ of the exponential distribution

Recall that if $X \sim Exp(\lambda)$,
then $E(X) = 1/\lambda$ and $V(X) = 1/\lambda^2$.

Outline

Statistical inference

Statistical inference is the process of drawing conclusions about a population or a scientific statement by analysis of data

The conclusion of a statistical inference is a statistical proposition such as

- parameter(s) estimation
- confidence intervals
- rejection (or not) of a hypothesis
- estimate of the relationship between two or more variables
- clustering or classification of data points into groups
- ....

Parameter(s) estimation

The term **parameter estimation** refers to the process of using sample data in order to estimate the parameter(s) of a probability distribution, such as the mean and standard deviation of a normal distribution.

The most commonly used estimators are

- the Method of Moments Estimator (MME)
- the Maximum Likelihood Estimators (MLE)
- the Least Square Method (LSM)

Method of moments estimator

Based on the assumption that sample moments provide GOOD ESTIMATES of the corresponding population moments.

The basic steps behind the method of moments are to

1. equate the first sample moment about the origin $M_1 = \frac{1}{n}\sum x_i = \bar{X}$ to the first theoretical moment $E(X)$.
2. equate the second sample moment about the origin $M_2 = \frac{1}{n}\sum x_i^2$ to the second theoretical moment $E(X^2)$.
3. continue equating sample moments about the origin, M_k, with the corresponding theoretical moments $E(X^k)$, $k = 3, 4, \ldots$ until the number of equations equals the number of parameters.
4. solve for the parameters.

Alternative: For $k \geq 2$ one could equate the sample moments about the mean with the corresponding theoretical moments about the mean.

Method of moments estimator – Example

The geometric distribution governs the number of trials needed to get the first success in a sequence of Bernoulli trials.

The probability density function is given by

$$g(x) = p(1-p)^{x-1}, \quad x \in N_+.$$

The mean of the distribution is $\mu = 1/p$.

Suppose now that $X = (X_1, X_2, \ldots, X_n)$ is a random sample of size n from the geometric distribution with parameter p, then the method of moments estimator $\tilde{p}$ of p is given by

$$\frac{1}{\tilde{p}} = \frac{1}{n} \sum X_i \; (= \bar{X})$$

and so

$$\tilde{p} = \frac{1}{\bar{X}}.$$

Method of moments estimator

Pros

- Easy to compute and can be always calculated
- MME is consistent (i.e. converges in probability to the true value of the parameter)

Cons

- MME are usually not the "best estimators" available.

Maximum likelihood estimators

The idea behind the Maximum likelihood estimators is to find the values of the parameters θ which maximizes the likelihood of getting the data that we, in fact, observed.

Suppose we have a random sample $X = (X_1, X_2, \ldots, X_n)$ for which the probability density (or mass) function of each X_i is $f(x_i; \theta)$. Then, the joint probability mass (or density) function of $X_1, X_2, \ldots, X_n$ denoted by $L(\theta)$ is given by

$$\begin{aligned} L(\theta) &= P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ &= f(x_1; \theta) \cdot f(x_2; \theta) \cdots f(x_n; \theta) \\ &= \prod_{i=1}^{n} f(x_i; \theta) \end{aligned}$$

The MLEs of θ are obtained by treating the likelihood function $L(\theta)$ as a function of θ, and find the value of θ that maximizes it or equivalently that maximizes the log-likelihood function

$$\ell(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

Maximum likelihood estimators – Example

Under some regular conditions the number of blood donations can be described by a Poisson process with constant (over time) rate λ.

As a result the time between consecutive blood donations of an individual follows an exponential distribution while the time until the r^{th} blood donation follows a Gamma distribution with pdf

$$f(x;\lambda) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)}.$$

- The likelihood function is given by $L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^r x_i^{r-1} e^{-\lambda x_i}}{\Gamma(r)}$.
- The log-likelihood function is given by

$$\ell(\lambda) = \sum_{i=1}^{n} \left(-x_i\lambda + (-1+r)\log x_i + r\log\lambda - \log\Gamma(r)\right)$$

- The MLE of λ is obtained by solving the equation $\frac{\partial \ell(\lambda)}{\partial \lambda} = 0$.
- The MLE $\hat{\lambda}$ of λ is given by $\hat{\lambda} = \frac{nr}{\sum x_i} = \frac{r}{\bar{X}}$.

Maximum likelihood estimators – Example

The following data represent the time (in days) until the 10^{th} blood donation of an individual.

1723	1387	648	1099	858	1059	1714	2082	1417	826
853	737	793	774	594	919	1123	897	1160	837
1187	651	868	1088	2268	813	493	1090	1016	796
910	1382	1202	952	954	711	1230	1482	739	977
1166	481	599	640	744	1150	678	1022	1202	877
1002	1309	689	1240	1130	1148	589	1835	1434	521
1036	1074	800	1475	518	1451	827	708	968	1033
941	941	835	981	1738	996	1946	899	727	708
847	549	738	930	1270	1211	1227	1391	1759	408
835	1324	1323	1191	716	1262	1302	1336	906	1011

$\bar{X} = 1039.03 \Rightarrow \hat{\lambda} = \frac{r}{\bar{X}} = \frac{10}{1039.03} \approx 0.0096$

Interpretation: Given the observed data the MLE of λ is 0.0096 which implies that we expect (in average) 0.0096 donation per day, or equivalent 1.15492 (in average) blood donations every 4 months ($0.0096 \times 4 \times 30 = 1.15492$).

This also implies that the expected time between blood donations is $\frac{1039.03}{10} = 103.903$ days.

Maximum likelihood estimators

Pros

When sample size n is large (n>30), MLE is

- unbiased (i.e. $E(\hat{\theta}) = \theta$)
- consistent (i.e. converges in probability to the true value of the parameter)
- normally distributed
- ...

As a result the MLEs are more useful in statistical inference and the most commonly used estimators.

Cons

- MLE can be highly biased for small samples
- In many cases the MLEs has no closed-form solution and we must use numerical methods in order to maximize the log-likelihood.

Least squares estimator

The method of least squares is a standard approach in regression analysis.

Although, it can also be used to estimate population parameters.

The objective consists of adjusting the parameters of a model function or distribution to best fit a data set. The least squares method finds the optimum values by minimizing

$$S = \sum_{i=1}^{n} (y_i - E(y_i))^2.$$

Least squares estimator - Example

Let $X_1, \ldots, X_n$ be a random sample from an $\text{Exp}(\lambda)$ population. The least square estimator $\hat{\lambda}$ of λ is obtained by minimizing

$$S(\lambda) = \sum_{i=1}^{n} \left(x_i - \frac{1}{\lambda} \right)^2 .$$

It is easy to show that

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{X}}.$$

Reading

- Biostatistics, A Foundation for Analysis in the Health Sciences, W.W. Daniel and C.L. Cross

  http://docshare02.docshare.tips/files/22448/224486444.pdf

  Chapters 2, 5 and 6

- Probability & Statistics for Engineers and Scientists R.E.Walpole, R.H. Myers, S.L.Myers, K.Ye

  https://fac.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineers_and_scientisst.pdf

  Chapters 1, 8 and 9