Probability and Statistics
MAT 271E

PART 6

Frequency Analysis and Parameter Estimation

Assist. Prof. Dr. Ümit KARADOĞAN

Course originally developed by : Prof. Dr. Mehmetçik BAYAZIT, Prof. Dr. Beyhan YEĞEN

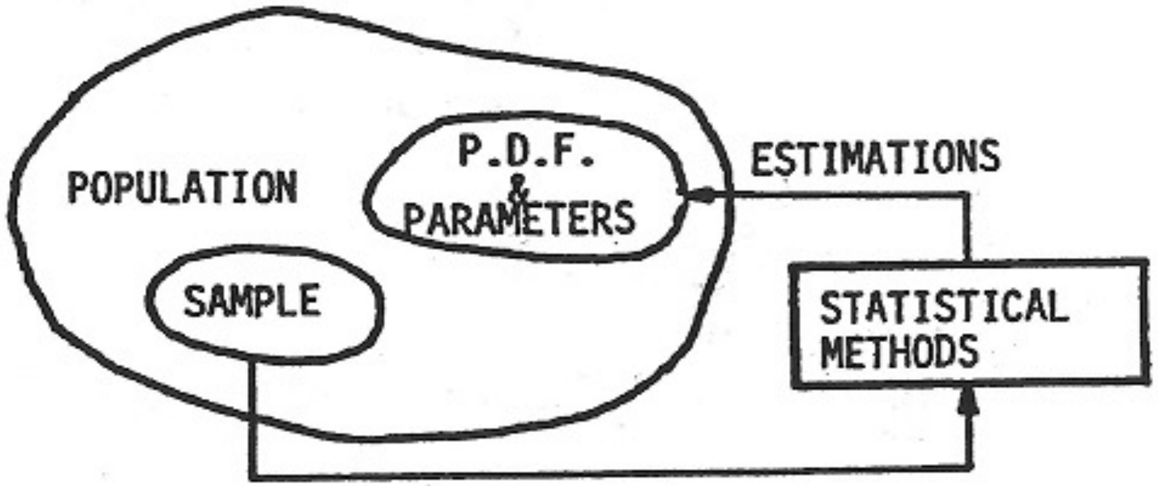
Statistical Sample

The **population** (consisting of **all** the observations belonging to a random variable) should be observed in order to determine exactly the **probability distribution** of the random variable.

However, in practice only a **statistical sample** (having **finite** number of elements) can be taken from the population.

FREQUENCY ANALYSIS AND PARAMETER ESTIMATION

A **sample** is a set of observations collected to determine the statistical properties of a random variable.
Each element in the sample is an event belonging to the random variable or is a value the random variable has taken.



Estimation of the distribution and parameters of a random variable from a statistical sample

FREQUENCY ANALYSIS AND PARAMETER ESTIMATION

A sample must be analyzed in an optimum way since the probability distribution function of the random variable and the parameters of this distribution can only be estimated depending on the **limited sample** in hand.

Statistics is the science which obtains all possible information from **samples** and arrives at conclusions about the statistical properties of the **population** using this information.

Statistics makes the **best estimates** of the properties of the population of the random variable by analyzing the information in the sample, depending on probability theory. It also evaluates the errors in these estimates.

FREQUENCY ANALYSIS AND PARAMETER ESTIMATION

The samples to be used in statistical studies should be adequate **qualitatively** and **quantitatively**.

The following conditions must be realized for a sample to be **qualitatively adequate**:

1- The data in the sample should be **homogeneous**, in other words all data should indeed be elements of the population of the **same random variable**.

Otherwise, the statistical calculations to be made will have no significance.

For example: In case the flow of a river is controlled by a dam, it would not be correct to evaluate the flows downstream of the dam measured **before** the dam construction together with the flows measured **after** the dam construction as one sample since these flows would not be **homogeneous**.

FREQUENCY ANALYSIS AND PARAMETER ESTIMATION

2- There should be **no systematic errors** in the measurement of the elements of the sample.

3- Random errors should be **minimized**.

Decrease of random measurement errors, which always exist, to an acceptable level will reduce errors in the results that will be obtained by the statistical analysis of the sample.

The sample being **quantitatively sufficient** means that the **number of elements in the sample is sufficiently large**.

Although an exact limit cannot be given for the sufficient number, it can be said that **more reliable** results can be obtained about the properties of the population **as the number of elements in the sample increases**.

In statistics, samples having **less than about 30 elements** are called **small samples** and in such analyses, using expressions valid for large samples is not correct.

FREQUENCY ANALYSIS AND PARAMETER ESTIMATION

Estimates made for the properties of the population (probability distribution function parameters) by statistical analysis of the sample are not equal to the **real** values of the population.
A part of the difference is due to: the **qualitative inadequacy** of the sample (*random errors, unnoticeable non-homogeneity and systematic errors*),
the other part is due to the **limited number of elements** of the sample (*sampling errors*)

There are small or large amounts of uncertainty in the estimated population properties, statistics gives expressions for the amount of uncertainty due to sampling errors.

Firstly, the estimation of the **probability distribution (frequency analysis)** of the random variable by statistical analysis of the sample and secondly the estimation of parameters should be conducted.

FREQUENCY ANALYSIS

Since it is not possible to observe the **whole population** of a random variable, it is assumed that the probability distribution is equivalent to the frequency distribution obtained by the analysis of the **sample** in hand.

The analysis of the sample with the aim of determining the **frequency distribution** is made by one of the following methods depending on the type of the random variable.

Frequency Analysis of Discrete Variables

Suppose we have a sample of **N** elements belonging to a **discrete** variable. If in this sample the **event $X=x_i$** occurs n_i times, the ***frequency*** of this event is defined as:

$$f(x_i)=\frac{n_i}{N}$$

The **frequency graph (histogram)** is obtained by drawing the calculated $f(x_i)$ values as vertical lines for abscissas x_i.

As the number of elements in the sample increases, the frequency graph approaches **probability mass function** because **frequency $f(x_i)$** converges to **probability $p(x_i)$**.

FREQUENCY ANALYSIS

Frequency Analysis of Discrete Variables

The **cumulative distribution function** is computed as:

$$F(x_i) = \frac{\sum_{j=1}^{i} n_j}{N} = \sum_{j=1}^{i} f(x_j)$$

The c.d.f. is the stepwise graph (see slide 12) obtained by drawing the calculated **F(x_i)** values as vertical lines for abscissas **x_i**.

As **N** gets larger, the **cumulative frequency distribution** approaches the **cumulative probability distribution.**

Example (M. Bayazıt, B. Oğuz, Example 3.1, pg 65)

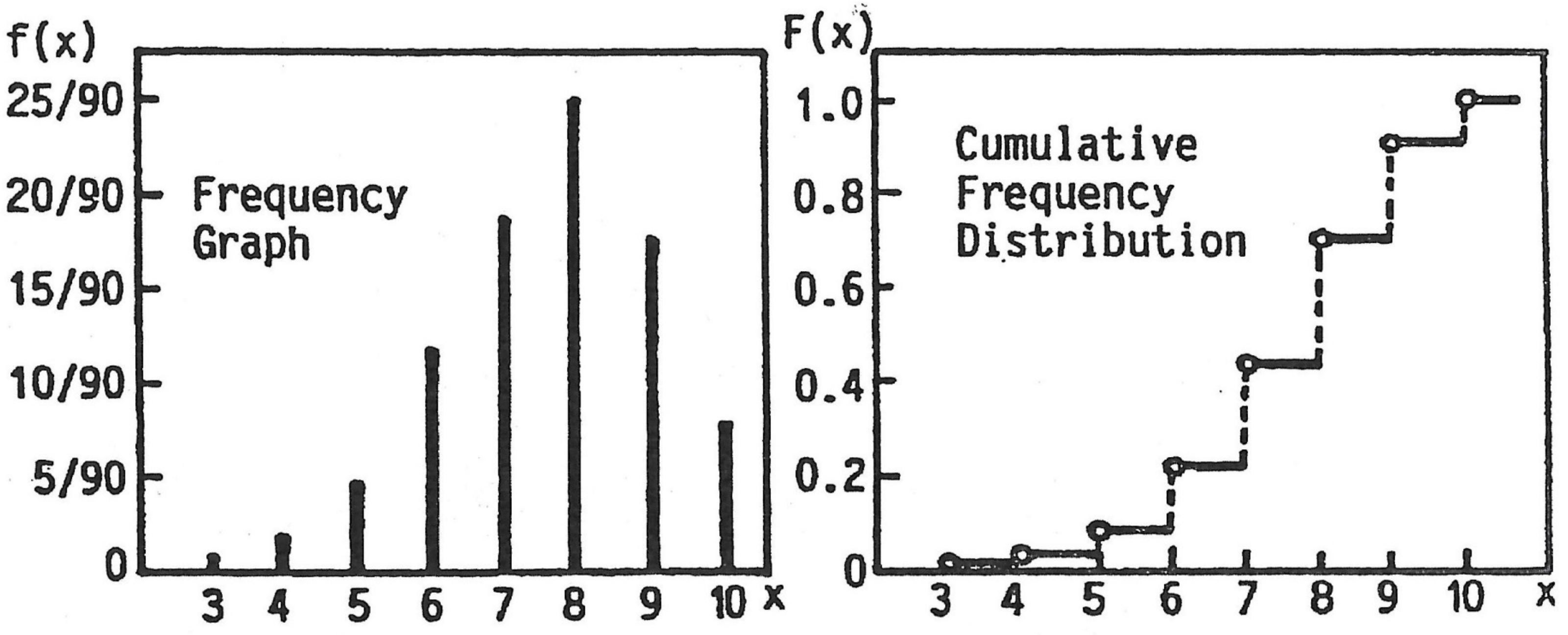
The distribution of the grades of an exam in a class of 90 students is given below. The **frequency graph** can be drawn by using the frequency values **f(x_i)** and the **cumulative distribution function** can be drawn using the frequency values **F(x_i)** calculated using the equations below.

$$f(x_i)=\frac{n_i}{N} \qquad F(x_i)=\frac{\sum_{j=1}^{i} n_j}{N}=\sum_{j=1}^{i} f(x_j)$$

x_i	3	4	5	6	7	8	9	10
n_i	1	2	5	12	19	25	18	8
$f(x_i)=n_i/N$	0.011	0.022	0.056	0.133	0.211	0.278	0.200	0.089
$F(x_i)=\sum_{j=1}^{i} f(x_j)$	0.011	0.033	0.089	0.222	0.433	0.711	0.911	1.000

1/90

Example (M. Bayazıt, B. Oğuz, Example 3.1, pg 65)



Frequency graph and cumulative frequency distribution of the random variable

Frequency Analysis of Continuous Variables

Frequency Analysis of Large Samples

The range of the random variable is divided into an appropriate number of **class intervals**. If the **number of observations** falling into the i-th class is $\mathbf{n_i}$, then the frequency of this class interval is:

$$f_i = \frac{n_i}{N}$$

The stepwise line, obtained by showing f_i values for the i-th class interval, is called ***frequency histogram***.
The cumulative frequency distribution is obtained similarly as in the discrete variable case .

Frequency Analysis of Continuous Variables

Frequency Analysis of Large Samples

An important point in frequency analysis is to choose the **number of class intervals**. This number should be increased as the number of elements in the sample increases.

Generally, the number of class intervals is kept between 5 and 20.

If **too few intervals** are used in the analysis, a large amount of information in the sample is to be lost.

On the other hand, **if too many class** intervals are used then both more effort than required will be needed and the histogram will have a quite irregular shape because very few or no observations will fall into some class intervals.

The following empirical formulas can be used to determine the number of class intervals.

Frequency Analysis of Continuous Variables

Frequency Analysis of Large Samples

Empirical formulas can be used to determine the **number of class intervals**:

$$m \cong 1 + 3.3 \log_{10} N \quad \text{or} \quad 2^m \geq N$$

The widths of the class intervals need **not be equal**; it might be appropriate to choose larger class intervals at both ends of the range of the random variable to let approximately equal number of observations to fall into each class.

The first and the last class intervals must be chosen so that the **minimum** value of the sample remains in the former and the **maximum** remains in the latter.

It is appropriate to choose the limits of class intervals as **round numbers**.

Frequency Analysis of Continuous Variables

Frequency Analysis of Small Samples

If the number of elements in the sample is **small**, it is not appropriate to analyze the data by dividing them into class intervals since in such a case a significant amount of the information will be lost and some class intervals may have no observation at all.
In this case, the objective is to determine the **cumulative frequency distribution** only.
The **ordered sample** is obtained by listing the elements of the sample from smaller values towards larger values:

$$x_1 \leq x_2 \leq \ldots\ldots \leq x_m < \ldots\ldots \leq x_N$$

Frequency Analysis of Continuous Variables

Frequency Analysis of Small Samples

The first expression to calculate the frequency of the random variable being equal to or smaller than x_m is:

$$F(x_m) = \frac{m}{N}$$

However, when this expression is used, the frequency of the random variable remaining equal to or smaller than the x_N value, the maximum value of the **sample**, is **1**.

Since elements greater than the value might exist in the **population** of the random variable, it is not correct to use this equation which implies that X would never exceed the value x_N.

Frequency Analysis of Continuous Variables

Frequency Analysis of Small Samples

Various empirical formulas called **plotting position formulas** have been proposed to eliminate this inconvenient aspect.

The most popular formula among these is known as the **Weibull formula**:

$$F(x_m) = \frac{m}{N+1}$$

The **frequency histogram** and the **cumulative frequency distribution** obtained in the frequency analysis have an irregular shape for small number of elements in the sample, because of the **sampling errors**. They become more regular as the number of elements in the sample increases.

Frequency Analysis of Continuous Variables

Frequency Analysis of Small Samples

Cumulative frequency distributions have more regular shapes because they are the integrals of histograms in a sense.
This is why in practice the **cumulative frequency distributions** are usually obtained directly.

The frequency analysis of **multivariable distributions** are made similarly. However, graphical representation becomes more difficult in this case since the frequency distribution should be expressed by multidimensional surfaces.

Example (M. Bayazıt, B. Oğuz, Example 3.2, pg 67)

The flood data *(ml/s)* of the Dicle river during 1956-75 are given below:

Year	1956	1957	1958	1959	1960	1961	1962	1963	1964	1965
X	2324	6300	2340	2080	2262	1250	3014	7910	4350	2630

Year	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975
X	8820	4516	4866	6450	2250	2250	3450	5300	963	2571

In the frequency analysis, firstly the elements are ordered according to their magnitudes since the sample is a **small** one ***(N=20 < 30)***.

The probability of being equal to or remaining smaller than x_m is calculated by the **Weibull formula**:

$$F(x_m) = \frac{m}{N+1}$$

Example (M. Bayazıt, B. Oğuz, Example 3.2, pg 67)

m	1	2	3	4	5	6	7	8	9	10
x_m	963	1250	2080	2250	2262	2340	2424	2571	2630	3014
$F(x_m)$	0.048	0.095	0.143	0.190	0.238	0.286	0.333	0.381	0.429	0.475

1/21

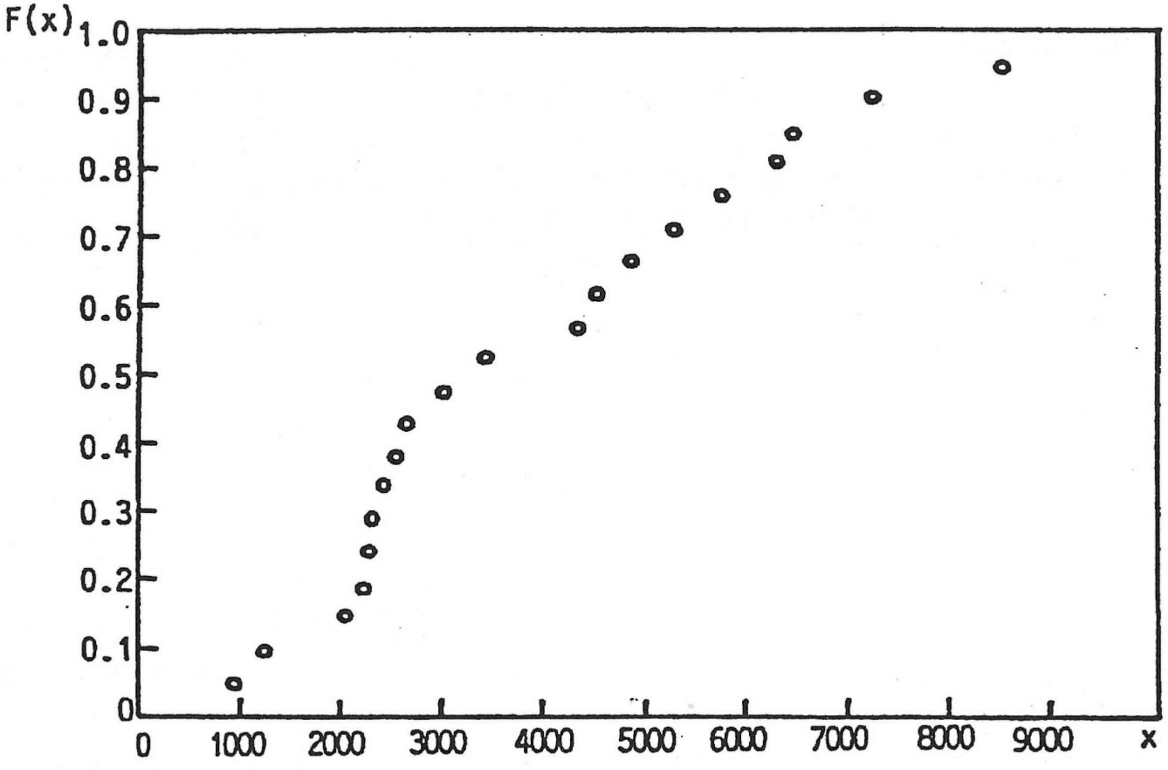
m	11	12	13	14	15	16	17	18	19	20
x_m	3450	4350	4516	4866	5300	5772	6300	6450	7910	8820
$F(x_m)$	0.524	0.571	0.619	0.667	0.714	0.762	0.810	0.857	0.905	0.952

Cumulative frequency distribution of floods is determined and plotted in a graph (next slide).

The probability that the flood discharge exceeds 8820 m^2/s can be seen in the table as:

1-F(8820)=1-0.952=0.048

Example (M. Bayazıt, B. Oğuz, Example 3.2, pg 67)



Cumulative frequency distribution of the flood flows

PARAMETER ESTIMATION

As the probability distribution of a random variable cannot be determined exactly from a **sample**, its parameters cannot be computed by the equations given in the former section (Parameters of Random Variables) because the probability distribution is not known.

The values of the **parameters estimated from a sample** are called **statistics**.

These **statistics** are **not** equal to the **population** values of the parameters. However, the differences between the parameters and statistics **(sampling errors)** can be minimized by using optimal **estimation methods**.

PARAMETER ESTIMATION

Properties of Parameter Estimates

The value estimated from the available **sample** for a parameter of the **population** of the **random variable** is also of a **random** character.
If we had **other samples of same size** drawn from the same population, the statistics calculated for the same parameter from these samples would be **different** from each other.

The appropriate choice of the **parameter estimation method** is important because this method affects the results obtained.

A desired property of the parameter estimation method is that the estimate to be obtained is **unbiased**.

PARAMETER ESTIMATION

Properties of Parameter Estimates

If the **expected value** (i.e. the **mean** of the statistics calculated from several samples) of the estimate ***a*** of the population parameter α is equal to α then this is an unbiased estimate and is shown by **â**.

$$E(\hat{a}) = \alpha$$

Another desired property in parameter estimation is that the estimated value (statistic) changes little from sample to sample, in other words its **sampling variance is small**.
The estimate with the smallest sampling variance is called the **efficient** estimate.

PARAMETER ESTIMATION

Estimation of the Parameters of a Random Variable

Statistical Moments: The parameters of a random variable that are of the **statistical moment** type can be estimated by the following equations:
The **mean**, which is the **expected value of the random variable**, is estimated as the arithmetic mean of the elements of a sample.
The **statistic of the mean (of the sample)** which is shown by $\bar{x}$ can be calculated as follows:

$$\bar{x} = \left(\sum_{i=1}^{N} x_i \right) / N$$

PARAMETER ESTIMATION

Estimation of the Parameters of a Random Variable

The variance can be estimated by the following formula since it is the **expected value of the square of the differences of the random variable from its mean**:

$$Var(x) = \sum_{i=1}^{N} (x_i - \bar{x})^2 / N = \left(\sum_{i=1}^{N} x_i^2 / N \right) - (\bar{x})^2$$

The expected value of the statistic calculated by the equation above is biased.

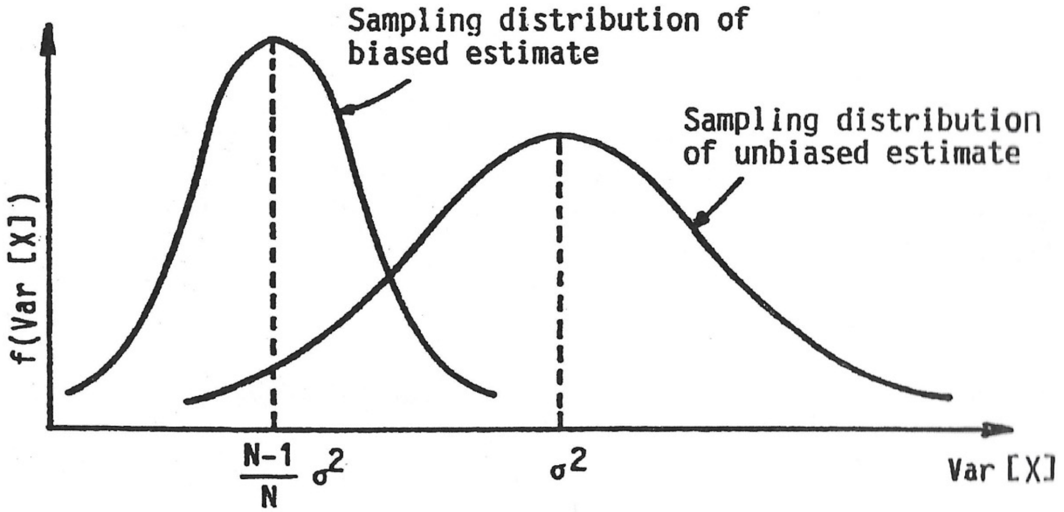
Therefore, we should divide it by ***N-1*** instead of ***N*** in order to obtain an unbiased estimate of variance ($V\hat{a}r(X)$):

$$V\hat{a}r(X) = \sum_{i=1}^{N} (x_i - \bar{x})^2 / (N - 1)$$

PARAMETER ESTIMATION

Estimation of the Parameters of a Random Variable

This correction has importance for **small** samples:
For ***N>30***, dividing by **N** instead of **N-1** would **<u>not</u>** lead to a significant difference.
The **sampling variance** of the **unbiased** estimate of the variance, is greater than that of the **biased** estimate.



Sampling distributions of the biased and unbiased estimates of the variance

PARAMETER ESTIMATION

Estimation of the Parameters of a Random Variable

The **statistic** of the **standard deviation** shown by ***Sx*** can be calculated as follows:

$$\boldsymbol{s_X} = \left[\sum_{i=1}^{N}(x_i - \bar{x})^2 \,/\boldsymbol{N}\right]^{1/2}$$

For small samples:

$$\widehat{\boldsymbol{s_X}} = \left[\sum_{i=1}^{N}(x_i - \bar{x})^2 \,/(\boldsymbol{N-1})\right]^{1/2}$$

PARAMETER ESTIMATION

Estimation of the Parameters of a Random Variable

Although the previously mentioned equation $\hat{Var}(X) = \sum_{i=1}^{N} (x_i - \bar{x})^2 / (N-1)$

gives an **unbiased** estimate of the variance, the estimate for the **standard deviation** provided by:

$$\widehat{s_X} = \left[\sum_{i=1}^{N} (x_i - \bar{x})^2 / (N-1) \right]^{1/2}$$

is **not quite unbiased.**

PARAMETER ESTIMATION

Estimation of the Parameters of a Random Variable

The statistics of **3rd** and **4th** order moments can be calculated similarly.

$$m_X^{(3)} = \sum_{i=1}^{N} (x_i - \bar{x})^3 \,/N$$

(To obtain an **unbiased** estimate for small samples, we must divide by *(N-1)(N-2)/N* instead of N.)

$$m_X^{(4)} = \sum_{i=1}^{N} (x_i - \bar{x})^4 \,/N$$

(To obtain an **unbiased** estimate for small samples, we must divide by *(N-1)(N-2)(N-3)/N*^{2} instead of N.)

Estimation of the Parameters of a Random Variable

Order Statistics:

Quartile: The **quartile X_q** is estimated from an **ordered sample** as follows. The **rank** of the quartile is computed as:

$$m = q(N+1)$$

taking the **closest integer value for *m*.**

Then, the observation of rank **m** in the ordered sample is the estimate of the quartile.

Median: If the number of elements in the sample is an **odd** number, the element in the middle,

if this number is an **even** number the average of the two elements in the middle gives the estimate for the **median**.

PARAMETER ESTIMATION

Estimation of the Parameters of a Random Variable

The estimates of the parameters such as median, interquartile range and quartile skewness coefficient (which are **not statistical moments**) are not much affected by the **outliers**.
*(Outlier: A data point that is **distinctly separate** from the rest of the data in terms of being either very small or large compared to the other values of the sample)*

Therefore, such parameters should be preferred in skewed distributions

Example (M. Bayazıt, B. Oğuz, Example 3.3, pg 70)

The mean of the grades in Example 3.1 can be calculated as follows:

x_i	3	4	5	6	7	8	9	10
n_i	1	2	5	12	19	25	18	8
$f(x_i)=n_i/N$	0.011	0.022	0.056	0.133	0.211	0.278	0.200	0.089
$F(x_i)=\sum_{j=1}^{i} f(x_j)$	0.011	0.033	0.089	0.222	0.433	0.711	0.911	1.000

$$\bar{x} = \left(\sum_{i=1}^{N} x_i\right)/N = \left(\sum_{i=1}^{m} x_i\, n_i\right)/N = \sum_{i=1}^{m} x_i\, f_i$$

Example (M. Bayazıt, B. Oğuz, Example 3.3, pg 70)

Here ***m*** is the number of **values the discrete variable can take** (in the example *m=8), **n**,* shows the **number of times each value is taken**.

$$\bar{x} = 3 \times 0.011 + 4 \times 0.022 + 5 \times 0.056 + 6 \times 0.133 + 7 \times 0.211 + 8 \times 0.278 + 9 \times 0.200 + 10 \times 0.089 = 7.59$$

The variance of the grades is:

$$Var(X) = \left(\sum_{i=1}^{N} x_i^2 / N\right) - \bar{x}^2 = \left(\sum_{i=1}^{m} x_i^2 n_i / N\right) - \bar{x}^2 = \left(\sum_{i=1}^{m} x_i^2 f_i\right) - \bar{x}^2$$

$$Var(X) = (3^2 \times 0.011 + 4^2 \times 0.022 + 5^2 \times 0.056 + 6^2 \times 0.133 + 7^2 \times 0.211 + 8^2 \times 0.278 + 9^2 \times 0.200 + 10^2 \times 0.089) - 7.59^2 = 2.26$$

Example (M. Bayazıt, B. Oğuz, Example 3.3, pg 70)

The standard deviation:

$$s_X = [Var(x)]^{1/2} = 1.50$$

Since the number of elements of the sample is **large** ***(N=90 >30)***, it is **not necessary** to divide by N-1 instead of N in the calculation of the variance and the standard deviation.

The coefficient of variation:

$$C_{vX} = s_X \ / \ \bar{x} = 1.50 / 7.59 = 0.198$$

PARAMETER ESTIMATION

Estimation of the Parameters of Probability Distribution Function

Each **probability density function** ***f(x)*** (or **cumulative distribution function** ***F(x)***) has a number of parameters **α, β,...** The estimates ***a, b,...*** of these parameters can be obtained from a sample by various methods.

Method of Moments: For each probability function, the parameters are related to the statistical moments of the variable by certain equations. Estimates of the parameters can be derived from these equations in terms of the computed statistics.

Number of equations to be solved equals the number of parameters of the function. Usually the first few moments are used.

Although in general it does not give efficient estimates, the **method of moments** is the method used most often because it is easy to apply.

PARAMETER ESTIMATION

Estimation of the Parameters of Probability Distribution Function

Maximum Likelihood Method: The previous method of moments does not generally give efficient estimates for the parameters of probability density functions.

It may be preferred to use the **maximum likelihood method** to obtain estimates with smaller sampling variances.

Suppose we have a sample of N elements; $x_1, x_2, \ldots, x_N$.

It is desired to estimate the **parameters** of the **probability density function** ***f(x;α, β,…)***.

The probability that the event $X = x_1$ will occur in an observation is proportional to $f(x_1;\alpha, \beta,\ldots)$.

PARAMETER ESTIMATION

Estimation of the Parameters of Probability Distribution Function

Similarly, the probabilities that the events $X= x_2, \ldots , X=x_N$ will occur are proportional to $\mathbf{f(x_2;\alpha, \beta,\ldots)}$, $\mathbf{f(x_N;\alpha, \beta,\ldots)}$.

Since these events are independent, the probability that events $X= x_1$, $X= x_2, \ldots ,$ $X=x_N$ will occur in N observations will be proportional to the multiplication of $\mathbf{f(x_1;\alpha, \beta,\ldots)},\ldots, \mathbf{f(x_N;\alpha, \beta,\ldots)}$:

$$L=\prod_{i=1}^{N} f(x_i;\alpha,\beta,\ldots)$$

L defined by the above equation is called the **likelihood function**.

In the **maximum likelihood method**, the **a,b,...** values which maximize the likelihood function are assumed to be the estimates of **α, β,...** parameters.

These estimates are obtained from the following set of equations.

PARAMETER ESTIMATION

Estimation of the Parameters of Probability Distribution Function

The estimates are obtained from the following set of equations:

$$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial b} = \ldots = 0$$

In practice, it is preferred to work with **In L** instead of **L** to change the product in the equation in the previous slide to a sum. Since **In L** is an increasing function of **L**, the a, b,... values, which make L maximum, make In L maximum as well. Therefore, the estimates a, b,... can be calculated by the following equations:

$$\frac{\partial(\ln L)}{\partial a} = \frac{\partial(\ln L)}{\partial b} = \ldots = 0$$

PARAMETER ESTIMATION

Estimation of the Parameters of Probability Distribution Function

<u>**Method of L-Moments (PWMs)**</u> is another parameter estimation method that has been proposed. *L-moments* are defined as functions of **probability-weighted moments** (PWMs) defined as:

$$\beta_r = E\left(X[F(x)]^r\right)$$

for *r=0,1,2,...* The first few L-moments are calculated in terms β_r from:

$$\lambda_1 = \beta_0$$
$$\lambda_2 = 2\beta_1 - \beta_0$$
$$\lambda_3 = 6\beta_2 - 6\beta_1 + \beta_0$$

It is seen that L-moments are **linear** functions of β_r, unlike the statistical moments which are functions of $E(X)$, $E(X^2)$ $E(X^3)$, ...

PARAMETER ESTIMATION

Estimation of the Parameters of Probability Distribution Function

The estimates b_r of β_r can be calculated by the following formulas:

$$b_0 = \bar{x}$$

$$b_1 = \sum_{j=1}^{N-1} \frac{(N-j)x_j}{N(N-1)}$$

$$b_2 = \sum_{j=1}^{N-2} \frac{(N-j)(N-j-1)x_j}{N(N-1)(N-2)}$$

where x_j is the j-th element of the ordered sample where the elements are ranked in the decreasing order ($x_1 \geq x_2 \geq \ldots \geq x_N$).

The parameters of a probability function can be estimated using the equations relating the parameters to the first few PWMs.