Statistics revision

Dr. Inna Namestnikova

inna.namestnikova@brunel.ac.uk

Introduction

Statistics is the science of collecting, analyzing and drawing conclusions from data.

Statistics

Descriptive **Inferential**

Descriptive statistics

Descriptive statistics:
Numerical, graphical and tabular methods for organizing and summarizing data.

- Organizing and summarizing the information.
- Compilation and presentation of data in effective meaningful forms.
- Tables, diagrams, graphs and numerical summaries allow increased understanding and provide an effective way to present data.

The object for research

- The entire collection of individuals or objects about which information is desired or required called the **population** of interest.
- A **sample** is a subset of the **population**, selected for study in some prescribed manner or a part of the population selected for study.

Inferential statistics

- Inferential statistics are used to draw inferences about a population from a sample.
  - We run the risk of an incorrect conclusion about the population will be reached on the basis of incomplete information.
  - There are two main methods used in inferential statistics
    - estimation
    - hypothesis testing

Types of data

- **Data**
  - **Categorical**
    - **Nominal**
    - **Ordinal**
  - **Numerical**
    - **Discrete**
    - **Continuous**

Types of data

- **Discrete numerical data** possible values are isolated points along the number line
- **Continuous numerical data** possible values form an interval along the number line
- **Nominal categorial data** are unordered data
- **Ordinal categorial data** are ordered data. All values or observations can be ranked or have a rating scale attached.

Coding data

The first step in analysing a questionnaire or any categorial data is to code responses to each question.

Where categorial data are used in a quantitative study, coding is employed to allow the researcher to count the occurrence of a given phenomena within the sample selected.

Question types

- Multiple choice questions
  - Single response
    Example: what age are you (please tick relevant category)
  - Multiple response
    Example: what is your normal mode of transport when coming to Brunel university (please tick those that apply)
    Bus, Train, Car, Walking.
- Likert scale questions
  The respondent indicates the amount of agreement or disagreement with issue.
  Example: Lecturers are nice people. We may have 5 points ranging from strongly agree to strongly disagree
- Free answer
- Combination question
  Example: what is your normal mode of transport when coming to Brunel university
  Bus, Train, Car, Walking, Other (please specify)

Evaluation Form

The information collected in this evaluation will be kept strictly confidential and no information will be passed to any Schools or course leaders.

About you

Name:						
Gender:	Male			Female		
Student Number:						
Brunel Email Address:						
Previous Maths Grade	GCSE:		AS:		A Level:	

School (circle one):	Arts	Busines s	Law	Eng & Desig n	Health Sciences and Social Care	ISCM	Social Sciences	Sport & Educati on
Level:	Foundation	L1	L2	L3	PG			

Please state your course (e.g. economics)

Please state/describe the maths problem you would like help with

Feed back about us

How useful did you find the advice/support given: (please circle one)

Very useful	**Useful**	**Undecided**	**Not useful**	**Not very useful**

How could the café be improved?

Any other comments

Evaluation Form (partly coded)

The information collected in this evaluation will be kept strictly confidential and no information will be passed to any Schools or course leaders.

About you

Name:						
Gender:	Male 0			Female 1		
Student Number:						
Brunel Email Address:						
Previous Maths Grade	GCSE:	1	AS:	2	A Level:	3

School (circle one):	Arts 1	Business 2	Law 3	Eng & Design 4	Health Sciences and Social Care 5	ISCM 6	Social Sciences 7	Sport & Education 8
Level:	Foundation 1	L1 2	L2 3	L3 4	PG 5			

Please state your course (e.g. economics)

Please state/describe the maths problem you would like help with

Feed back about us

How useful did you find the advice/support given: (please circle one)

Very useful	Useful	undecided	Not useful	Not very useful
-2	-1	0	1	2

How could the café be improved?

Any other comments

Frequency

- The **frequency** for particular category is the number of times the category appears in the data set.
- The **relative frequency** for particular category is the fraction or proportion of the time that the category appears in the data set.
  - It is calculated as

$$\boxed{\text{Relative frequency} = \frac{\text{frequency}}{\text{total number of observation in the data set}}}$$

Frequency distribution

- A **frequency table** or **frequency distribution** is a way of summarizing a set of data.
- It is a record of **how often each value (or set of values) of the variable in question occurs**. The table displays the possible categories along with the associated frequencies or relative frequencies.
- A frequency table can be used to summarize all types of data.
- When the table includes relative frequencies, it is sometimes referred to as a **relative frequency distribution**.

Example 1

The reasons that college seniors leave their college programs before graduating were examined. Forty two college seniors at a large American University who dropped out prior to graduation were interviewed and asked the main reason of leave. The results are given in the table below.

Reason for leaving the University	**Code**	**Frequency**
Academic problems	1	7
Poor advising or teaching	2	3
Needed a break	3	2
Economic reasons	4	11
Family responsibilities	5	4
To attend another school	6	9
Personal problems	7	3
Other	8	3

Frequency distribution

Reason for leaving the University	**Frequency**	**Relative freq.**
Academic problems	7	0.167
Poor advising or teaching	3	0.071
Needed a break	2	0.048
Economic reasons	11	0.262
Family responsibilities	4	0.095
To attend another school	9	0.214
Personal problems	3	0.071
Other	3	0.071
Total	42	1

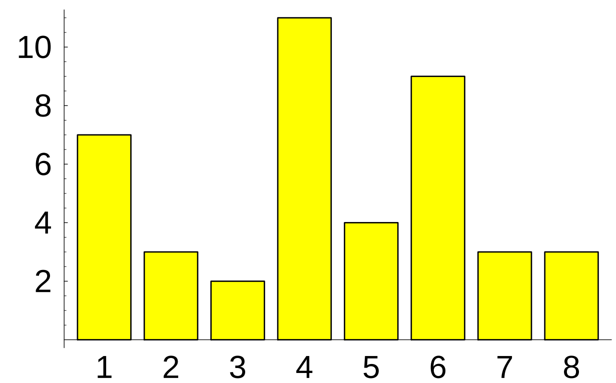
Graphs

- A **bar chart** is a graph of the frequency distribution of categorical data. Each category in the frequency distribution is presented by a bar or rectangle.
- In a **pie chart**, a circle is used to represent the whole data set with "slices" of the pie representing the possible categories.
- A **histogram** for discrete numerical data is a graph of the frequency distribution that is very similar to the bar chart for categorical data.

Bar Charts

- Draw a horizontal line, and write the category names or labels below the line at regularly spaced intervals.
- Draw a vertical line, and label the scale using either frequency or relative frequency.
- Place a rectangular bar above each category label. The hight is determined by the category's frequency or relative frequency, and all bars should have the same width. With the same width, both the height and the area of the bar are proportional to the relative frequency.
-

Reason for leaving the University	Frequency	Relative freq.
Academic problems	7	0.167
Poor advising or teaching	3	0.071
Needed a break	2	0.048
Economic reasons	11	0.262
Family responsibilities	4	0.095
To attend another school	9	0.214
Personal problems	3	0.071
Other	3	0.071



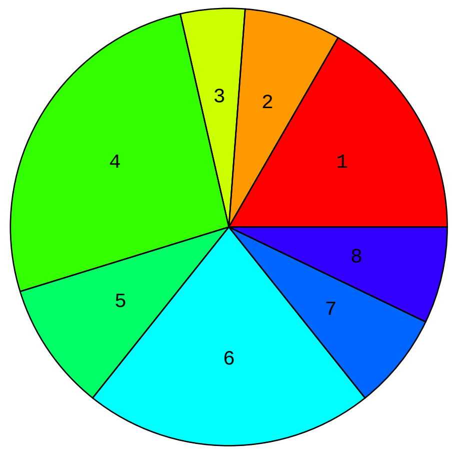
Pie Charts

- Draw a circle to represent the entire data set.
- For each category, calculate the "slice" size.

$$\boxed{\text{"slice" size} = \text{category relative frequency} \times 360^o}$$

(since there are 360 degrees in a circle)

- Draw a slice of appropriate size for each category.



-

Code	Reason for leaving the University	Frequency	Relative freq.
1	Academic problems	7	0.167
2	Poor advising or teaching	3	0.071
3	Needed a break	2	0.048
4	Economic reasons	11	0.262
5	Family responsibilities	4	0.095
6	To attend another school	9	0.214
7	Personal problems	3	0.071
8	Other	3	0.071

Discrete data set

We can

- Display the data in tabular form.
- Provide suitable statistical chart(s)/diagram(s) to summarize and present the data.
- Calculate suitable statistics to describe the data.
- Comment on their interpretation.

Mode and Median

- The **mode** is the most frequently occurring value in a set of discrete data.
  There can be more than one mode if two or more values are equally common.
- The **median** is the value halfway through the ordered data set, below and above which there lies an equal number of data values.

Median

$$\mathbf{2,\ 3,\ |5|, 6,\ 7}$$

The **median** (middle score) is **5**.

$$\mathbf{2,\ 3,\ 5, ||\ 6,\ 7,\ 9}$$

The **median** (middle score) is $\mathbf{\frac{5+6}{2}}$ =**5.5**.

Mode and Median

Suppose the results of an end of term Statistics exam were distributed as follows:

Student	1	2	3	4	5	6	7	8	9
Score	94	81	56	**90**	70	65	**90**	**90**	30
Ordered Score	30	56	65	70	**81**	90	90	90	94

Then the **mode** (most common score) is **90**.
The **median** (middle score) is **81**.

Box and Whisker Plots

Box Plots is a way of summarising data based on the median and interquartile range which contains 50% of the value.

Example: For the following data set construct a box plot

$$9, 3, 3, 4, 11, 7, 2, 3$$

Ordered data: $2, 3, | 3, 3, 4, 7, | 9, 11$

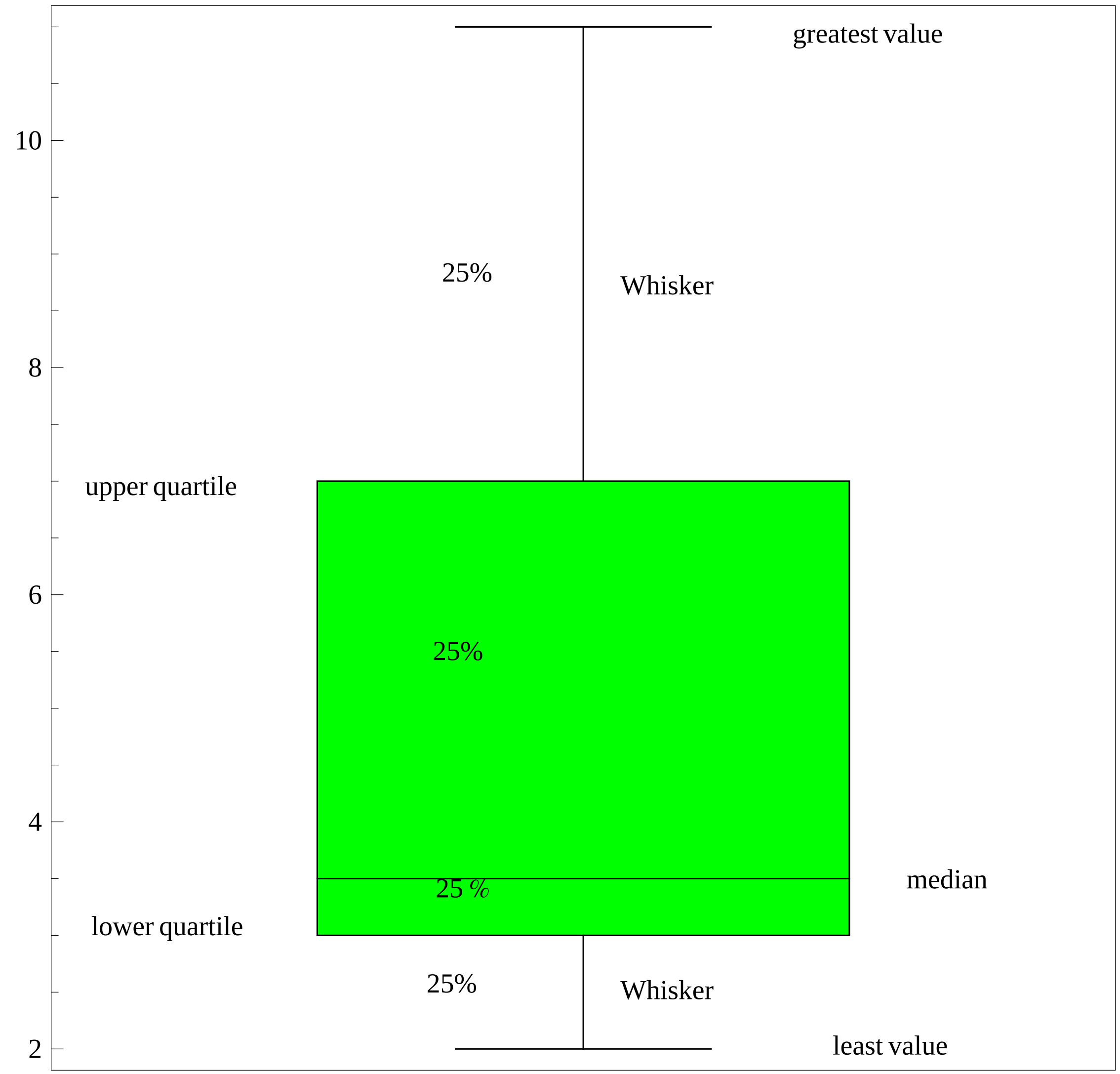
Lower Quartile Q_2 is at

$$\frac{n}{4} = \frac{8}{4} = 2, \quad \Rightarrow \quad Q_2 = 3$$

Upper Quartile Q_3 is at

$$3 \times \frac{n}{4} = 3 \times \frac{8}{4} = 6, \quad \Rightarrow \quad Q_2 = 7$$

Box and Whisker Plots



Example 2 (discrete data set)

In a survey of the size of families in a certain neighbourhood the following set of data of the number of persons in each family was obtained

$$\{2, 2, 5, 6, 3, 3, 7, 4, 7, 5, 2, 2, 2, 4, 3, 5, 9\}$$

A table of frequency and relative frequency distribution of family size is constructed.

Example 2 (discrete data set)

Family size	Tally	Frequency	Cumulative freq.	Relative freq.
2	𝍸	5	5	0.294
3	𝍷𝍷𝍷	3	8	0.176
4	𝍷𝍷	2	10	0.118
5	𝍷𝍷𝍷	3	13	0.176
6	𝍷	1	14	0.059
7	𝍷𝍷	2	16	0.118
8		0	16	0
9	𝍷	1	17	0.059
Total		17		**1**

Example 2 (discrete data set)

Data set

$$\{2, 2, 5, 6, 3, 3, 7, 4, 7, 5, 2, 2, 2, 4, 3, 5, 9\}$$

Ordered data set

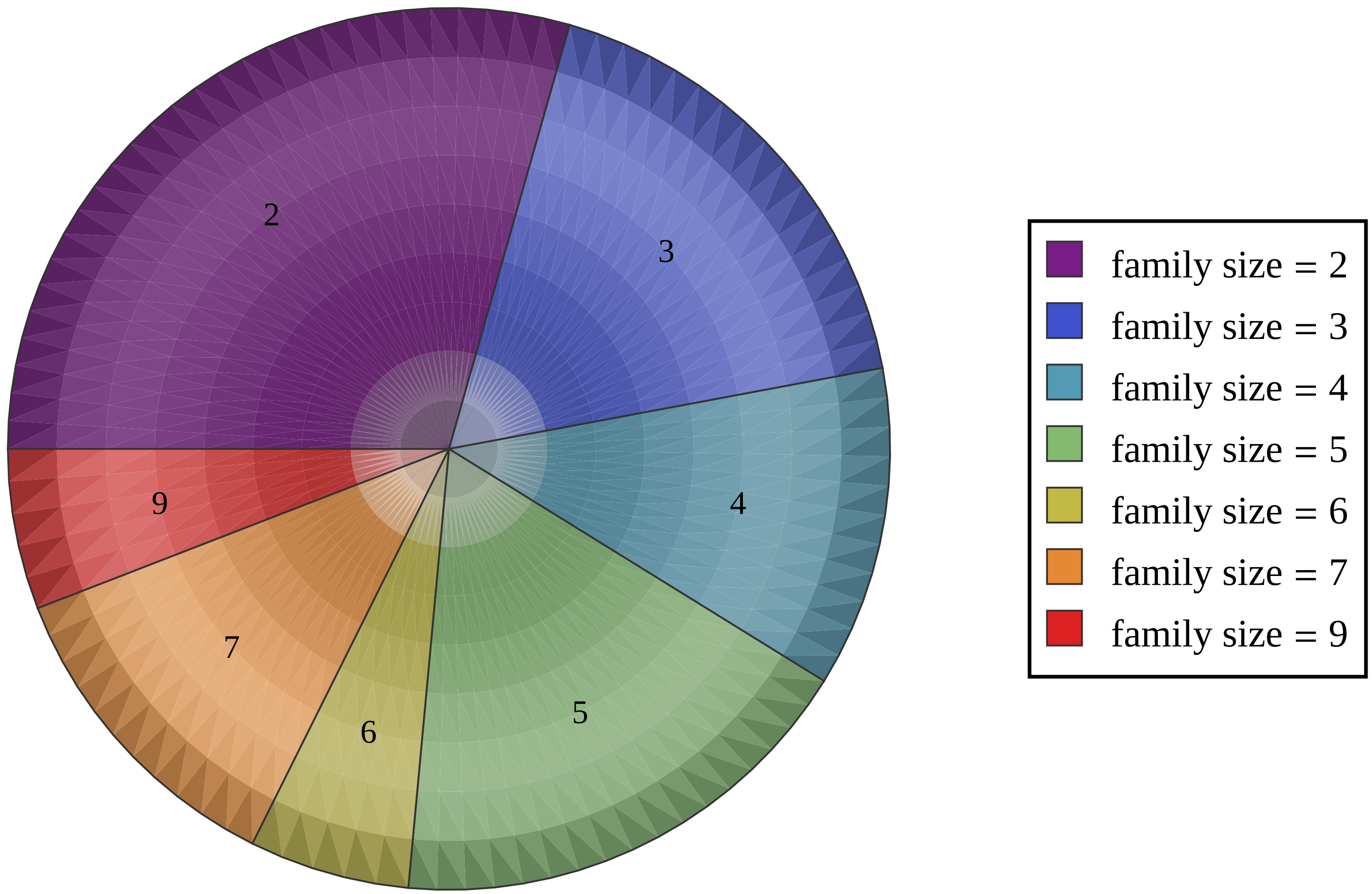
$$\{2, 2, 2, 2, 2, 3, 3, 3, \ \mathbf{4}, 4, 5, 5, 5, 6, 7, 7, 9\}$$

Mode is 2

Median is 4

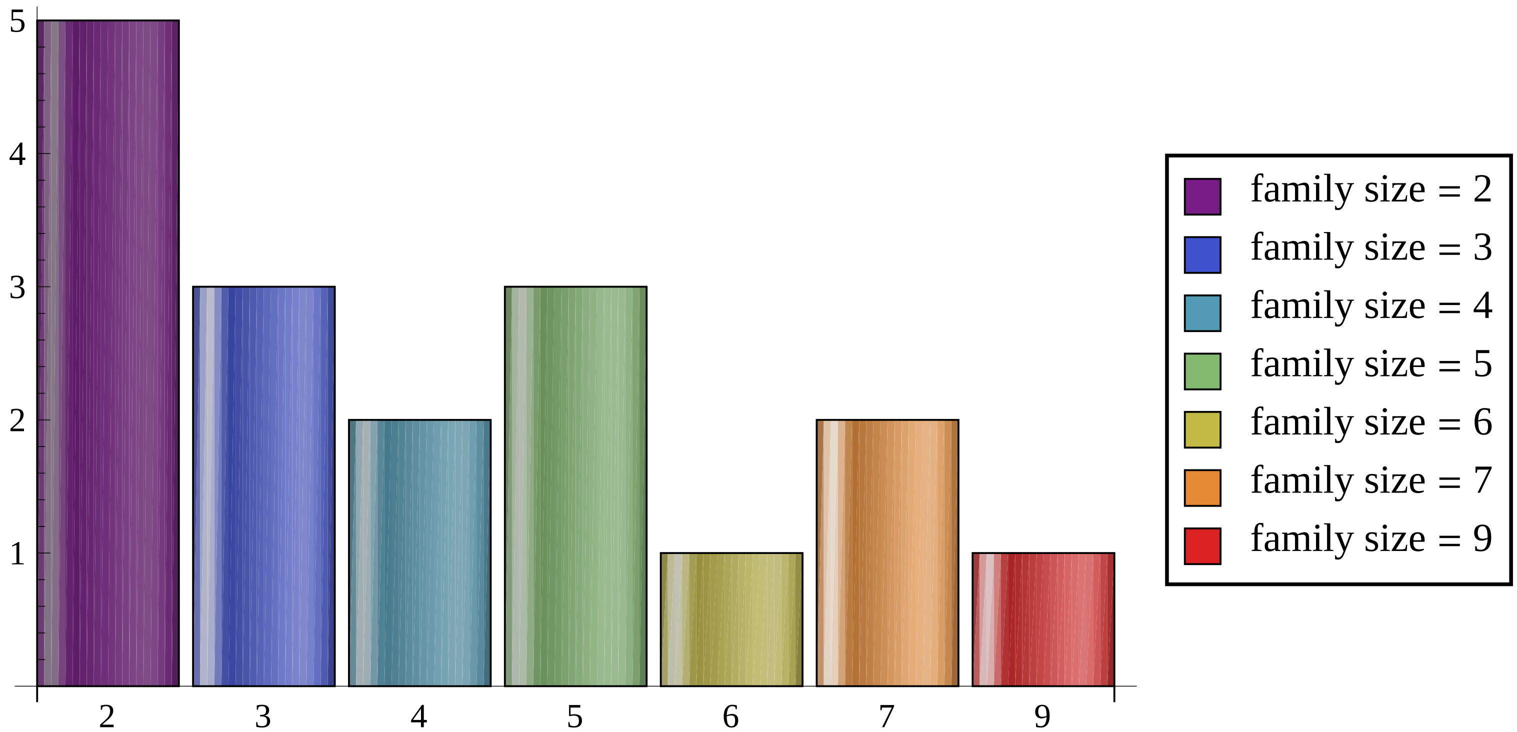
Pie Chart

Out[18]=

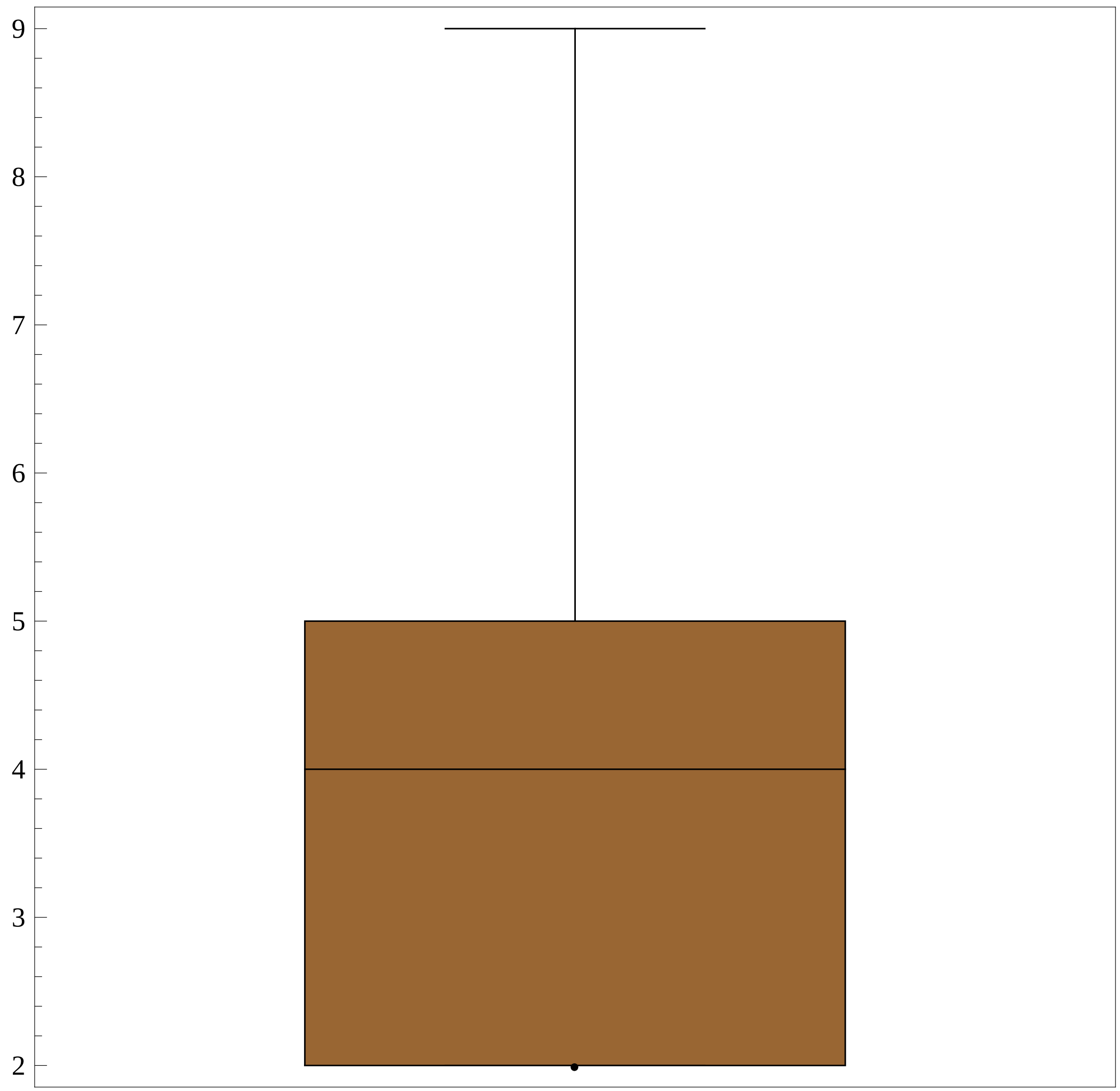


Bar Chart

Out[32]=



Box and Whisker Plots



Example 3 (discrete data set)

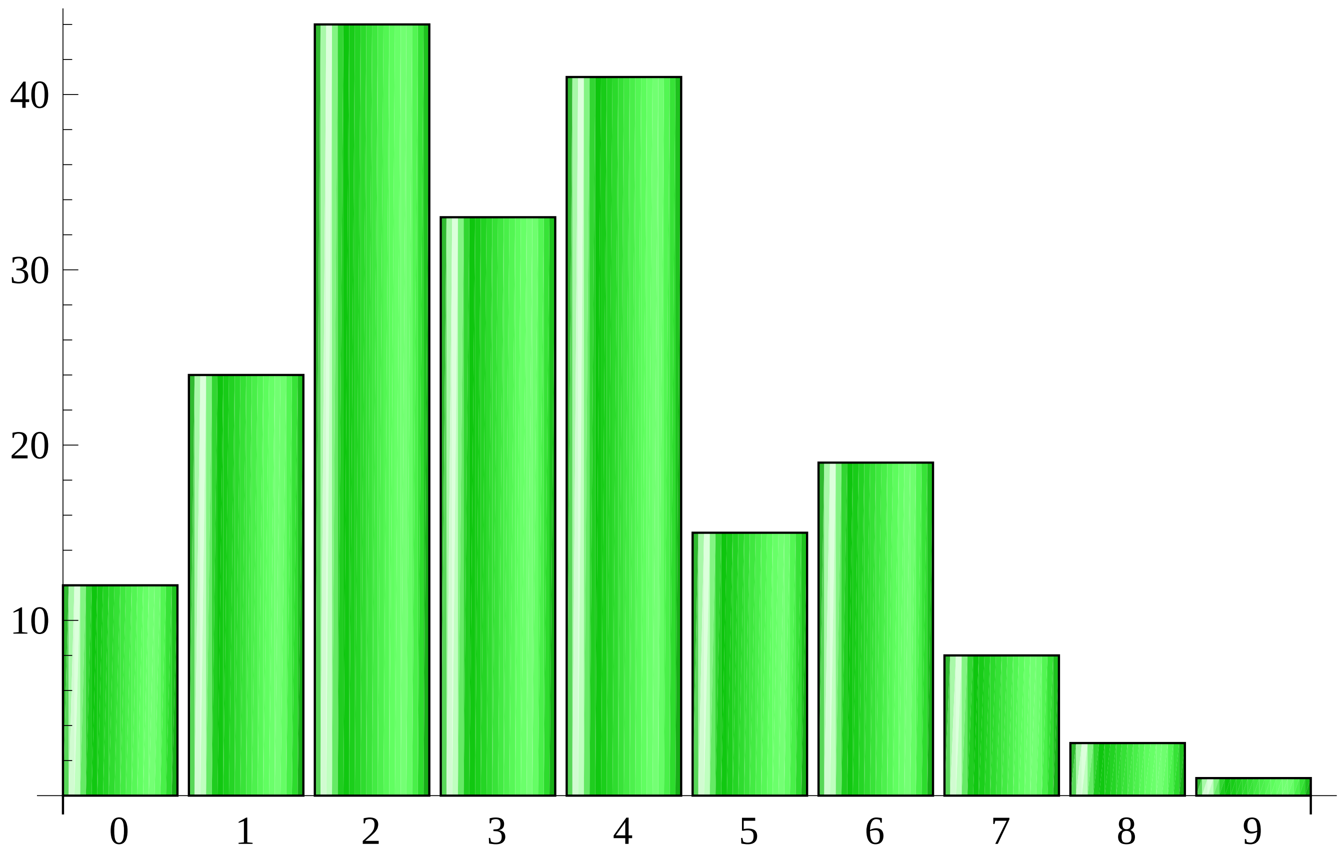
The data represent the number of accident claims per day processed by a certain insurance company on a random sample of 200 days.

3	3	2	5	6	2	2	7	2	1	4	5	5	6	6	1	4	2	4	2
1	3	3	6	4	6	2	0	4	4	6	1	3	4	2	2	4	4	2	1
0	3	3	6	6	1	1	0	2	1	5	9	3	3	6	6	8	5	4	4
2	1	3	3	2	4	5	4	3	3	5	4	2	3	6	4	4	7	7	4
4	1	2	7	2	0	5	2	0	2	8	4	3	4	2	1	3	2	2	3
4	2	2	4	6	2	0	4	3	2	2	3	3	5	2	4	6	1	0	4
3	4	4	2	5	2	3	3	6	1	3	4	2	6	2	2	5	1	7	3
5	0	6	7	2	2	2	4	3	0	4	2	3	6	2	4	2	0	1	2
2	6	1	4	3	6	2	5	1	3	1	0	4	3	2	4	1	4	8	1
7	4	5	4	4	4	4	7	1	5	3	1	0	2	3	1	2	4	1	3

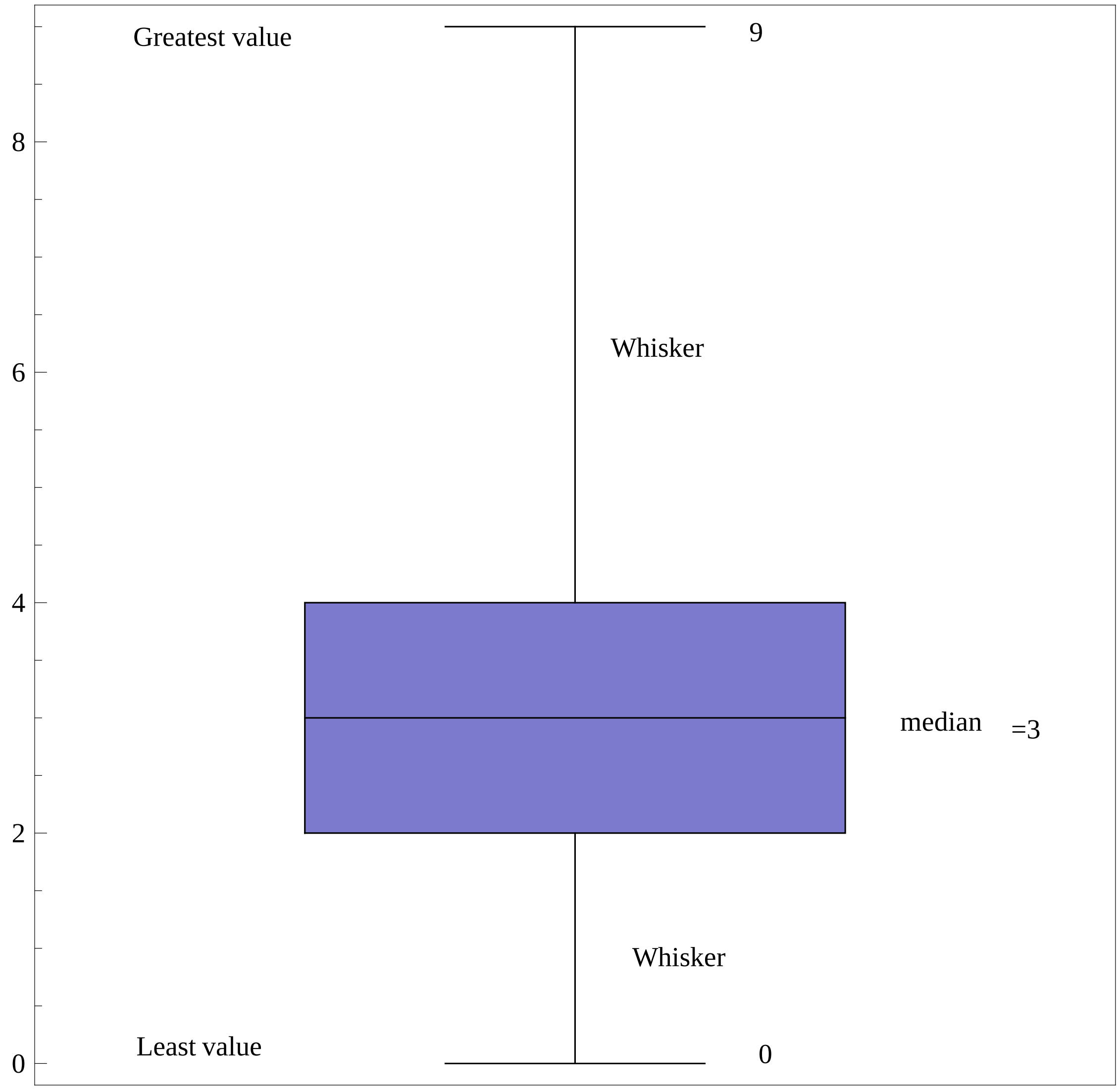
Frequency Table

Number	Frequency	Relative freq.
0	12	0.060
1	24	0.120
2	44	0.220
3	33	0.165
4	41	0.205
5	15	0.075
6	19	0.095
7	8	0.040
8	3	0.015
9	1	0.005
Total	**200**	**1**

Bar Chart



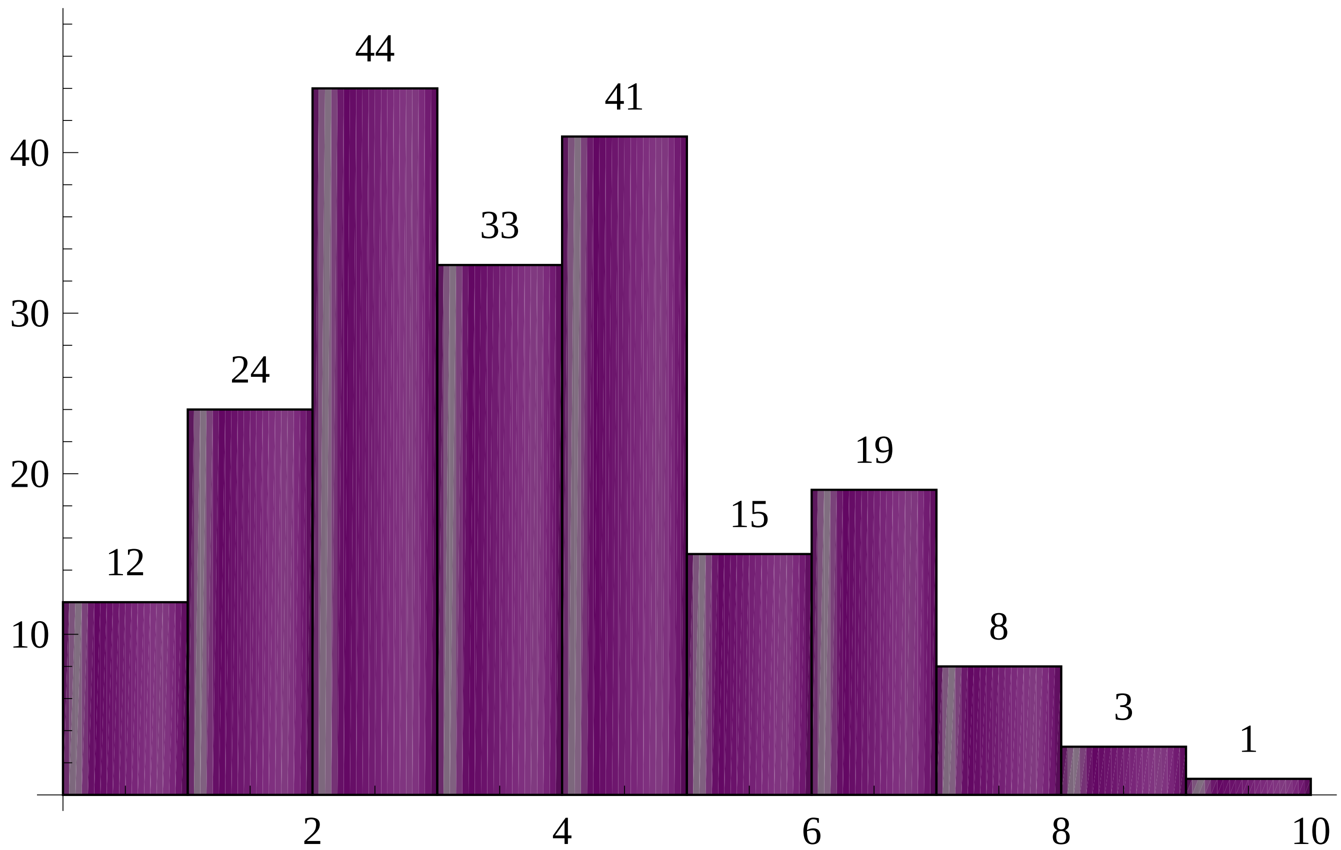
Box and Whisker Plots



Histogram

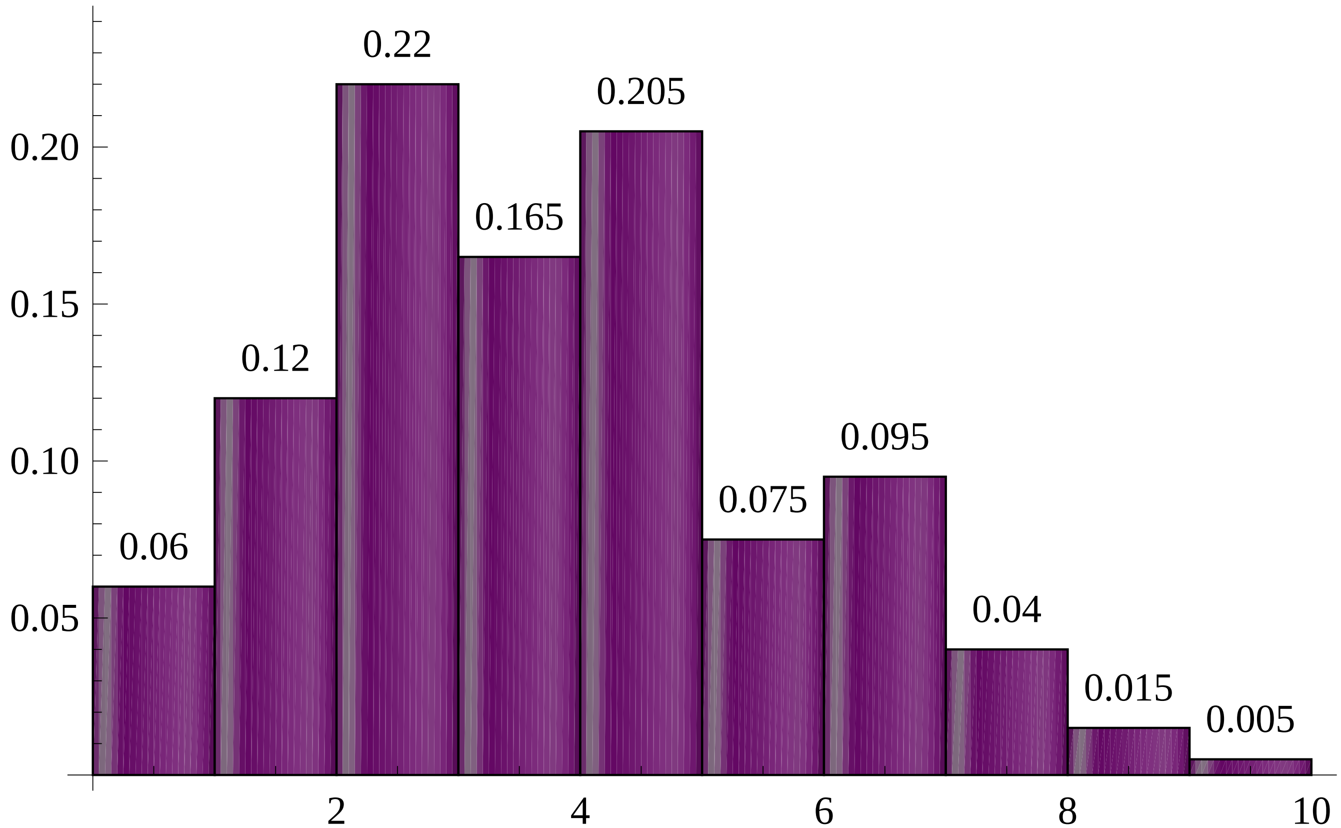
- A **histogram** is a way of summarising data that are measured on an **interval scale** (either discrete or continuous).
- It divides up the range of possible values in a data set into **classes or groups**.
- The histogram is only appropriate for variables whose values are numerical and measured on an interval scale. It is generally used when dealing with **large data sets**
- A histogram can also help detect any unusual observations (outliers), or any gaps in the data set.

Histogram



Intervals: $0 \leqslant x < 1, \quad 1 \leqslant x < 2, \quad 2 \leqslant x < 3, \quad 3 \leqslant x < 4, \quad 4 \leqslant x < 5, \ldots$
Class mid-points: $0.5, \quad 1.5, \quad 2.5, \quad 3.5, \quad 4.5, \quad 5.5, \quad 6.5, \ldots$

Histogram


0.22
0.205
0.165
0.12
0.095
0.075
0.06
0.04
0.015
0.005
0.20
0.15
0.10
0.05
2
4
6
8
10

Sample Mean

- The **sample mean** is the sum of all the observations divided by the total number of observations.
- It is a measure of location, commonly called the **average**
- The **sample mean** is an estimator available for estimating the population mean.
- For **sample** $\{x_1, x_2, x_3, ..., x_n\}$ with observed frequencies $\{f_1, f_2, f_3, ..., f_n\}$, the **sample mean** $\bar{x}$ can be calculated by

$$\boxed{\bar{x} = \frac{\sum_i x_i}{n} = \frac{\sum_i f_i x_i}{\sum_i f_i}}$$

Example 2: Frequency Table

Family size	Frequency	Relative freq.	
x	f		fx
2	5	0.294	10
3	3	0.176	9
4	2	0.118	8
5	3	0.176	15
6	1	0.059	6
7	2	0.118	14
8	0	0.	0
9	1	0.059	9
Total	**17**	**1**	**71**

$\bar{x} = \frac{71}{17} \approx$ **4.18**

Example 3: Frequency Table

Number of accident claims	Frequency	Relative freq.	
x	f		fx
0	12	0.060	0
1	24	0.120	24
2	44	0.220	88
3	33	0.165	99
4	41	0.205	164
5	15	0.075	75
6	19	0.095	114
7	8	0.040	56
8	3	0.015	24
9	1	0.005	9
Total	**200**	**1**	**653**

$$\bar{x} = \frac{653}{200} \approx \mathbf{3.27}$$

Sample Variance

- We can measure **dispersion** relative to the scatter of the values about their mean.

- For **data** $\{x_1, x_2, x_3, ...x_n\}$

$$\textbf{Sample variance}, \quad \sigma^2 = \frac{\sum_i x_i^2}{n} - (\bar{x})^2$$

For **frequency distribution**

x	x_1	x_2	x_3	...	x_i	...	x_n
freq	f_1	f_2	f_3	...	f_i	...	f_n

$$\textbf{Sample variance}, \quad \sigma^2 = \frac{\sum_i f_i x_i^2}{\sum_i f_i} - (\bar{x})^2$$

Sample Standard Deviation

- **Standard deviation** is a measure of the spread or dispersion of a set of data.
- The more widely the values are spread out, the larger the standard deviation is.
- For **data** $\{x_1, x_2, x_3, ...x_n\}$

$$\boxed{\textbf{Standard Deviation}, \quad \sigma = \sqrt{\frac{\sum_i x_i^2}{n} - (\bar{x})^2}}$$

- For **frequency distribution**

x	x_1	x_2	x_3	...	x_i	...	x_n
freq	f_1	f_2	f_3	...	f_i	...	f_n

$$\boxed{\textbf{Standard Deviation}, \quad \sigma = \sqrt{\frac{\sum_i f_i x_i^2}{\sum_i f_i} - (\bar{x})^2}}$$

Example 2: Frequency Table

Family size	Frequency	Relative freq.		
x	f		fx	fx^2
2	5	0.294	10	20
3	3	0.176	9	27
4	2	0.118	8	32
5	3	0.176	15	75
6	1	0.059	6	36
7	2	0.118	14	58
8	0	0.	0	0
9	1	0.059	9	81
Total	**17**	**1**	**71**	**369**

$\bar{x} = \frac{71}{17} \approx$ **4.18**

$$\sigma = \sqrt{\frac{369}{17} - (4.18)^2} \approx \mathbf{2.1}$$

Example 3: Frequency Table

Number of accident claims	Frequency	Relative freq.		
x	f		fx	fx^2
0	12	0.060	0	0
1	24	0.120	24	24
2	44	0.220	88	176
3	33	0.165	99	297
4	41	0.205	164	656
5	15	0.075	75	375
6	19	0.095	114	684
7	8	0.040	56	392
8	3	0.015	24	192
9	1	0.005	9	81
Total	**200**	**1**	**653**	**2877**

$\bar{x} \approx$ **3.27** $\qquad \sigma = \sqrt{\frac{2877}{200} - (3.27)^2} \approx$ **1.92**

Example 4 (continuous data set)

The concentration of suspended solids in the river water is an important environmental characteristics. In a paper reported on concentration (in parts per million, or ppm) for several different rivers. Suppose that the accompanying 50 observations had been obtained for a particular river.

55.80	60.90	37.00	91.30	65.80	42.30	33.80	60.60	76.00	69.00
45.90	39.10	35.50	56.00	44.60	71.70	61.20	61.50	47.20	74.50
83.20	40.00	31.70	36.70	62.30	47.30	**94.60**	56.30	30.00	68.20
75.30	71.40	65.20	52.60	58.20	48.00	61.80	78.80	39.80	65.00
60.70	77.10	59.10	49.50	69.30	69.80	64.90	**27.10**	66.30	87.10

$$\text{Mean} = \frac{\mathbf{55.8 + 45.9 + 83.2 + ... + 65 + 87.1}}{\mathbf{50}} = \mathbf{58.5}$$

Class intervals

- **maximum value** $= 94.6$
- **minimum value** $= 27.1$
- **Class intervals**

$$\begin{array}{lll} 20 \leqslant x < 30, & 30 \leqslant x < 40, & 40 \leqslant x < 50, \\ 50 \leqslant x < 60, & 60 \leqslant x < 70, & 70 \leqslant x < 80, \\ 80 \leqslant x < 90, & 90 \leqslant x < 100 & \end{array}$$

- Use **class mid-points** as **estimates of the class means**

$$25,\ 35,\ 45,\ 55,\ 65,\ 75,\ 85,\ 95$$

Frequency Table

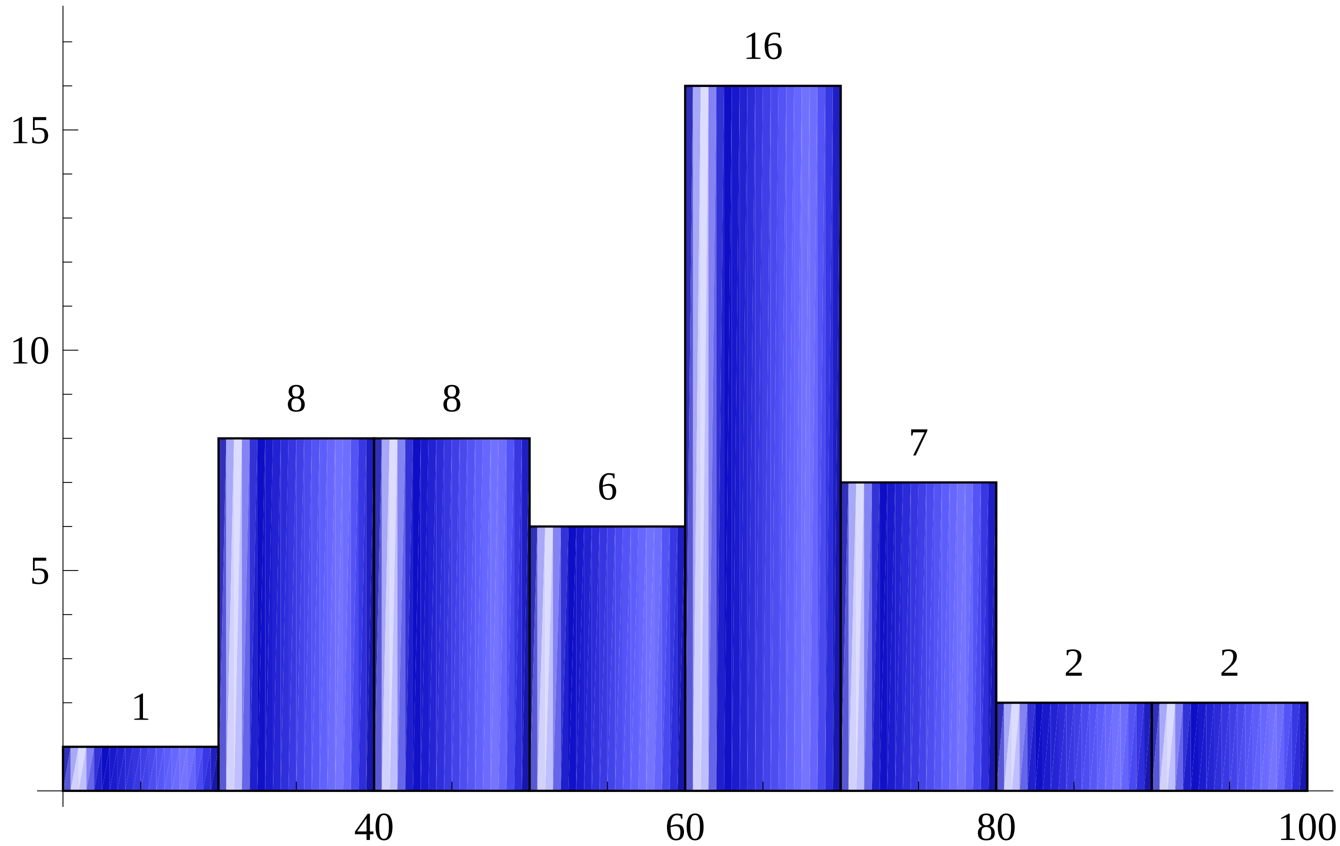
Concentration	Tally	Frequency	Relative freq.
$20 \leqslant x < 30$	𝍷	1	0.02
$30 \leqslant x < 40$	𝍸 𝍷𝍷𝍷	8	0.16
$40 \leqslant x < 50$	𝍸 𝍷𝍷𝍷	8	0.16
$50 \leqslant x < 60$	𝍸 𝍷	6	0.12
$60 \leqslant x < 70$	𝍸 𝍸 𝍸 𝍷	16	0.32
$70 \leqslant x < 80$	𝍸 𝍷𝍷	7	0.14
$80 \leqslant x < 90$	𝍷𝍷	2	0.04
$90 \leqslant x < 100$	𝍷𝍷	2	0.04
Class intervals	Total	50	1

Frequency Table

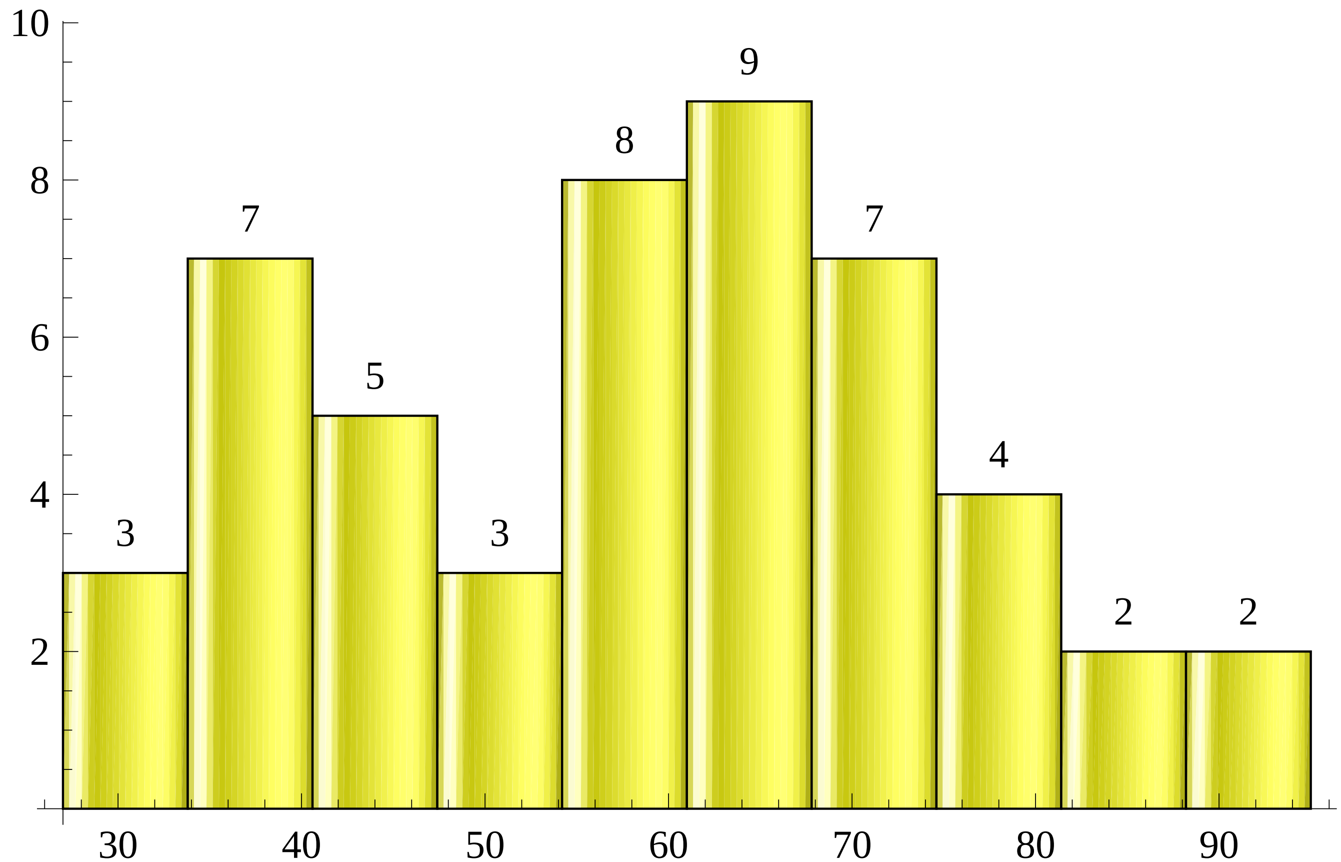
class mid-points	Frequency	
x	f	fx
25	1	25
35	8	280
45	8	360
55	6	330
65	16	1040
75	7	525
85	2	170
95	2	190
Total	**50**	**2920**

hence $\bar{x} = \frac{2920}{50} =$ **58.4**

Histogram



Histogram



Example 5 (continuous data set)

Data were collected on the blood glucose (in mmol/l) measured in the blood of **100** subjects during a research study at a certain nutrition department.

3.27792	3.37444	4.97057	4.02437	4.40855	4.69663	3.34397	5.22305	3.55060	2.98057
5.81152	4.58240	5.08875	4.04497	3.87288	4.67210	4.90091	4.31757	5.20679	3.25989
3.90416	5.37304	4.64384	4.38037	3.94797	2.76160	6.02717	5.29289	2.84805	4.780400
4.11426	3.73694	5.20243	1.79561	3.71626	3.24735	5.51044	3.26583	4.46252	5.460610
5.48467	3.60436	2.98056	5.53549	3.89788	4.14706	2.96069	5.37283	5.05862	3.67263
3.25160	6.63551	3.18142	5.22402	3.37358	3.15472	3.21479	3.44678	4.93306	4.31728
4.14319	1.77422	4.25183	2.84643	4.89365	3.56778	3.23527	6.17919	4.35063	5.11706
4.85987	4.20730	2.88155	5.59583	3.94908	4.02062	5.03695	4.35373	5.44498	4.20769
3.53962	5.20128	5.23739	4.37652	3.65423	3.42377	4.31031	5.73569	4.61766	3.85986
5.74499	3.64311	2.21657	3.69019	5.70689	4.24800	4.63107	4.74557	3.68453	5.15948

Class intervals

- **maximum value** $= 6.63551$
- **minimum value** $= 1.77422$
- **Class intervals**

$$
\begin{array}{lll}
1.5 \leqslant x < 2.0, & 2.0 \leqslant x < 2.5, & 2.5 \leqslant x < 3.0, \\
3.0 \leqslant x < 3.5, & 3.5 \leqslant x < 4.0, & 4.0 \leqslant x < 4.5, \\
4.5 \leqslant x < 5.0, & 5.0 \leqslant x < 5.5, & 5.5 \leqslant x < 6.0, \\
6.0 \leqslant x < 6.5, & 6.5 \leqslant x < 7.0 &
\end{array}
$$

- Use **class mid-points** as **estimates of the class means**

$$
\begin{array}{l}
1.75, 2.25, 2.75, 3.25, 3.75, 4.25, \\
4.75, 5.25, 5.75, 6.25, 6.75
\end{array}
$$

Class intervals

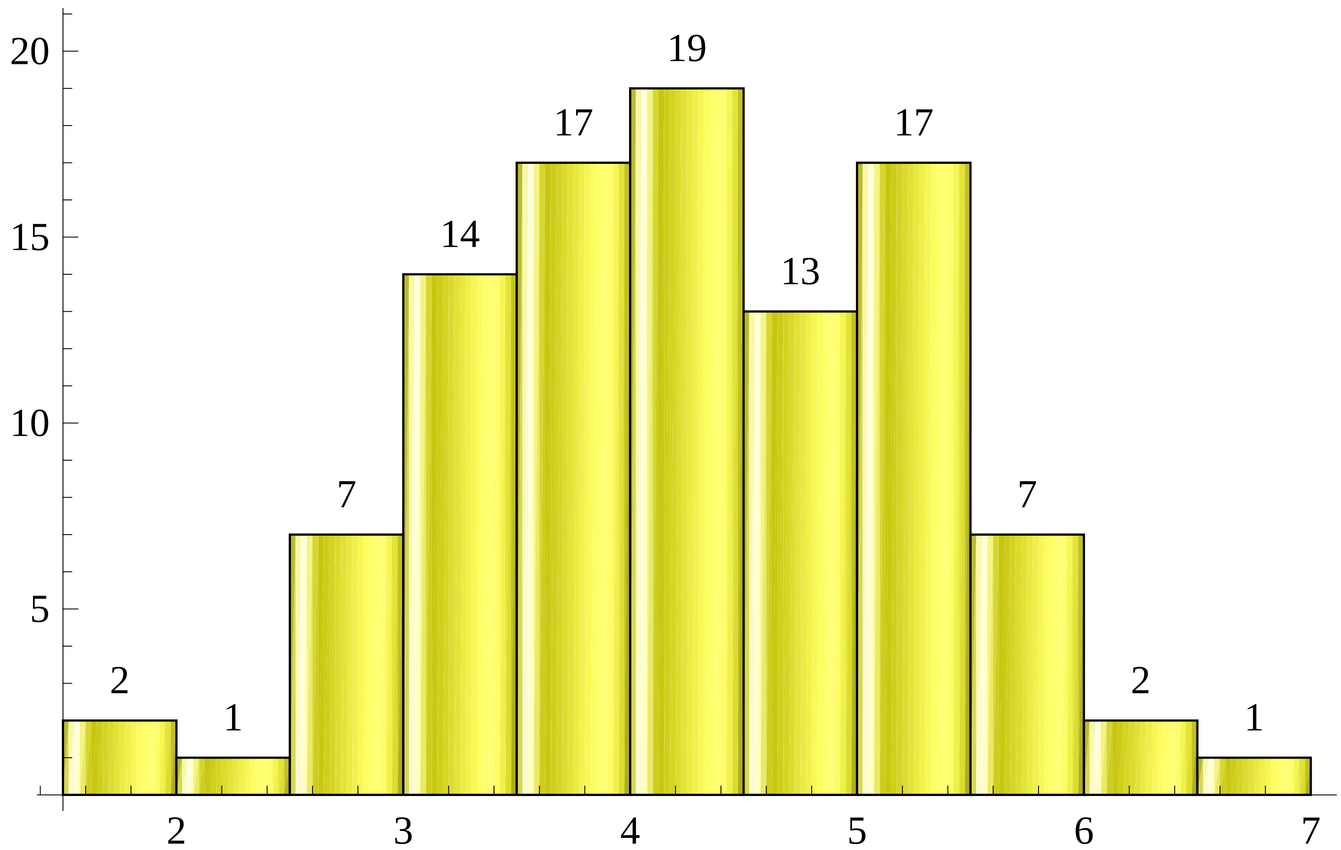
$$\begin{aligned}
1.5 \leqslant x < 2.0, &\quad 1.79561, 1.77422\\
2.0 \leqslant x < 2.5, &\quad 2.21657\\
2.5 \leqslant x < 3.0, &\quad 2.98057, 2.76160, 2.84805, 2.98056,\\
&\quad 2.96069, 2.84643, 2.88155\\
3.0 \leqslant x < 3.5, &\quad 3.27792, 3.37444, 3.34397,\\
&\quad 3.25989, 3.24735, 3.26583,\\
&\quad 3.2516, 3.18142, 3.37358,\\
&\quad 3.15472, 3.21479, 3.44678,\\
&\quad 3.23527, 3.42377
\end{aligned}$$

Frequency Table

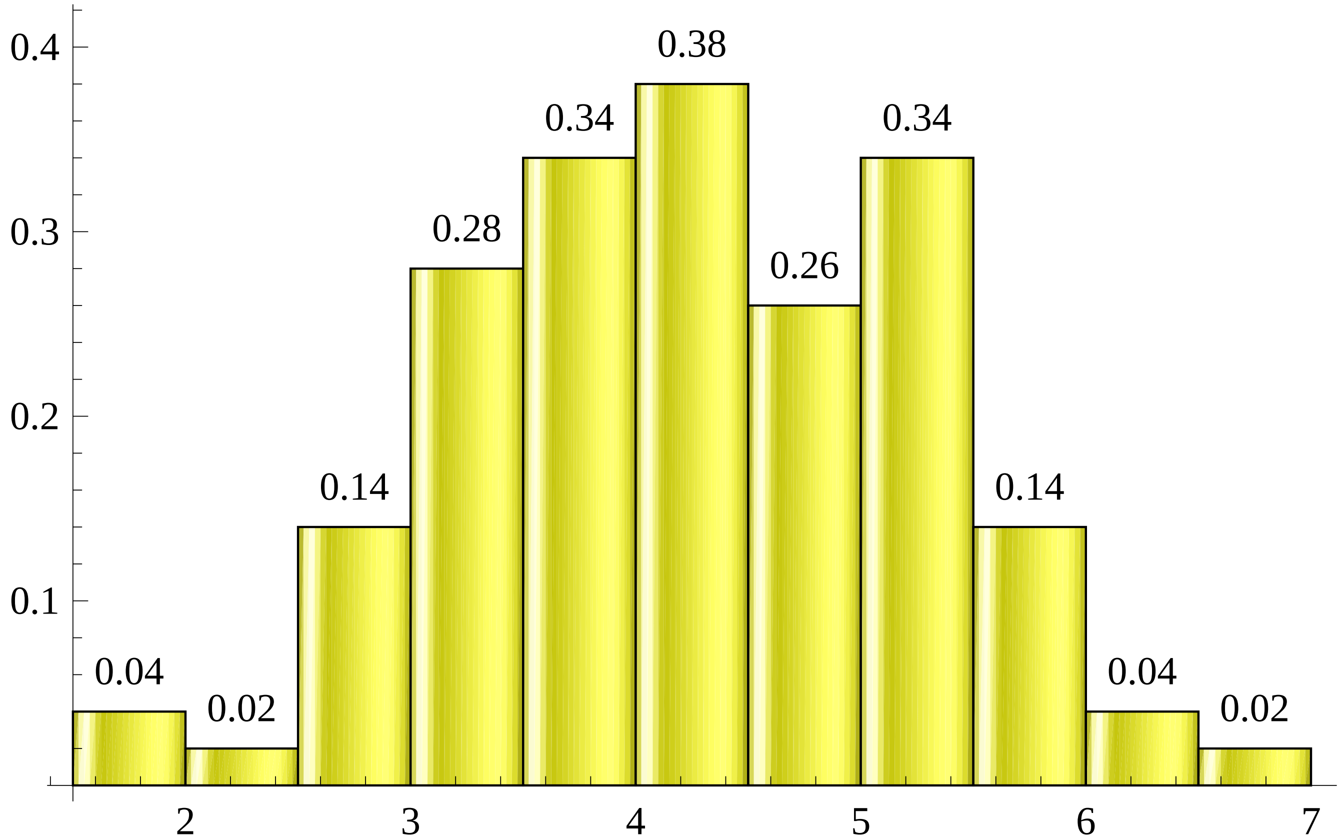
class mid-points	Frequency	
x	f	fx
1.75	2	3.5
2.25	1	2.25
2.75	7	19.25
3.25	14	45.5
3.75	17	63.75
4.25	19	80.75
4.75	13	61.75
5.25	17	89.25
5.75	7	40.25
6.25	2	12.5
6.75	1	6.75
Total	**100**	**425.5**

hence $\bar{x} = \frac{425.5}{100} =$ **4.255**

Histogram



Histogram



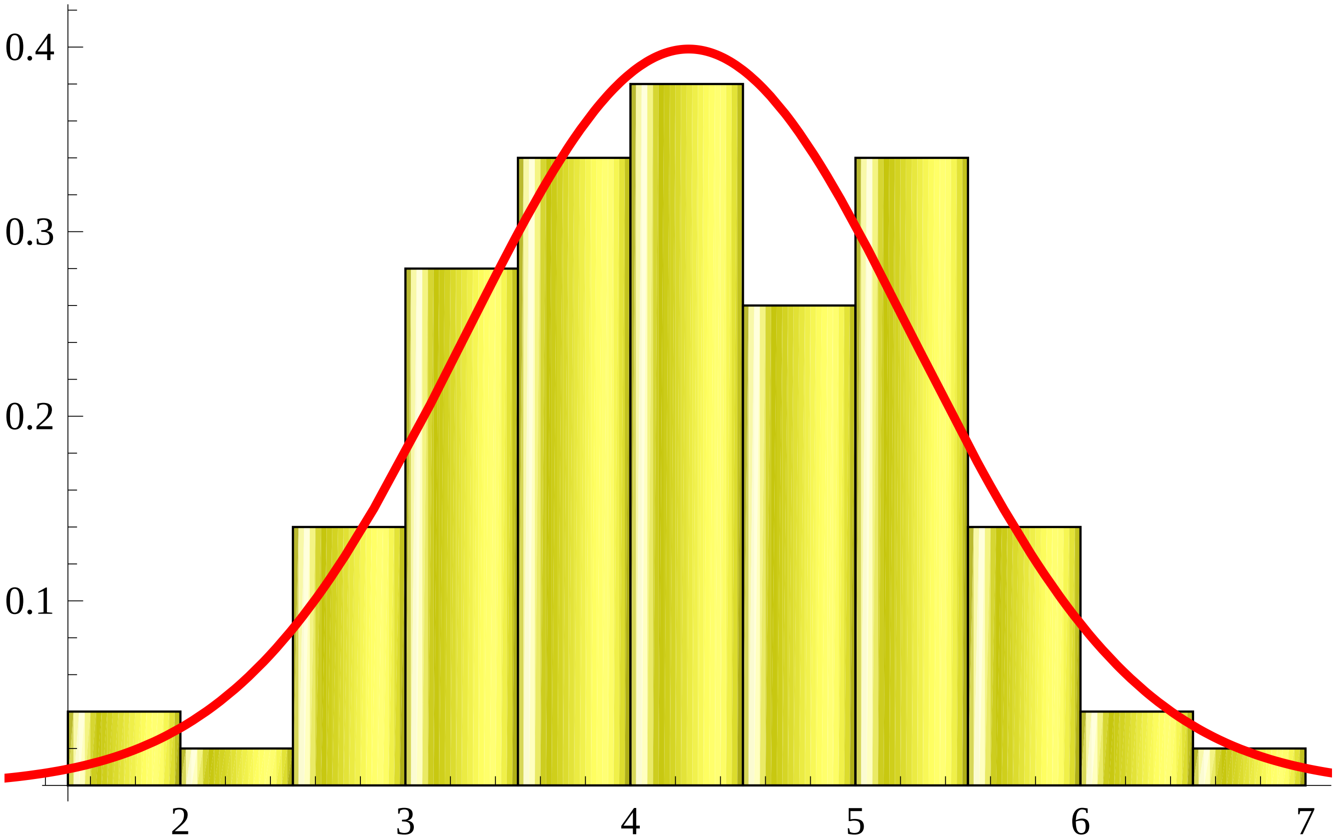
Frequency Table

class mid-points	Frequency		
x	f	fx	fx^2
1.75	2	3.5	6.125
2.25	1	2.25	5.0625
2.75	7	19.25	52.9375
3.25	14	45.5	147.875
3.75	17	63.75	239.063
4.25	19	80.75	343.188
4.75	13	61.75	293.313
5.25	17	89.25	468.563
5.75	7	40.25	231.438
6.25	2	12.5	78.125
6.75	1	6.75	45.5625
Total	**100**	**425.5**	**1911.25**

$$\bar{x} = \frac{425.5}{100} = \mathbf{4.255}$$

$$\sigma = \sqrt{\frac{1911.25}{100} - (4.255)^2} = \mathbf{1.00373}$$

Histogram and N(4.26, 1.004)



Normal distribution: mean, median and mode are identical in value.

Inferential statistics

Statistical inference

Problems of estimation | **Testing of hypothesis**

If we use the value of a statistics to estimate a population parameter, this value is a **point estimator** of the parameter.

The statistic, whose value is used as the point estimate of a parameter, is called an **estimator**.

$$\bar{x}\text{(sample)} \Rightarrow \mu\text{(population)}$$

$$s\text{(sample)} \Rightarrow \sigma\text{(population)}$$

Point and interval estimators

Estimator

Point estimator
(one number)

Interval estimator
(two numbers)

A statistic $\hat{\theta}$ is an **unbiased estimator** of the parameter θ if the expected value of an estimator equals to the parameter which it is supposed to estimate

$$E[\hat{\theta}] = \theta$$

Confidence interval

Based on the sampling distribution of θ we can assert with a given **probability** whether such an interval will actually contain the parameter it is supposed to estimate,

$$\boxed{P(\bar{\theta}_1 < \theta < \bar{\theta}_2) = \gamma}$$

Such an interval $\bar{\theta}_1 < \theta < \bar{\theta}_2$, computed for a particular sample, is called a **confidence interval**.

The number γ is the **confidence coefficient** or **degree of confidence**.

$\bar{\theta}_1$ is **lower confidence limit**;
$\bar{\theta}_2$ is **upper confidence limit**;

Confidence Interval for Population Mean

The general formula for a confidence interval for a population mean μ when

- $\bar{x}$ is the sample mean from a random sample;
- s is the sample standard deviation from a random sample;
- the population distribution is normal, or the sample size n is large (generally $n \geqslant 30$);
- σ, the **population standard deviation**, is **unknown**

is

$$\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

One sample Confidence Interval for Population Mean

$$\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

where $\alpha = 1 - \gamma$ is statistical significance.

$t_{\alpha/2,n-1}$ critical value of **Student distribution**, is based on $n-1$ **degrees of freedom**.

The corresponding table gives critical values appropriate for each of the **confidence levels** $\gamma = 90\%$, 95%, and 99% ($\alpha = 10\%$, 5%, and 1%)

Example 6

A set of 25 data values has a mean of 2.3 and a standard deviation of 0.1. Calculate 99% and 95% confidence limits and compare the results.

Solution

For confidence level 95% $\quad t_{0.025,24} = 2.064$

$$\bar{x} \pm (t \quad \text{critical value}) \left(\frac{s}{\sqrt{n}} \right) =$$

$$2.064 \cdot \frac{0.1}{\sqrt{25}} = 0.04128$$

Hence the confidence limits: 2.3 ± 0.04128

Example 6

For confidence level 99% $\quad t_{0.005,24} = 2.797$

$$\bar{x} \pm (t \quad \text{critical value}) \left(\frac{s}{\sqrt{n}}\right) =$$

$$2.797 \cdot \frac{0.1}{\sqrt{25}} = 0.05594$$

Hence the confidence limits: 2.3 ± 0.05594

confidence level 95% $\quad 2.259 < \mu < 2.341$

confidence level 99% $\quad 2.224 < \mu < 2.356$

Example 7

A manufacturer wants to determine the average drying time of a new outdoor paint. If for 20 areas of equal size he obtained a mean drying time of 83.2 minutes and standard deviation of 7.3 minutes, construct a 95% confidence interval for the true mean μ.

Solution:
Substituting $\bar{x} = 83.2, \quad s = 7.3$ and $t_{0.025,19} = 2.093$ (from table for t-distribution), the 95% confidence interval for μ becomes

$$83.2 - 2.093\frac{7.3}{\sqrt{20}} < \mu < 83.2 + 2.093\frac{7.3}{\sqrt{20}}$$

or simply $79.8 < \mu < 86.6$

This means that we can assert with a 95% degree of confidence that the interval from 79.8 minutes to 86.6 minutes contains the true average drying time of the paint.

Hypothesis Testing

Hypothesis testing is used when we are testing the **validity** of some claim or theory that has been made about a population.

A **hypothesis** is simply a statement about one or more of the population parameters (e.g. mean, variance).

The purpose of hypothesis testing is to determine the validity of a hypothesis by examining a **random sample** of data taken from the population.

Statistical hypothesis

Null hypothesis H_0

Alternative hypothesis H_A

Hypothesis Testing

The **null hypothesis**, denoted by $\boldsymbol{H_0}$, is a claim about a population characteristic that is initially assumed to be **true**.

The **alternative hypothesis**, denoted by $\boldsymbol{H_A}$, is the competing claim.

$\boldsymbol{P(\theta > \theta_{critical}) = \alpha}$ - **one-sided test**
(one tailed test)

$\boldsymbol{P(\theta = \theta_{critical}) = \alpha}$ - **two-sided test**
(two tailed test)

Example 8

The mean length of time required to perform a certain task on an assembly line is 15.5 minutes. A new method is taught and after the training period, a random sample of times is taken and is found to have mean 13.5 minutes. There are three possible questions we could ask here:

1. Has the mean time changed?
2. Has the mean time increased?
3. Has the mean time decreased?

- In (1) we are testing
  $H_0 =$\{ the mean time changed\}
- In (2) we are testing
  $H_0 =$\{ the mean time increased\}
- In (3) we are testing
  $H_0 =$\{ the mean time decreased\}

In (1) we are performing a two-tailed test; in (2) and (3) we are performing a one-tailed test.

Hypothesis Testing

Statistical hypothesis

Simple hypothesis

Example:

(for normal distribution)

If σ is ***known***

$$H_0 : \bar{x} = 3$$

Composite hypothesis

Example:

(for normal distribution)

If σ is ***unknown***

$$H_0 : \bar{x} = 3$$

$$\sigma = A$$

The Structure of a Hypothesis Test

All hypothesis tests have the following components:

1. a statement of the **NULL** and **ALTERNATIVE** hypotheses;
2. a significance level, denoted by α;
3. a test statistic;
4. a rejection region;
5. calculations;
6. a conclusion.

Regression Analysis

Regression analysis is used to model and analyse numerical data consisting of values of an **independent variable** X (the variable that we fix or choose deliberately) and **dependent variable** Y.

The main purpose of finding a **relationship** is that the knowledge of the relationship may enable events to be predicted and perhaps controlled.

Correlation Coefficient

To measure the strength of the linear relationship between X and Y the **sample correlation coefficient** r is used.

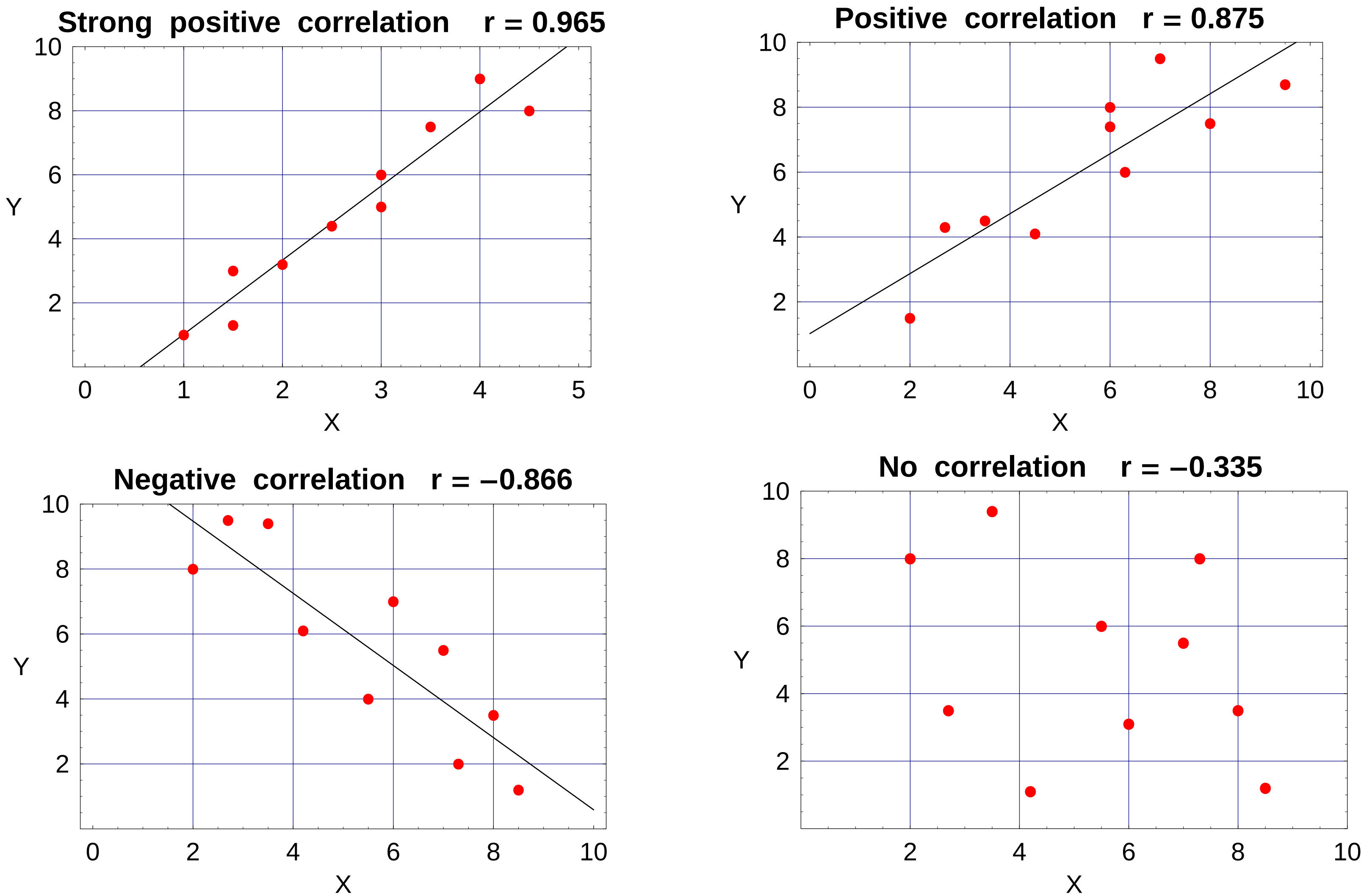
$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{xy}}},$$

$$S_{xy} = n\sum xy - \sum x \sum y,$$

$$S_{xx} = n\sum x^2 - \left(\sum x\right)^2, \quad S_{yy} = n\sum y^2 - \left(\sum y\right)^2$$

Where x and y observed values of variables X and Y respectively.

Correlation Coefficient



Linear Regression Analysis

When a scatter plot indicates that there is a strong linear relationship between two variables (confirmed by high correlation coefficient), we can fit a **straight line** to this data

This regression line may be used to predict a value of the dependent variable, given the value of the independent variable.

Linear Regression Analysis

The equation of a **regression line** is

$$y = a + bx$$

$$b = \frac{S_{xy}}{S_{xx}} \qquad a = \bar{y} - b\bar{x} = \frac{\sum_i y_i - b\sum_i x_i}{n}$$

Example 9

Suppose that we had the following results from an experiment in which we measured the growth of a cell culture (as optical density) at different pH levels.

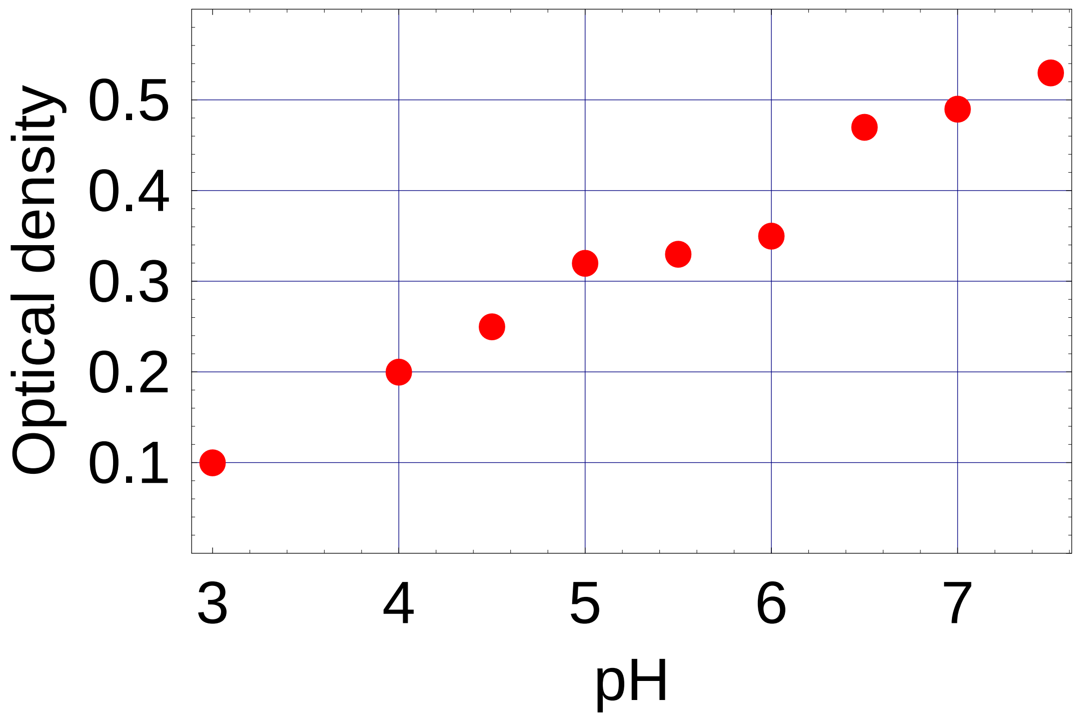
pH	3	4	4.5	5	5.5	6	6.5	7	7.5
Optical density	0.1	0.2	0.25	0.32	0.33	0.35	0.47	0.49	0.53

Find the equation to fit these data.

Solution of example 9

The data set consists of $n = 9$ observations.

Step 1. To construct the scatter diagram for the given data set to see any correlation between two sets of data.



These results suggest a **linear relationship**.

Solution of example 9

Step 2. Set out a table as follows and calculate all required values $\sum x, \quad \sum y, \quad \sum x^2, \quad \sum y^2, \quad \sum xy.$

pH (x)	Optical density(y)	x^2	y^2	xy
3	0.1	9	0.01	0.3
4	0.2	16	0.04	0.8
4.5	0.25	20.25	0.0625	1.125
5	0.32	25	0.1024	1.6
5.5	0.33	30.25	0.1089	1.815
6	0.35	36	0.1225	2.1
6.5	0.47	42.25	0.2209	3.055
7	0.49	49	0.240	3.43
7.5	0.53	56.25	0.281	3.975
$x = 49$	$y = 3.04$	$x^2 = 284$	$y^2 = 1.1882$	$xy = 18.2$
$\bar{x} = 5.444$	$\bar{y} = 0.3378$			

Solution of example 9

Step 3.
Calculate

$$S_{xy} = n\sum xy - \sum x \sum y = 9 \times 18.2 - 49 \times 3.04$$
$$= 163.8 - 148.96 = 14.84.$$

$$S_{xx} = n\sum x^2 - \left(\sum x\right)^2 = 2556 - 2401 = 155.$$

$$S_{yy} = n\sum y^2 - \left(\sum y\right)^2 = 10.696 - 9.242 = 1.454$$

Step 4.
Finally we obtain **correlation coefficient** r

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{xy}}} = \frac{14.84}{\sqrt{155 \times 1.454}} = 0.989$$

Solution of example 9

The correlation coefficient is closed to 1 therefore it is likely that the linear relationship exists between the two variables. To verify the correlation r we can run a hypothesis test.

Step 5. **A hypothesis test**

• **Hypothesis** about the **population** correlation coefficient ρ

1. The **null hypothesis** $H_0 : \rho = 0$.
2. The **alternative hypothesis** $H_A : \rho \neq 0$.

Solution of example 9

- **Distribution of test statistic.**

When H_0 is true ($\rho = 0$) and the assumptions are met, the appropriate test statistic $t = r\sqrt{\dfrac{n-2}{1-r^2}}$ with $n-2$ degrees of freedom is distributed as **Student's t distribution** .

The number of degrees of freedom is $9 - 2 \equiv 7$

- **Decision rule.**

If we let $\alpha = 0.025, \quad 2\alpha = 0.05$, the critical values of t in the present example are ± 2.365

(e.g. see John Murdoch, "Statistical tables for students of science, engineering, psychology, business, management, finance", 1998, Macmillan, 79 p., Table 7).

Solution of example 9

- **Calculation of test statistic.**

$$t = 0.989\sqrt{\frac{7}{1 - 0.989^2}} = 17.69$$

- **Statistical decision.** Since the computed value of the test statistic exceed the critical value of t, we **reject** the null hypothesis.

- **Conclusion.** We conclude that there is a **very highly significant positive correlation** between pH and growth as measured by optical density of the cell culture.

Solution of example 9

Step 6.

Now we use **regression analysis** to find the line of best fit to the data.

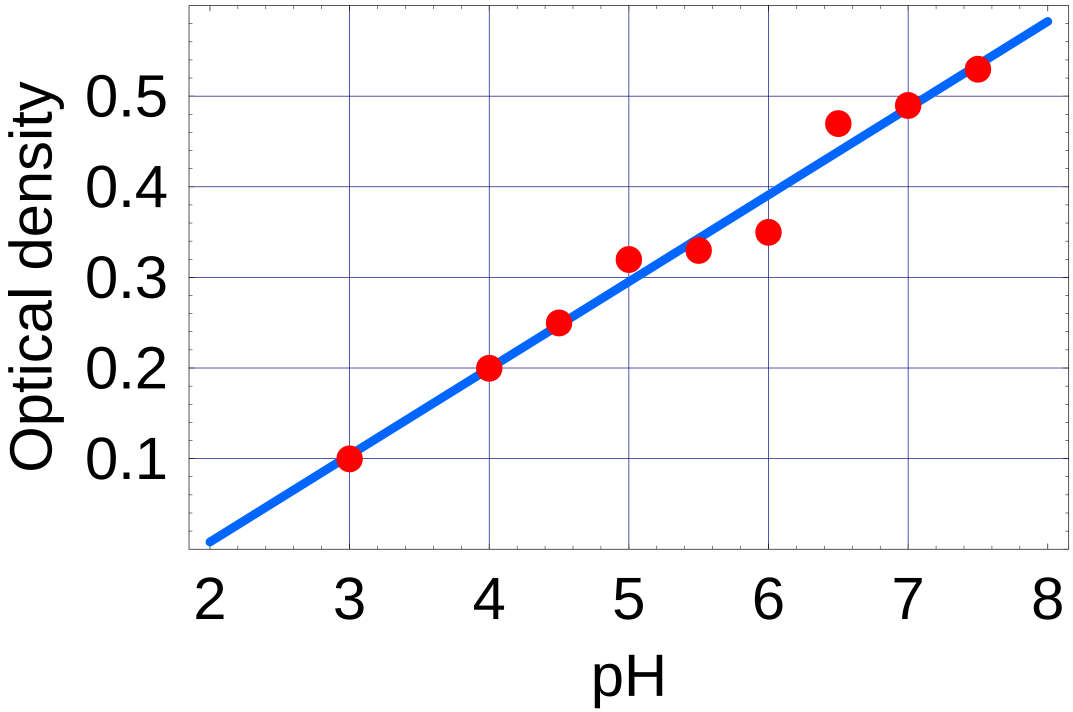
The **regression equation** is

$$y = bx + a$$

where

$$\begin{aligned} b &= \frac{S_{xy}}{S_{xx}} = \frac{14.84}{155} = 0.096 \\ a &= \bar{y} - b\bar{x} = 0.3378 - 0.096 \cdot 5.444 = -0.184 \end{aligned}$$

Regression Line



$$
\begin{aligned}
r &= 0.989 \\
y &= 0.096x - 0.184
\end{aligned}
$$

Chi-Square Goodness-of-Fit Test

Question: Can we assume that the distribution of a sample is valid for the whole population?

The **Pearson's chi-square test** (χ^2-test) is used to test if a sample of data came from a population with a specific distribution.

Advantage : Can be used for discrete distributions such as the binomial and the Poisson and continuous distributions such as normal distribution.

Disadvantage:

- the value of χ^2-test statistic are dependent on how the data is binned.
- χ^2-test requires a sufficient sample size in order for χ^2 approximation to be valid.

Chi-Square Goodness-of-Fit Test

For the χ^2 goodness-of-fit computation, the data are divided into k bins and the **test statistic** is defined as

$$\boxed{\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}}$$

If the computed test statistic is large, then the observed and expected values are not close and the model is a poor fit to the data.

The chi-square test is defined for the **hypothesis**:

H_0: The data follow a **specified distribution**.
H_a: The data do not follow the **specified distribution**.

Chi-Square Goodness-of-Fit Test

The hypothesis that the data are from a population with the specified distribution H_0 is **rejected** if

$$\boxed{\chi^2 > \chi^2_{\alpha,n-c}}$$

where α is the desired level of significance and $\chi^2_{\alpha,k-c}$ is the chi-square percent point function with $n - c$ degrees of freedom.

Example 10 (Chi-Square Test)

Number	Frequency	Relative freq.	
x	f	f^*	fx
0	25	0.031	0
1	81	0.101	0.101
2	124	0.155	0.310
3	146	0.183	0.549
4	175	0.219	0.876
5	106	0.132	0.660
6	80	0.100	0.600
7	35	0.044	0.308
8	16	0.020	0.160
9	6	0.008	0.072
10	6	0.008	0.080
Total	**800**	**1**	**3.716**

$$\bar{x} = 3.716.$$

Poisson Distribution

H_0: The data follow Poisson distribution.
H_a: The data do not follow Poisson distribution.

The probability that there are exactly k occurrences of an event is equal to

$$p_k = \frac{\lambda^k e^{-\lambda}}{k!} \quad k = 0, 1, 2, \ldots$$

where

- k is the number of occurrences of an event.
- λ is a positive real number, equal to the expected number of occurrences that occur during the given interval.

Example 10

Number	Probability	Frequency
x	p	np
0	0.0243	19.44
1	0.0904	72.32
2	0.1680	134.4
3	0.2081	166.48
4	0.1933	154.64
5	0.1437	114.96
6	0.0890	71.2
7	0.0472	37.76
8	0.0219	17.52
9	0.0091	7.28
10	0.0033	2.64

Let $\alpha = 0.1$ (confidence level is 99%)
We assume that $\lambda = 3.716$

$$\begin{aligned}\chi^2 &= \sum_i^{11} \frac{(f_i - np_i)^2}{np_i} \\ &= \frac{(25 - 19.44)^2}{19.44} + \frac{(81 - 72.32)^2}{72.32} + \ldots \\ &= 15.26\end{aligned}$$

We have two constrains:

- $\sum_i^{11} f_i^* = 1$
- $\bar{x} = \lambda$

Therefore degrees of freedom is $11 - 2 = 9$

From the table: $\chi^2_{0.05,9} = 14.68.$
Hence H_0 is rejected and the data do not follow Poisson distribution.