Biostatistics

Luonan Chen

Chinese Academy of Sciences

Lecture 2

Strategies for understanding the meanings of Data

- Key words：

frequency table, bar chart ,range

width of interval , mid-interval

Histogram , Polygon

多边形

Important Characteristics of Data

1. Center: A representative or average value that indicates where the middle of the data set is located

2. Variation: A measure of the amount that the values vary among themselves

3. Distribution: The nature or shape of the distribution of data (such as bell-shaped, uniform, or skewed)

4. Outliers: Sample values that lie very far away from the vast majority of other sample values

5. Time: Changing characteristics of the data over time

Descriptive Statistics

Frequency Distribution for Discrete Random Variables

Example:

Suppose that we take a sample of size 16 from children in a primary school and get the following data about the number of their decayed teeth,

3,5,2,4,0,1,3,5,2,3,2,3,3,2,4,1

To construct a frequency table:

1- Order the values from the smallest to the largest.

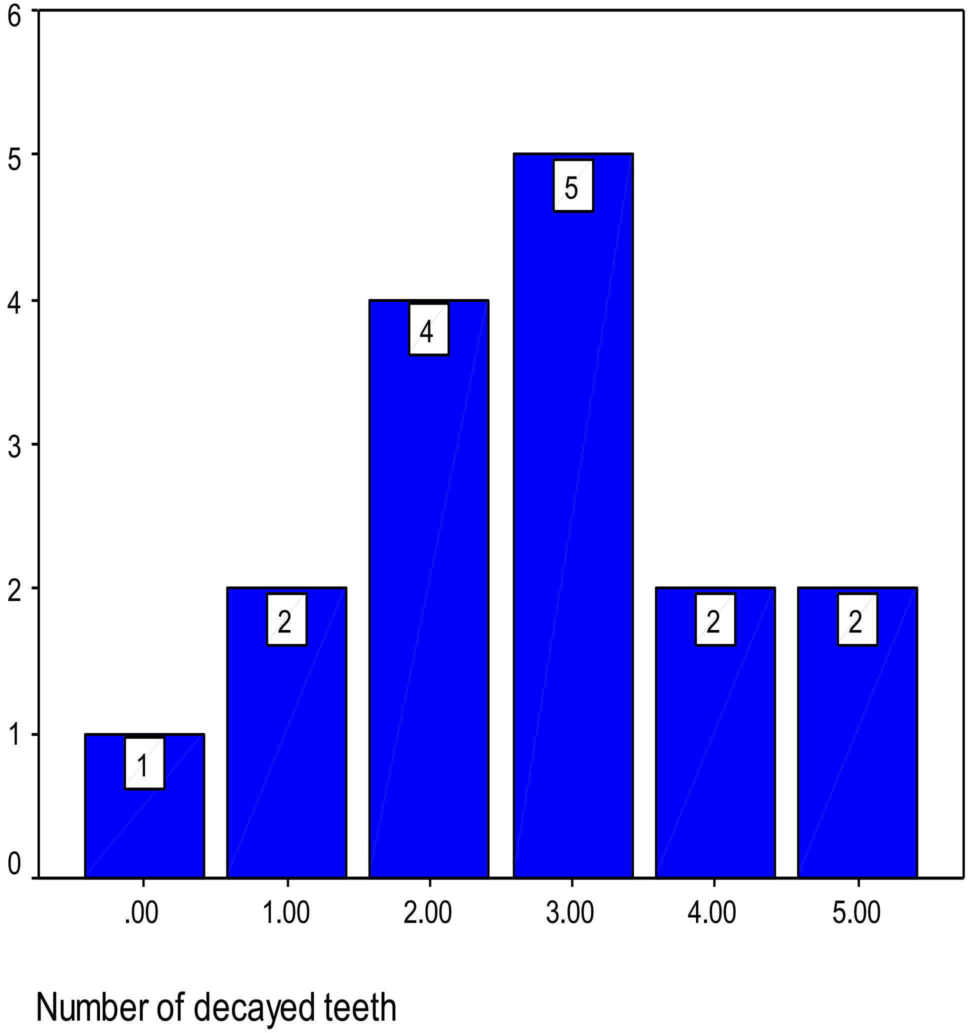
0,1,1,2,2,2,2,3,3,3,3,3,4,4,5,5

2- Count how many numbers are the same.

No. of decayed teeth	Frequency	Relative Frequency
0	1	0.0625
1	2	0.125
2	4	0.25
3	5	0.3125
4	2	0.125
5	2	0.125
Total	16	1

Representing the simple frequency table using the bar chart

We can represent the above simple frequency table using the bar chart.



Frequency Distribution for Continuous Random Variables

For large samples, we can’t use the simple frequency table to represent the data.

We need to divide the data into groups or intervals or classes.

So, we need to determine:

1- The number of intervals (k).

Too few intervals are not good because information will be lost.

Too many intervals are not helpful to summarize the data.

A commonly followed rule is that $6 \le k \le 15$,

or the following formula may be used,

$k = 1 + 3.322 (\log n)$

10为底

2- The range (R).

It is the difference between the largest and the smallest observation in the data set.

3- The Width of the interval (w).

Class intervals generally should be of the same width. Thus, if we want k intervals, then w is chosen such that

$w \geq R / k$.

Example:

Assume that the number of observations equal 100, then

$k = 1+3.322(\log 100)$

$= 1 + 3.3222\ (2) = 7.6 \cong 8.$

Assume that the smallest value = 5 and the largest one of the data = 61, then

$R = 61 – 5 = 56$ and

$w = 56 / 8 = 7.$

To make the summarization more comprehensible, the class width may be 5 or 10 or the multiples of 10.

Relative Frequency Distribution

$$\text{relative frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}}$$

The Cumulative Frequency:

It can be computed by adding successive frequencies.

The Cumulative Relative Frequency:

It can be computed by adding successive relative frequencies.

The Mid-interval:

It can be computed by adding the lower bound of the interval plus the upper bound of it and then divide over 2.

Class interval (age)	Frequency
30 – 39	11
40 – 49	46
50 – 59	70
60 – 69	45
70 – 79	16
80 – 89	1
Total	189

Sum of frequency =sample size=n

For the above example, the following table represents the cumulative frequency, the relative frequency, the cumulative relative frequency and the mid-interval.

R.f= freq/n

Class interval	Mid – interval	Frequency Freq (f)	Cumulative Frequency	Relative Frequency R.f	Cumulative Relative Frequency
30 – 39	34.5	**11**	11	0.0582	0.0582
40 – 49	44.5	**46**	57	0.2434	-
50 – 59	54.5	**70**	127	-	0.6720
60 – 69	64.5	**45**	-	0.2381	0.9101
70 – 79	74.5	**16**	188	0.0847	0.9948
80 – 89	84.5	**1**	189	0.0053	1
Total		189		1	

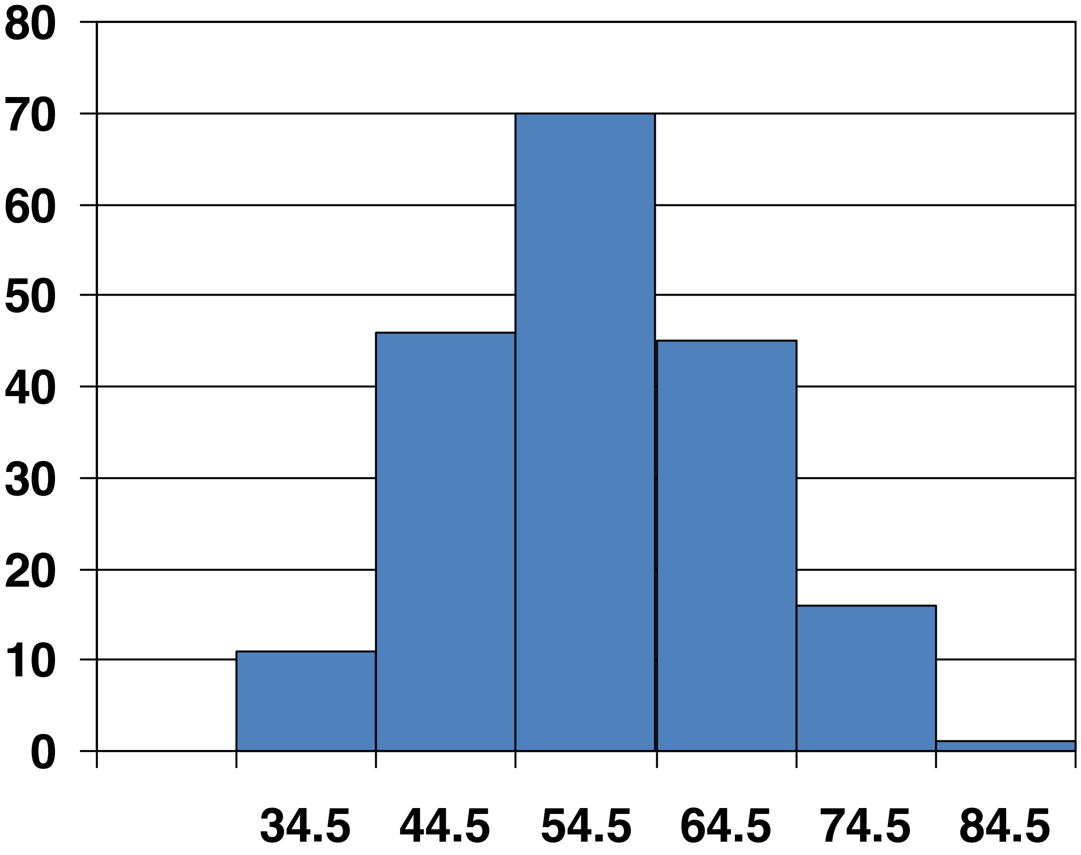
Example

- From the above frequency table, complete the table then answer the following questions:
- 1-The number of objects with age less than 50 years ?
- 2-The number of objects with age between 40-69 years ?
- 3-Relative frequency of objects with age between 70-79 years ?
- 4-Relative frequency of objects with age more than 69 years ?
- 5-The percentage of objects with age between 40-49 years ?
- 6- The percentage of objects with age less than 60 years ?
- 7-The Range (R) ?
- 8- Number of intervals (K)?
- 9- The width of the interval (W) ?

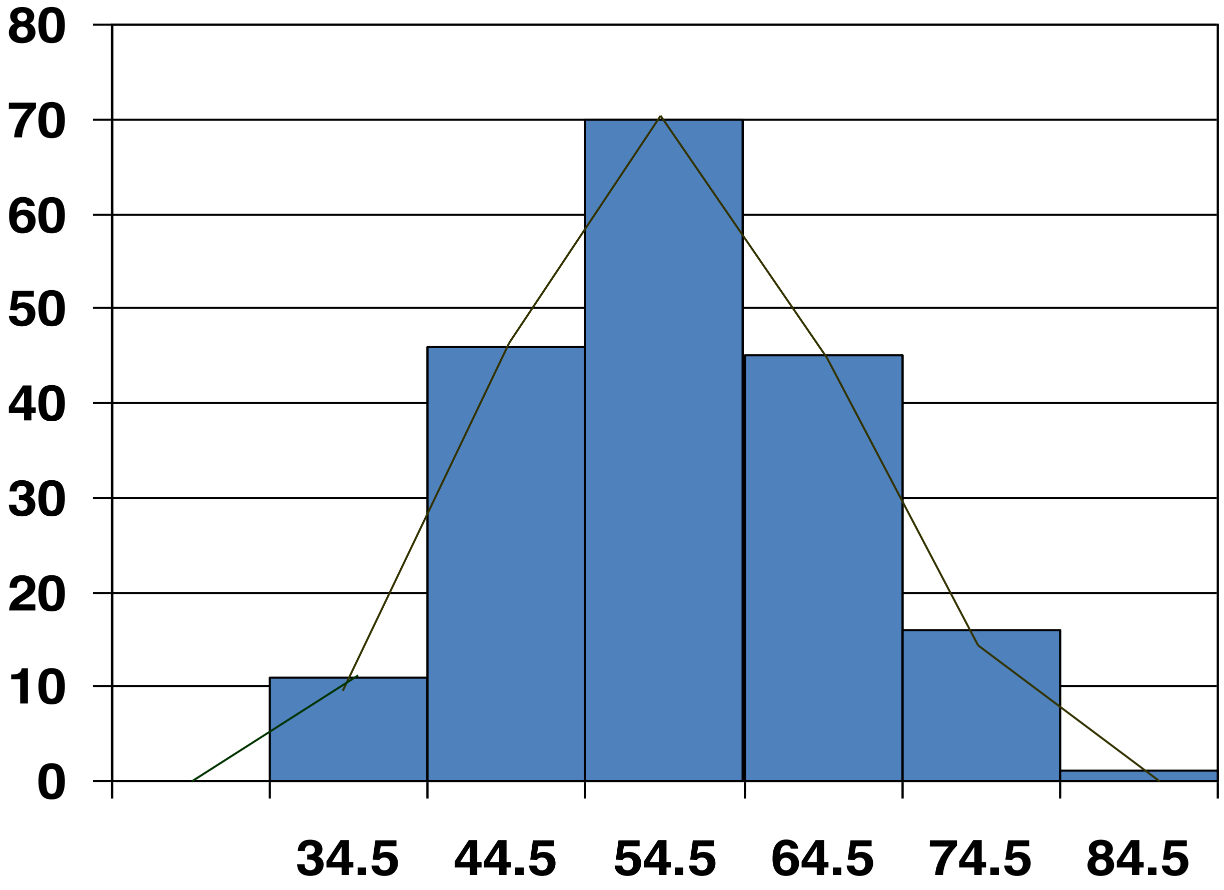
Representing the grouped frequency table using the histogram

To draw the histogram, the true classes limits should be used. They can be computed by subtracting 0.5 from the lower limit and adding 0.5 to the upper limit for each interval.

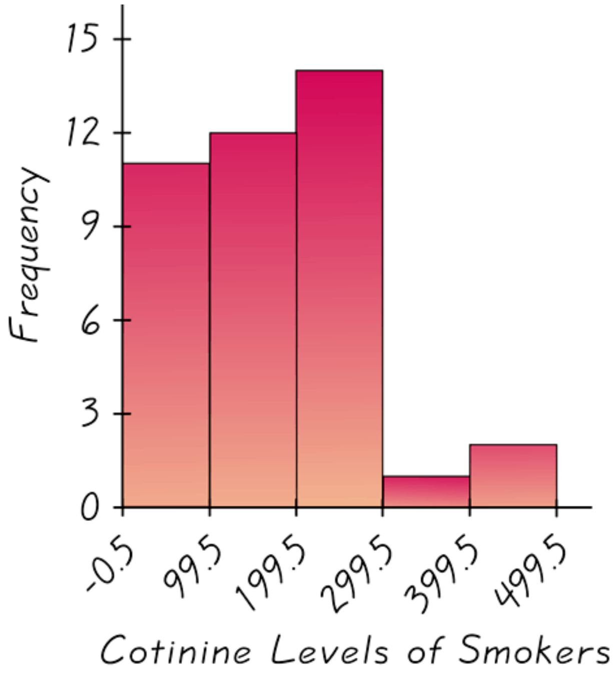
True class limits	Frequency
29.5 – <39.5	**11**
39.5 – < 49.5	**46**
49.5 – < 59.5	**70**
59.5 – < 69.5	**45**
69.5 – < 79.5	**16**
79.5 – < 89.5	**1**
Total	189

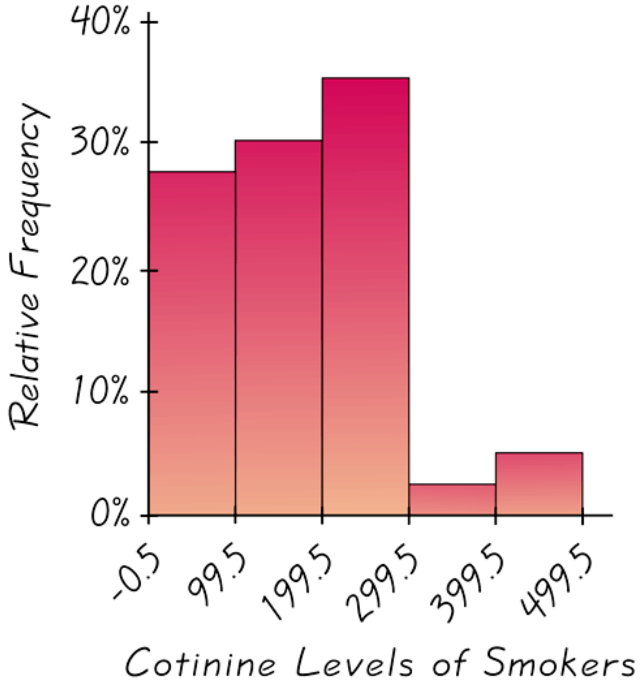


Representing the grouped frequency table using the Polygon



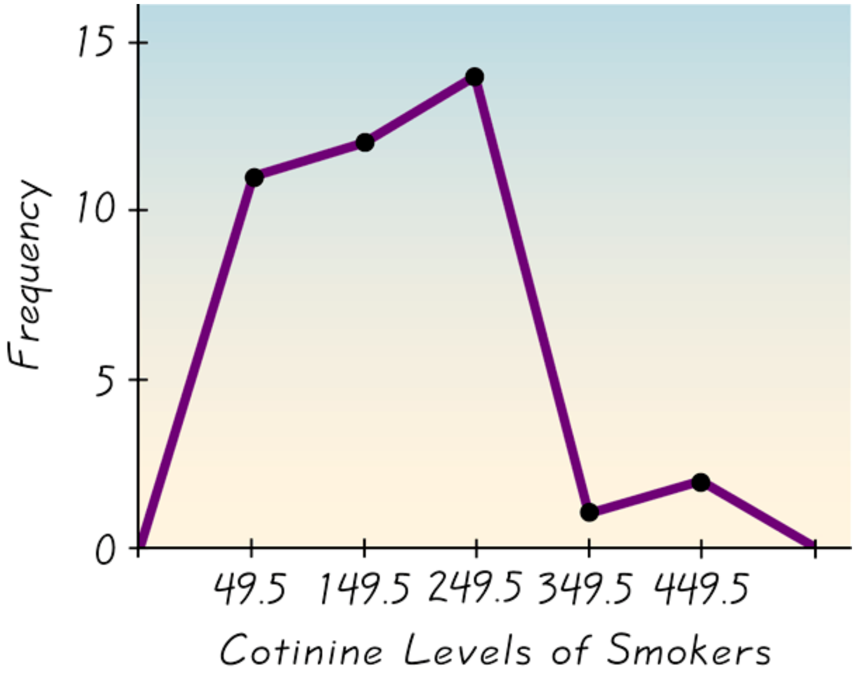
Histogram and Relative Frequency Histogram





Frequency Polygon

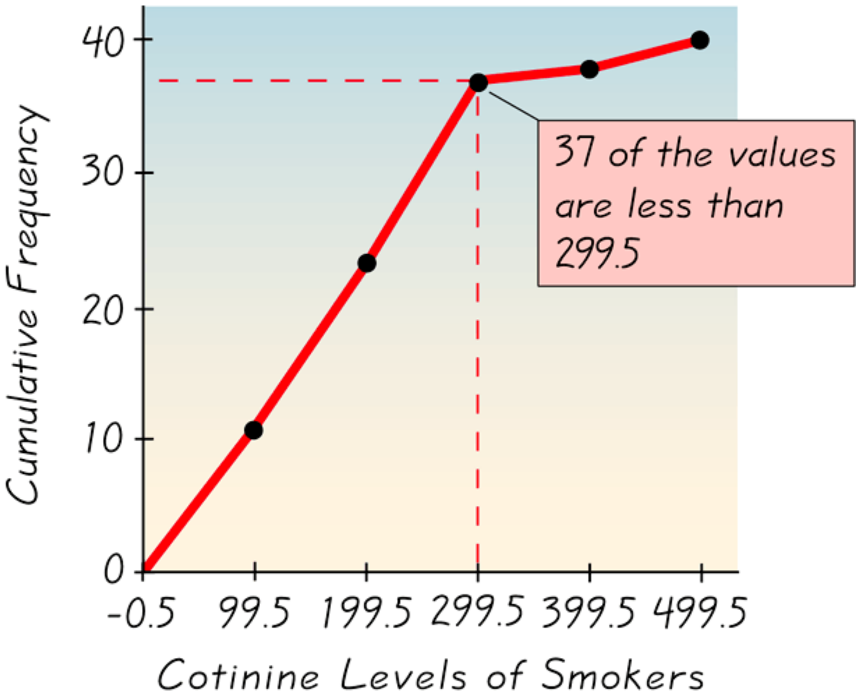
Uses line segments connected to points directly above class midpoint values



Ogive

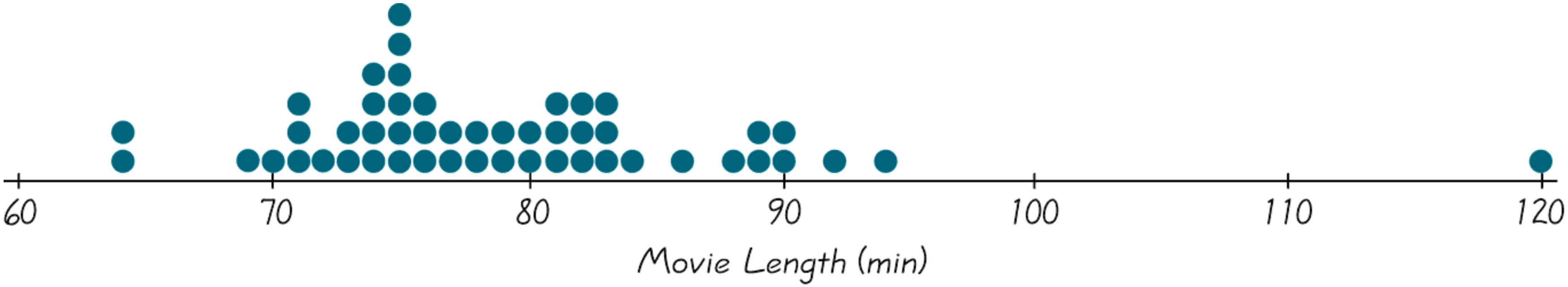
穹形图

A line graph that depicts cumulative frequencies



Dot Plot

Consists of a graph in which each data value is plotted as a point along a scale of values



Pareto Chart

排列图

A bar graph for qualitative data, with the bars arranged in order according to frequencies

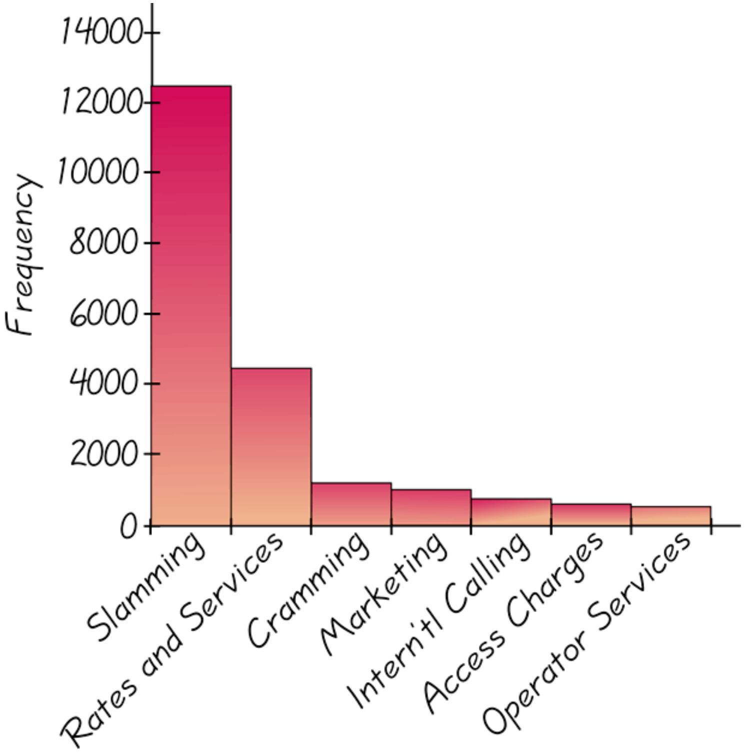
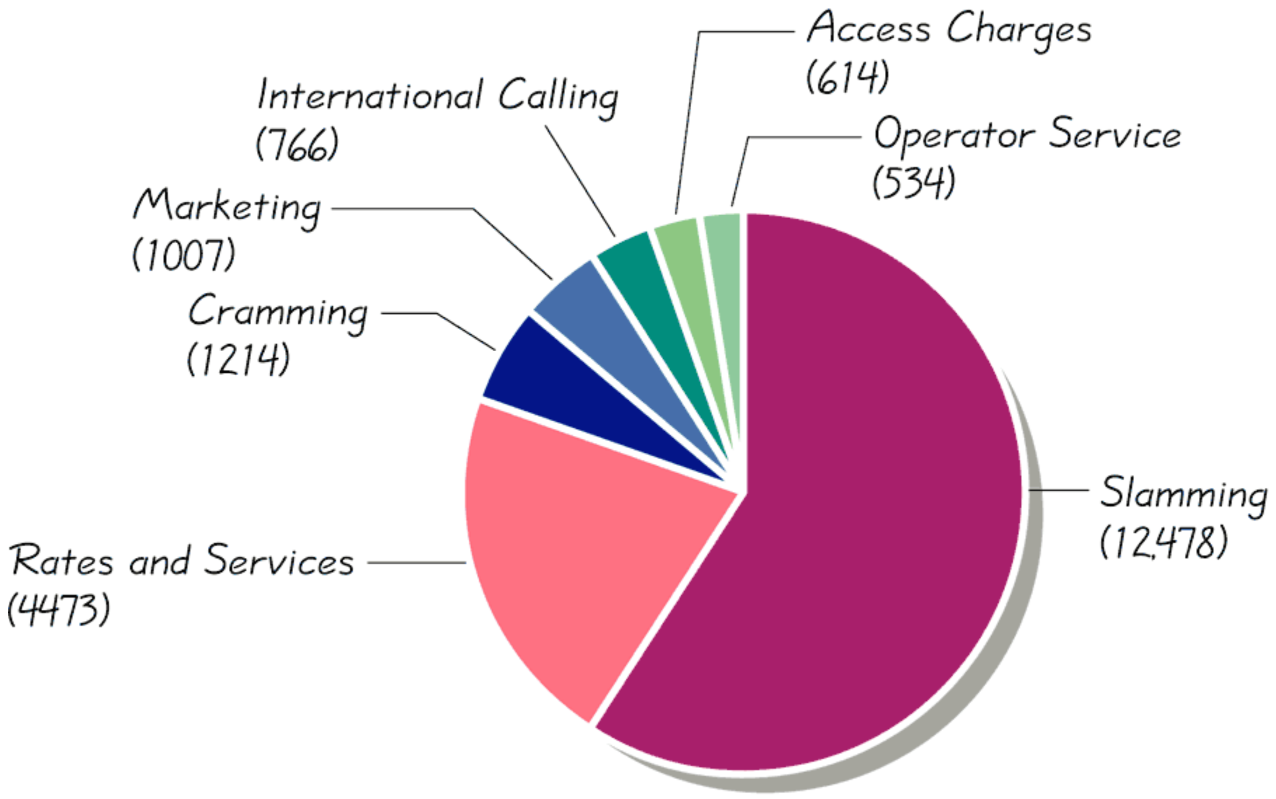


Figure 2-6

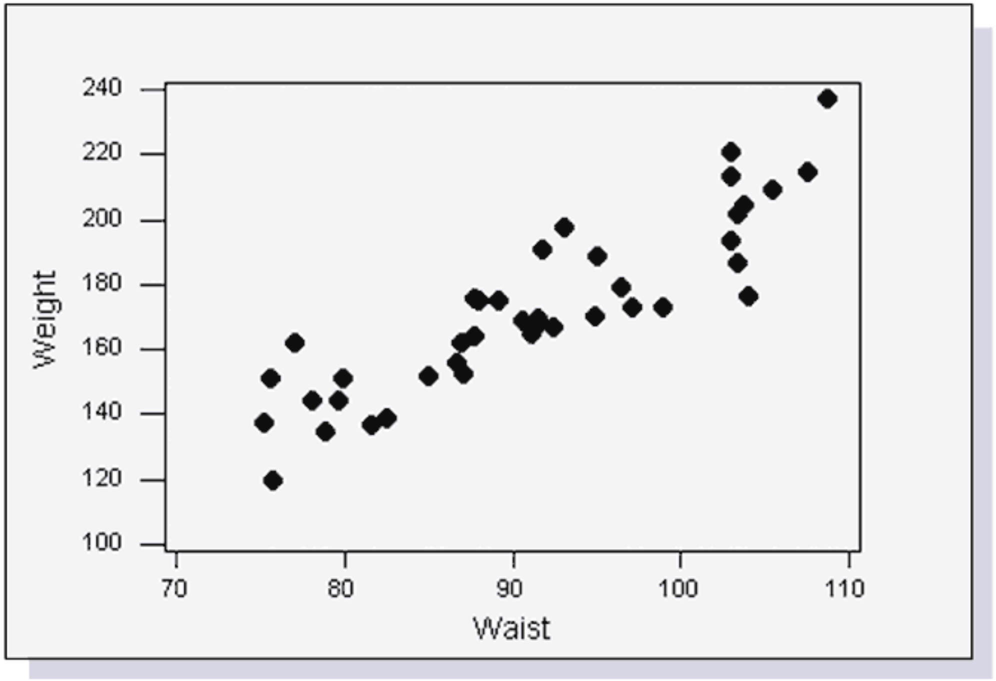
Pie Chart

A graph depicting qualitative data as slices pf a pie



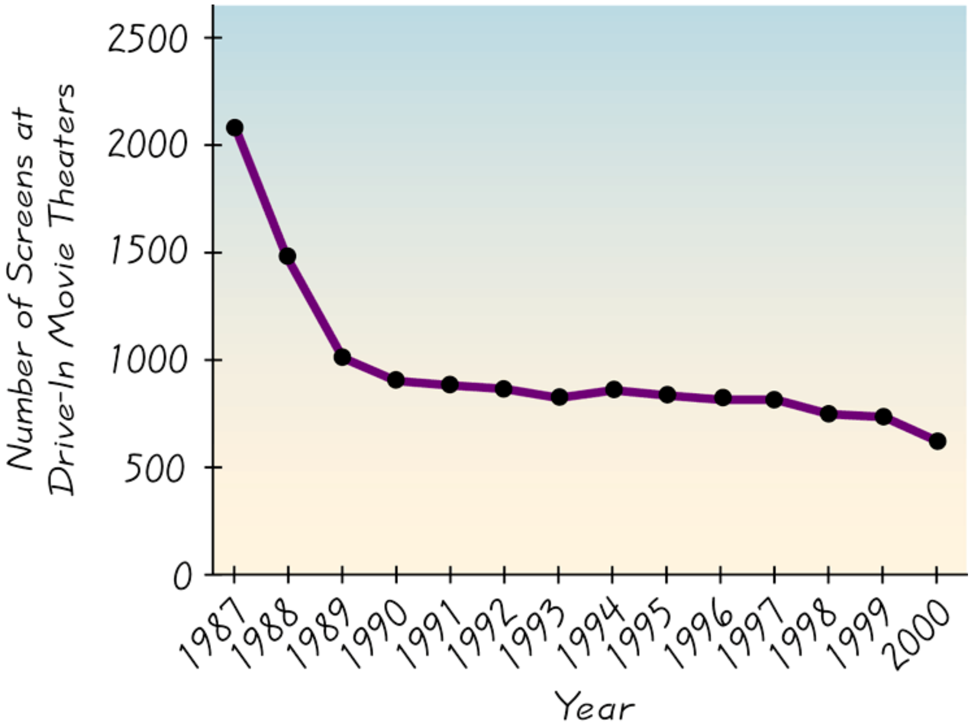
Scatter Diagram

A plot of paired (x,y) data with a horizontal x-axis and a vertical y-axis



Time-Series Graph

Data that have been collected at different points in time



Lecture 3

Descriptive Statistics Measures of Central Tendency

key words:

Descriptive Statistic, measure of central tendency ,statistic, parameter, mean (μ) ,median, mode.

The Statistic and The Parameter

- ***A Statistic:***

It is a descriptive measure computed from the data of a sample. (e.g., average value)

- ***A Parameter:***

It is a a descriptive measure computed from the data of a population. (e.g., total samples = n)

Since it is difficult to measure a parameter from the population, a sample is drawn of size n, whose values are $\chi_1, \chi_2, \ldots, \chi_n$. From this data, we measure the statistic.

Notation

Σ denotes the addition of a set of values

x is the variable usually used to represent the individual data values

n represents the number of values in a sample

N represents the number of values in a population

Measures of Central Tendency

A measure of central tendency is a measure which indicates where the middle of the data is.

The three most commonly used measures of central tendency are:

The Mean（平均数）, the Median（中位数）, and the Mode（众数）, Skewness

The Mean:

It is the average of the data.

The Population Mean:

$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$ which is usually unknown, then we use the

sample mean to estimate or approximate it.

The Sample Mean:

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

Example:

Here is a random sample of size 10 of ages, where

$\chi_1 = 42$, $\chi_2 = 28$, $\chi_3 = 28$, $\chi_4 = 61$, $\chi_5 = 31$,
$\chi_6 = 23$, $\chi_7 = 50$, $\chi_8 = 34$, $\chi_9 = 32$, $\chi_{10} = 37$.

$\bar{\chi}$ = (42 + 28 + … + 37) / 10 = 36.6

Properties of the Mean:

- **Uniqueness.** For a given set of data there is one and only one mean.
- **Simplicity.** It is easy to understand and to compute.
- **Affected by extreme values.** Since all values enter into the computation.

Example: Assume the values are 115, 110, 119, 117, 121 and 126. The mean = 118.

But assume that the values are 75, 75, 80, 80 and 280. The mean = 118, a value that is not representative of the set of data as a whole.

The Median:

When ordering the data, it is the observation that divide the set of observations into two equal parts such that half of the data are before it and the other are after it.

* If n is odd, the median will be the middle of observations. It will be the (n+1)/2 th ordered observation.

When n = 11, then the median is the 6th observation.

* If n is even, there are two middle observations. The median will be the mean of these two middle observations. It will be the (n+1)/2 th ordered observation.

When n = 12, then the median is the 6.5th observation, which is an observation halfway between the 6th and 7th ordered observation.

Example:

For the same random sample, the ordered observations will be as:

23, 28, 28, 31, 32, 34, 37, 42, 50, 61.

Since n = 10, then the median is the 5.5th observation, i.e. = (32+34)/2 = 33.

Properties of the Median:

- **Uniqueness.** For a given set of data there is one and only one median.
- **Simplicity.** It is easy to calculate.
- **It is not affected by extreme values** as is the mean.

The Mode:

It is the value which occurs most frequently.

If all values are different there is no mode.

Sometimes, there are more than one mode.

Bimodal
Multimodal
No Mode

Example:

For the same random sample, the value 28 is repeated two times, so it is the mode.

Properties of the Mode:

- Sometimes, it is not unique.
- It may be used for describing qualitative data.

Definitions

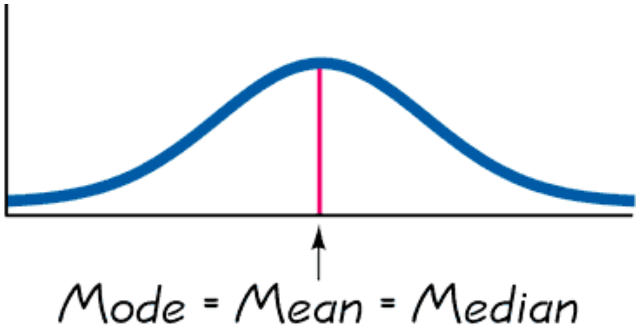
- Symmetric

  Data is symmetric if the left half of its histogram is roughly a mirror image of its right half.

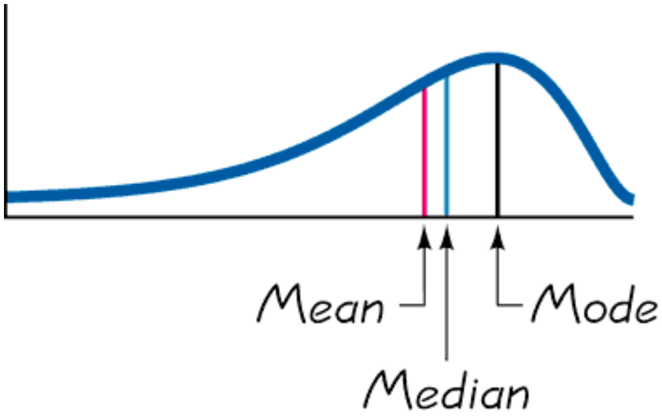
- Skewed

  Data is skewed if it is not symmetric and if it extends more to one side than the other.

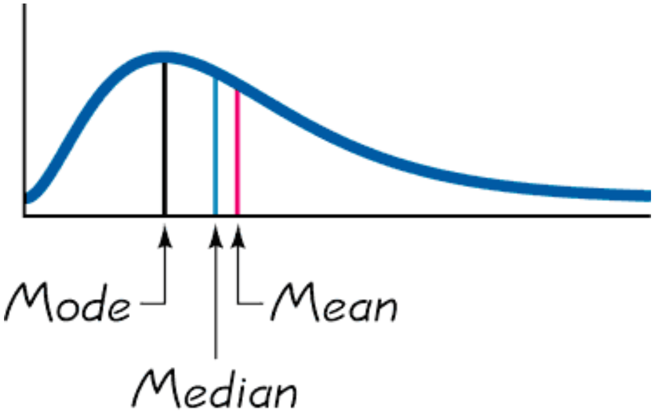
Skewness



(b) Symmetric



(a) Skewed to the Left
(Negatively)



(c) Skewed to the Right
(Positively)

Descriptive Statistics
Measures of Dispersion

key words:

Descriptive Statistic, measure of dispersion , range ,variance, coefficient of variation.

变异系数

2.5. Descriptive Statistics – Measures of Dispersion:

- A measure of dispersion 离差 conveys information regarding the amount of variability present in a set of data.

Note:

1. If all the values are the same

   → There is no dispersion .

2. If all the values are different

   → There is a dispersion:

3. If the values close to each other

   →The amount of Dispersion small.

4. If the values are widely scattered

   → The Dispersion is greater.

- **Measures of Dispersion are :**

1. Range (R).
2. Variance.
3. Standard deviation.
4. Coefficient of variation (C.V).

1.The Range (R):

- Range =Largest value- Smallest value =

$$x_L - x_S$$

Note:

- Range concern only onto two values

- Data （age）：
  43,66,61,64,65,38,59,57,57,50.
- Find Range?
  Range=66-38=28

2.The Variance:

- It measure dispersion relative to the scatter of the values about the mean.

Sample Variance : S^2

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$, **where $\bar{x}$ is sample mean**

- If the mean is unknown (and is computed as the sample mean), then the sample variance is a biased estimator: it underestimates the variance by a factor of (n − 1) / n; correcting by this factor (dividing by n − 1 instead of n) is called Bessel's correction. The resulting estimator is unbiased, and is called the **(corrected) sample variance** or **unbiased sample variance**.

- **Find Sample Variance of ages , $\bar{x}$ = 56**
- **Solution:**
- S^2= [(43-56) ^{2} +(66-56) ^{2}+…..+(50-56) ^{2}]/ (10-1) = 900/9 = 100

Population Variance : σ^2

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$ **is Population mean**

3.The Standard Deviation:

is the square root of variance= $\sqrt{Varince}$

a) Sample Standard Deviation = S = $\sqrt{S^2}$

b) Population Standard Deviation = σ = $\sqrt{\sigma^2}$

Derivation of Sample variance

Directly taking the variance of the sample gives the average

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Since the y_i are selected randomly, and both the above two variables are random variables. Their expected values S can be evaluated by summing over the ensemble of all possible samples $\{y_i\}$ from the population.

$$\begin{aligned}
E[\sigma_y^2] &= E\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{n}\sum_{j=1}^{n} y_j\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left[y_i^2 - \frac{2}{n} y_i \sum_{j=1}^{n} y_j + \frac{1}{n^2}\sum_{j=1}^{n} y_j \sum_{k=1}^{n} y_k\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n-2}{n}E[y_i^2] - \frac{2}{n}\sum_{j\neq i} E[y_i y_j] + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k\neq j} E[y_j y_k] + \frac{1}{n^2}\sum_{j=1}^{n} E[y_j^2]\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n-2}{n}(\sigma^2 + \mu^2) - \frac{2}{n}(n-1)\mu^2 + \frac{1}{n^2}n(n-1)\mu^2 + \frac{1}{n}(\sigma^2 + \mu^2)\right] \\
&= \frac{n-1}{n}\sigma^2.
\end{aligned}$$

Hence, unbiased sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$

4.The Coefficient of Variation (C.V):

is a measure use to compare the dispersion in two sets of data which is independent of the unit of the measurement .

$$C.V = \frac{S}{\overline{X}}(100)$$

where S: Sample standard deviation.

$\overline{X}$: Sample mean.

Example

- Suppose two samples of human males yield the following data:

	Sampe1	Sample2
Age	25-year-olds	11year-olds
Mean weight	145 pound	80 pound
Standard deviation	10 pound	10 pound

- We wish to know which is more variable.

Solution:

- c.v (Sample1)= (10/145)*100= 6.9

- c.v (Sample2)= (10/80)*100= 12.5

- Then age of 11-years old(sample2) is more variation

Lecture 3

Probability

The Basis of the Statistical inference

- Key words:

- Probability, objective Probability,

subjective Probability, equally likely,

Mutually exclusive, multiplicative rule,

Conditional Probability, independent events, Bayes theorem

3.1 Introduction

- The concept of probability is frequently encountered in everyday communication. **For example**, a physician may say that a patient has a 50-50 chance of surviving a certain operation, and another may say that she is 95 percent certain that a patient has a particular disease.

- Most people express probabilities in terms of percentages.

- But, it is more convenient to express probabilities as fractions. Thus, we may measure the probability of the occurrence of some event by a number between 0 and 1.

- The more likely the event, the closer the number is to one. An event that cannot occur has a probability of zero, and an event that is certain to occur has a probability of one.

3.2 Two views of Probability

$\cap$

- **Some definitions:**

1.Equally likely outcomes:

The outcomes have the same chance of occurring.

2.Mutually exclusive:

Two events are said to be mutually exclusive if they cannot occur simultaneously such that $A \cap B = \Phi$.

- **The universal Set** (S): The set all possible outcomes.
- **The empty set** Φ : Contain no elements.
- **The event** E : is a set of outcomes in S which has a certain characteristic.
- **Classical Probability** : If an event can occur in N mutually exclusive and equally likely ways, and if m of these possess a triat E, the probability of the occurrence of event E is equal to m/ N .
- **For Example:** in the rolling of the die , each of the six sides is equally likely to be observed . So, the probability that a 4 will be observed is equal to 1/6.

- **Relative Frequency Probability:**
- **Definition:** If some processe is repeated a large number of times, n, and if some resulting event E occurs m times , the relative frequency of occurrence of E , m/n will be approximately equal to probability of E . P(E) = m/n .

- **Subjective Probability** :
- Probability measures the confidence that a particular individual has in the truth of a particular proposition.
- **For Example** : the probability that a cure for cancer will be discovered within the next 10 years.

3.3 Elementary Properties of Probability:

- Given some process (or experiment) with n mutually exclusive events E_1, E_2, E_3,…………., E_n, then
- 1- $P(E_i) >= 0$, i= 1,2,3,……n
- 2- $P(E_1) + P(E_2) + \ldots\ldots + P(E_n) = 1$
- 3- $P(E_i + E_J) = P(E_i) + P(E_J)$, where E_i ,E_J are mutually exclusive

Rules of Probability

- 1- Addition Rule

  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- 2- If A and B are mutually exclusive (disjoint) ,then

  $P(A \cap B) = 0$

  Then , addition rule is

  $P(A \cup B) = P(A) + P(B)$.

- 3- Complementary Rule

  $P(A') = 1 - P(A)$

where, A' = complement event

Example

Family history of Mood Disorders (情绪障碍)	Early = 18 (E)	Later >18 (L)	Total
Negative(A)	28	35	63
Bipolar Disorder(B)	19	38	57
Unipolar (C)	41	44	85
Unipolar and Bipolar(D)	53	60	113
Total	141	177	318

Answer the following questions:

Suppose we pick a person at random from this sample.

1-The probability that this person will be 18-years old or younger?

2-The probability that this person has family history of mood orders Unipolar(C)?

3-The probability that this person has no family history of mood orders Unipolar($\overline{C}$)?

4-The probability that this person is 18-years old or younger or has no family history of mood orders Negative (A)?

5-The probability that this person is more than18-years old and has family history of mood orders Unipolar and Bipolar(D)?

Conditional Probability:

P(A\B) is the probability of A assuming that B has happened.

- P(A\B)= $\frac{P(A\cap B)}{P(B)}$, P(B)≠ 0
- P(B\A)= $\frac{P(A\cap B)}{P(A)}$, P(A)≠ 0

Example

From previous example , answer

- suppose we pick a person at random and find he is 18 years or younger (E),what is the probability that this person will be one who has no family history of mood disorders (A)?
- suppose we pick a person at random and find he has family history of mood (D) what is the probability that this person will be 18 years or younger (E)?

Calculating a joint Probability :

- Example

- Suppose we pick a person at random from the 318 subjects. Find the probability that he will early (E) and has no family history of mood disorders (A).

Multiplicative Rule:

- $P(A \cap B) = P(A \backslash B)P(B)$
- $P(A \cap B) = P(B \backslash A)P(A)$
- Where,
- $P(A)$: marginal probability of A.
- $P(B)$: marginal probability of B.
- $P(B \backslash A)$:The conditional probability.

Example

- From previous example, we wish to compute the joint probability of Early age at onset(E) and a negative family history of mood disorders(A) from a knowledge of an appropriate marginal probability and an appropriate conditional probability.

Independent Events:

- If A has no effect on B, we said that A,B are independent events.
- Then,
- 1- $P(A \cap B)= P(B)P(A)$

A、B是独立事件

- 2- $P(A\backslash B)=P(A)$
- 3- $P(B\backslash A)=P(B)$

Example

- In a certain high school class consisting of 60 girls and 40 boys, it is observed that 24 girls and 16 boys wear eyeglasses . If a student is picked at random from this class ,the probability that the student wears eyeglasses , P(E), is 40/100 or 0.4 .
- What is the probability that a student picked at random wears eyeglasses given that the student is a boy?
- What is the probability of the joint occurrence of the events of wearing eye glasses and being a boy?

Example

- Suppose that of 1200 admission to a general hospital during a certain period of time,750 are private admissions. If we designate these as a set A, then compute P(A) , P($\overline{A}$).

Marginal Probability:

- **Definition:**
- Given some variable that can be broken down into m categories designated

By $A_1, A_2, \ldots\ldots, A_i, \ldots\ldots, A_m$ and another jointly occurring variable that is broken down into n categories designated by $B_1, B_2, \ldots\ldots, B_j, \ldots\ldots, B_n$, the marginal probability of A_i with all the categories of B . That is,

$P(A_i) = \sum P(A_i \cap B_j),$ for all value of j

Baye's Theorem

Definition.1

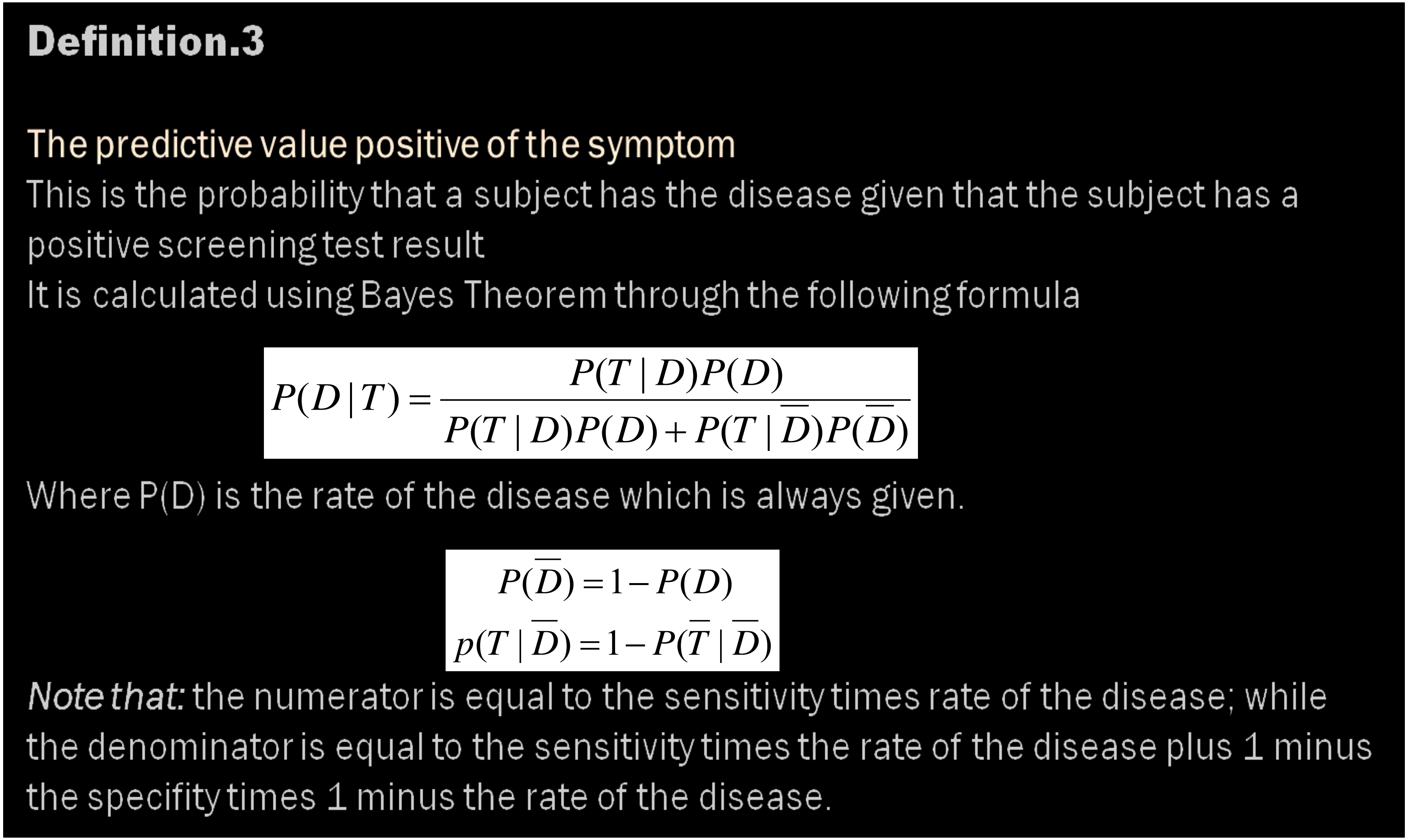
The sensitivity of the symptom

This is the probability of a positive result given that the subject has the disease. It is denoted by P(T|D)

Definition.2

The specificity of the symptom

This is the probability of negative result given that the subject does not have the disease. It is denoted by $P(\overline{T}|\overline{D})$



Definition.3

The predictive value positive of the symptom

This is the probability that a subject has the disease given that the subject has a positive screening test result

It is calculated using Bayes Theorem through the following formula

$$P(D|T)=\frac{P(T\mid D)P(D)}{P(T\mid D)P(D)+P(T\mid \overline{D})P(\overline{D})}$$

Where P(D) is the rate of the disease which is always given.

$$P(\overline{D})=1-P(D)$$
$$p(T\mid \overline{D})=1-P(\overline{T}\mid \overline{D})$$

Note that: the numerator is equal to the sensitivity times rate of the disease; while the denominator is equal to the sensitivity times the rate of the disease plus 1 minus the specifity times 1 minus the rate of the disease.

Definition.4

The predictive value negative of the symptom

This is the probability that a subject does not have the disease given that the subject has a negative screening test result
It is calculated using Bayes Theorem through the following formula

$$P(\overline{D}\mid\overline{T})=\frac{P(\overline{T}\mid\overline{D})P(\overline{D})}{P(\overline{T}\mid\overline{D})P(\overline{D})+P(\overline{T}\mid D)P(D)}$$

where,

$$p(\overline{T}\mid D)=1-P(T\mid D)$$

Example

A medical research team wished to evaluate a proposed screening test for Alzheimer’s disease. The test was given to a random sample of 450 patients with Alzheimer’s disease and an independent random sample of 500 patients without symptoms of the disease. The two samples were drawn from populations of subjects who were 65 years or older. The results are as follows.

Test Result	Yes (D)	No ($\overline{D}$)	Total
Positive(T)	436	5	441
Negativ($\overline{\mathrm{T}}$)	14	495	509
Total	450	500	950

In the context of this example

a)What is a false positive?

A false positive is when the test indicates a positive result (T) when the person does not have the disease $\overline{D}$

b) What is the false negative?

A false negative is when a test indicates a negative result *(* $\overline{\mathrm{T}}$ *)* when the person has the disease (D).

c) Compute the sensitivity of the symptom.

$$P(T \mid D) = \frac{436}{450} = 0.9689$$

d) Compute the specificity of the symptom.

$$P(\overline{T} \mid \overline{D}) = \frac{495}{500} = 0.99$$

e) Suppose it is known that the rate of the disease in the general population is 11.3%. What is the predictive value positive of the symptom and the predictive value negative of the symptom 先验概率

The predictive value positive of the symptom is calculated as

$$P(D|T)=\frac{P(T|D)P(D)}{P(T|D)P(D)+P(T|\overline{D})P(\overline{D})}$$

$$=\frac{(0.9689)(0.113)}{(0.9689)(0.113)+(.01)(1-0.113)}=0.925$$

The predictive value negative of the symptom is calculated as

$$P(\overline{D}|\overline{T})=\frac{P(\overline{T}|\overline{D})P(\overline{D})}{P(\overline{T}|\overline{D})P(\overline{D})+P(\overline{T}|D)P(D)}$$

$$=\frac{(0.99)(0.887)}{(0.99)(0.887)+(0.0311)(0.113)}=0.996$$