

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

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.

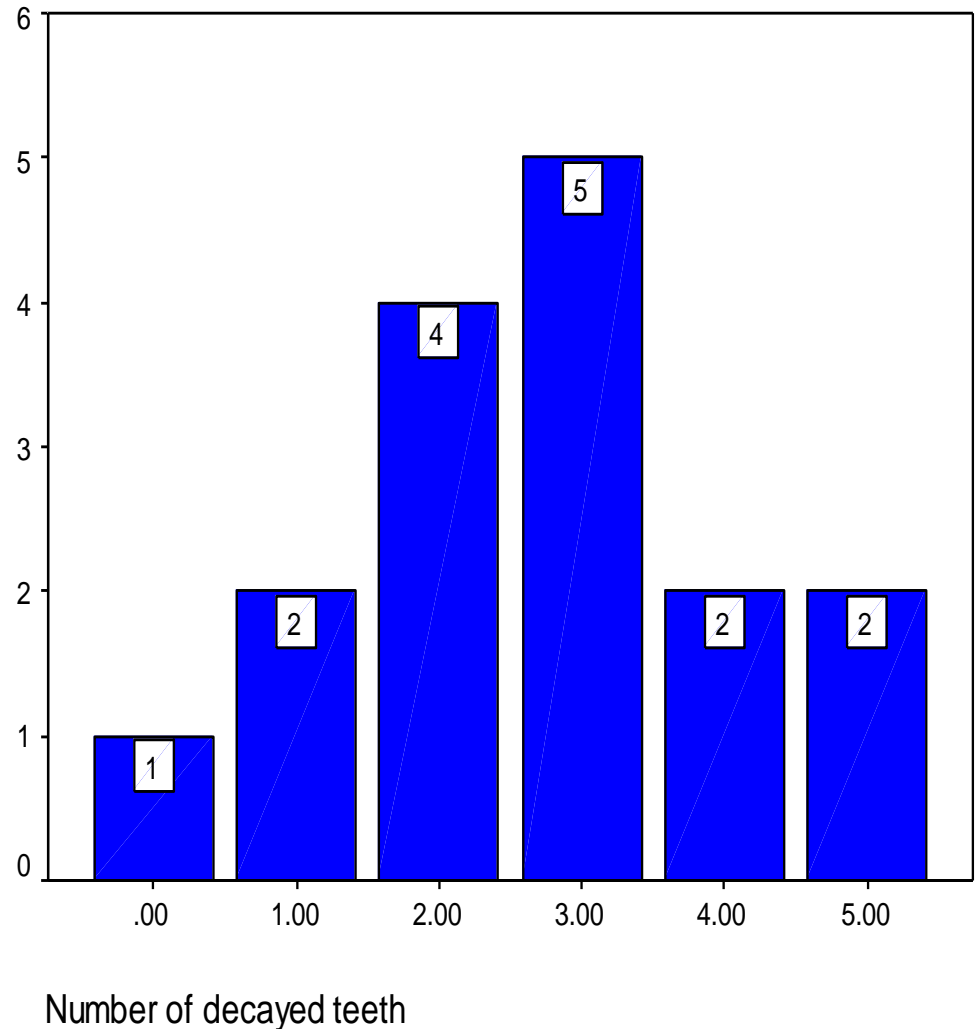
0,1,1,2,2,2,2,3,3,3,3,3,4,4,5,5

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 $6 \leq k \leq 15$,
or the following formula may be used,

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

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

$$R = 61 - 5 = 56 \text{ and}$$

$$w = 56 / 8 = 7.$$

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.

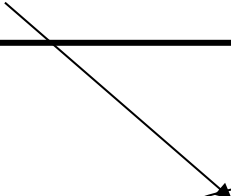
The Cumulative Relative Frequency:

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.



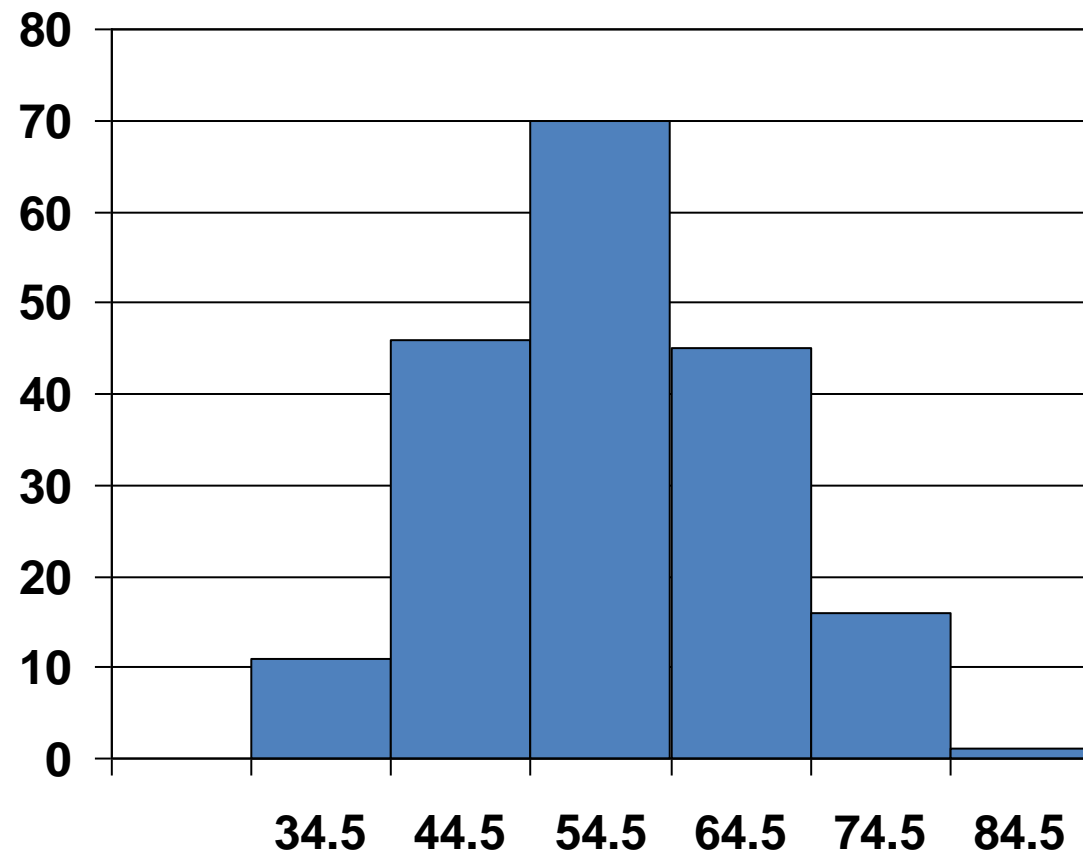
R.f= freq/n

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	13

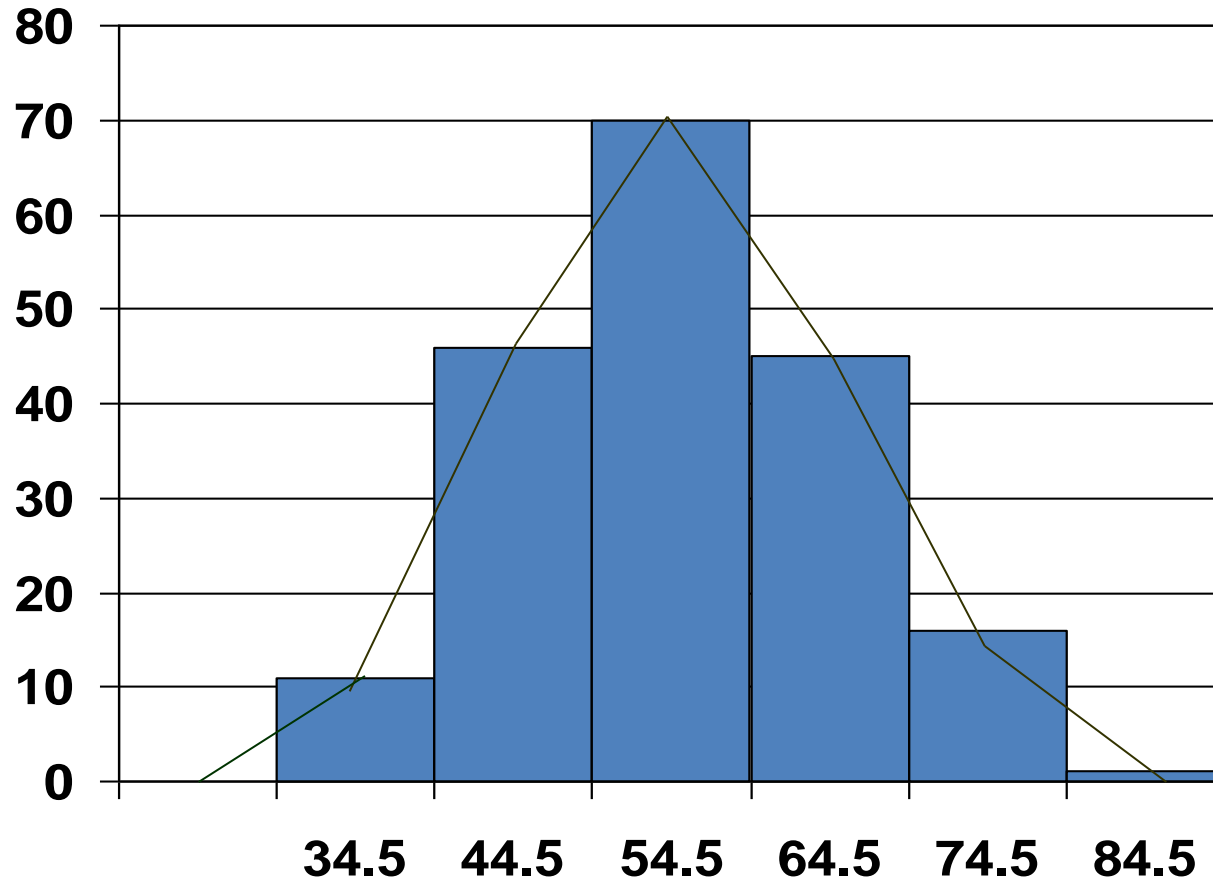
Representing the grouped frequency table using the histogram

To draw the histogram, the true class 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.

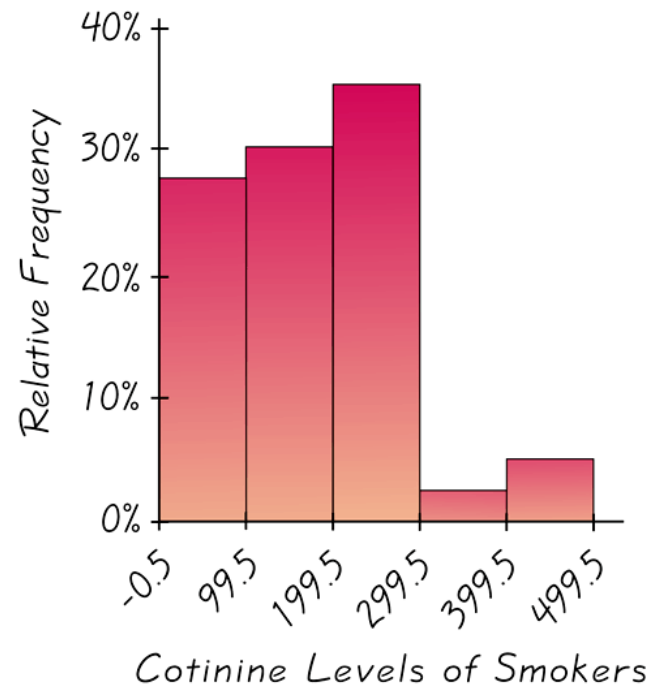
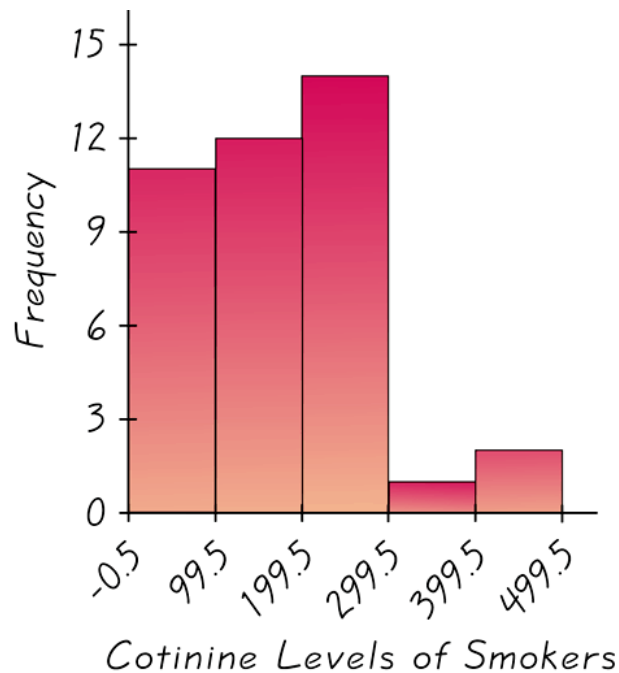
True class limits	Frequency
29.5 – <39.5	11
39.5 – < 49.5	46
49.5 – < 59.5	70
59.5 – < 69.5	45
69.5 – < 79.5	16
79.5 – < 89.5	1
Total	189



Representing the grouped frequency table using the Polygon

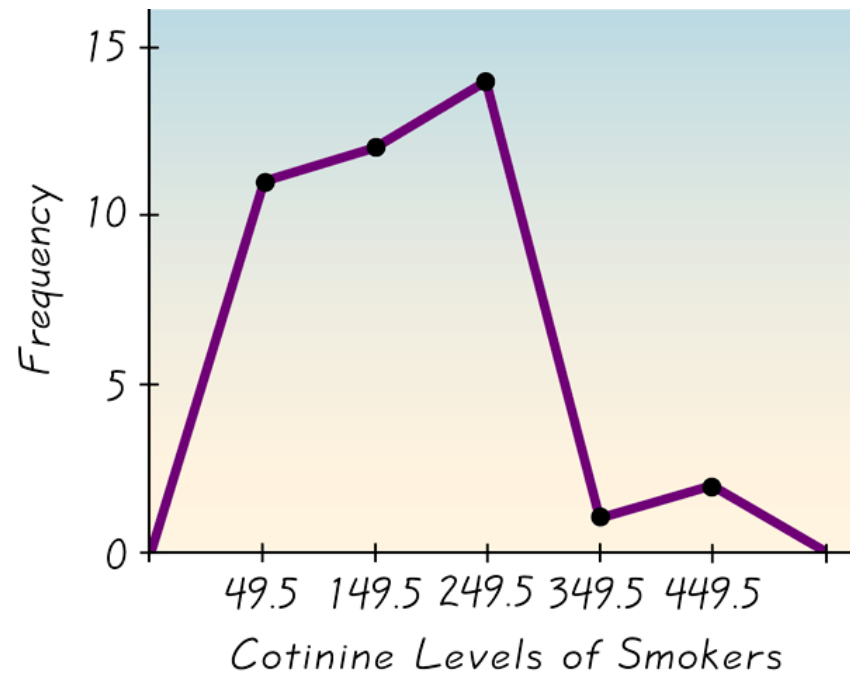


Histogram and Relative Frequency Histogram



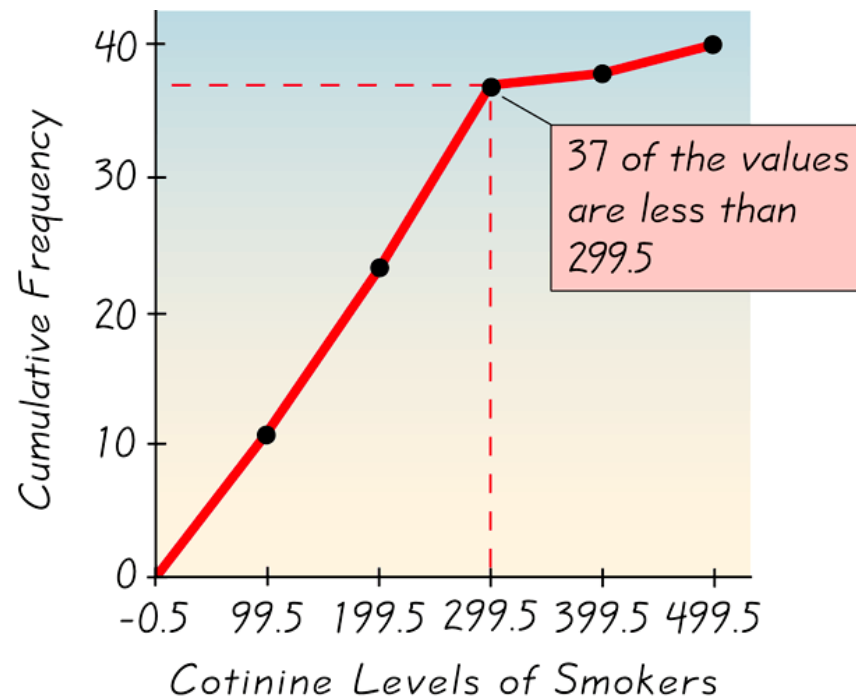
Frequency Polygon

Uses line segments connected to points directly above class midpoint values



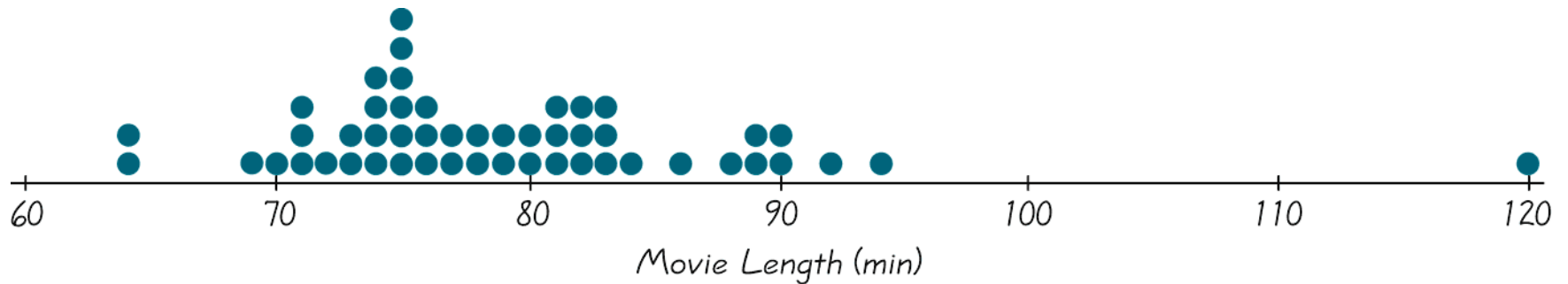
Ogive

A line graph that depicts cumulative frequencies



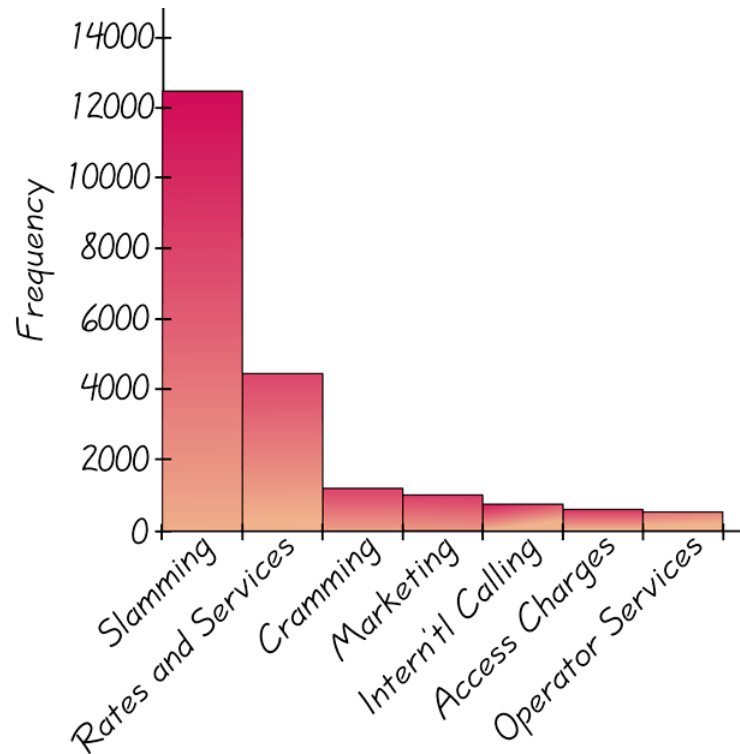
Dot Plot

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



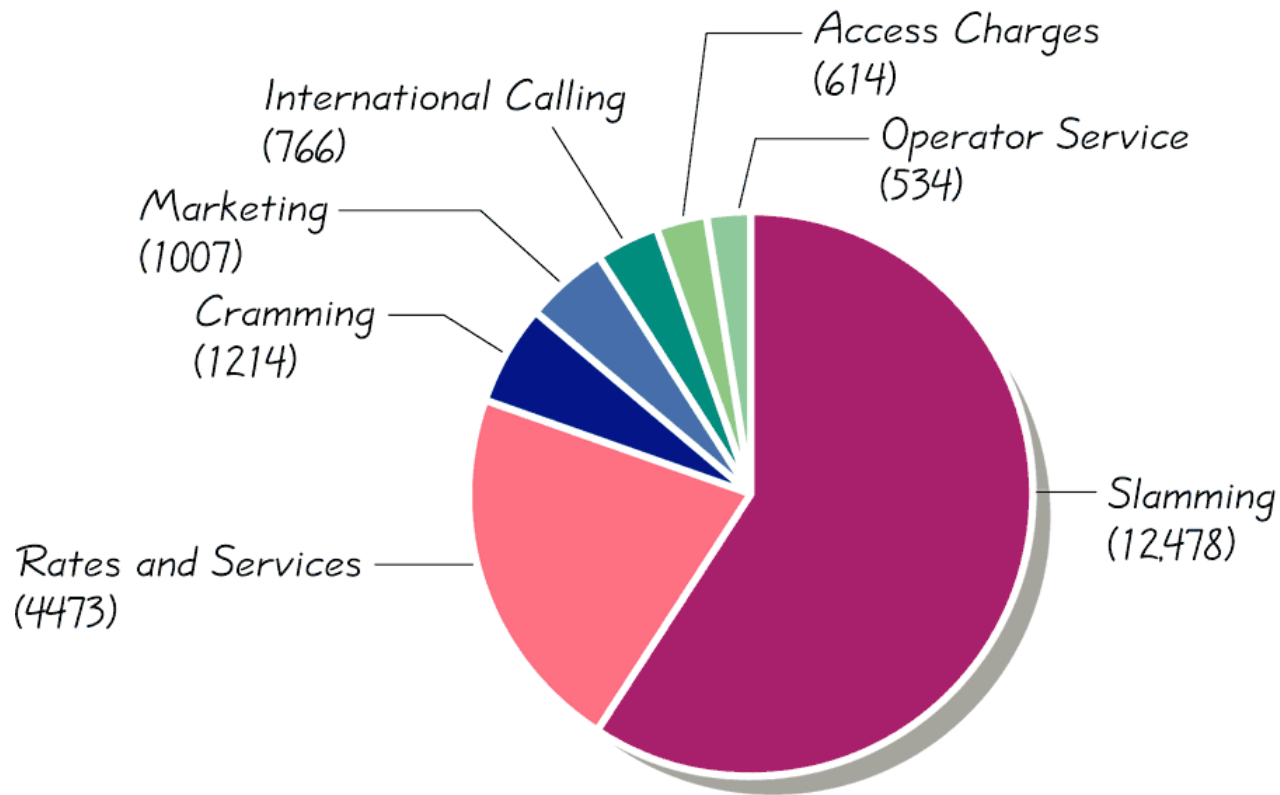
Pareto Chart

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



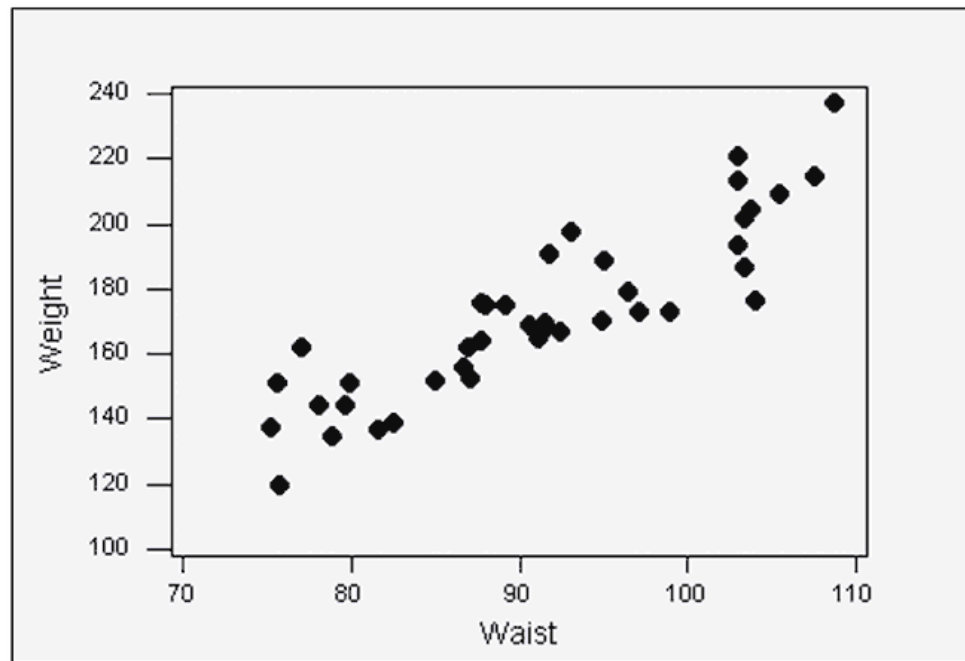
Pie Chart

A graph depicting qualitative data as slices of a pie



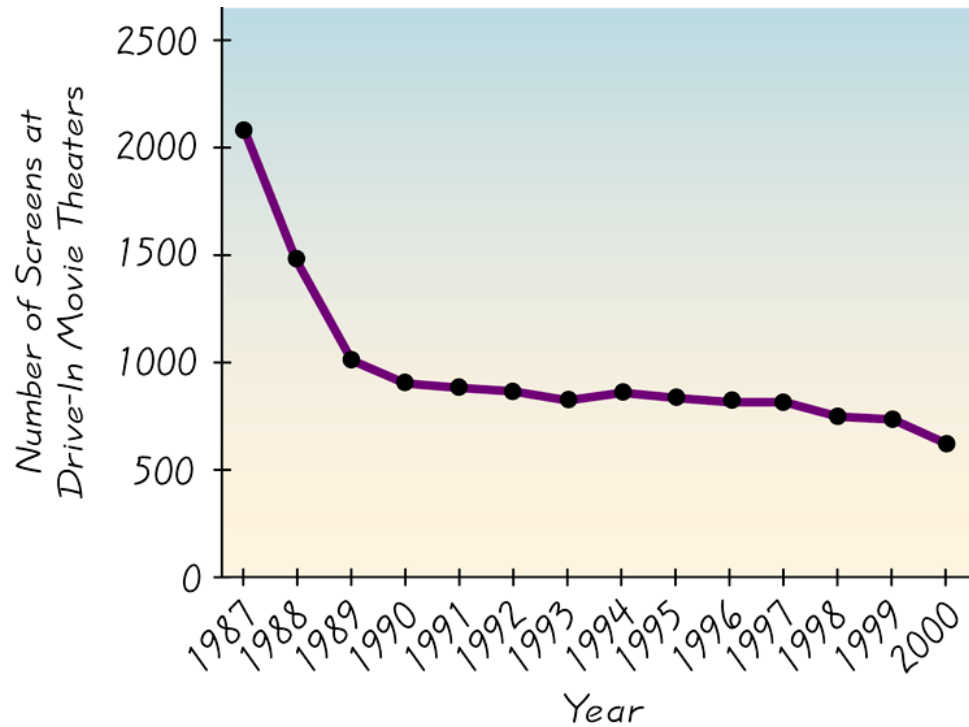
Scatter Diagram

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 $\chi_1, \chi_2, \dots, \chi_n$. 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:

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

Example:

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

$$\begin{aligned} x_1 &= 42, x_2 = 28, x_3 = 28, x_4 = 61, x_5 = 31, \\ x_6 &= 23, x_7 = 50, x_8 = 34, x_9 = 32, x_{10} = 37. \end{aligned}$$

$$\bar{x} = (42 + 28 + \dots + 37) / 10 = 36.6$$

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**.

Example:

Bimodal
Multimodal
No Mode

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

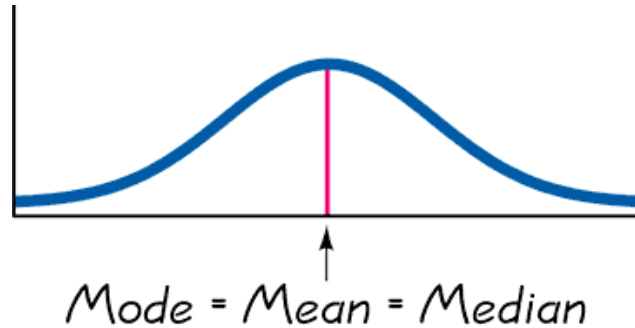
❖ Symmetric

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

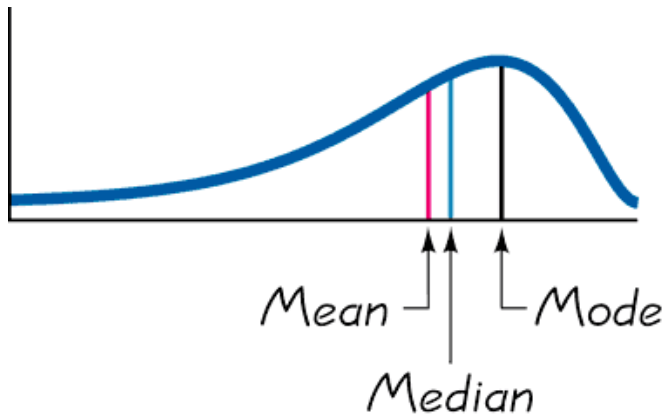
❖ Skewed

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

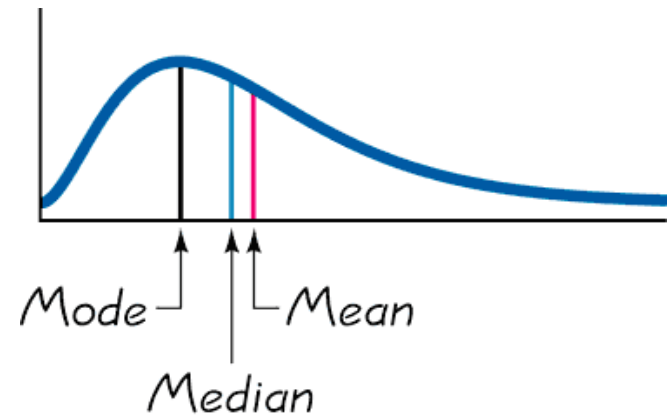
Skewness



(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.

2.5. Descriptive Statistics – Measures of Dispersion:

- A measure of dispersion 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.

Population Variance : σ^2

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

μ is Population mean

3.The Standard Deviation:

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

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

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

Sample Variance : S^2

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

- 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:**
- $S^2 = [(43-56)^2 + (66-56)^2 + \dots + (50-56)^2] / (10-1) = 900/9 = 100$

Average value and mean

$$E(X) = \sum x_i p_i = \frac{1}{n} \sum x_i$$

$$E(X) = \int x f(x) dx$$

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 .

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

where S: Sample standard deviation.

\bar{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:

- $\text{c.v (Sample1)} = (10/145) * 100 = 6.9$
- $\text{c.v (Sample2)} = (10/80) * 100 = 12.5$
- 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, Marginal 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

- 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 is 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 possesses 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 process 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.

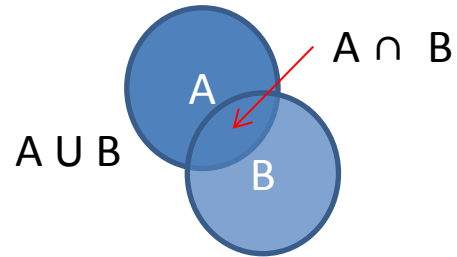
3.3 Elementary Properties of Probability:

- Given some process (or experiment) with n mutually exclusive events $E_1, E_2, E_3, \dots, E_n$, then
 - 1. $P(E_i) \geq 0, i = 1, 2, 3, \dots, n$
 - 2. $P(E_1) + P(E_2) + \dots + P(E_n) = 1$
 - 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

$$P(A \cap B) = 0$$

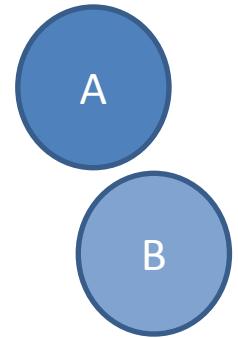
Then, addition rule is

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

- 3- Complementary Rule

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

where, A' = complement event



Note, we define $P(A, B) = P(AB) = P(A \cap B)$, $P(A \cup B) = P(A + B)$

Calculation of mean (first moment)

Discrete random variables and continuous random variables for average value μ

$$E[X] = \sum_{i=1}^{\infty} x_i p_i, \quad E[X] = \int_{-\infty}^{\infty} x f(x) dx.$$

The expected value operator (or **expectation operator**) E is [linear](#) in the sense that

$$\begin{aligned} E[X + c] &= E[X] + c \\ E[X + Y] &= E[X] + E[Y] \\ E[aX] &= a E[X] \end{aligned}$$

Calculation of variance (second moment)

The variance of a random variable X is its second [central moment](#), the [expected value](#) of the squared deviation from the mean $\mu = E[X]$:

$$\text{Var}(X) = E[(X - \mu)^2].$$

$$\text{Var}(X) = \sum_{i=1}^n p_i \cdot (x_i - \mu)^2 = \sum_{i=1}^n (p_i \cdot x_i^2) - \mu^2 \quad \text{Var}(X) = \sigma^2 = \int (x - \mu)^2 f(x) dx = \int x^2 f(x) dx - \mu^2$$

Variance is non-negative because the squares are positive or zero.

$$\text{Var}(X) \geq 0.$$

$$\text{Var}(X + a) = \text{Var}(X).$$

$$\text{Var}(aX) = a^2 \text{Var}(X).$$

$$\begin{aligned} \text{Var}(X) &= E[X^2 - 2X E[X] + (E[X])^2] \\ &= E[X^2] - 2E[X]E[X] + (E[X])^2 \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

Higher moments

The n th moment (first moment = mean)

$$E(X^n) = \sum x_i^n p_i$$

$$E(X^n) = \int x^n f(x) dx$$

The n th central moment (second central moment = variance)

$$E((X - \mu)^n) = \sum (x_i - \mu)^n p_i$$

$$E((X - \mu)^n) = \int (x - \mu)^n f(x) dx$$

Normalized the n th moment

(third normalized moment = skewness, fourth normalized moment = kurtosis)

$$E((X - \mu)^n) / \sigma^n$$

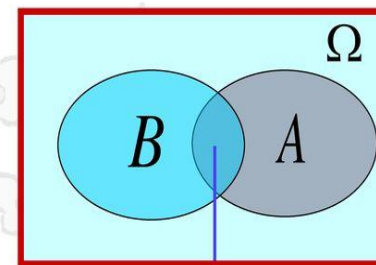
→ Cumulant

Conditional Probability:

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

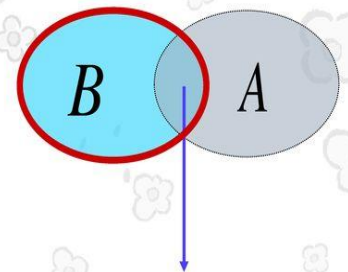
- $P(A|B) = \frac{P(A \cap B)}{P(B)}$, $P(B) \neq 0$
- $P(B|A) = \frac{P(A \cap B)}{P(A)}$, $P(A) \neq 0$

条件概率 $P(A|B)$ 的样本空间 (红色领域)



Sample space

$P(AB)$



Reduced sample space
given event B

$P(A|B)$

Note: $P(A \cap B) = P(AB)$

乘法公式 **Multiplicative Rule**

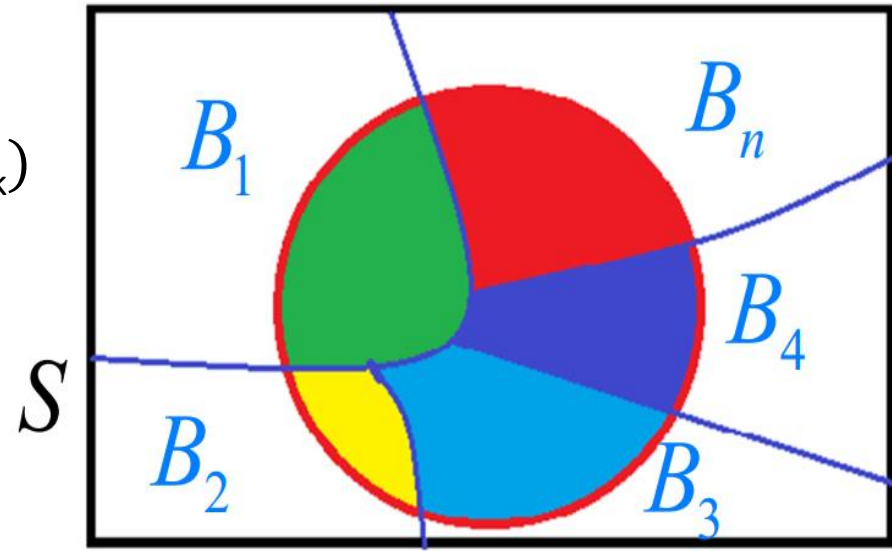
$$P(AB) = P(B) \cdot P(A | B) = P(A) \cdot P(B | A)$$



$$P(A_1 A_2 \dots A_{n-1} A_n) = P(A_1) P(A_2 | A_1) P(A_3 | A_1 A_2) \dots P(A_n | A_1 A_2 \dots A_{n-1})$$

全概率 formula of total probability

$$P(A) = \sum_{k=1}^n P(AB_k) = \sum_{k=1}^n P(A|B_k)P(B_k)$$



$$B_i \cap B_j = \emptyset \text{ for } i \neq j, B_1 \cup B_2 \cup \dots = S = \Omega$$

Note: decomposition $A = AB_1 + AB_2 + \dots + AB_n$

全概率: 由所有已知(n个)原因B推断结果A

Multiplicative Rule:

- $P(A \cap B) = P(A | B)P(B)$
- $P(A \cap B) = P(B | A)P(A)$

Where,

- $P(A)$: marginal probability of A.
- $P(B)$: marginal probability of B.
- $P(B | A)$: The conditional probability.

Independent Events:

- If A has no effect on B, we said that A,B are independent events.

Then,

- 1. $P(A \cap B) = P(B)P(A)$
- 2. $P(A | B) = P(A)$
- 3. $P(B | A) = P(B)$

Marginal Probability:

- Definition:

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

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

That is,

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

贝叶斯公式 Bayes formula

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^n P(B_j)P(A|B_j)}$$

全概率公式“由原因推结果”；而贝叶斯公式“由结果推原因”

贝叶斯公式：由结果A推断各个原因 B_i

$$P(B|A) = P(\text{因}|\text{果}) = P(\text{因})P(\text{果}|\text{因})/P(\text{果})$$

Baye's Theorem

$$P(S \mid R) = \frac{P(R \mid S)P(S)}{P(R)} = \frac{P(R \mid S)P(S)}{P(R \mid S)P(S) + P(R \mid N)P(N)}$$

Posterior likelihood Prior
normalizing constant

$$P(R) = P(X=R, Y=S) + P(X=R, Y=N) = P(X=R \mid Y=S)P(Y=S) + P(X=R \mid Y=N)P(Y=N)$$

S: south, N: north, R: rice

$$P(W \mid L) = \frac{P(L \mid W)P(W)}{P(L)} = \frac{P(L \mid W)P(W)}{P(L \mid W)P(W) + P(L \mid M)P(M)}$$

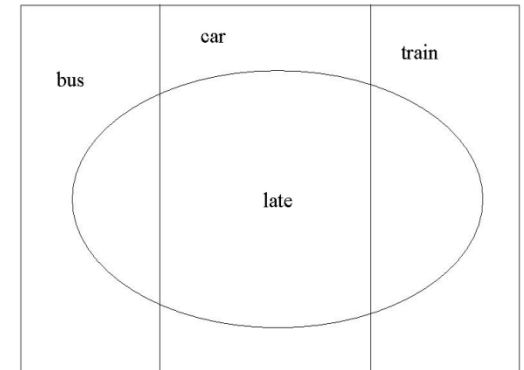
L: long, M: man, W: woman

Example of Bayes Theorem

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\sum_{k=1}^n P(A \mid B_k)P(B_k)}, i = 1, 2, \dots, n.$$

Suppose that Bob can decide to go to work by one of three modes of transportation, car, bus, or commuter train. Because of high traffic, if he decides to go by car, there is a 50% chance he will be late. If he goes by bus, which has special reserved lanes but is sometimes overcrowded, the probability of being late is only 20%. The commuter train is almost never late, with a probability of only 1%, but is more expensive than the bus.

- (a) Suppose that Bob is late one day, and his boss wishes to estimate the probability that he drove to work that day by car. Since he does not know which mode of transportation Bob usually uses, he gives a prior probability of 1/3 to each of the three possibilities. What is the boss' estimate of the probability that Bob drove to work?
- (b) Suppose that a coworker of Bob's knows that he almost always takes the commuter train to work, never takes the bus, but sometimes, 10% of the time, takes the car. What is the coworkers probability that Bob drove to work that day, given that he was late?



(a) We have the following information given in the problem:

$$P(\text{bus}) = P(\text{car}) = P(\text{train}) = 1/3$$

$$P(\text{late} \mid \text{car}) = 0.5, P(\text{late} \mid \text{train}) = 0.01, P(\text{late} \mid \text{bus}) = 0.2$$

We want to calculate $\Pr\{\text{car} \mid \text{late}\}$. By Bayes Theorem, this is

$$P(\text{car} \mid \text{late})$$

$$= P(\text{late} \mid \text{car})P(\text{car}) / [P(\text{late} \mid \text{car})P(\text{car}) + P(\text{late} \mid \text{bus})P(\text{bus}) + P(\text{late} \mid \text{train})P(\text{train})]$$

$$= 0.5 \times 1/3 / [0.5 \times 1/3 + 0.2 \times 1/3 + 0.01 \times 1/3]$$

$$= 0.7042$$

(b) Repeat the identical calculations as the above, but instead of the prior probabilities being $1/3$, we use $P(\text{bus}) = 0$, $P(\text{car}) = 0.1$, and $P(\text{train}) = 0.9$. Plugging in to the same equation with these three changes, we get

$$P(\text{car} \mid \text{late}) = 0.8475$$

Lung cancer rate (D) in population, 0.1 %.

The positive rate (P) for lung cancer patients, 99.9%;

the positive rate measurement for normal person (N), 10%.

What is the probability for a positive patient to have the lung cancer ?

$P(D) = 0.001$, $P(P | D) = .999$, $P(P | N) = 0.1$, $P(N) = 1 - P(D) = 0.999$.

$$P(D | P) = \frac{P(P | D)P(D)}{P(P | D)P(D) + P(P | N)P(N)} = \frac{0.999 * 0.001}{0.999 * 0.001 + 0.1 * 0.999}$$
$$= 0.0099 = 1\%$$

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 $P(\bar{T}|\bar{D})$

p-value and two probabilities

- Observed evidence: E , (e.g., data)
- Hypothesis Rule: R , (e.g. gene A regulates gene B)
- p-value: $p(E | \bar{R})$, where \bar{R} is null hypothesis
- But we want to know: $p(R | E)$

$$p(R | E) = 1 - \frac{p(\bar{R})}{p(E)} p(E | \bar{R})$$