
ACADGILD
Mastering Data Science

Statistics

Session 6 – Introduction to Statistics

Agenda

Statistics

- Statistics is the science of collecting, organizing, presenting, analyzing, and interpreting data to help in making more effective decisions.
- Statistical Analysis is implemented to manipulate, summarize and investigate data, so that useful decision-making information results are obtained.

Types of Statistics

- Descriptive Statistics is a method of organizing, summarizing, and presenting data in an informative way.
- Inferential Statistics is a method which is used in determining something about a population on the basis of a sample.
  - Population - The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest.
  - Sample – A portion, or part, of the population of interest.

Introduction to Basic Terms

➢ Population - A collection/set of individuals/objects/events whose properties are to be analyzed. There are two kinds:

- Finite
- Infinite

➢ Sample - A population subset.

Introduction to Basic Terms

- ➢ Variable - A characteristic about each individual element of a population/sample.
- ➢ Data (singular) - A value of the associated variable with one element of a population/sample. This value may be a number, a word, or a symbol.
- ➢ Data (plural) - A set of values collected for the variable from each of the elements belonging to the sample.
- ➢ Experiment - A planned activity whose results yield a set of data.
- ➢ Parameter - A numerical value which summarizes the entire population data.
- ➢ Statistics - A numerical value which summarizes the sample data.

Two Kinds of Variables

Qualitative, or Attribute, or Categorical, Variable

➢ A variable that categorizes or describes a population element.

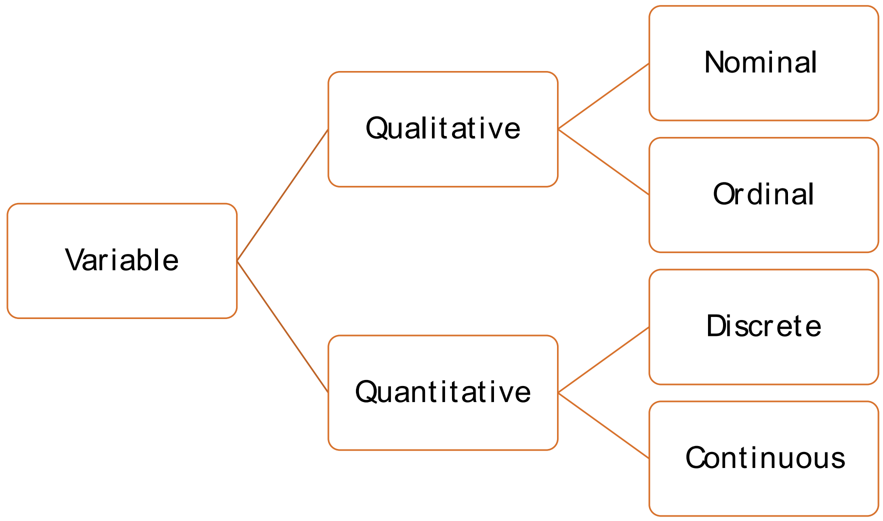
Note: Arithmetic operations such as addition and averaging, are not meaningful for data resulting from a qualitative variable.

Quantitative, or Numerical, Variable

➢ A variable that quantifies a population element.

Note: Arithmetic operations such as addition and averaging, are meaningful for data resulting from a quantitative variable.

Two Kinds of Variables



Two Kinds of Variables

- Nominal Variable - A qualitative variable that categorizes (or describes, or names) a population element.
- Ordinal Variable - A qualitative variable that incorporates an ordered position or ranking.
- Discrete Variable - A quantitative variable that can assume a countable number of values.
  - This can assume values corresponding to the isolated points along a line interval.
  - There is a gap between any two values
- Continuous Variable - A quantitative variable that can assume an uncountable number of values.
  - This can assume any value along a line interval
  - Including every possible value between any two values

➢ Let x1, x2, x3,.., xn be the realized values of a random variable 'X', from a sample of size 'n'.

The sample arithmetic mean is defined as:

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

Example

➢ The systolic blood pressure of seven middle aged men were as follows:

151, 124, 132, 170, 146, 124 and 113.

The Mean is $\bar{x} = \dfrac{(151+124+132+170+146+124+113)}{7}$

$= 137.14$

Median and Mode

➢ The median for the sample data arranged in an increasing order is defined as:

i. If "n" is an odd number - Middle value

ii. If "n" is an even number - Midway between the two middle values

➢ The mode is the most commonly occurring value.

Median and Mode

Example - n is odd

The re-ordered systolic blood pressure data seen earlier are:

113, 124, 124, 132, 146, 151, and 170.

➢ The Median is the middle value of the ordered data, i.e. 132.

➢ Two individuals have systolic blood pressure = 124 mm Hg, so the Mode is 124.

Median and Mode

Example – n is even

Six men with high cholesterol participated in a study to investigate the effects of diet on cholesterol level. At the beginning of the study, their cholesterol levels (mg/dL) were as follows:

366, 327, 274, 292, 274 and 230

Rearrange the data in numerical order as follows:

230, 274, 274, 292, 327 and 366.

- ➢ The Median is half way between the middle two readings, i.e. (274+292) / 2 = 283.
- ➢ The mode between the two men having the same cholesterol level = 274.

Mean Vs. Median

➢ If the histogram of the data is right-skewed then large sample values tend to inflate the mean.

➢ If the distribution is skewed then the median is not influenced by large sample values and is a better measure of centrality.

Note - If mean = median = mode then the data are said to be symmetrical.

For example,

➢ In the CK measurement study, the sample mean = 98.28.

➢ The median = 94.5, i.e. mean is larger than median indicating that mean is inflated by two large data values 201 and 203.

Measures of Dispersion

➢ The concept Measures of Dispersion characterize how to spread out the distribution, i.e., how variable the data are.

➢ The commonly used dispersion measures include:

- Range
- Variance and Standard Deviation

Range

➢ The Range is the difference between the largest and the smallest observations in the sample.

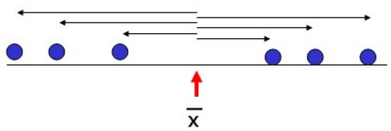
➢ For example, the minimum and maximum blood pressure is 113 and 170 respectively. Hence the range is 57 mmHg

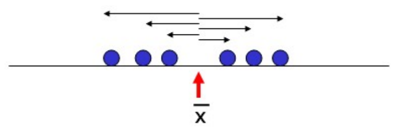
- Easy to calculate;
- Implemented for both “best” or “worst” case scenarios
- Too sensitive for extreme values

Sample Variance

➢ The sample variance, s2, is the arithmetic mean of the squared deviations from the sample mean:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$





Standard Deviation

➢ The sample standard deviation (s) is the square-root of the variance.

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

➢ The sample standard deviation has an advantage of being in the same units as the original variable (x).

Population Vs. Sample

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Vs.

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

Population Mean

Sample Mean

Population Vs. Sample

	Population	Sample
Size	N	n
Mean		

Population Vs. Sample

	Population	Sample
Size	N	n
Mean		
Variance		

Population Vs. Sample

➢ The variance of a population is:

Population Mean

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

Population Size

➢ The variance of a sample is:

Sample Mean

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Note! the denominator is sample size (n) minus one !

Population Vs. Sample

➢ The square root of the variance is termed as the Standard Deviation, thus:

- The population Standard Deviation = $\sigma = \sqrt{\sigma^2}$
- The Sample Standard deviation = $s = \sqrt{s^2}$

Chebysheff's Theorem

- A more general interpretation of the standard deviation is derived from Chebysheff's Theorem, which applies to all shapes of histograms (except bell shaped).
- The proportion of observations in any sample that lie within k standard deviations of the mean is at least:

$$1-\frac{1}{k^2} \quad for \quad k>1$$

For k=2 (say), the theorem states that at least 3/4 of all observations lie within 2 standard deviations of the mean. This is a "lower bound" compared to Empirical Rule's approximation (95%).

Set operations

- ➢ Union (A U B)
- ➢ Intersection (A B)

Venn diagrams

- ➢ Basic operations on Venn diagrams

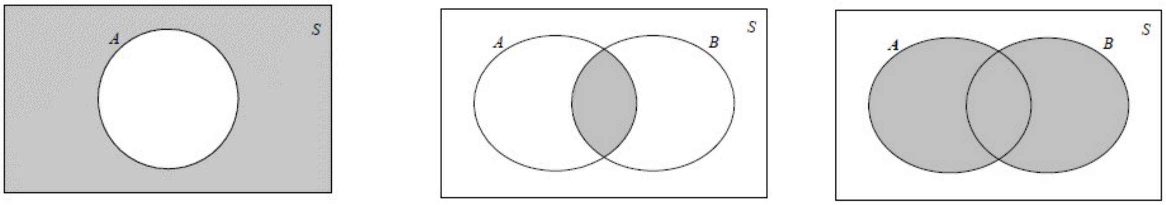
Basic probability axioms

- ➢ P (S) = 1
- ➢ P (A) >= 0 for all A S
- ➢ P (A U B) = P(A) + P(B) – P (A B)

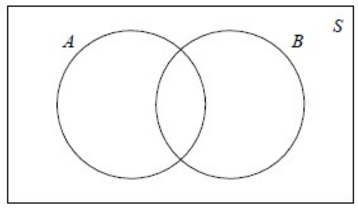
Conditional probability

- ➢ P(A| B) = P (A B)/ P(B)

Bayes theorem

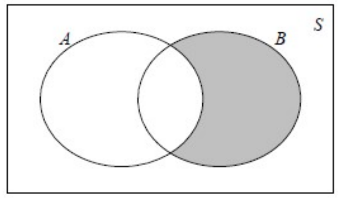


For the Venn diagram below, use shading to identify the following regions:

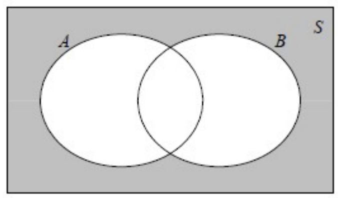


(i) $A' \cap B$ (ii) $A' \cap B'$ (iii) $(A \cap B)'$ (iv) $(A \cup B)'$

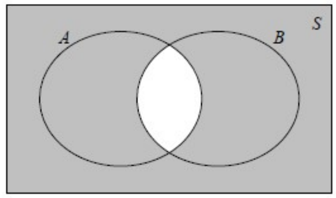
(i) $A' \cap B$ is everything not in set A **and** in set B:



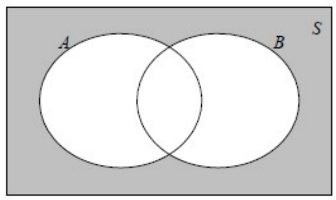
(ii) $A' \cap B'$ is everything not in set A **and** not in set B:



(iii) $(A \cap B)'$ is everything not in set A **and** set B (*ie* everything outside of the intersection of A and B).



(iv) $(A \cup B)'$ is everything not in set A **or** set B (*ie* everything outside of the union of A and B).



Probability Fundamentals

- Strength of belief
- A number between 0 and1 that expresses an opinion about the likelihood of an event
- Probability of an event that is certain to occur is 1
- Probability of an event that is certain to NOT occur is 0

Summation Principle

Probability that an event will occur plus probability that it will not occur equals 1

Probability of all possible outcomes of a chance event is always equal to 1

- Blood type: What is p[AB] given p[O]=0.46, p[A]=.40, and p[B]=.10?
- Fraternal triplets: What is the probability of at least one boy and one girl?

3.a. Probability

- Probability is a numerical way of describing how likely something is to happen.
- One of the fundamental methods of calculating probability is by using set theory.
- A set is defined as a collection of objects and each individual object is called an element of that set.
- Example from number of credit cards data, the distinct number of credit cards owned form a set

  # Cards = {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}

  - Numbers present on a dice form a set:

    Dice = {1, 2, 3, 4, 5, 6}

  The sample space (S) is the set of all possible outcomes that might be observed for an event/experiment.

- If each of the elements in the sample space are equally likely, then we can define the probability of event A as

  P(A) = (# elements in A)/(# elements in sample space)

  e.g. P(# Cards = 1) = (# of customers having 1 card)/(Total number of customers) = 100/1000 = 0.10 = 10%

  e.g. Probability of rolling an even number on a dice

  Sample space (S) = {1, 2, 3, 4, 5, 6}

  Event (A) = {2, 3, 4}

  P(A) = 3/6 = 0.5 = 50%

Why is it important from analytics perspective?

- What we do: analyze historical data to find pattern under assumption that past is a reflection of future.
- By means of probability theory, predict the future using historical patterns.

Find the probability of rolling an even number on an ordinary dice

Solution

We have a sample space of S = {1, 2, 3, 4, 5, 6}.

Defining "throwing an even number" as event A, we have A = {2, 4, 6}.

So the probability of throwing an even number is given by:

P(A) = [number of elements in A]/ [number of elements in S] = 3/ 6 = 0.5

One card is picked from an ordinary pack of 52 playing cards. What is the probability of obtaining:

(i) a diamond

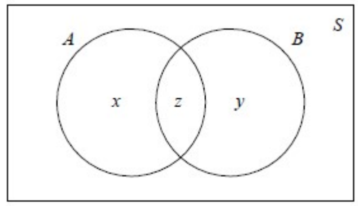
(ii) an ace

(iii) the ace of diamonds

(iv) a jack, queen or king.

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

On a Venn diagram we have:



It can be seen that $P(A) = x + z$, and $P(B) = y + z$

$P(A \cup B)$ can also be expressed as $P(A' \cap B) + P(A \cap B') + P(A \cap B)$
$= x + y + z$

$= (x + z) + (y + z) - z$

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

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

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

$$P(A \mid B) = \frac{P(A \cap B)}{P(A \cap B) + P(A' \cap B)}$$

A and B are independent if $P(A) = P(A \mid B) = P(A \mid B')$

Given that:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Then if A and B are independent:

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

Example

Two dice are thrown. Find the probability of rolling a 5 on both dice.

Solution

$$A = \{\text{roll a 5 on the 1st dice}\} \quad \Rightarrow \quad P(A) = \tfrac{1}{6}$$
$$B = \{\text{roll a 5 on the 2nd dice}\} \quad \Rightarrow \quad P(B) = \tfrac{1}{6}$$

Since these events are independent:

$$P(A \cap B) = P(A)P(B) = \tfrac{1}{6} \times \tfrac{1}{6} = \tfrac{1}{36}$$

Consider our previous example of picking a card from an ordinary pack of cards:

$A = \{\text{pick a spade}\}$ $B = \{\text{pick an } 8\}$

Calculate the probability of picking a spade *given* that we have picked an 8, *ie* calculate $P(A \mid B)$.

Solution

We obtain the same answer as before – but our calculation is much simpler:

$$A \cap B = \{8 \text{ of spades}\} \quad \Rightarrow \quad P(A \cap B) = \tfrac{1}{52}$$

$$B = \{\text{pick an } 8\} \quad \Rightarrow \quad P(B) = \tfrac{4}{52}$$

$$\Rightarrow \quad P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{52}}{\frac{4}{52}} = \frac{1}{4}$$

Bayes Theorem

Most common form:

$$P(F \mid E) = \frac{P(E \mid F)P(F)}{P(E)}$$

Expanded form (using law of total probability):

$$P(F \mid E) = \frac{P(E \mid F)P(F)}{P(E \mid F)P(F) + P(E \mid F^c)P(F^c)}$$

Proof:

$$P(F \mid E) = \frac{P(EF)}{P(E)} = \frac{P(E \mid F)P(F)}{P(E)}$$

Question

Suppose an HIV test is 98% effective in detecting HIV, i.e., its "false negative" rate = 2%. Suppose furthermore, the test's "false positive" rate = 1%.

- 0.5% of population has HIV
- Let E = you test positive for HIV
- Let F = you actually have HIV

What is P(F|E) ?

Solution

Suppose an HIV test is 98% effective in detecting HIV, i.e., its "false negative" rate = 2%. Suppose furthermore, the test's "false positive" rate = 1%.

- 0.5% of population has HIV
- Let E = you test positive for HIV
- Let F = you actually have HIV

What is P(F|E) ?

Solution:

$$
\begin{aligned}
P(F \mid E) &= \frac{P(E \mid F)P(F)}{P(E \mid F)P(F) + P(E \mid F^c)P(F^c)} \\
&= \frac{(0.98)(0.005)}{(0.98)(0.005) + (0.01)(1 - 0.005)} \\
&\approx 0.330
\end{aligned}
$$

3.b. Random Variables

I. Definition

II. Types of Random Variables

1. Discrete
2. Continuous

III. Distribution and Probability Density functions of Random Variables

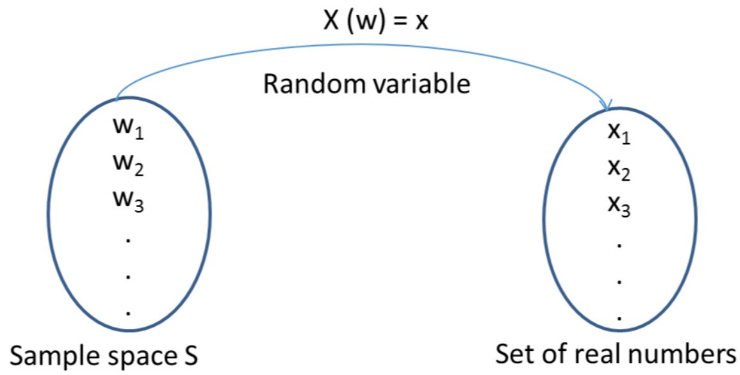
IV. Expected value (or Mean) of Random Variables

V. Variance of Random Variables

VI. Coefficient of skewness of Random Variables

3.b. Random variables- Definition

➢ A random variable is a function or a rule which maps each event in a sample space to real numbers.



➢ So, if w is an element of the sample space S (i.e. w is one of the possible outcomes of the experiment concerned) and the number x is associated with this outcome, then X(w) = x .

➢ Convention:

   ➢ Denote random variable by capital letter "X"

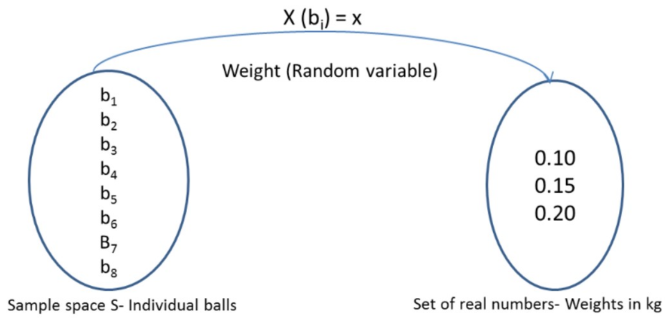
   ➢ Denote the outcome or possible values by small letter "x" i.e. X(w) = x

Example:

➢ Suppose there are 8 balls in a bag. The random variable X is the weight, in kg, of a ball selected at random. Balls 1, 2 and 3 weigh 0.1kg, balls 4 and 5 weigh 0.15kg and balls 6, 7 and 8 weigh 0.2kg. Using the notation above, write down this information.

Solution:

➢ X(b1) = 0.10 kg, X(b2) = 0.10 kg, X(b3) = 0.1 kg
X(b4) = 0.15 kg, X(b5) = 0.15 kg
X(b6) = 0.2 kg, X(b7) = 0.2 kg



3.b. Types of Random variables

There are two types of Random Variables

1. Discrete Random Variables
2. Continuous Random Variables

3.b. Discrete Random variables

Definition:

➢ The set of all possible values of the outcome (or x) takes discrete values

- ➢ e.g. Outcome of rolling a dice= {1, 2, 3, 4, 5, 6}
- ➢ Or # credit cards owned by an individual = {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}

Probabilities:

➢ Probabilities are defined on events (subsets of the sample space S).

So what is meant by " P(X = x) " ?

- ➢ Suppose sample space consists of eight events {s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8}
- ➢ Let the outcome for
  - ➢ E1 = {s_1, s_2, s_3} be associated with number x1
  - ➢ E2 = {s_4, s_5} be associated with number x2
  - ➢ E3 = {s_6, s_7, s_8} be associated with number x3
- ➢ P(X = x1) is meant P(E1)
- ➢ P(X = x2) is meant P(E2)
- ➢ P(X = x3) is meant P(E3)

Probability functions

- The function $f_X(x) = P(X = x)$ for each x in the range of X is the probability function (PF) of X
- It specifies how the total probability of 1 is divided up amongst the possible values of X
- Thus, gives the probability distribution of X.
- Also known as "probability distribution functions" (pdf)

Following are the requirements for a function to qualify as the probability function of a discrete random variable:

- $f_X(x) >= 0$ for all x within the range of X
- $\sum f_X(x) = 1$

Cumulative distribution functions

- Gives the probability that X assumes a value that does not exceed x.
- Denoted as $F_X(x) = P(X <= x)$ where max $(F_X(x)) = 1$

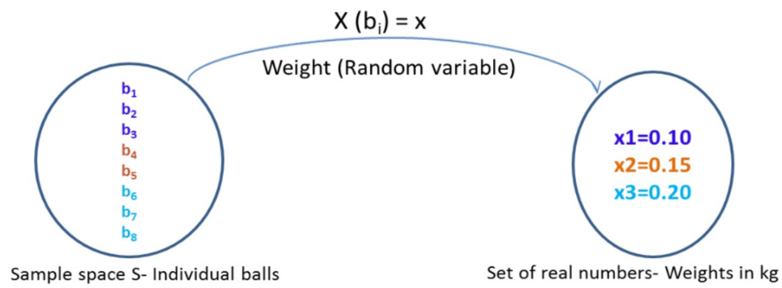
Example:

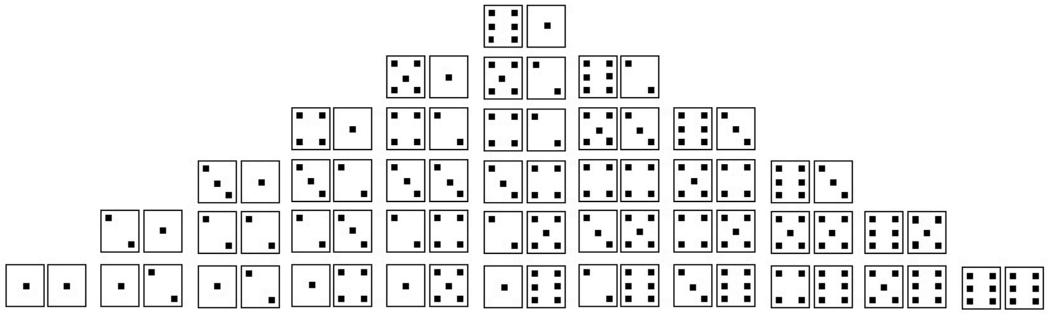
- Suppose there are 8 balls in a bag. The random variable X is the weight, in kg, of a ball selected at random. Balls 1, 2 and 3 weigh 0.1kg, balls 4 and 5 weigh 0.15kg and balls 6, 7 and 8 weigh 0.2kg. Write down the different probability distribution functions.

Solution:

- $f_X(0.10) = P(X=0.10)$ = probability the ball b1 or b2 or b3 is selected out of 8 balls = 3/8
- $f_X(0.15) = P(X=0.15)$ = probability the ball b4 or b5 is selected out of 8 balls = 2/8
- $f_X(0.20) = P(X=0.20)$ = probability the ball b6 or b7 or b8 is selected out of 8 balls = 3/8

- $F_X(0.10) = P(X <= 0.10) = P(X=0.10) = 3/8$
- $F_X(0.15) = P(X<=0.15) = P(X=0.10)+ P(X=0.15) = 2/8 + 3/8 = 5/8$
- $F_X(0.20) = P(X<=0.20) = P(X=0.10) + P(X=0.15) + P(X=0.20) = 3/8 + 2/8 + 3/8 = 8/8 = 1$

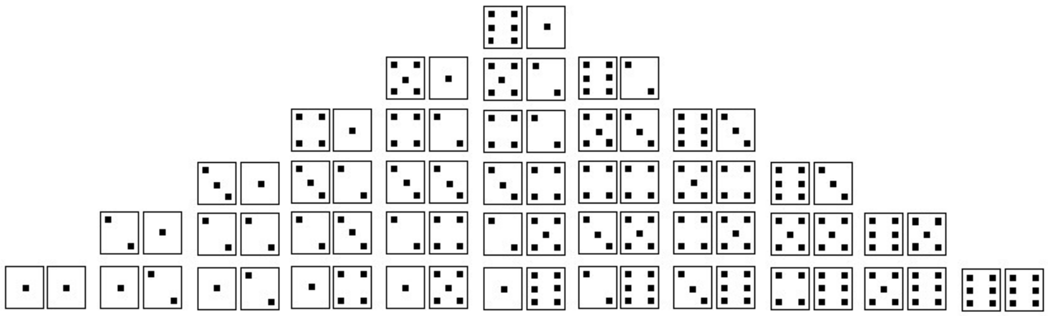




Value (x_i)	2	3	4	5	6	7	8	9	10	11	12
Freq (n_i)	1	2	3	4	5	6	5	4	3	2	1
Prob (p_i)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Cumulative Distribution

- The cumulative distribution represents the summation of the probabilities.
- The number 2 occurs 1/36 of the time, the number 3 occurs 2/36 of the time.
- Therefore a number equal to 3 or less will occur 3/36 of the time.



Value	2	3	4	5	6	7	8	9	10	11	12
Prob (p_i)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
Cdf	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$

Probability Distribution Function (pdf)

- The probabilities form a pdf. The sum of the probabilities must sum to 1.
- The distribution can be characterized by two variables, its mean and standard deviation

$$\sum_{i=1}^{11} p_i = 1$$

Mean

- The mean is simply the expected value from rolling the dice, this is calculated by multiplying the probabilities by the possible outcomes (values).
- In this case it is also the value with the highest frequency (mode)

$$E(x) = \sum_{i=1}^{11} p_i x_i = \frac{252}{36} = 7$$

3.b. Continuous Random variables

Definition:

➢ The set of possible values taken by a continuous random variable falls in an interval (or a collection of intervals) on the real line:

  ➢ e.g. Salary of a set of individuals

  ➢ Mathematically examples $\{x: x > 0\}$ or $\{x: -\infty < x < \infty\}$ or $\{x: 0 < x < 1\}$

Probability Density Function

➢ First define the range or the interval in which the probability has to be determined.

➢ Say its (a, b).

➢ The probability associated is represented as $P(a < X < b)$ or $P(a \le X \le b)$.

➢ Also, it is the area under the curve of the probability density function (PDF) from a to b.

➢ So probabilities can be evaluated by integrating the PDF $f_X(x)$.

This relationship defines the PDF.

Mathematically

  ➢ $P(a < X < b) = \int_a^b f_X(x)\,dx$

The conditions for a function to serve as PDF are

  ➢ $f_X(x) \ge 0 \quad -\infty \le x \le \infty$

  ➢ $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$

Cumulative distribution function:

- The cumulative distribution function (CDF) is defined to be the function:
  - $F_X(x) = P(X \le x)$
- For a continuous random variable, FX (x) is a continuous, non-decreasing function, defined for all real values of x.
  - $F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$

3. b. Random variables- Expected values

Definition:

- Expected values are numerical summaries of important characteristics of the distributions of random variables.
- Expected values of a Random Variable "X" is denoted as E[X]
- Important Expected values are
    - Mean
    - Variance and Standard deviation

Mean:

- E[X] is a measure of central location
- For discrete case calculated as $E[X] = \sum(x_i * P_i)$ OR $E[X] = (\sum x * f_X(x))$
- For continuous case calculated as $E[X] = \int_{-\infty}^{\infty} x * f_X(x)\, dx$
- Usually denoted by μ

Variance:

- $Var[X] = E[\{X - E[X]\}^2]$
- $Var[X] = E[X^2] - E^2[X]$

Population mean (or expectation)

$$\mu = E(X) = \sum_x xP(X = x) \quad \text{or} \quad \int_{-\infty}^{\infty} x f_X(x)\, dx$$

$$E[g(X)] = \sum_x g(x)P(X = x) \quad \text{or} \quad \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$$

Population variance

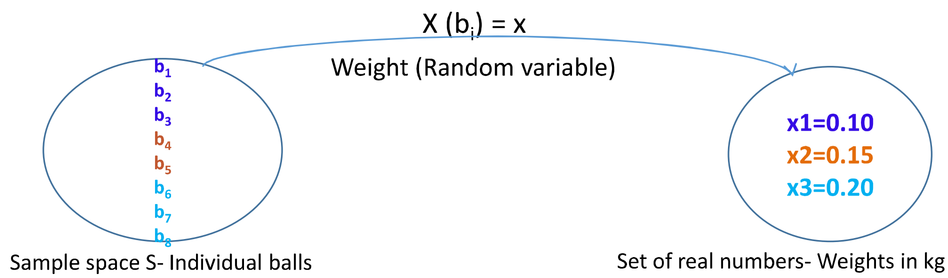
$$\sigma^2 = \mathrm{var}(X) = E[(X - \mu)^2] = E(X^2) - E^2(X)$$

Example:

- Suppose there are 8 balls in a bag. The random variable X is the weight, in kg, of a ball selected at random. Balls 1, 2 and 3 weigh 0.1kg, balls 4 and 5 weigh 0.15kg and balls 6, 7 and 8 weigh 0.2kg. Find mean and variance of weight.

Solution:

- $f_X(0.10) = P(X=0.10) = 3/8$
- $f_X(0.15) = P(X=0.15) = 2/8$
- $f_X(0.20) = P(X=0.20) = 3/8$
- $E[X] = \sum P_i * x_i = 3/8 * 0.10 + 2/8 * 0.15 + 3/8 * 0.20 = 1.2/8 = 0.15$ kg
- $Var[X] = E[X^2] - E^2[X] = 0.024375 - 0.0225 = 0.001875$ kg^2



Two Types of Random Variables

Discrete Random Variable

- Takes on a countable number of values
- For example, values on the roll of dice: 2, 3, 4, .., 12

Continuous Random Variable

- Values are not discrete, not countable
- For example, time (30.1 minutes? 30.10000001 minutes?)

Analogy

- Integers are discrete, while Real Numbers are Continuous

Laws of Expected Value

- $E(C) = C$
  - The expected Value of a Constant is just the value of the constant.

- $E(X + C) = E(X) + C$

- $E(CX) = cE(X)$
  - We can “pull” a constant out of the expected value expression (either as part of a sum with a random variable X or as a coefficient of random variable X).

Laws of Variance

➢ $V(c) = 0$

- The Variance of constant (c) is zero.

➢ $V(X + c) = V(X)$

- The Variance of random variable and a constant is just the variance of the random variable (per 1 above).

➢ $V(cX) = c^2 V(X)$

- The Variance of a random variable and a constant co-efficient is the co-efficient squared times in the variance of the random variable.

Probability Density Functions

Unlike a discrete random variable, a continuous random variable is one that can assume an uncountable number of values.

- ➢ We cannot list the possible values because there is an infinite number of them.
- ➢ The probability of each individual value is virtually 0 as there is an infinite number of values

Point Probabilities are Zero

If the probability of each individual value is virtually 0 then there is an infinite number of values.

Thus, we can determine the probability of a range of values only.

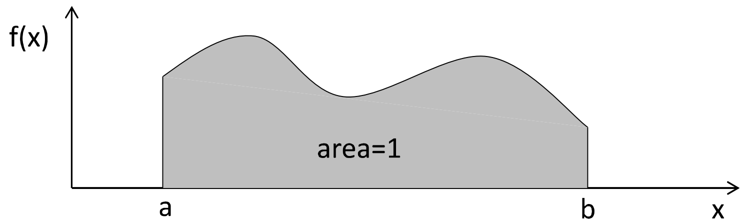
➢ For example, with a discrete random variable like tossing a die, it is meaningful to talk about P(X=5)

➢ In a continuous setting (e.g. with time as a random variable), the probability the random variable of interest say task length, takes exactly 5 minutes is infinitely small, hence P(X=5) = 0.

Probability Density Function

A function f(x) is called a Probability Density Function over the range $a \le x \le b$ if it meets the following requirements:

1. $f(x) \ge 0$ for all x between a and b, and



2. The total area under the curve between a and b is 1.0

Email us - support@acadgild.com