IAM 530
ELEMENTS OF PROBABILITY AND STATISTICS

LECTURE 3-RANDOM VARIABLES

VARIABLE

- Studying the behavior of random variables, and more importantly functions of random variables is essential for both the theory and practice of statistics.

  Variable: A characteristic of population or sample that is of interest for us.

  Random Variable: A function defined on the sample space *S* that associates a real number with each outcome in *S*. In other words, a numerical value to each outcome of a particular experiment.

  - For each element of an experiment's sample space, the random variable can take on exactly one value

TYPES OF RANDOM VARIABLES

We will start with univariate random variables.

- **Discrete Random Variable:** A random variable is called discrete if its range consists of a countable (possibly infinite) number of elements.
- **Continuous Random Variable:** A random variable is called continuous if it can take on any value along a continuum (but may be reported "discretely"). In other words, its outcome can be any value in an interval of the real number line.

Note:

- Random Variables are denoted by upper case letters (X)
- Individual outcomes for RV are denoted by lower case letters (x)

DISCRETE RANDOM VARIABLES

EXAMPLES

- A random variable which takes on values in {0,1} is known as a Bernoulli random variable.
- Discrete Uniform distribution:

$$P(X = x) = \frac{1}{N}; \ x = 1,2,\ldots, N; \quad N = 1,2,\ldots$$

- Throw a fair die. P(X=1)=…=P(X=6)=1/6

DISCRETE RANDOM VARIABLES

- **Probability Distribution:** Table, Graph, or Formula that describes values a random variable can take on, and its corresponding probability (discrete random variable) or density (continuous random variable).
- **Discrete Probability Distribution:** Assigns probabilities (masses) to the individual outcomes.

PROBABILITY MASS FUNCTION (PMF)

- **Probability Mass Function**
  - $0 \le p_i \le 1$ and $\sum_i p_i = 1$
  - **Probability :**

$$P(X = x_i) = p_i$$

Example

Consider tossing three fair coins.

- Let X=number of heads observed.

- S={TTT, TTH, THT, HTT, THH, HTH, HHT, HHH}

- $P(X=0)=P(X=3)=1/8$; $P(X=1)=P(X=2)=3/8$

CUMULATIVE DISTRIBUTION FUNCTION (CDF)

Cumulative Distribution Function (CDF):

$$F(y) = P(Y \le y)$$

$$F(b) = P(Y \le b) = \sum_{y=-\infty}^{b} p(y)$$

$$F(-\infty) = 0 \quad F(\infty) = 1$$

$F(y)$ is monotonically increasing in y

Example

X= Sum of the up faces of the two die. Table gives value of *y* for all elements in *S*

1st\2nd	**1**	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

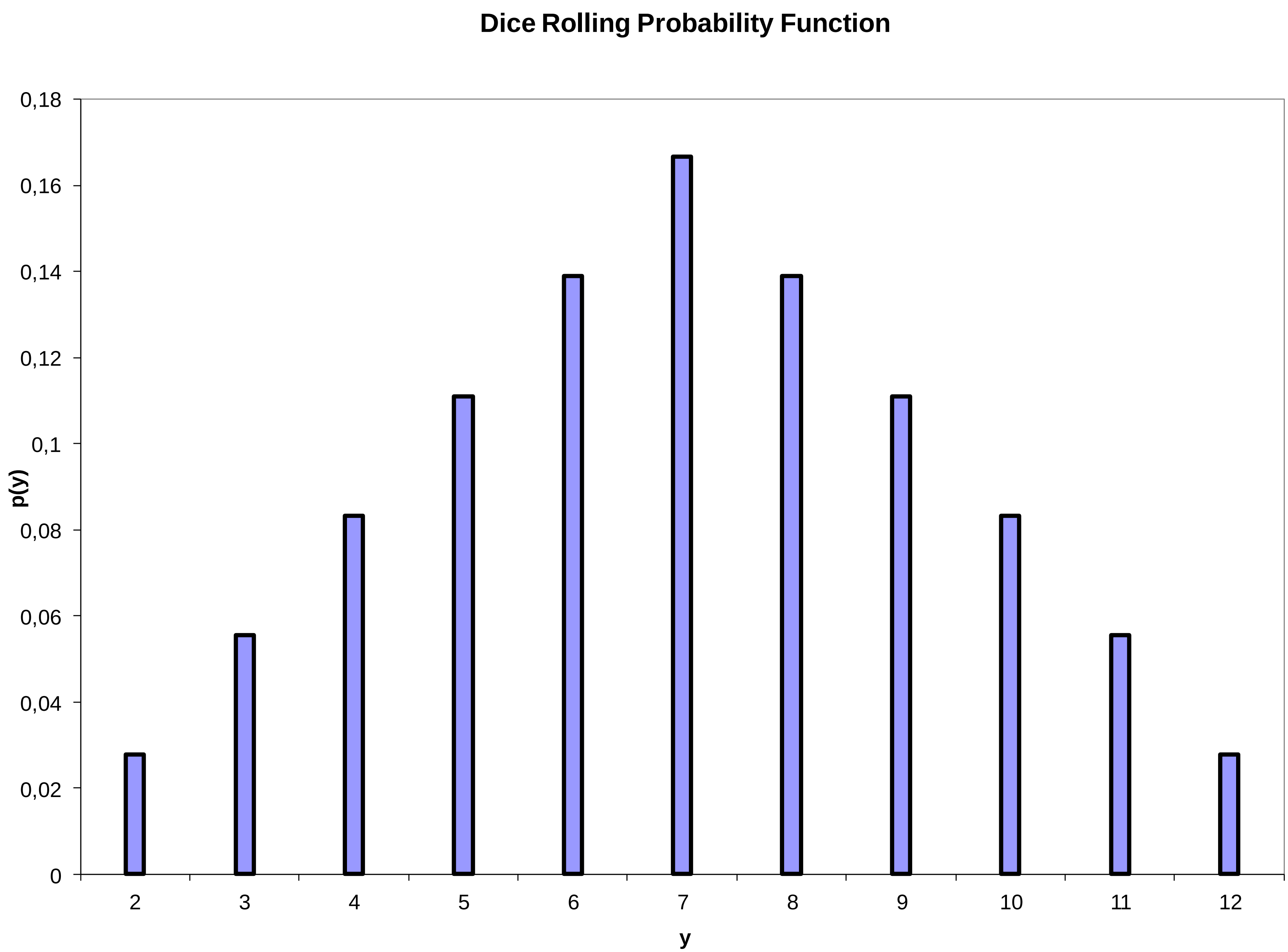
PMF and CDF

y	p(y)	F(y)
2	1/36	1/36
3	2/36	3/36
4	3/36	6/36
5	4/36	10/36
6	5/36	15/36
7	6/36	21/36
8	5/36	26/36
9	4/36	30/36
10	3/36	33/36
11	2/36	35/36
12	1/36	36/36

$$p(y) = \frac{\text{\# of ways 2 die can sum to } y}{\text{\# of ways 2 die can result in}}$$

$$F(y) = \sum_{t=2}^{y} p(t)$$

PMF-Graph


Dice Rolling Probability Function
0,18
0,16
0,14
0,12
0,1
0,08
0,06
0,04
0,02
0
p(y)
2
3
4
5
6
7
8
9
10
11
12
y

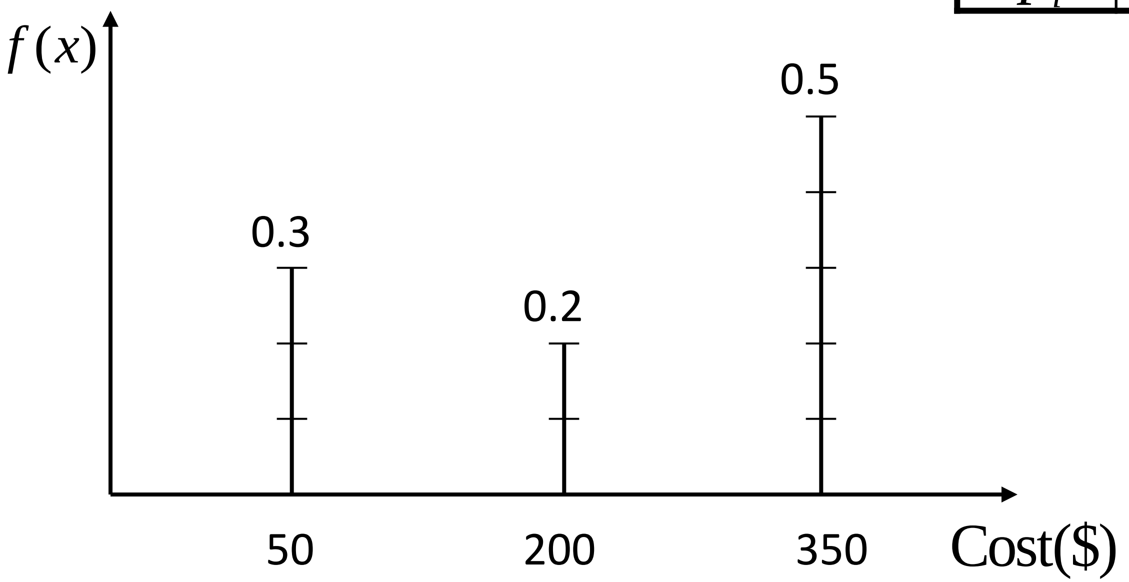
Example 2

- Machine Breakdowns
  - Sample space : $S = \{electrical, mechanical, misuse\}$
  - Each of these failures may be associated with a repair cost
  - State space : $\{50, 200, 350\}$
  - Cost is a random variable : 50, 200, and 350

– P (cost=50)=0.3, P (cost=200)=0.2, P (cost=350)=0.5

– 0.3 + 0.2 + 0.5 =1

x_i	50	200	350
p_i	0.3	0.2	0.5



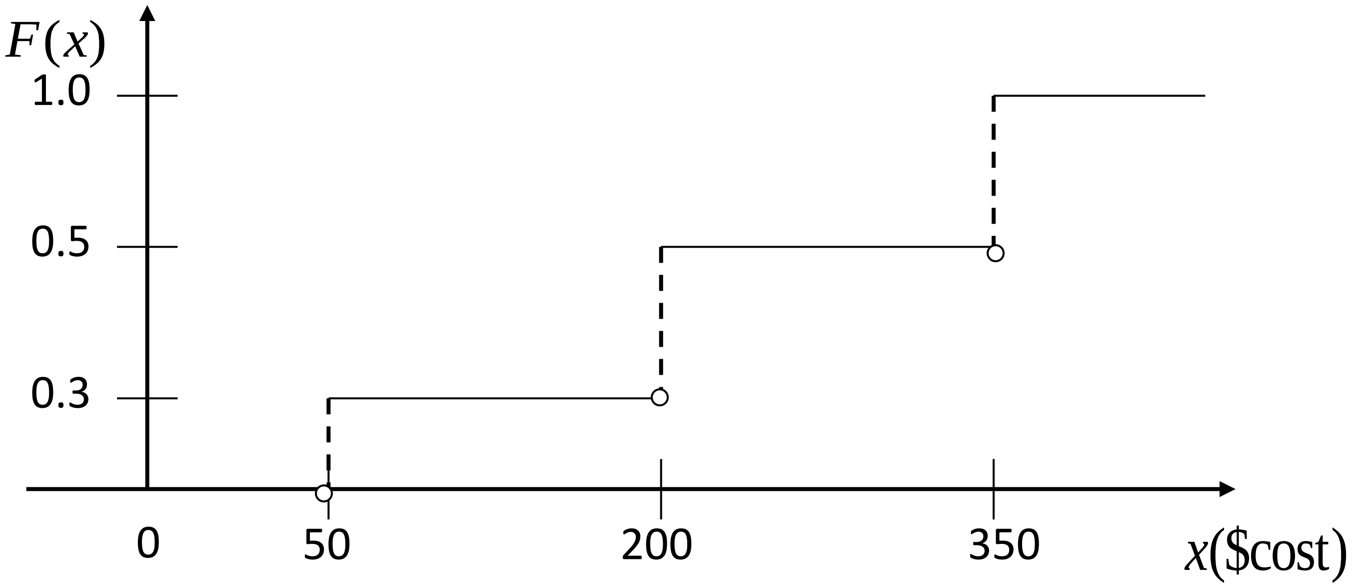
- Cumulative Distribution Function

$$-\infty < x < 50 \Rightarrow F(x) = P(\text{cost} \le x) = 0$$

$$50 \le x < 200 \Rightarrow F(x) = P(\text{cost} \le x) = 0.3$$

$$200 \le x < 350 \Rightarrow F(x) = P(\text{cost} \le x) = 0.3 + 0.2 = 0.5$$

$$350 \le x < \infty \Rightarrow F(x) = P(\text{cost} \le x) = 0.3 + 0.2 + 0.5 = 1.0$$



CONTINUOUS RANDOM VARIABLES

- When sample space is uncountable (continuous)
- For a continuous random variable *P(X = x) = 0*.

Examples:

- Continuous Uniform(a,b)

$$f(X) = \frac{1}{b-a} \quad a \le x \le b.$$

- Suppose that the random variable X is the diameter of a randomly chosen cylinder manufactured by the company.

PROBABILITY DENSITY FUNCTION (PDF)

- Probability Density Function
  - Probabilistic properties of a continuous random variable

$$f(x) \geq 0$$

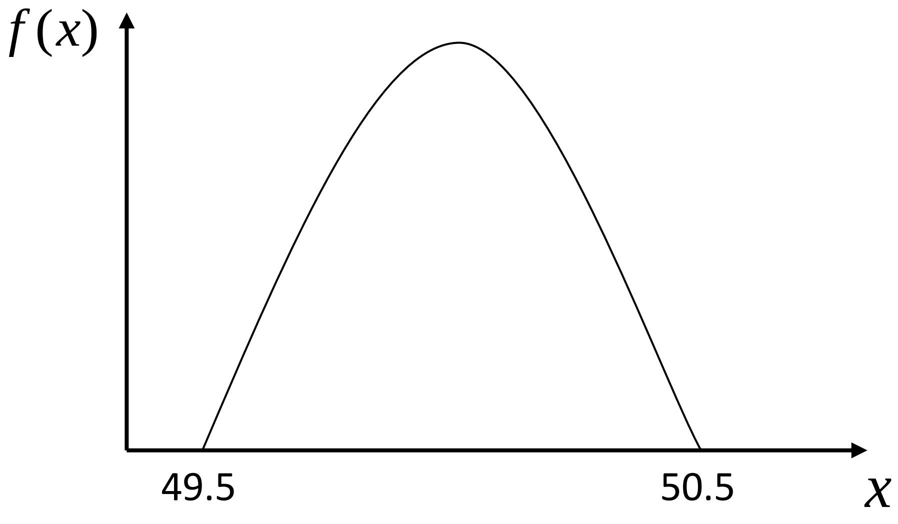
$$\int_{statespace} f(x)dx = 1$$

Example

– Suppose that the diameter of a metal cylinder has a p.d.f

$$f(x) = 1.5 - 6(x - 50.2)^2 \quad \text{for} \quad 49.5 \le x \le 50.5$$

$$f(x) = 0, \quad \text{elsewhere}$$



- This is a valid p.d.f.

$$\int_{49.5}^{50.5} (1.5 - 6(x-50.0)^2)\,dx = [1.5x - 2(x-50.0)^3]_{49.5}^{50.5}$$

$$= [1.5 \times 50.5 - 2(50.5 - 50.0)^3]$$

$$-[1.5 \times 49.5 - 2(49.5 - 50.0)^3]$$

$$= 75.5 - 74.5 = 1.0$$

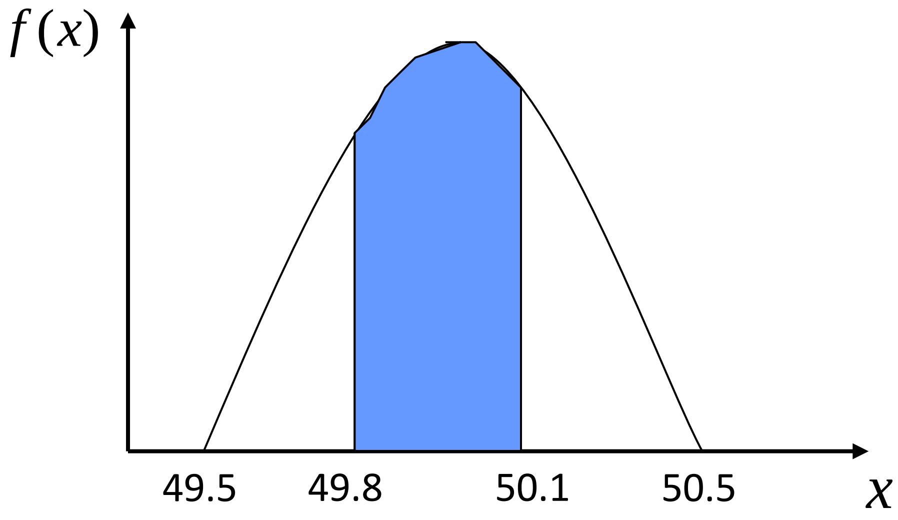
- The probability that a metal cylinder has a diameter between 49.8 and 50.1 mm can be calculated to be

$$\int_{49.8}^{50.1}(1.5-6(x-50.0)^2)dx=[1.5x-2(x-50.0)^3]_{49.8}^{50.1}$$

$$=[1.5\times 50.1-2(50.1-50.0)^3]$$

$$-[1.5\times 49.8-2(49.8-50.0)^3]$$

$$=75.148-74.716=0.432$$



CUMULATIVE DISTRIBUTION FUNCTION (CDF)

- $F(x) = P(X \le x) = \int_{-\infty}^{x} f(y)dy$
- $f(x) = \dfrac{dF(x)}{dx}$

- $$\begin{aligned} P(a < X \le b) &= P(X \le b) - P(X \le a) \\ &= F(b) - F(a) \end{aligned}$$

- $P(a \le X \le b) = P(a < X \le b)$

$$F(x) = P(X \le x) = \int_{49.5}^{x} (1.5 - 6(y - 50.0)^2)dy$$

$$= [1.5y - 2(y - 50.0)^3]_{49.5}^{x}$$

$$= [1.5x - 2(x - 50.0)^3] - [1.5 \times 49.5 - 2(49.5 - 50.0)^3]$$

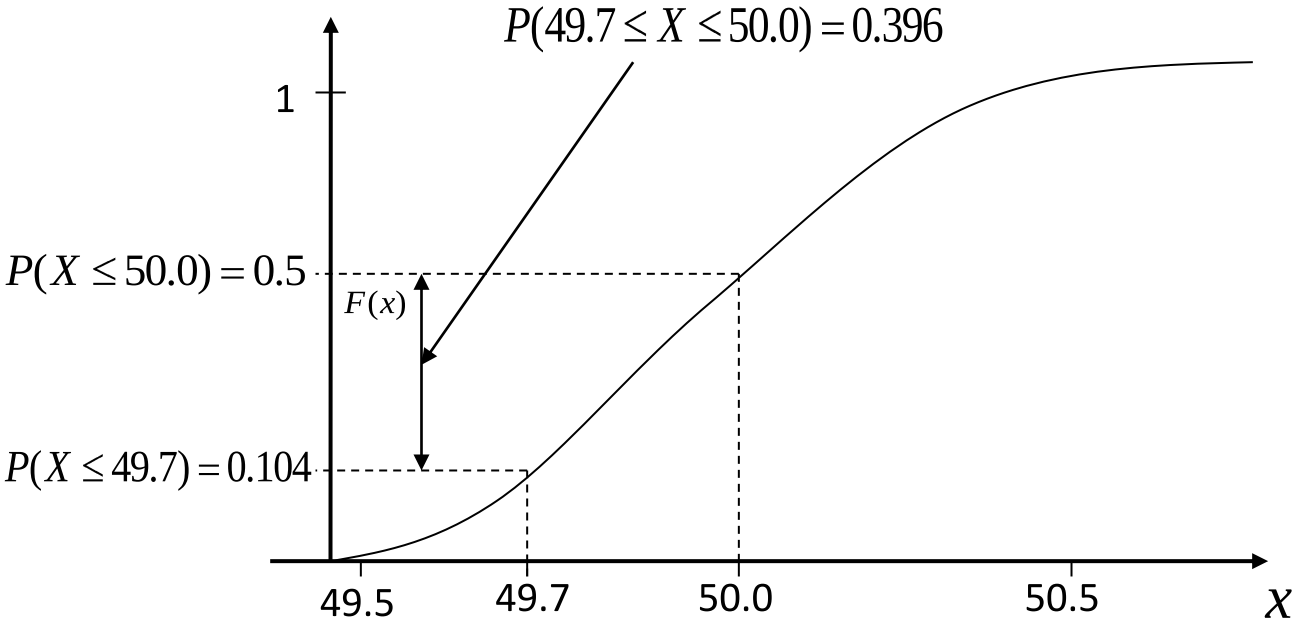
$$= 1.5x - 2(x - 50.0)^3 - 74.5$$

$$P(49.7 \le X \le 50.0) = F(50.0) - F(49.7)$$

$$= (1.5 \times 50.0 - 2(50.0 - 50.0)^3 - 74.5)$$

$$-(1.5 \times 49.7 - 2(49.7 - 50.0)^3 - 74.5)$$

$$= 0.5 - 0.104 = 0.396$$



$P(49.7 \leq X \leq 50.0) = 0.396$
1
$P(X \leq 50.0) = 0.5$
$F(x)$
$P(X \leq 49.7) = 0.104$
49.5
49.7
50.0
50.5
x

Example

- Suppose cdf of the random variable X is given as: $F(\mathrm{x}) = 4\mathrm{x}^3 - 6x^2 + 3x$

  Find the pdf for X.

$$\frac{dF(\mathrm{x})}{dx} = 12x^2 - 12x + 3 = 12\left(x^2 - x + \frac{1}{4}\right) = 12\left(\mathrm{x} - \frac{1}{2}\right)^2$$

THE EXPECTED VALUE

Let X be a rv with pdf $f_X(x)$ and $g(X)$ be a function of X. Then, the expected value (or the mean or the mathematical expectation) of $g(X)$

$$E\left[g(X)\right] = \begin{cases} \sum_x g(x) f_X(x), \ if\ X \ is\ discrete \\ \int_{-\infty}^{\infty} g(x) f_X(x)\, dx, \ if\ X \ is\ continuous \end{cases}$$

providing the sum or the integral exists, i.e., $-\infty < E[g(X)] < \infty$.

EXPECTED VALUE (MEAN) AND VARIANCE OF A DISCERETE RANDOM VARIABLE

- Given a discrete random variable X with values x_i, that occur with probabilities $p(x_i)$, the population mean of X is

$$E(X) = \mu = \sum_{all \ x_i} x_i \cdot p(x_i)$$

- Let X be a discrete random variable with possible values x_i that occur with probabilities $p(x_i)$, and let $E(x_i) = \mu$. The variance of X is defined by

$$V(X) = \sigma^2 = E\left[(X-\mu)^2\right] = \sum_{all \ x_i} (x_i - \mu)^2 p(x_i)$$

$$The \ s\tan dard \ deviation \ is$$

$$\sigma = \sqrt{\sigma^2}$$

Mean: $E(X) = \mu$

$$V(X) = \sigma^2 = E\left[(X - E(X))^2\right] = E\left[(X - \mu)^2\right] =$$

$$= \sum_{\text{all } x} (x-\mu)^2 p(x) = \sum_{\text{all x}} \left(x^2 - 2x\mu + \mu^2\right) p(x) =$$

$$= \sum_{\text{all x}} x^2 p(x) - 2\mu \sum_{\text{all } x} xp(x) + \mu^2 \sum_{\text{all x}} p(x) =$$

$$= E\left[X^2\right] - 2\mu(\mu) + \mu^2(1) = E\left[X^2\right] - \mu^2$$

Example – Rolling 2 Dice

y	p(y)	yp(y)	y^2p(y)
2	1/36	2/36	4/36
3	2/36	6/36	18/36
4	3/36	12/36	48/36
5	4/36	20/36	100/36
6	5/36	30/36	180/36
7	6/36	42/36	294/36
8	5/36	40/36	320/36
9	4/36	36/36	324/36
10	3/36	30/36	300/36
11	2/36	22/36	242/36
12	1/36	12/36	144/36
Sum	36/36= 1.00	252/36 =7.00	1974/36=5 4.833

$$\mu = E(Y) = \sum_{y=2}^{12} yp(y) = 7.0$$

$$\sigma^2 = E\left[Y^2\right] - \mu^2 = \sum_{y=2}^{12} y^2 p(y) - \mu^2$$

$$= 54.8333 - (7.0)^2 = 5.8333$$

$$\sigma = \sqrt{5.8333} = 2.4152$$

Example 2

- The pmf for the number of defective items in a lot is as follows

$$p(x) = \begin{cases} 0.35, x = 0 \\ 0.39, x = 1 \\ 0.19, x = 2 \\ 0.06, x = 3 \\ 0.01, x = 4 \end{cases}$$

Find the expected number and the variance of defective items.

Results: E(X)=0.99, Var(X)=0.8699

EXPECTED VALUE (MEAN) AND VARIANCE OF A CONTINUOUS RANDOM VARIABLE

- The expected value or mean value of a continuous random variable X with pdf $f(x)$ is

$$\mu = E(X) = \int_{all\ x} x f(x) dx$$

- The variance of a continuous random variable X with pdf $f(x)$ is

$$\sigma^2 = Var(X) = E(X-\mu)^2 = \int_{\text{all x}} (x-\mu)^2 f(x) dx$$

$$= E(X^2) - \mu^2 = \int_{\text{all x}} (x)^2 f(x) dx - \mu^2$$

Example

- In the flight time example, suppose the probability density function for X is

$$f(x)=\frac{4}{3},\quad 0\le x\le 0.5;\quad f(x)=\frac{2}{3},\quad 0.5<x\le 1.$$

- Then, the expected value of the random variable X is

$$E(X)=\int_0^1 xf(x)dx=\int_0^{0.5} x\cdot\frac{4}{3}dx+\int_{0.5}^1 x\cdot\frac{2}{3}dx=\frac{x^2}{2}\cdot\frac{4}{3}\Big|_0^{0.5}+\frac{x^2}{2}\cdot\frac{2}{3}\Big|_{0.5}^1$$

$$=\left(\frac{0.5^2}{2}\cdot\frac{4}{3}-\frac{0^2}{2}\cdot\frac{4}{3}\right)+\left(\frac{1^2}{2}\cdot\frac{2}{3}-\frac{0.5^2}{2}\cdot\frac{2}{3}\right)=\frac{5}{12}$$

- Variance

$$Var(X) = E \quad X - E(X) \quad ^{2} = \int_{0}^{1} \left(x - \frac{5}{12} \right)^{2} f(x)dx$$

$$= \int_{0}^{0.5} \left(x - \frac{5}{12} \right)^{2} \cdot \frac{4}{3} dx + \int_{0.5}^{1} \left(x - \frac{5}{12} \right)^{2} \cdot \frac{2}{3} dx$$

$$= \left(\frac{x^{3}}{3} - \frac{5x^{2}}{12} + \frac{25x}{144} \right) \cdot \frac{4}{3} \Big|_{0}^{0.5} + \left(\frac{x^{3}}{3} - \frac{5x^{2}}{12} + \frac{25x}{144} \right) \cdot \frac{2}{3} \Big|_{0.5}^{1} = \frac{11}{144}$$

Example 2

- Let X be a random variable. Its pdf is

$$f(x)=2(1-x),\ 0< x < 1$$

Find $E(X)$ and $Var(X)$.

CHEBYSHEV’S INEQUALITY

- Chebyshev’s Inequality
  - If a random variable has a mean μ and a variance σ^2, then

$$P(\mu - c\sigma \leq X \leq \mu + c\sigma) \geq 1 - \frac{1}{c^2}$$

$$\text{for} \quad c \geq 1$$

  - For example, taking c=2 gives

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \geq 1 - \frac{1}{2^2} = 0.75$$

- Proof

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \geq \int_{|x-\mu|>c\sigma} (x-\mu)^2 f(x)dx \geq c^2\sigma^2 \int_{|x-\mu|>c\sigma} f(x)dx.$$

$$\Rightarrow P(|x-\mu|>c\sigma) \leq 1/c^2$$

$$\Rightarrow P(|x-\mu|\leq c\sigma) = 1 - P(|x-\mu|>c\sigma) \geq 1-1/c^2$$

LAWS OF EXPECTED VALUE AND VARIANCE

Let X be a random variable and c be a constant.

Laws of Expected Value

- $E(c) = c$
- $E(X + c) = E(X) + c$
- $E(cX) = cE(X)$

Laws of Variance

- $V(c) = 0$
- $V(X + c) = V(X)$
- $V(cX) = c^2V(X)$

LAWS OF EXPECTED VALUE

- Let X be a random variabke and a, b, and c be constants. Then, for any two functions $g_1(x)$ and $g_2(x)$ whose expectations exist,

$$a) E\left(ag_1(X) + bg_2(X) + c\right) = aE\left(g_1(X)\right) + bE\left(g_2(X)\right) + c$$

$$b)\ If\ g_1(x) \geq 0\ for\ all\ x,\ then\ E\left(g_1(X)\right) \geq 0.$$

$$c)\ If\ g_1(x) \leq g_2(x)\ \ for\ all\ x, then\ E\left(g_1(x)\right) \leq E\left(g_2(x)\right).$$

$$d)\ If\ a \leq g_1(x) \leq b\ for\ all\ x, then\ a \leq E\left(g_1(X)\right) \leq b$$

LAWS OF EXPECTED VALUE (Cont.)

$$E\left(\sum_{i=1}^{k} a_i X_i\right) = \sum_{i=1}^{k} a_i E\left(X_i\right)$$

If X and Y are independent,

$$E\left[g\left(X\right)h\left(Y\right)\right] = E\left[g\left(X\right)\right]E\left[h\left(Y\right)\right]$$

THE COVARIANCE

- The covariance between two real-valued random variables *X* and *Y*, is

$$Cov(X, Y) = E((X - E(X)).(Y - E(Y))) =$$
$$= E(X.Y - E(X)Y - E(Y)X + E(X)E(Y))$$
$$= E(X.Y) - E(X)E(Y) - E(Y)E(X) + E(Y)E(X)$$
$$= E(X.Y) - E(Y)E(X)$$

- Cov(X, Y) can be negative, zero, or positive
- We can show Cov(X,Y) as $\sigma_{X,Y}$
- Random variables with covariance is zero are called **uncorrelated** or **independent**

- If the two variables move in the same direction, (both increase or both decrease), the covariance is a large positive number.
- If the two variables move in opposite directions, (one increases when the other one decreases), the covariance is a large negative number.
- If the two variables are unrelated, the covariance will be close to zero.

Example

$P(X_iY_i)$	Economic condition	Investment: Passive Fund X	Investment: Aggressive Fund Y
0.2	Recession	- 25	- 200
0.5	Stable Economy	+ 50	+ 60
0.3	Expanding Economy	+ 100	+ 350

$$E(X) = \mu_X = (-25)(.2) + (50)(.5) + (100)(.3) = 50$$

$$E(Y) = \mu_Y = (-200)(.2) + (60)(.5) + (350)(.3) = 95$$

$$\begin{aligned}\sigma_{X,Y} &= (-25-50)(-200-95)(.2) + (50-50)(60-95)(.5) \\ &\quad + (100-50)(350-95)(.3) \\ &= 8250\end{aligned}$$

Properties

- If *X* and *Y* are real-valued random variables and ***a*** and ***b*** are constants ("constant" in this context means non-random), then the following facts are a consequence of the definition of covariance:

$$Cov(X, a) = 0$$

$$Cov(X, X) = Var(X)$$

$$Cov(X, Y) = Cov(Y, X)$$

$$Cov(aX, bY) = abCov(X, Y)$$

$$Cov(X + a, Y + b) = Cov(X, Y)$$

$$Cov(X + Y, X) = Cov(X, X) + Cov(\mathrm{Y}, X)$$

If X and Y are independent,

$$Cov(X,Y) = 0$$

The reverse is usually not correct! It is only correct under normal distribution.

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X_1, X_2)$$

If X_1 and X_2 are independent, so that then

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$$

MOMENTS

- Moments:

$$\mu_k^* = E\left[X^k\right] \rightarrow \textit{ the k-th moment}$$

$$\mu_k = E\left[(X-\mu)^k\right] \rightarrow \textit{ the k-th central moment}$$

- Population Mean: μ = E(X)
- Population Variance:

$$\sigma^2 = E\left[(X-\mu)^2\right] = E\left[X^2\right] - \mu^2 \geq 0$$

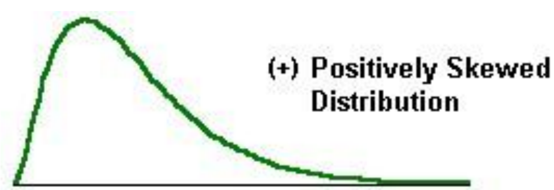
SKEWNESS

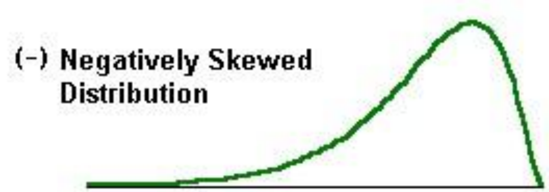
- Measure of lack of symmetry in the pdf.

$$Skewness = \frac{E\left(X-\mu\right)^{3}}{\sigma^{3}} = \frac{\mu_3}{\mu_2^{3/2}}$$

If the distribution of X is symmetric around its mean μ,

$$\mu_3 = 0 \rightarrow \text{Skewness} = 0$$

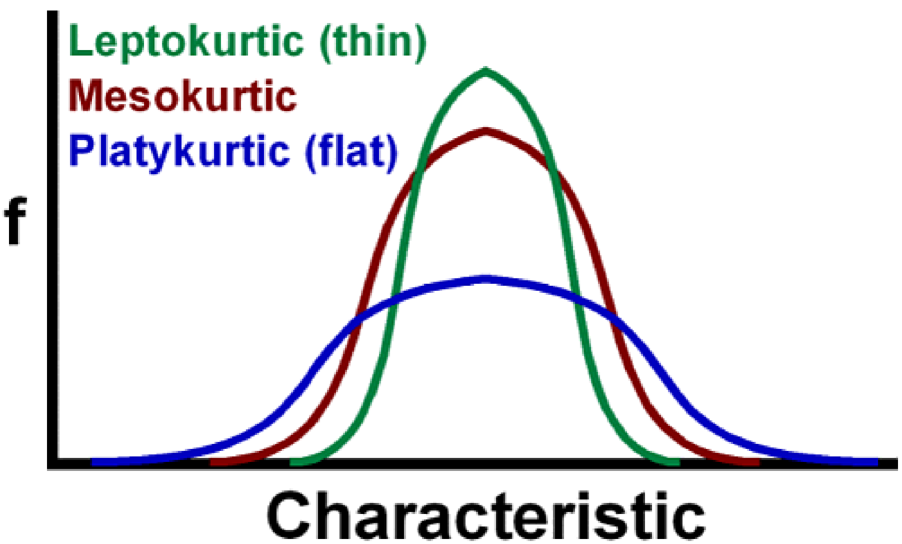




KURTOSIS

- Measure of the peakedness of the pdf. Describes the shape of the distribution.

$$Kurtosis = \frac{E\left[(X-\mu)^4\right]}{\sigma^4} = \frac{\mu_4}{\mu_2^2}$$



Kurtosis=3 $\rightarrow$ Normal

Kurtosis >3 $\rightarrow$ Leptokurtic
(peaked and fat tails)

Kurtosis<3 $\rightarrow$ Platykurtic
(less peaked and thinner tails)

QUANTILES OF RANDOM VARIABLES

- Quantiles of Random variables
  - The p th quantile of a random variable X $F(x) = p$
  - A probability of that the random variable takes a value less than the p^{th} quantile
- Upper quartile
  - The 75th percentile of the distribution
- Lower quartile
  - The 25th percentile of the distribution
- Interquartile range
  - The distance between the two quartiles

- Example

$$F(x) = 1.5x - 2(x - 50.0)^3 - 74.5 \text{ for } 49.5 \le x \le 50.5$$

  - Upper quartile : $F(x) = 0.75$ $\quad x = 50.17$
  - Lower quartile : $F(x) = 0.25$ $\quad x = 49.83$
  - Interquartile range : $50.17 - 49.83 = 0.34$

CENTRAL TENDENCY MEASURES

- In statistics, the term **central tendency** relates to the way in which quantitative data tend to cluster around a “central value”.
- A **measure of central tendency** is any of a number of ways of specifying this "central value.“
- There are three important descriptive statistics that gives measures of the central tendency of a variable:
  - The Mean
  - The Median
  - The Mode

THE MEAN

- The **arithmetic mean** is the most commonly-used type of average.
- In mathematics and statistics, the **arithmetic mean** (or simply the **mean**) of a list of numbers is the sum of all numbers in the list divided by the number of items in the list.

  - If the list is a statistical population, then the mean of that population is called a **population mean**.
  - If the list is a statistical sample, we call the resulting statistic a **sample mean**.

- If we denote a set of data by $X = (x_1, x_2, \ldots, x_n)$, then the **sample mean** is typically denoted with a horizontal bar over the variable ($\bar{X}$)

- The Greek letter μ is used to denote the arithmetic mean of an entire population.

THE SAMPLE MEAN

- In mathematical notation, the sample mean of a set of data denoted as $X = (x_1, x_2, \ldots, x_n)$ is given by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

- Suppose daily asset price are:
- 67.05, 66.89, 67.45, 68.39, 67.45,70.10, 68.39

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{7}(67.05 + 66.89 + \ldots + 68.39) = 67.96$$

THE MEDIAN

- In statistics, a **median** is described as the numeric value separating the higher half of a sample or population from the lower half.
- The *median* of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one.
- If there is an even number of observations, then there is no single middle value, so we take the mean of the two middle values.
- Organize the price data in the previous example

  67.05, 66.89, 67.45, **67.45**, 68.39, 68.39, 70.10
- The median of this price series is **67.45**

THE MODE

- In statistics, the **mode** is the value that occurs the most frequently in a data set.
- The mode is not necessarily unique, since the same maximum frequency may be attained at different values.
- Organize the price data in the previous example in ascending order

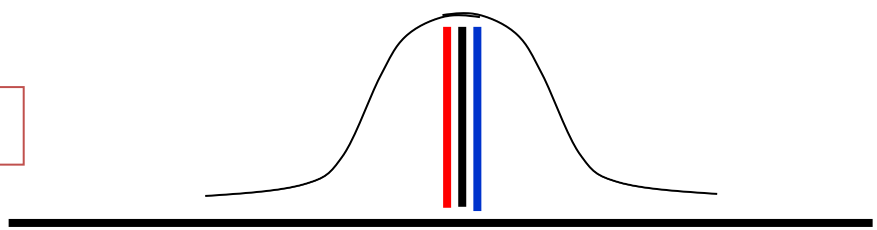
  67.05, 66.89, **67.45, 67.45**, **68.39, 68.39**, 70.10

- There are two modes in the given price data – 67.45 and 68.39
- The sample price dataset may be said to be **bimodal**
- A population or sample data may be unimodal, bimodal, or multimodal

RELATIONSHIP AMONG MEAN, MEDIAN, AND MODE

- If a distribution is from a bell shaped symmetrical one, then the mean, median and mode coincide

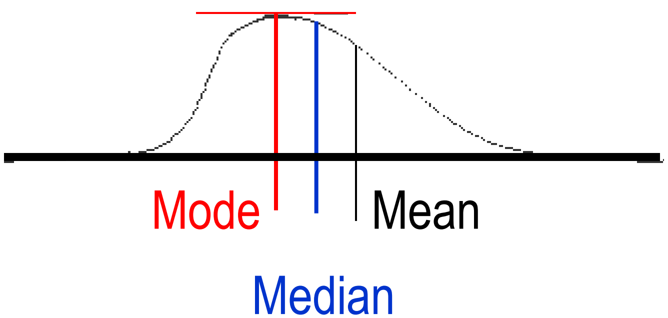
Mean = Median = Mode



- If a distribution is non symmetrical, and skewed to the left or to the right, the three measures differ.

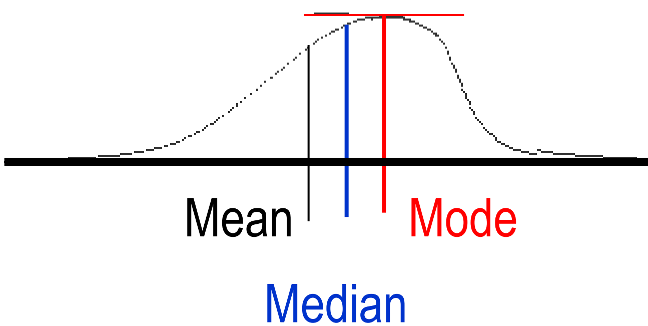
Mode < Median < Mean

A positively skewed distribution
("skewed to the right")



Mean < Median < Mode

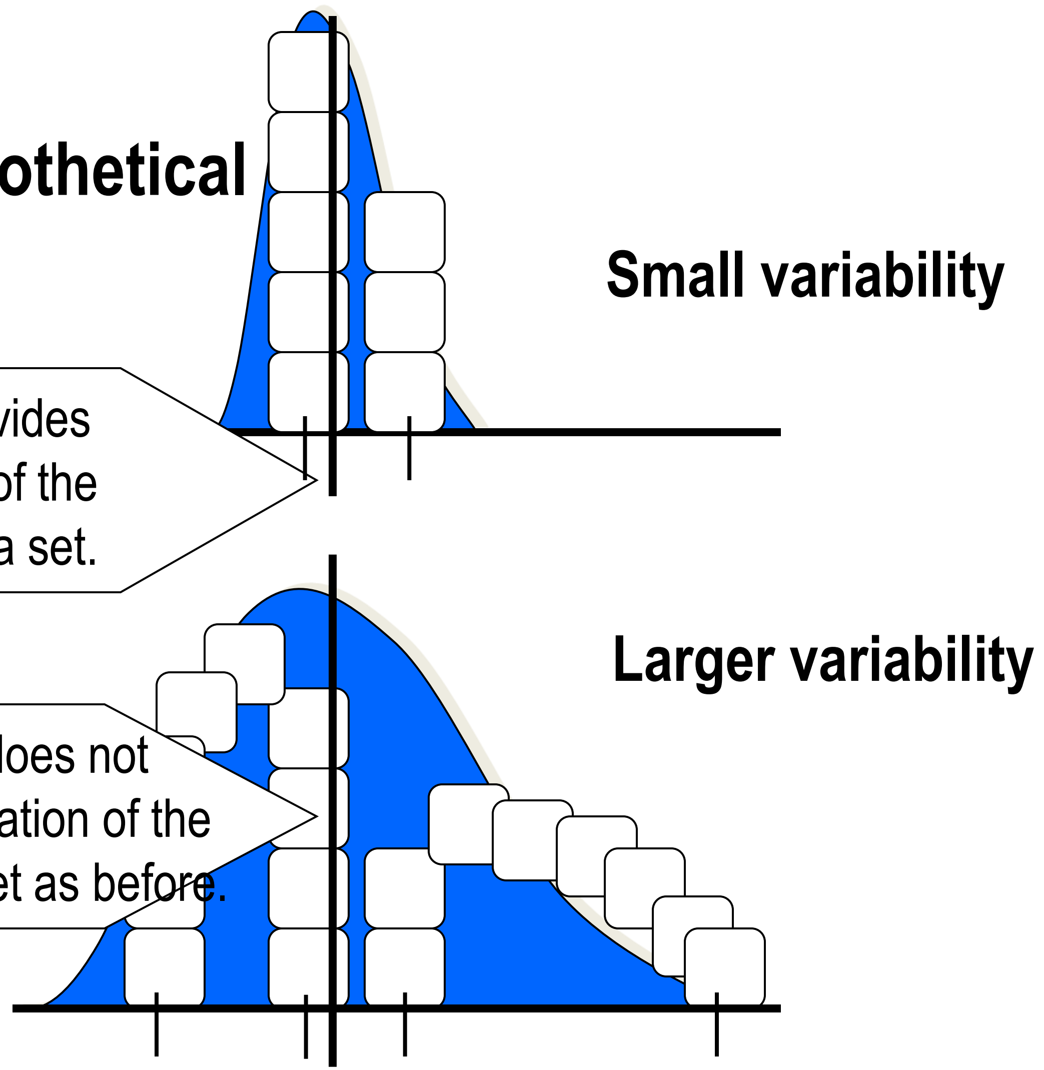
A negatively skewed distribution
("skewed to the left")



STATISTICAL DISPERSION

- In statistics, **statistical dispersion** (also called **statistical variability** or **variation**) is the variability or spread in a variable or probability distribution.
- Common measures of statistical dispersion are
  - The Variance, and
  - The Standard Deviation
- Dispersion is contrasted with location or central tendency, and together they are the most used properties of distributions

Observe two hypothetical data sets:



The average value provides a good representation of the observations in the data set.

The same average value does not provide as good representation of the observations in the data set as before.

THE VARIANCE

- In statistics, the **variance** of a random variable or distribution is the expected (mean) value of the square of the deviation of that variable from its expected value or mean.
- Thus the variance is a measure of the amount of variation within the values of that variable, taking account of all possible values and their probabilities.
- If a random variable X has the expected (mean) value $E[X]=\mu$, then the variance of X can be given by:

$$Var(X) = E[(X-\mu)^2] = \sigma_x^2$$

THE VARIANCE

- The above definition of variance encompasses random variables that are discrete or continuous. It can be expanded as follows:

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

THE SAMPLE VARIANCE

- If we have a series of *n* measurements of a random variable *X* as X_i, where *i* = 1, 2, ..., *n*, then the *sample variance* can be calculated as

$$S_x^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2}{n-1}$$

The denominator, (*n*-1) is known as the **degrees of freedom.** Intuitively, only n-1 observation values are free to vary, one is predetermined by mean. When *n* = 1 the variance of a single sample is obviously zero

THE SAMPLE VARIANCE

- For the hypothetical price data 67.05, 66.89, 67.45, 67.45, 68.39, 68.39, 70.10, the ***sample variance*** can be calculated as

$$S_x^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}$$

$$= \frac{1}{7-1}\left[\left(67.05-67.96\right)^2 + \ldots + \left(70.10-67.96\right)^2\right]$$

$$= 1.24$$

THE STANDARD DEVIATION

- In statistics, the **standard deviation** of a random variable or distribution is the **square root of its variance**.
- If a random variable *X* has the expected value (mean) $E[X]=\mu$, then the standard deviation of *X* can be given by:

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{E[(X-\mu)^2]}$$

- If we have a series of *n* measurements of a random variable *X* as X_i, where i = 1, 2, ..., *n*, then the *sample standard deviation*, can be used to estimate the *population standard deviation* of $X = (x_1, x_2, ..., x_n)$. The sample standard deviation is calculated as

$$S_x = \sqrt{S_x^2} = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}} = \sqrt{1.24} = 1.114$$

SAMPLE COVARIANCE

- If we have a series of *n* measurements of a random variable *X* as X_i, where *i* = 1, 2, ..., *n* and a series of *n* measurements of a random variable *Y* as Y_i, where *i* = 1, 2, ..., *n*, then the sample covariance is calculated as:

$$S_{X,Y} = Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}\left[X - \bar{X}\right]\left[Y - \bar{Y}\right]$$

Example

- Compare the following three sets

x_i	y_i	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$
2 6 7	13 20 27	-3 1 2	-7 0 7	21 0 14
$\bar{x}=5$	$\bar{y}=20$			Cov(x,y)=17.5

x_i	y_i	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$
2 6 7	27 20 13	-3 1 2	7 0 -7	-21 0 -14
$\bar{x}=5$	$\bar{y}=20$			Cov(x,y)=-17.5

CORRELATION COEFFICIENT

- If we have a series of *n* measurements of *X* and *Y* written as X_i and Y_i, where *i* = 1, 2, ..., *n*, then the *sample correlation coefficient*, can be used to estimate the *population correlation coefficient* between *X* and *Y*. The sample correlation coefficient is calculated as

$$r_{x,y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_x S_y} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left[n\sum_{i=1}^{n} X^2{}_i - (\sum_{i=1}^{n} X_i)^2\right]\left[n\sum_{i=1}^{n} Y^2{}_i - (\sum_{i=1}^{n} Y_i)^2\right]}}$$

Population coefficient of correlation

$$\rho = \frac{COV(X,Y)}{\sigma_{\mathrm{x}}\sigma_y}$$

- The value of correlation coefficient falls between −1 and 1:

$$-1 \leq r_{x,y} \leq 1$$

- $r_{x,y}$= 0 => X and Y are **uncorrelated**
- $r_{x,y}$= 1 => X and Y are **perfectly positively correlated**
- $r_{x,y}$=−1=>X and Y are **perfectly negatively correlated**

Example

	X	Y	X*Y	X^2	Y^2
A	43	128	5504	1849	16384
B	48	120	5760	2304	14400
C	56	135	7560	3136	18225
D	61	143	8723	3721	20449
E	67	141	9447	4489	19881
F	70	152	10640	4900	23104
Sum	**345**	**819**	**47634**	**20399**	**112443**

- Substitute in the formula and solve for *r*:

$$r = \frac{6*47634 - 345*819}{\sqrt{\left[6*20399 - (345)^2\right]\left[6112443 - (819)^2\right]}} = 0.897$$

- The correlation coefficient suggests a strong positive relationship between X and Y.

CORRELATION AND CAUSATION

- Recognize the difference between correlation and causation — just because two things occur together, that does not necessarily mean that one causes the other.
- For random processes, causation means that if *A occurs, that causes a change in the* probability that *B occurs*.

- Existence of a statistical relationship, no matter how strong it is, does not imply a cause-and-effect relationship between X and Y. for ex, let X be size of vocabulary, and Y be writing speed for a group of children. There most probably be a positive relationship but this does not imply that an increase in vocabulary causes an increase in the speed of writing. Other variables such as age, education etc will affect both X and Y.
- Even if there is a causal relationship between X and Y, it might be in the opposite direction, i.e. from Y to X. For eg, let X be thermometer reading and let Y be actual temperature. Here Y will affect X.