# What is Statistics?

**Definition of Statistics**

- **<u>Statistics</u>** is the science of collecting, organizing, analyzing, and interpreting data in order to make a decision.

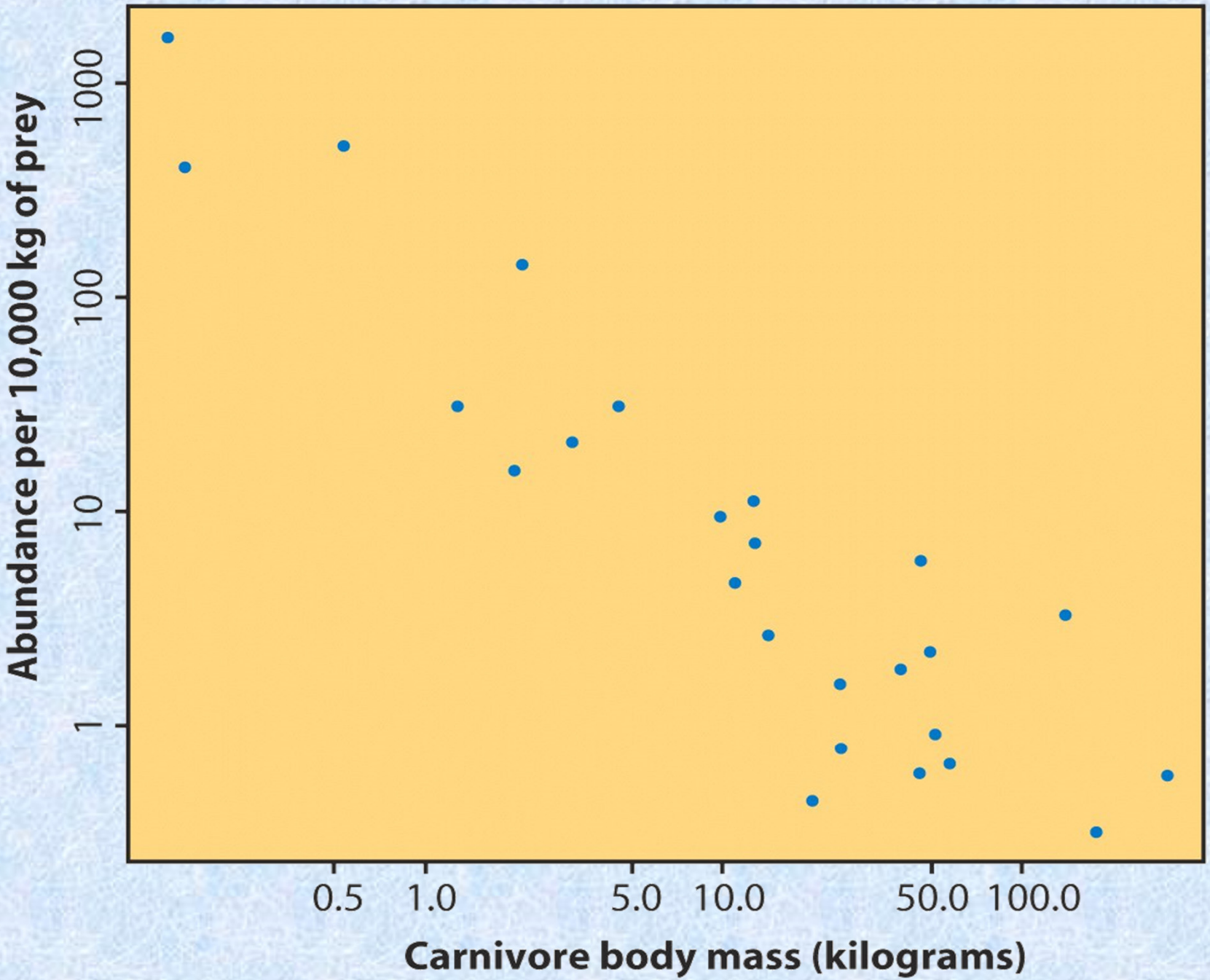
- **Branches of Statistics**
  - The study of statistics has two major branches – descriptive(exploratory) statistics and inferential statistics.
    - **Descriptive statistics** is the branch of statistics that involves the organization, summarization, and display of data.
    - **Inferential statistics** is the branch of statistics that involves using a sample to draw conclusions about population. A basic tool in the study of inferential statistics is probability.

# Scatterplots and Correlation

- **Displaying relationships: Scatterplots**
- **Interpreting scatterplots**
- **Adding categorical variables to scatterplots**
- **Measuring linear association: correlation *r***
- **Facts about correlation**

- Response variable measures **an outcome of a study**.
- An explanatory variable explains, **influences or cause changes in a response variable**.
- Independent variable and dependent variable.
- WARNING: The relationship between two variables can be strongly influenced by other variables that are lurking in the background.
- **Note:** There is not necessary to have a cause-and-effect relationship between explanatory and response variables.
- Example. Sales of personal computers and athletic shoes

# Example - 1


Abundance per 10,000 kg of prey
1000
100
10
1
0.5
1.0
5.0
10.0
50.0
100.0
Carnivore body mass (kilograms)

# Definitions

- **Sample space**: the set of all possible outcomes. We denote S
- **Event**: an outcome or a set of outcomes of a random phenomenon. An event is a subset of the sample space.
- **Probability** is the proportion of success of an event.
- **Probability model**: a mathematical description of a random phenomenon consisting of two parts: S and a way of assigning probabilities to events.

# Probability distributions

- **Probability distribution of a random variable X**: it tells what values X can take and how to assign probabilities to those values.
  - Probability of discrete random variable: list of the possible value of X and their probabilities
  - Probability of continuous random variable: density curve.

# Measuring linear association: correlation *r*

**(**The *Pearson Product-Moment Correlation Coefficient or Correlation Coefficient*)

- The **correlation r** measures the <u>strength</u> and <u>direction</u> of the ***<u>linear association</u>*** between two quantitative variables, usually labeled X and Y.

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

# Facts about correlation

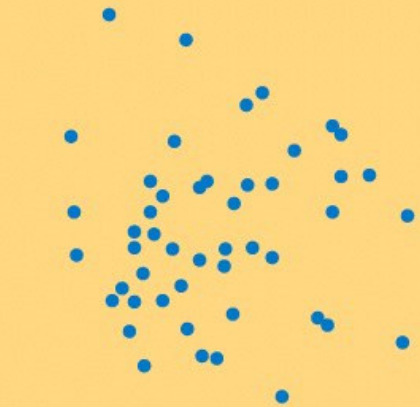
- What kind of variables do we use?
  - 1. No distinction between explanatory and response variables.
  - 2. Both variables should be quantitative
- Numerical properties
  - 1. $-1 \le r \le 1$
  - 2. r>0: positive association between variables
  - 3. r<0: negative association between variables
  - 4. If r =1or r = - 1, it indicates perfect linear relationship
  - 5. As |r| is getting close to 1, much stronger relationship

$$\triangleleft - negative\ relationship - \triangleright \triangleleft - positive\ relationship - \triangleright$$

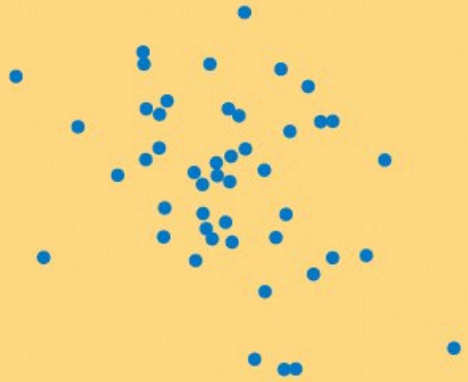
$$-1 \qquad\qquad 0 \qquad\qquad 1$$

$$\triangleleft - - - - stronger \qquad\qquad stronger - - - - \triangleright$$

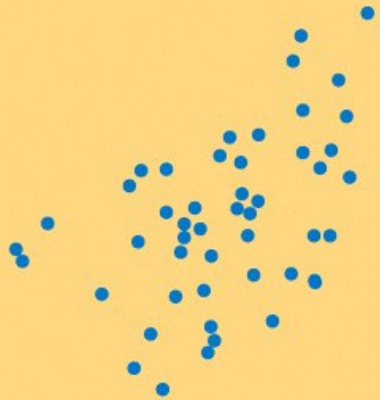
  - 6. Effected by a few outliers →not resistant.
  - 7. It doesn't describe curved relationships
  - 8. Not easy to guess the value of r from the appearance of a scatter plot



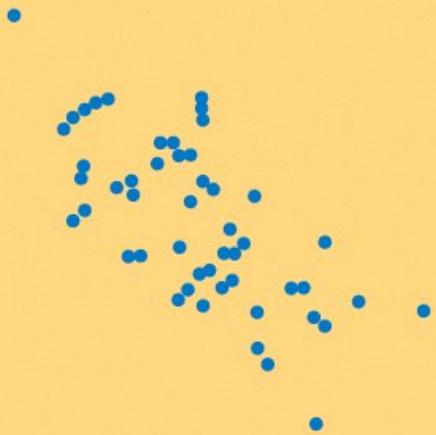
Correlation $r = 0$



Correlation $r = -0.3$

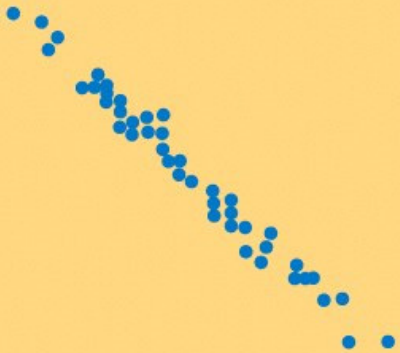


Correlation $r = 0.5$



Correlation $r = -0.7$

Correlation $r = 0.9$



Correlation $r = -0.99$

# Some necessary elements of

# Probability theory and Statistics

# The NORMAL DISTRIBUTION

**The normal (or Gaussian) distribution, is a very commonly used (occurring) function in the fields of probability theory, and has wide applications in the fields of:**

- **Pattern Recognition;**
- **Machine Learning;**
- **Artificial Neural Networks and Soft computing;**
- **Digital Signal (image, sound , video etc.) processing**
- **Vibrations, Graphics etc.**

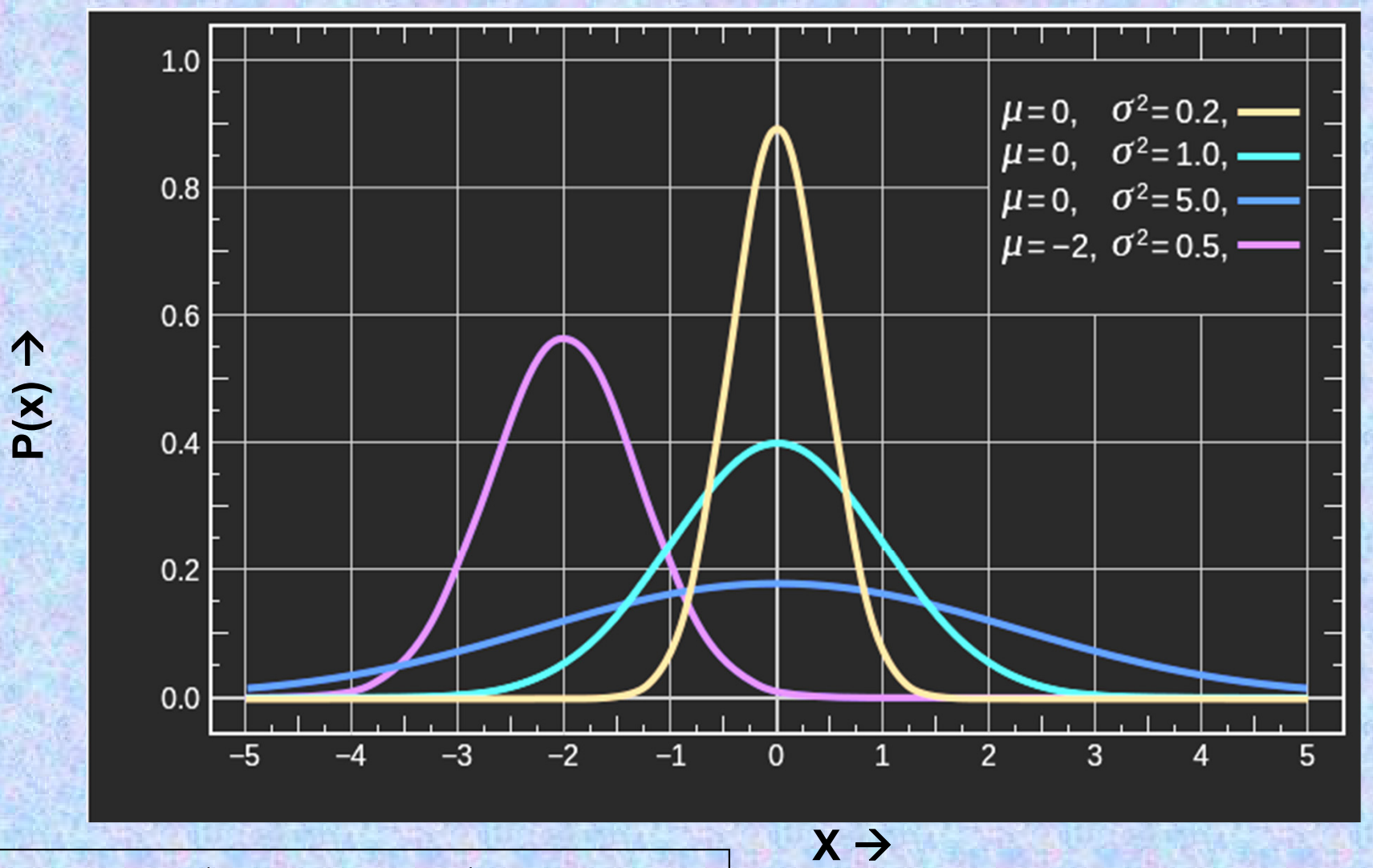
**Its also called a BELL function/curve.**

**The formula for the normal distribution is:**

$$\boxed{p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]}$$

**The parameter $\mu$ is called the mean or expectation (or median or mode) of the distribution.**

**The parameter $\sigma$ is the standard deviation;**
**and variance is thus $\sigma^2$.**




$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]$$

https://en.wikipedia.org/wiki/File:Normal_Distribution_PDF.svg
***(2013)***

**The normal distribution $p(x)$, with any mean $\mu$ and any positive deviation $\sigma$, has the following properties:**

- **It is symmetric around the mean ($\mu$) of the distribution.**
- **It is unimodal: its first derivative is positive for $x < \mu$, negative for $x > \mu$, and zero only at $x = \mu$.**
- **It has two inflection points (where the second derivative of $f$ is zero and changes sign), located one standard deviation away from the mean, $x = \mu - \sigma$ and $x = \mu + \sigma$.**
- **It is log-concave.**
- **It is infinitely differentiable, indeed supersmooth of order 2.**

**Also, the standard normal distribution $p$ (with $\mu = 0$ and $\sigma = 1$) also has the following properties:**

- **Its first derivative $p'(x)$ is: $-x.p(x)$.**
- **Its second derivative $p''(x)$ is: $(x^2 - 1).p(x)$**
- **More generally, its $n$-th derivative :**

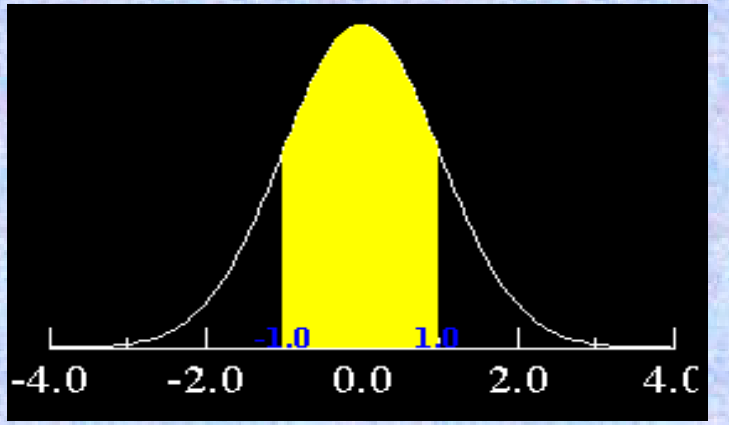
$$p^{(n)}(x) \text{ is: } \quad (-1)^n H_n(x) p(x),$$

**where, $H_n$ is the Hermite polynomial of order $n$.**

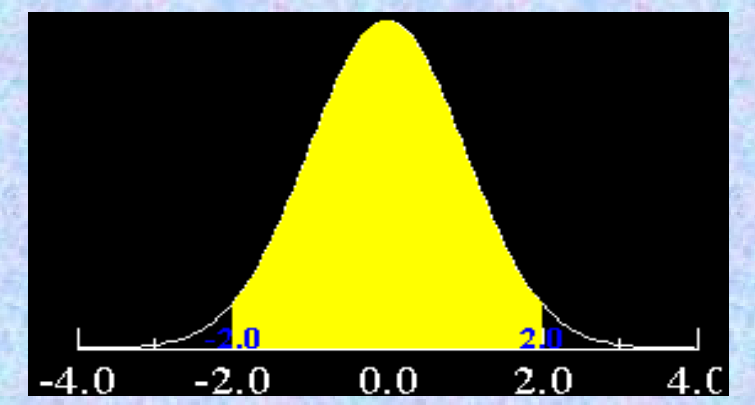
**The 68 – 95 - 99.7% Rule:**

**All normal density curves satisfy the following property which is often referred to as the Empirical Rule:**

**- 68% of the observations fall within 1 standard deviation of the mean, that is, between** $(\mu-\sigma)$ and $(\mu+\sigma)$

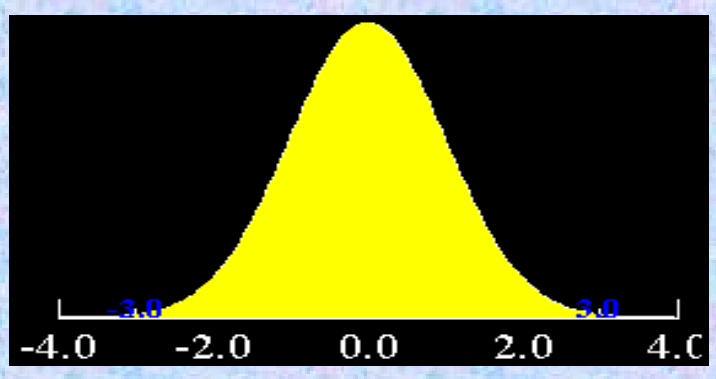


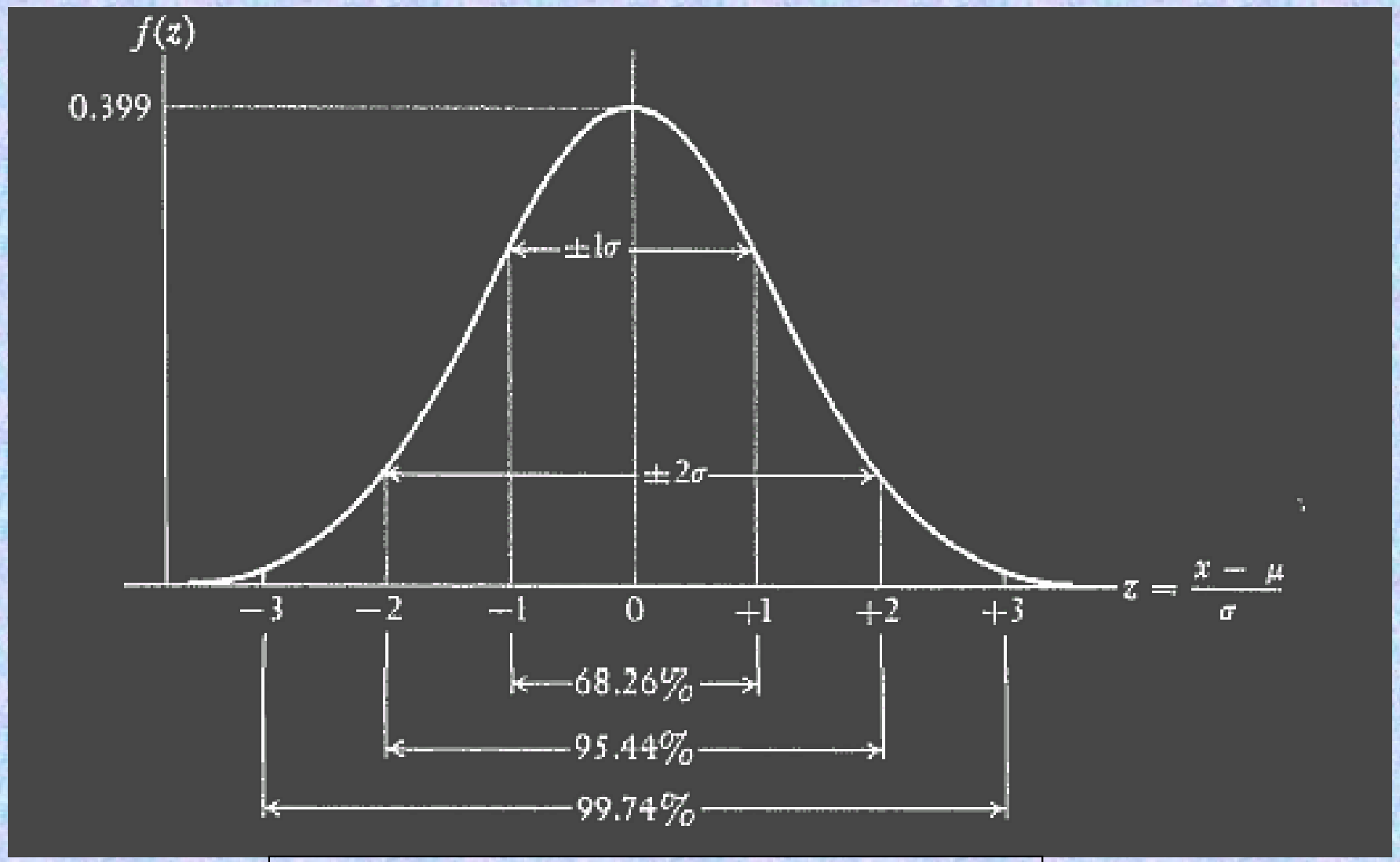

**- 95% of the observations fall within 2 standard deviations of the mean, that is, between** $(\mu-2\sigma)$ and $(\mu+2\sigma)$




**- 99.7% of the observations fall within 3 standard deviations of the mean, that is, between**

$$(\mu-3\sigma) \text{ and } (\mu+3\sigma)$$



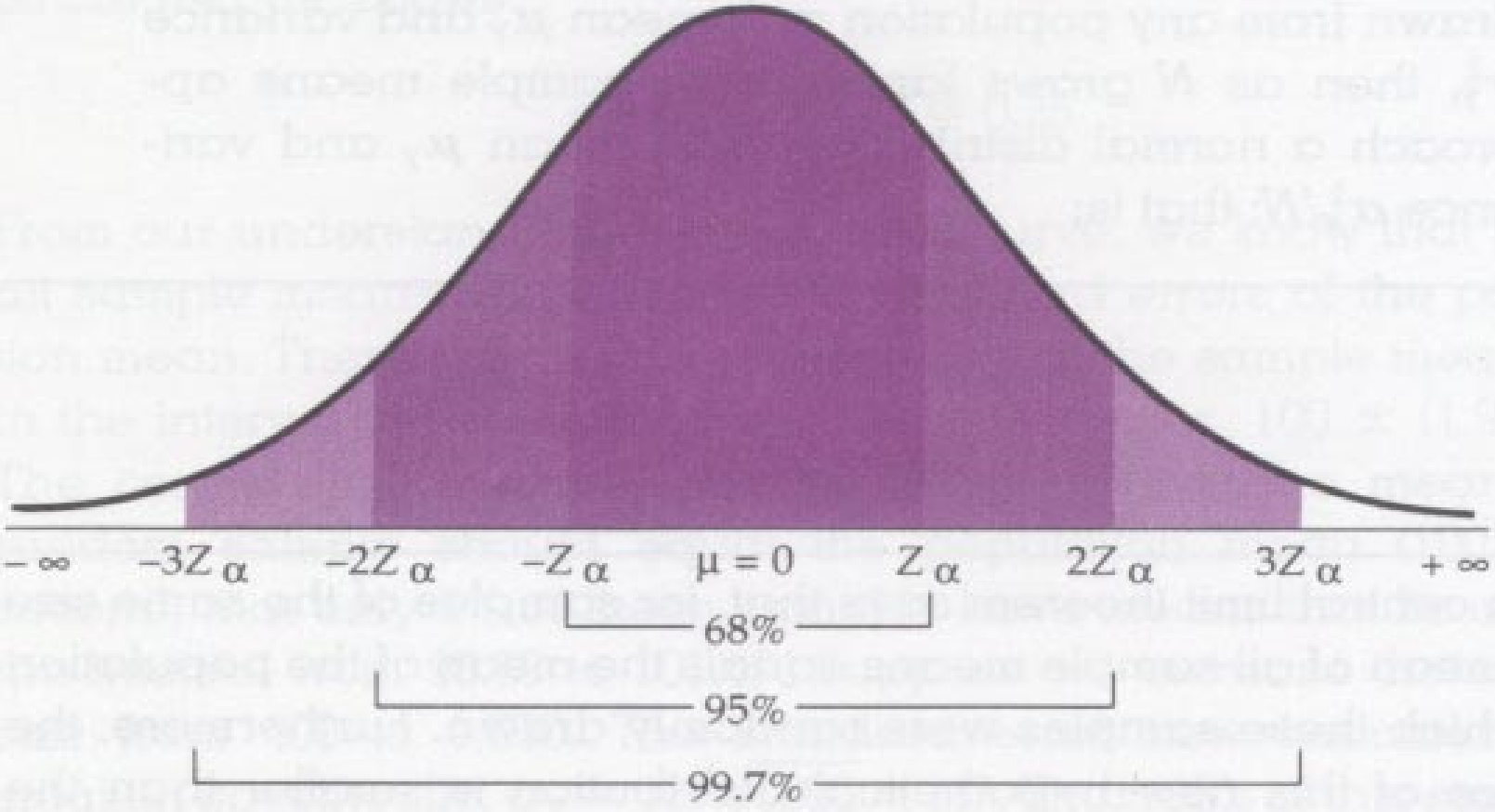



$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]$$

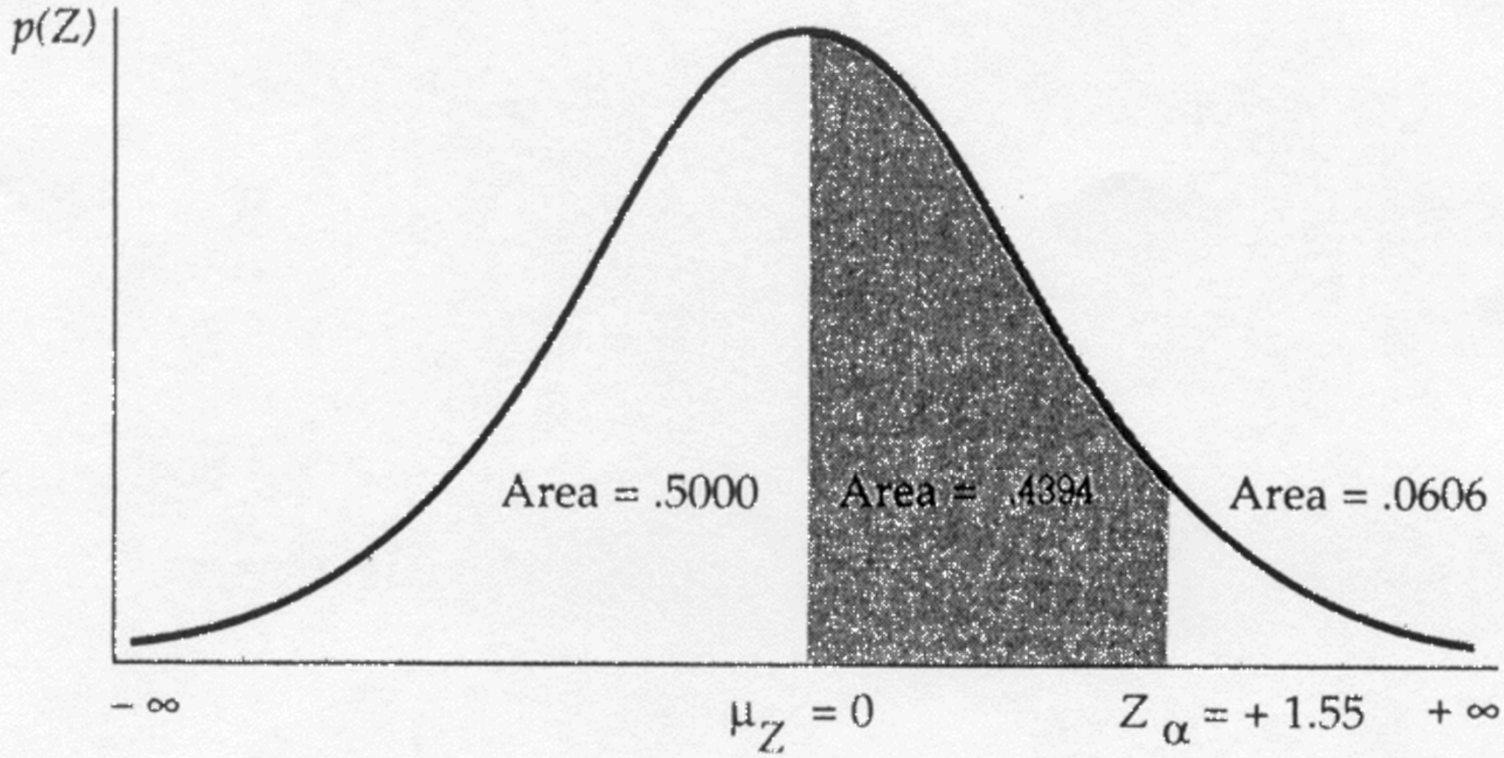
A normal distribution:

1. is **symmetrical** (both halves are *identical*);
2. is **asymptotic** (its *tails never touch* the underlying x-axis; the curve reaches to $-\infty$ and $+\infty$ and thus must be truncated);
3. has **fixed** and **known** *areas under the curve* (these fixed areas are marked off by units along the x-axis called **z-scores**; imposing truncation, the normal curve ends at + 3.00 z on the right and - 3.00 z on the left).

## Areas Under the Normal Curve for Various Z Scores



# Example of the Probability of Observing an Outcome in a Standar Distribution



# Conditional Distribution

The *conditional probability mass function* of Y given X is:

$$p(y|x) = P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)} = \frac{p(x, y)}{p(x)}.$$

For continuous random variables, we can define the *conditional probability density function*:

$$f(y|x) = \frac{f(x, y)}{f(x)}.$$

Conditional probability: $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$

Multiplication rule: $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B) = \mathbb{P}(B \mid A)\mathbb{P}(A).$

Rewriting the above equation yields:

$$f(x, y) = f(x) \cdot f(y|x).$$

The marginal density of Y can then be obtained from:

$$f(y) = \int_{-\infty}^{\infty} f(x) \cdot f(y|x) dx.$$

conditional probability which is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ when } P(B) > 0.$$

Any other formula regarding conditional probability can be derived from the above formula. Specifically, if you have two random variables $X$ and $Y$, you can write

$$P(X \in C|Y \in D) = \frac{P(X \in C, Y \in D)}{P(Y \in D)}, \text{ where } C, D \subset \mathbb{R}.$$

[redacted] the **conditional PMF**. Specifically, the conditional PMF of $X$ given event $A$, is defined as

$$\begin{aligned} P_{X|A}(x_i) &= P(X = x_i|A) \\ &= \frac{P(X = x_i \text{ and } A)}{P(A)}. \end{aligned}$$

Similarly, we define the **conditional CDF** of $X$ given $A$ as

$$F_{X|A}(x) = P(X \leq x|A).$$

Two discrete random variables $X$ and $Y$ are independent if

$$P_{XY}(x, y) = P_X(x)P_Y(y), \quad \text{for all } x, y.$$

Equivalently, $X$ and $Y$ are independent if

$$F_{XY}(x, y) = F_X(x)F_Y(y), \quad \text{for all } x, y.$$

For discrete random variables $X$ and $Y$, the **conditional PMFs** of $X$ given $Y$ and vice versa are defined as

$$P_{X|Y}(x_i|y_j) = \frac{P_{XY}(x_i, y_j)}{P_Y(y_j)},$$

$$P_{Y|X}(y_j|x_i) = \frac{P_{XY}(x_i, y_j)}{P_X(x_i)}$$

for any $x_i \in R_X$ and $y_j \in R_Y$.

So, if $X$ and $Y$ are independent, we have

$$\begin{aligned} P_{X|Y}(x_i|y_j) &= P(X = x_i|Y = y_j) \\ &= \frac{P_{XY}(x_i, y_j)}{P_Y(y_j)} \\ &= \frac{P_X(x_i)P_Y(y_j)}{P_Y(y_j)} \\ &= P_X(x_i). \end{aligned}$$

As we expect, for independent random variables, the conditional PMF is equal to the marginal PMF. In other words, knowing the value of $Y$ does not provide any information about $X$.

# Expected Value of Random Variables

The expected value of a random variable is the weighted average of all possible values of the variable. The weight here means the probability of the random variable taking a specific value.

$$E[X] = \sum x_i p(x_i)$$

$x_i$ = The values that X takes

$p(x_i)$ = The probability that X takes the value $x_i$

$$\begin{array}{r} 4.21875 \\ 4.21875 \\ 1.40625 \\ +\ 0.15625 \\ \hline 10. \end{array}$$

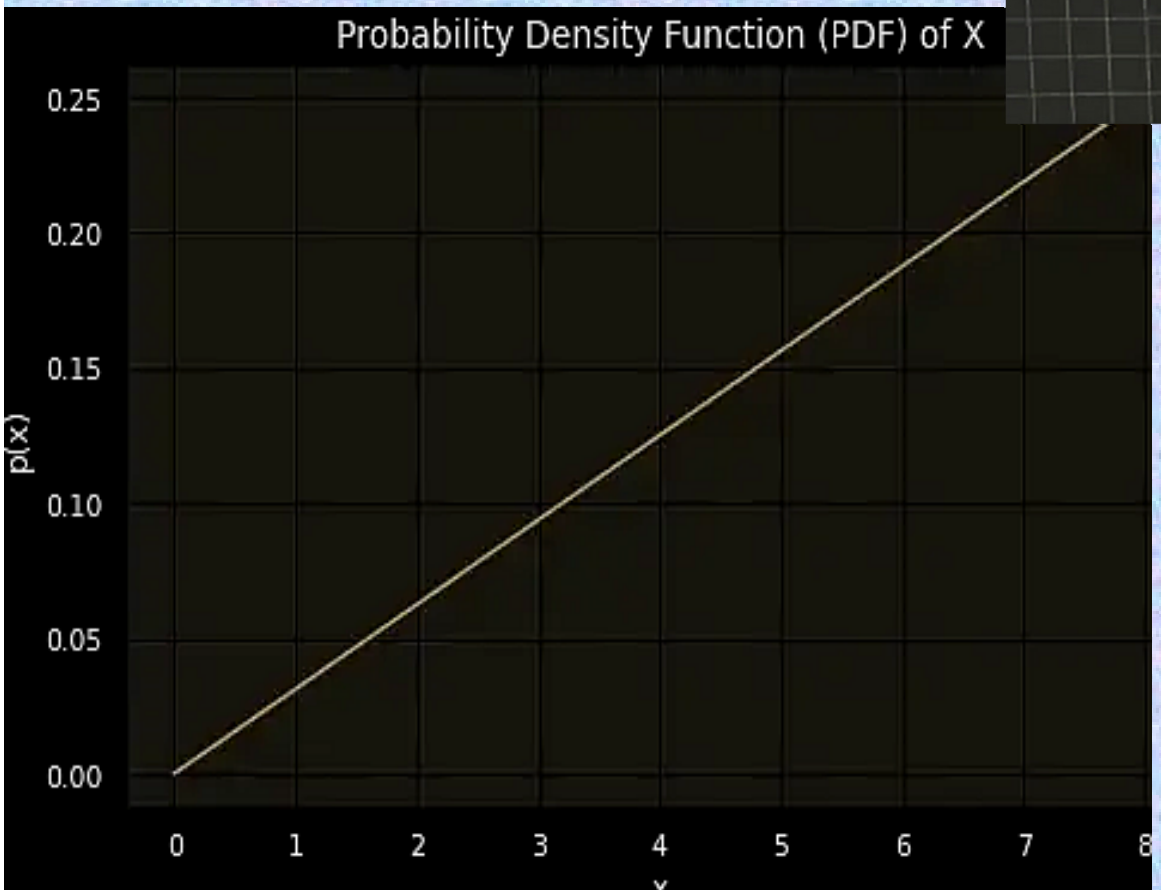
| # of correct answers | Probability | Point |
|---|---|---|
| 0 | $(3/4)^4$ | 0 |
| 1 | $(1/4)^1 \cdot 4 \cdot (3/4)^3$ | 10 |
| 2 | $(1/4)^2 \cdot 6 \cdot (3/4)^2$ | 20 |
| 3 | $(1/4)^3 \cdot 4 \cdot (3/4)^1$ | 30 |
| 4 | $(1/4)^4$ | 40 |

$$E[X] = \int_{xmin}^{xmax} x f(x) dx$$

*f(x)* is the PDF of X

$$E[X] = \int_5^{10} x \, 0.2 \, dx = 0.2 \frac{x^2}{2} \Big|_5^{10}$$

$$= 0.1 x^2 \Big|_5^{10} = 0.1 (100 - 25)$$




$$E[X] = \int_0^8 x \, 0.03125 x \, dx$$

$$= 0.03125 \frac{x^3}{3} \Big|_0^8$$

$$= 0.03125 \frac{(8)^3}{3}$$

$$= 5.33$$

his

**Example** Let $X$ be a continuous random variable with support $R_X = [0, \infty)$ and probability density function

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \in [0, \infty) \\ 0 & \text{otherwise} \end{cases}$$

where $\lambda > 0$. Its expected value is

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_{0}^{\infty} x \lambda \exp(-\lambda x) dx \end{aligned}$$

$$\begin{aligned} \mathrm{E}[Y] &= \sum_{x \in R_X} (a + bx) p_X(x) && \text{(by the transformation theorem)} \\ &= \sum_{x \in R_X} a p_X(x) + \sum_{x \in R_X} b x p_X(x) \\ &= a \sum_{x \in R_X} p_X(x) + b \sum_{x \in R_X} x p_X(x) \\ &= a + b \sum_{x \in R_X} x p_X(x) && \text{(because probabilities sum up to 1)} \\ &= a + b\mathrm{E}[X] && \text{(by the definition of } \mathrm{E}[X]\text{)} \end{aligned}$$

## Expectation of $g(X)$

Let $g(X)$ be a function of $X$. We can imagine a long-term average of $g(X)$ just as we can imagine a long-term average of $X$. This average is written as $\mathbb{E}(g(X))$. Imagine observing $X$ many times ($N$ times) to give results $x_1, x_2, \ldots, x_N$. Apply the function $g$ to each of these observations, to give $g(x_1), \ldots, g(x_N)$. The mean of $g(x_1), g(x_2), \ldots, g(x_N)$ approaches $\mathbb{E}(g(X))$ as the number of observations $N$ tends to infinity.

*Definition:* Let $X$ be a <u>continuous</u> random variable, and let $g$ be a function. The expected value of $g(X)$ is

$$\mathbb{E}\Big(g(X)\Big) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx.$$

*Definition:* Let $X$ be a <u>discrete</u> random variable, and let $g$ be a function. The expected value of $g(X)$ is

$$\mathbb{E}\Big(g(X)\Big) = \sum_x g(x) f_X(x) = \sum_x g(x) \mathbb{P}(X = x).$$

Let $X$ and $Y$ be <u>independent</u> random variables, and $g, h$ be functions. Then

$$\begin{aligned} \mathbb{E}(XY) &= \mathbb{E}(X)\mathbb{E}(Y) \\ \mathbb{E}\Big(g(X)h(Y)\Big) &= \mathbb{E}\Big(g(X)\Big)\mathbb{E}\Big(h(Y)\Big). \end{aligned}$$

## Probability as a conditional expectation

Define the indicator random variable: $I_A = \begin{cases} 1 & \text{if event } A \text{ occurs,} \\ 0 & \text{otherwise.} \end{cases}$

Then $\mathbb{E}(I_A) = \mathbb{P}(I_A = 1) = \mathbb{P}(A)$.

$$\mathbb{P}(A) = \mathbb{E}_Y\Big(\mathbb{E}(I_A \,|\, Y)\Big) = \mathbb{E}_Y\Big(\mathbb{P}(A \,|\, Y)\Big)$$

Law of Total Probability:

$$P(X \in A) = \sum_{y_j \in R_Y} P(X \in A | Y = y_j) P_Y(y_j), \quad \text{for any set } A.$$

Law of Total Expectation:

1. If $B_1, B_2, B_3, \ldots$ is a partition of the sample space $S$,

$$EX = \sum_i E[X|B_i] P(B_i) \tag{5.3}$$

2. For a random variable $X$ and a discrete random variable $Y$,

$$EX = \sum_{y_j \in R_Y} E[X|Y = y_j] P_Y(y_j) \tag{5.4}$$

# Conditional Distribution and Conditional Expectation

The *conditional probability mass function* of Y given X is:

Conditional probability: $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.

Multiplication rule: $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B) = \mathbb{P}(B \mid A)\mathbb{P}(A)$.

$$p(y|x) = P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)} = \frac{p(x, y)}{p(x)}.$$

For continuous random variables, we can define the *conditional probability density function*:

$$f(y|x) = \frac{f(x, y)}{f(x)}.$$

The *conditional expectation* of a random variable $Y$ is the expected value of $Y$ given $[X=x]$, and is denoted: $E[Y|X=x]$ or $E[Y|x]$. If the conditional probability density function is known, then the conditional expectation can be found using:

$$E[Y|X = x] = \begin{cases} \int_{-\infty}^{\infty} y \cdot f(y|x)dy & \text{if } Y \text{ is continuous} \\ \sum_y y \cdot p(y|x) & \text{if } Y \text{ is discrete} \end{cases} \tag{38}$$

To obtain the unconditional expectation of $Y$, we can take the expectation of $E[Y|X]$. The result is the *theorem of total expectation*:

$$E[Y] = \begin{cases} \int_{-\infty}^{\infty} E[Y|X = x]f(x)dx & \text{if } X \text{ is continuous} \\ \sum_x E[Y|X = x]p(x) & \text{if } X \text{ is discrete.} \end{cases} \tag{39}$$

## Conditional Expectation of $X$:

$$E[X|A] = \sum_{x_i \in R_X} x_i P_{X|A}(x_i),$$

$$E[X|Y = y_j] = \sum_{x_i \in R_X} x_i P_{X|Y}(x_i|y_j)$$

## Iterated Expectations:

Let us look again at the law of total probability for expectation. Assuming $g(Y) = E[X|Y]$, we have

$$\begin{aligned} E[X] &= \sum_{y_j \in R_Y} E[X|Y = y_j] P_Y(y_j) \\ &= \sum_{y_j \in R_Y} g(y_j) P_Y(y_j) \\ &= E[g(Y)] \\ &= E[E[X|Y]]. \end{aligned}$$

**Theorem 1** *Let $X, Y, Z$ be random variables, $a, b \in \mathbb{R}$, and $g : \mathbb{R} \to \mathbb{R}$. Assuming all the following expectations exist, we have*

*(i)* $E[a|Y] = a$

*(ii)* $E[aX + bZ|Y] = aE[X|Y] + bE[Z|Y]$

*(iii)* $E[X|Y] \geq 0$ *if* $X \geq 0$.

*(iv)* $E[X|Y] = E[X]$ *if* $X$ *and* $Y$ *are independent.*

*(v)* $E[E[X|Y]] = E[X]$

*(vi)* $E[Xg(Y)|Y] = g(Y)E[X|Y]$. *In particular,* $|E[g(Y)|Y] = g(Y)$.

*(vii)* $E[X|Y, g(Y)] = E[X|Y]$

*(viii)* $E[E[X|Y, Z]|Y] = E[X|Y]$

**Theorem 2** *For any function $h : \mathbb{R} \to \mathbb{R}$,*

$$E[(X - E[X|Y])^2] \leq E[(X - h(Y))^2]$$

*and we have equality if and only if $h(Y) = E[X|Y]$.*

This follows immediately from the law of total expectation:

$$\mathbb{E}(X) = \mathbb{E}_Y\Big\{\mathbb{E}(X \,|\, Y)\Big\} = \sum_y \mathbb{E}(X \,|\, Y = y)\mathbb{P}(Y = y).$$

## Laws of Total Expectation and Variance

If all the expectations below are finite, then for ANY random variables $X$ and $Y$, we have:

i) $\boxed{\mathbb{E}(X) = \mathbb{E}_Y\Big(\mathbb{E}(X \,|\, Y)\Big)}$ *Law of Total Expectation.*

*Note that we can pick <u>any</u> r.v. $Y$, to make the expectation as easy as we can.*

ii) $\mathbb{E}(g(X)) = \mathbb{E}_Y\Big(\mathbb{E}(g(X) \,|\, Y)\Big)$ *for any function $g$.*

we can give a proof of (1) in the special case where $(X, Y, Z)$ are jointly continuous with a pdf $f(x, y, z)$:

$$E[X \mid Y = y, Z = z] = \frac{\int x \cdot f(x, y, z)\, dx}{\int f(x, y, z)\, dx},$$

$$\Downarrow$$

$$\begin{aligned}
E[E[X \mid Y, Z = z] \mid Z = z] &= \iint \frac{\int x \cdot f(x, y, z)\, dx}{\int f(x, y, z)\, dx} \cdot f(x, y, z)\, dx\, dy \\
&= \int \left( \int x \cdot f(x, y, z)\, dx \right) \frac{\int f(x, y, z)\, dx}{\int f(x, y, z)\, dx}\, dy \\
&= \iint x \cdot f(x, y)\, dx\, dy \\
&= E[X \mid Z = z]
\end{aligned}$$

You can give a similar proof in the case where $X, Y, Z$ are jointly discrete, with a joint probability mass function $f(x, y, z) = P(X = x, Y = y, Z = z)$, for $(x, y, z)$ ranging over some countable support set. Basically, you do this by replacing $\int$ with $\sum$ in the proof above.

One thing you can say is that

$$E[E[X \mid Y, Z] \mid Z] = E[X \mid Z] \tag{1}$$

$E[E[X|Y;Z]|Y=y]$. $E[X|Y;Z]$ is a random variable. Given that $Y=y$, its possible values are $E[X|Y=y;Z=z]$ where $z$ varies over the range of $Z$. Given that $Y=y$, the probability that $E[X|Y;Z]=E[X|Y=y;Z=z]$ is just $P(Z=z|Y=y)$. Hence,

$$
\begin{aligned}
E[E[X|Y;Z]|Y=y] &= \sum_z E[X|Y=y,Z=z]P(Z=z|Y=y) \\
&= \sum_z \sum_x x\,P(X=x|Y=y,Z=z)P(Z=z|Y=y) \\
&= \sum_{z,x} x \frac{P(X=x,Y=y,Z=z)}{P(Y=y,Z=z)} \frac{P(Z=z,Y=y)}{P(Y=y)} \\
&= \sum_{z,x} x \frac{P(X=x,Y=y,Z=z)}{P(Y=y)} \\
&= \sum_{x} x \frac{P(X=x,Y=y)}{P(Y=y)} \\
&= \sum_{x} x\,P(X=x|Y=y) \\
&= E[X|Y=y]
\end{aligned}
$$

This follows immediately from the law of total expectation:

$$\mathbb{E}(X) = \mathbb{E}_Y\Big\{\mathbb{E}(X \,|\, Y)\Big\} = \sum_y \mathbb{E}(X \,|\, Y = y)\mathbb{P}(Y = y).$$

## Laws of Total Expectation and Variance

If all the expectations below are finite, then for ANY random variables $X$ and $Y$, we have:

i) $$\boxed{\mathbb{E}(X) = \mathbb{E}_Y\Big(\mathbb{E}(X \,|\, Y)\Big)}$$ *Law of Total Expectation.*

*Note that we can pick <u>any</u> r.v. $Y$, to make the expectation as easy as we can.*

ii) $\mathbb{E}(g(X)) = \mathbb{E}_Y\Big(\mathbb{E}(g(X) \,|\, Y)\Big)$ *for any function $g$.*

iii) $$\boxed{Var(X) = \mathbb{E}_Y\Big(Var(X \,|\, Y)\Big) + Var_Y\Big(\mathbb{E}(X \,|\, Y)\Big)}$$

*Law of Total Variance.*

(i) is a special case of (ii), so we just need to prove (ii). Begin at RHS:

$$\text{RHS} = \mathbb{E}_Y\left[\mathbb{E}(g(X)\,|\,Y)\right] = \mathbb{E}_Y\left[\sum_x g(x)\mathbb{P}(X=x\,|\,Y)\right]$$

$$= \sum_y \left[\sum_x g(x)\mathbb{P}(X=x\,|\,Y=y)\right]\mathbb{P}(Y=y)$$

(iii) Wish to prove $\mathrm{Var}(X) = \mathbb{E}_Y[\mathrm{Var}(X\,|\,Y)] + \mathrm{Var}_Y[\mathbb{E}(X\,|\,Y)]$. Begin at RHS:

$$\mathbb{E}_Y[\mathrm{Var}(X\,|\,Y)] + \mathrm{Var}_Y[\mathbb{E}(X\,|\,Y)]$$

$$= \mathbb{E}_Y\left\{\mathbb{E}(X^2\,|\,Y) - (\mathbb{E}(X\,|\,Y))^2\right\} + \left\{\mathbb{E}_Y\left\{[\mathbb{E}(X\,|\,Y)]^2\right\} - \Big[\underbrace{\mathbb{E}_Y(\mathbb{E}(X\,|\,Y))}_{\mathbb{E}(X)\text{ by part (i)}}\Big]^2\right\}$$

$$= \underbrace{\mathbb{E}_Y\{\mathbb{E}(X^2\,|\,Y)\}}_{\mathbb{E}(X^2)\text{ by part (i)}} - \mathbb{E}_Y\{[\mathbb{E}(X\,|\,Y)]^2\} + \mathbb{E}_Y\{[\mathbb{E}(X\,|\,Y)]^2\} - (\mathbb{E}X)^2$$

$$= \mathbb{E}(X^2) - (\mathbb{E}X)^2$$

$$= \mathrm{Var}(X) = \text{LHS}. \qquad \square$$

## Theorem 2.4: The Partition Theorem (Law of Total Probability)

*Let $B_1, \ldots, B_m$ form a partition of $\Omega$. Then for any event $A$,*

$$\boxed{\mathbb{P}(A) = \sum_{i=1}^{m} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{m} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}$$

## Proof of partition formula

$$\begin{aligned}
\sum_i \mathrm{E}(X \mid A_i)\,\mathrm{P}(A_i) &= \sum_i \int_\Omega X(\omega)\,\mathrm{P}(d\omega \mid A_i) \cdot \mathrm{P}(A_i) \\
&= \sum_i \int_\Omega X(\omega)\,\mathrm{P}(d\omega \cap A_i) \\
&= \sum_i \int_\Omega X(\omega) I_{A_i}(\omega)\,\mathrm{P}(d\omega) \\
&= \sum_i \mathrm{E}(X I_{A_i}),
\end{aligned}$$

where $I_{A_i}$ is the indicator function of the set $A_i$.

If the partition $\{A_i\}_{i=0}^{n}$ is finite, then, by linearity, the previous expression becomes

$$\mathrm{E}\left(\sum_{i=0}^{n} X I_{A_i}\right) = \mathrm{E}(X),$$

**5.2. Expectation and Variance of Standard Normal Distribution.** Assume $X \backsim \mathcal{N}(0,1)$. Then

$$\mathbf{E}X = \int_{-\infty}^{+\infty} xe^{-x^2/2}\,dx = 0,$$

because the function inside the integral is odd. We can also say that $X$ is symmetric with respect to zero, so $\mathbf{E}X = 0$. Now,

$$\mathbf{E}X^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2 e^{-x^2/2}\,dx = 1.$$

Why is this? We know that

$$\int_{-\infty}^{+\infty} e^{-x^2/2}\,dx = \sqrt{2\pi}.$$

Let $u = e^{-x^2/2}$, $v = x$. Integrate by parts: note that $uv = xe^{-x^2/2} = 0$ for $x = \pm\infty$. So

$$\int_{-\infty}^{+\infty} e^{-x^2/2}\,dx = \int_{-\infty}^{+\infty} u\,dv = uv\Big|_{x=-\infty}^{x=+\infty} - \int_{-\infty}^{+\infty} v\,du$$

$$= -\int_{-\infty}^{+\infty} x\,de^{-x^2/2} = -\int_{-\infty}^{+\infty} x\,(-x)e^{-x^2/2}\,dx = \int_{-\infty}^{+\infty} x^2 e^{-x^2/2}\,dx.$$

This is equal to $\sqrt{2\pi}$, which proves $\mathbf{E}X^2 = 1$. So $\operatorname{Var} X = \mathbf{E}X^2 - (\mathbf{E}X)^2 = 1$. This proves that

$$\boxed{X \sim \mathcal{N}(0,1) \;\Rightarrow\; \mathbf{E}X = 0,\ \operatorname{Var} X = 1}$$

**Normal Density:** $$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]$$

**Bivariate Normal Density:**

$$p(x,y) = \frac{e^{-\frac{1}{2(1-\rho_{xy}^2)}[(\frac{x-\mu_x}{\sigma_x})^2 - \frac{2\rho_{xy}(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + (\frac{y-\mu_y}{\sigma_y})^2]}}{2\pi\sigma_x\sigma_y\sqrt{(1-\rho_{xy}^2)}}$$

$\mu$ - Mean; $\sigma$ - S.D.; $\rho_{xy}$ - Correlation Coefficient

**Visualize ρ as equivalent to the orientation of tilted asymmetric Gaussian filter.**

**For x as a discrete random variable, the expected value of x:**

$$E(x) = \sum_{i=1}^{n} x_i P(x_i) = \mu_x$$

**E(x) is also called the first moment of the distribution.**
**The $k^{th}$ moment is defined as:**

$$E(x^k) = \sum_{i=1}^{n} x_i^k P(x_i)$$

**$P(x_i)$ is the probability of $x = x_i$.**

**Covariance of x and y, is defined as:** $\sigma_{xy} = E[(x-\mu_x)(y-\mu_y)]$

**Covariance indicates how much x and y vary together. The value depends on how much each variable tends to deviate from its mean, and also depends on the degree of association between x and y.**

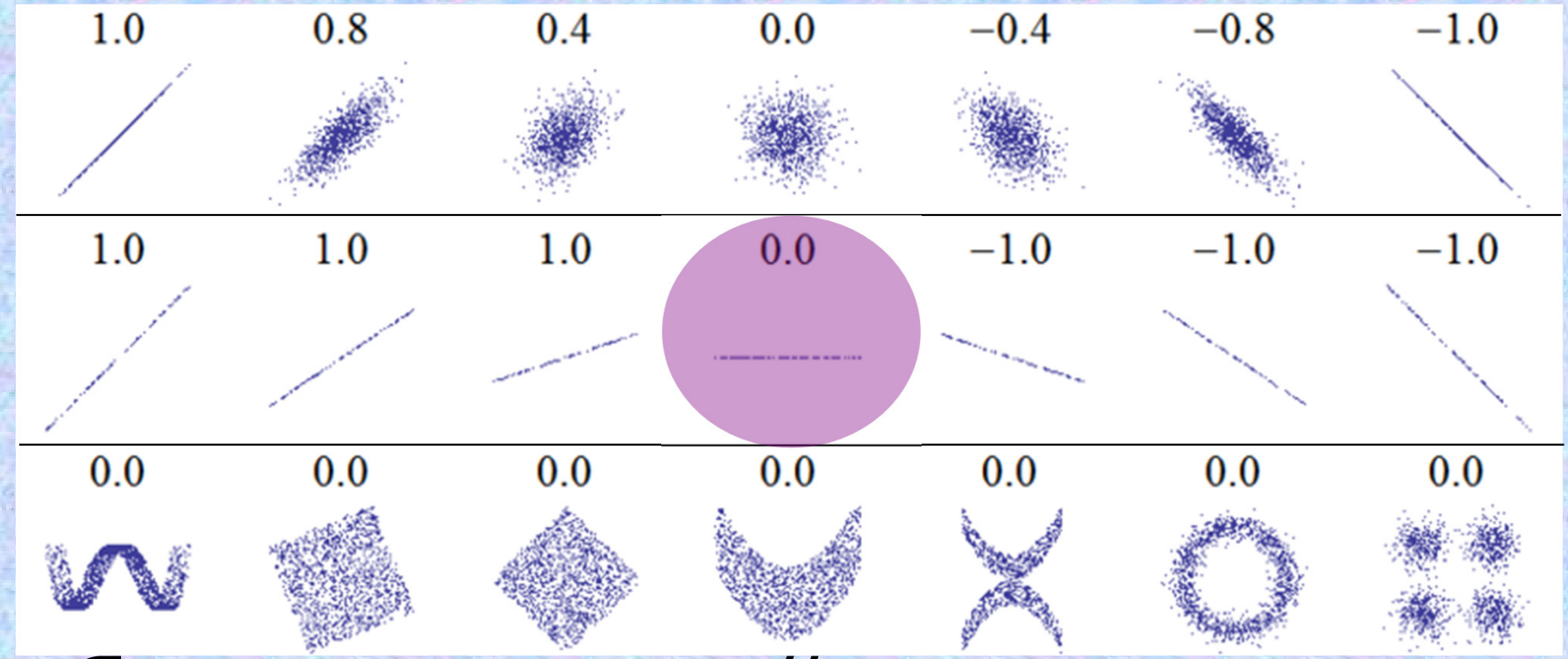
**Correlation between x and y:** $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = E[(\frac{x-\mu_x}{\sigma_x})(\frac{y-\mu_y}{\sigma_y})]$

**Property of correlation coefficient:** $-1 \le \rho_{xy} \le 1$

**For Z = ax + by ;**

$$E[(z-\mu_z)^2] = a^2\sigma_x^2 + 2ab\sigma_{xy} + b^2\sigma_y^2;$$

$$If \quad \sigma_{xy} = 0, \quad \sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2$$




$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = E[(\frac{x-\mu_x}{\sigma_x})(\frac{y-\mu_y}{\sigma_y})]$$

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}}$$

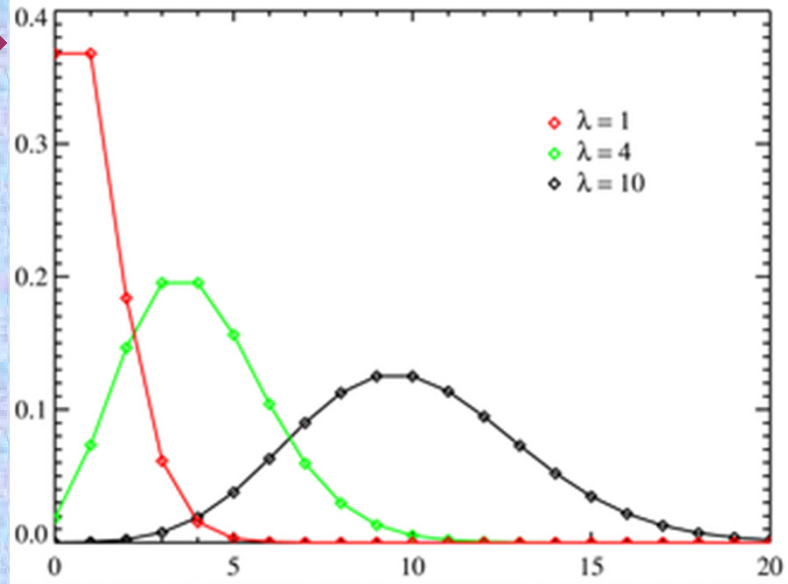
**The correlation coefficient can also be viewed as the cosine of the angle between the two vectors ($\mathbb{R}^D$) of samples drawn from the two random variables.**

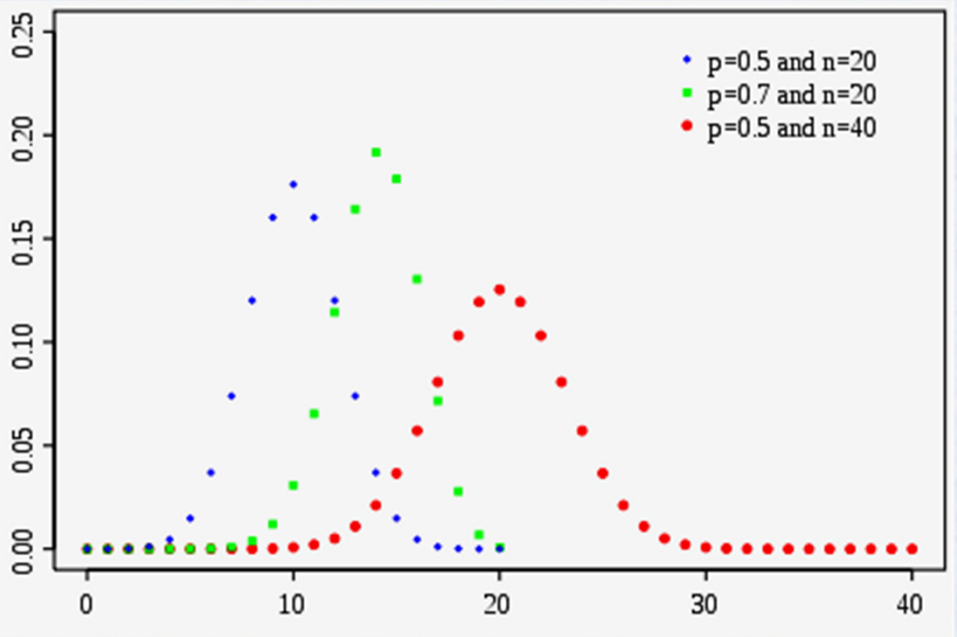
**This method only works with centered data, i.e., data which have been shifted by the sample mean so as to have an average of zero.**

## Other PDFs:

Poisson →

$$P(x) = \frac{\lambda^x}{x!} e^{-\lambda;} \quad \lambda > 0$$



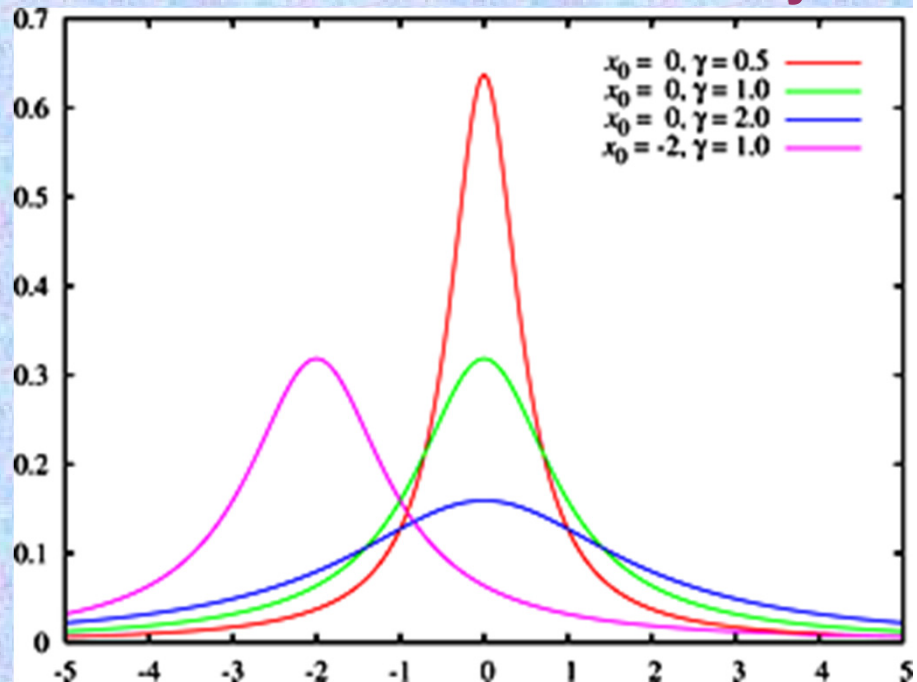

$$\Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

← Binomial

Cauchy

$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma \left[1 + \left(\frac{x - x_0}{\gamma}\right)^2\right]}$$

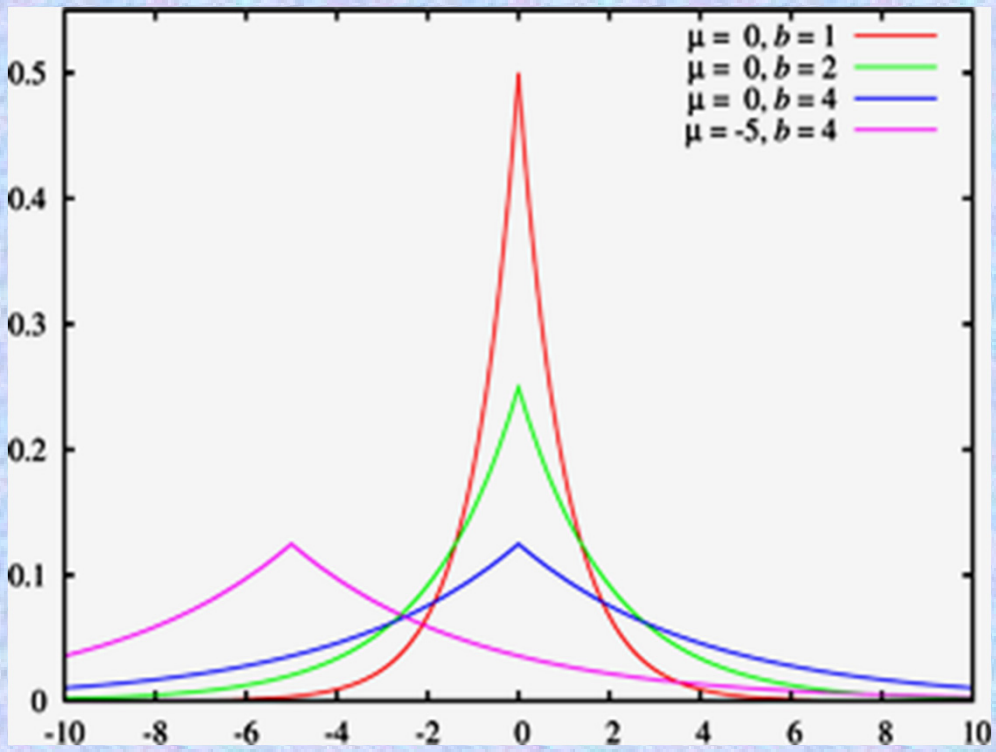
$$= \frac{1}{\pi} \left[\frac{\gamma}{(x - x_0)^2 + \gamma^2}\right]$$



## LAPLACE:

$$f(x|\mu,b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$

$$= \frac{1}{2b}\begin{cases}\exp\left(-\frac{\mu-x}{b}\right) & \text{if } x < \mu \\ \exp\left(-\frac{x-\mu}{b}\right) & \text{if } x \geq \mu\end{cases}$$




## Double Exponential Density:

$$P(x) = \frac{1}{2b}e^{-\left|x-a/b\right|};$$

Read about:

- **Central Limit Theorem**
- **Uniform Distribution**
- **Geometric Distribution**
- **Quantile-Quantile (QQ) Plot**
- **Probability-Probability (P-P) Plot**

| Name of the probability distribution | Probability distribution function | Mean | Variance |
|---|---|---|---|
| Binomial distribution | $\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ | $np$ | $np(1-p)$ |
| Geometric distribution | $\Pr(X = k) = (1-p)^{k-1} p$ | $\frac{1}{p}$ | $\frac{(1-p)}{p^2}$ |
| Normal distribution | $f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ | $\mu$ | $\sigma^2$ |
| Uniform distribution (continuous) | $f(x \mid a, b) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$ | $\frac{a+b}{2}$ | $\frac{(b-a)^2}{12}$ |
| Exponential distribution | $f(x \mid \lambda) = \lambda e^{-\lambda x}$ | $\frac{1}{\lambda}$ | $\frac{1}{\lambda^2}$ |
| Poisson distribution | $f(x \mid \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$ | $\lambda$ | $\lambda$ |

The variance of a random variable $X$ is the expected value of the squared deviation from the mean of $X$, $\mu = \mathrm{E}[X]$:

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

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

In other words, the variance of $X$ is equal to the mean of the square of $X$ minus the square of the mean of $X$.

A formula for calculating the variance of an entire population of size $N$ is:

$$\sigma^2 = \overline{(x^2)} - \bar{x}^2 = \frac{\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2 / N}{N}.$$

Using Bessel's correction to calculate an unbiased estimate of the population variance from a finite sample of $n$ observations

$$s^2 = \left(\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2\right) \cdot \frac{n}{n-1}.$$

## Discrete random variable [ edit ]

If the generator of random variable $X$ is discrete with probability mass function $x_1 \mapsto p_1, x_2 \mapsto p_2, \ldots, x_n \mapsto p_n$, then

$$\operatorname{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2,$$

or equivalently,

$$\operatorname{Var}(X) = \left(\sum_{i=1}^{n} p_i x_i^2\right) - \mu^2,$$

where $\mu$ is the expected value. That is,

$$\mu = \sum_{i=1}^{n} p_i x_i.$$

(When such a discrete weighted variance is specified by weights whose sum is not 1, then one divides by the sum of the weights.)

The variance of a collection of $n$ equally likely values can be written as

$$\operatorname{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \mu^2,$$

where $\mu$ is the average value. That is,

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

If the random variable $X$ has a probability density function $f(x)$, and $F(x)$ is the corresponding cumulative distribution function, then

$$\begin{aligned}
\operatorname{Var}(X) = \sigma^2 &= \int_{\mathbb{R}} (x-\mu)^2 f(x)\,dx \\
&= \int_{\mathbb{R}} x^2 f(x)\,dx - 2\mu \int_{\mathbb{R}} x f(x)\,dx + \mu^2 \int_{\mathbb{R}} f(x)\,dx \\
&= \int_{\mathbb{R}} x^2\,dF(x) - 2\mu \int_{\mathbb{R}} x\,dF(x) + \mu^2 \int_{\mathbb{R}} dF(x) \\
&= \int_{\mathbb{R}} x^2\,dF(x) - 2\mu \cdot \mu + \mu^2 \cdot 1 \\
&= \int_{\mathbb{R}} x^2\,dF(x) - \mu^2,
\end{aligned}$$

or equivalently,

$$\operatorname{Var}(X) = \int_{\mathbb{R}} x^2 f(x)\,dx - \mu^2,$$

where $\mu$ is the expected value of $X$ given by

$$\mu = \int_{\mathbb{R}} x f(x)\,dx = \int_{\mathbb{R}} x\,dF(x).$$

*Variance* ( $\sigma^2$ ) of a discrete random variable $X$ is defined as

$$\sigma^2 = V(X) = E(X - E(X))^2$$

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

$$\begin{aligned}
\mathrm{Var}(X) = \sigma^2 &= \int_{\mathbb{R}} (x-\mu)^2 f(x)\,dx \\
&= \int_{\mathbb{R}} x^2 f(x)\,dx - 2\mu \int_{\mathbb{R}} x f(x)\,dx + \mu^2 \int_{\mathbb{R}} f(x)\,dx \\
&= \int_{\mathbb{R}} x^2\,dF(x) - 2\mu \int_{\mathbb{R}} x\,dF(x) + \mu^2 \int_{\mathbb{R}} dF(x) \\
&= \int_{\mathbb{R}} x^2\,dF(x) - 2\mu \cdot \mu + \mu^2 \cdot 1 \\
&= \int_{\mathbb{R}} x^2\,dF(x) - \mu^2,
\end{aligned}$$

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

$$\sum_x x^2 p(x) - \mu^2,$$

$$2E(X)\cdot X + E^2(X)\}$$

$$= E(X^2) - 2\cdot E(X)E(X) + E^2(X) = E(X^2) - E^2(X).$$

## Definition [ edit ]

Throughout this article, boldfaced unsubscripted $\mathbf{X}$ and $\mathbf{Y}$ are used to refer to random vectors, and unboldfaced subscripted $X_i$ and $Y_i$ are used to refer to scalar random variables.

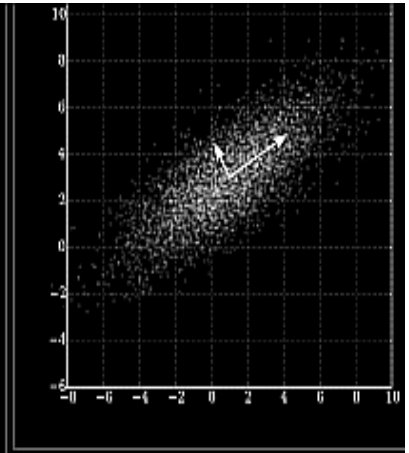
If the entries in the column vector

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)^{\mathrm{T}}$$

are random variables, each with finite variance and expected value, then the covariance matrix $\mathrm{K}_{\mathbf{XX}}$ is the matrix whose $(i, j)$ entry is the covariance[1]:p. 177

$$\mathrm{K}_{X_i X_j} = \operatorname{cov}[X_i, X_j] = \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])]$$

where the operator $\mathrm{E}$ denotes the expected value (mean) of its argument.



Sample points from a bivariate Gaussian distribution with a standard deviation of 3 in roughly the lower left-upper right direction and of 1 in the orthogonal direction. Because the *x* and *y* components co-vary, the variances of $x$ and $y$ do not fully describe the distribution. A $2 \times 2$ covariance matrix is needed; the directions of the arrows correspond to the eigenvectors of this covariance matrix and their lengths to the square roots of the eigenvalues.

### Conflicting nomenclatures and notations [ edit ]

Nomenclatures differ. Some statisticians, following the probabilist William Feller in his two-volume book *An Introduction to Probability Theory and Its Applications*,[2] call the matrix $\mathrm{K}_{\mathbf{XX}}$ the **variance** of the random vector $\mathbf{X}$, because it is the natural generalization to higher dimensions of the 1-dimensional variance. Others call it the **covariance matrix**, because it is the matrix of covariances between the scalar components of the vector $\mathbf{X}$.

$$\operatorname{var}(\mathbf{X}) = \operatorname{cov}(\mathbf{X}, \mathbf{X}) = \mathrm{E}\left[(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\mathbf{X} - \mathrm{E}[\mathbf{X}])^{\mathrm{T}}\right].$$

Both forms are quite standard, and there is no ambiguity between them. The matrix $\mathrm{K}_{\mathbf{XX}}$ is also often called the *variance-covariance matrix*, since the diagonal terms are in fact variances.

By comparison, the notation for the cross-covariance matrix *between* two vectors is

$$\operatorname{cov}(\mathbf{X}, \mathbf{Y}) = \mathrm{K}_{\mathbf{XY}} = \mathrm{E}\left[(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\mathbf{Y} - \mathrm{E}[\mathbf{Y}])^{\mathrm{T}}\right].$$

## Basic properties

For $\mathrm{K_{XX}} = \mathrm{var}(\mathbf{X}) = \mathrm{E}\left[(\mathbf{X} - \mathrm{E}[\mathbf{X}])\,(\mathbf{X} - \mathrm{E}[\mathbf{X}])^{\mathrm{T}}\right]$ and $\mu_{\mathrm{X}} = \mathrm{E}[\mathbf{X}]$, where $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$ is a $n$-dimensional random variable, the following basic properties apply:[4]

1. $\mathrm{K_{XX}} = \mathrm{E}(\mathbf{X}\mathbf{X}^{\mathrm{T}}) - \mu_{\mathrm{X}}{\mu_{\mathrm{X}}}^{\mathrm{T}}$
2. $\mathrm{K_{XX}}$ is positive-semidefinite, i.e. $\mathbf{a}^T\,\mathrm{K_{XX}}\,\mathbf{a} \geq 0 \quad \text{for all } \mathbf{a} \in \mathbb{R}^n$
3. $\mathrm{K_{XX}}$ is symmetric, i.e. $\mathrm{K_{XX}^{T}} = \mathrm{K_{XX}}$
4. For any constant (i.e. non-random) $m \times n$ matrix $\mathbf{A}$ and constant $m \times 1$ vector $\mathbf{a}$, one has $\mathrm{var}(\mathbf{A}\mathbf{X} + \mathbf{a}) = \mathbf{A}\ \mathrm{var}(\mathbf{X})\,\mathbf{A}^{\mathrm{T}}$
5. If $\mathbf{Y}$ is another random vector with the same dimension as $\mathbf{X}$, then $\mathrm{var}(\mathbf{X} + \mathbf{Y}) = \mathrm{var}(\mathbf{X}) + \mathrm{cov}(\mathbf{X}, \mathbf{Y}) + \mathrm{cov}(\mathbf{Y}, \mathbf{X}) + \mathrm{var}(\mathbf{Y})$ where $\mathrm{cov}(\mathbf{X}, \mathbf{Y})$ is the cross-covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$.

For random vectors $\mathbf{X}$ and $\mathbf{Y}$, each containing random elements whose expected value and variance exist, the **cross-covariance matrix** of $\mathbf{X}$ and $\mathbf{Y}$ is defined by[1]:p.336

$$\boxed{\mathrm{K_{XY}} = \mathrm{cov}(\mathbf{X}, \mathbf{Y}) \stackrel{\mathrm{def}}{=} \mathrm{E}[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{Y} - \mu_{\mathbf{Y}})^{\mathrm{T}}] \quad \textbf{(Eq.1)}}$$

where $\mu_{\mathbf{X}} = \mathrm{E}[\mathbf{X}]$ and $\mu_{\mathbf{Y}} = \mathrm{E}[\mathbf{Y}]$ are vectors containing the expected values of $\mathbf{X}$ and $\mathbf{Y}$. The vectors $\mathbf{X}$ and $\mathbf{Y}$ need not have the same dimension, and either might be a scalar value.

The cross-covariance matrix is the matrix whose $(i, j)$ entry is the covariance

$$\mathrm{K}_{X_i Y_j} = \mathrm{cov}[X_i, Y_j] = \mathrm{E}[(X_i - \mathrm{E}[X_i])(Y_j - \mathrm{E}[Y_j])]$$

For the cross-covariance matrix, the following basic properties apply:[2]

1. $\operatorname{cov}(\mathbf{X},\mathbf{Y}) = \operatorname{E}[\mathbf{X}\mathbf{Y}^{\mathrm{T}}] - \mu_{\mathbf{X}}\mu_{\mathbf{Y}}{}^{\mathrm{T}}$
2. $\operatorname{cov}(\mathbf{X},\mathbf{Y}) = \operatorname{cov}(\mathbf{Y},\mathbf{X})^{\mathrm{T}}$
3. $\operatorname{cov}(\mathbf{X}_1 + \mathbf{X}_2, \mathbf{Y}) = \operatorname{cov}(\mathbf{X}_1, \mathbf{Y}) + \operatorname{cov}(\mathbf{X}_2, \mathbf{Y})$
4. $\operatorname{cov}(A\mathbf{X} + \mathbf{a}, B^{\mathrm{T}}\mathbf{Y} + \mathbf{b}) = A\operatorname{cov}(\mathbf{X},\mathbf{Y})B$
5. If $\mathbf{X}$ and $\mathbf{Y}$ are independent (or somewhat less restrictedly, if every random variable in $\mathbf{X}$ is uncorrelated with every random variable in $\mathbf{Y}$), then $\operatorname{cov}(\mathbf{X},\mathbf{Y}) = 0_{p\times q}$

where $\mathbf{X}$, $\mathbf{X}_1$ and $\mathbf{X}_2$ are random $p \times 1$ vectors, $\mathbf{Y}$ is a random $q \times 1$ vector, $\mathbf{a}$ is a $q \times 1$ vector, $\mathbf{b}$ is a $p \times 1$ vector, $A$ and $B$ are $q \times p$ matrices of constants, and $0_{p\times q}$ is a $p \times q$ matrix of zeroes.

Given a sample consisting of $n$ independent observations $x_1,\ldots, x_n$ of a $p$-dimensional random vector $X \in \mathbf{R}^{p\times 1}$ (a $p\times 1$ column-vector), an unbiased estimator of the ($p\times p$) covariance matrix

$$\Sigma = \operatorname{E}\left[(X - \operatorname{E}[X])\,(X - \operatorname{E}[X])^{\mathrm{T}}\right]$$

is the sample covariance matrix

$$\mathrm{Q} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^{\mathrm{T}},$$

where $x_i$ is the $i$-th observation of the $p$-dimensional random vector, and the vector

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

is the sample mean. This is true regardless of the distribution of the random variable $X$, provided of course that the theoretical means and covariances exist. The reason

# Which matrices are covariance matrices?

let $\mathbf{b}$ be a $(p \times 1)$ real-valued vector, then

$$\operatorname{var}(\mathbf{b}^{\mathrm{T}}\mathbf{X}) = \mathbf{b}^{\mathrm{T}} \operatorname{var}(\mathbf{X})\mathbf{b},$$

which must always be nonnegative, since it is the variance of a real-valued random variable, so a covariance matrix is always a positive-semidefinite matrix.

The above argument can be expanded as follows:

$$w^{\mathrm{T}} \operatorname{E}\left[(\mathbf{X} - \operatorname{E}[\mathbf{X}])(\mathbf{X} - \operatorname{E}[\mathbf{X}])^{\mathrm{T}}\right] w = \operatorname{E}\left[w^{\mathrm{T}}(\mathbf{X} - \operatorname{E}[\mathbf{X}])(\mathbf{X} - \operatorname{E}[\mathbf{X}])^{\mathrm{T}} w\right]$$
$$= \operatorname{E}\left[\left(w^{\mathrm{T}}(\mathbf{X} - \operatorname{E}[\mathbf{X}])\right)^2\right] \geq 0,$$

where the last inequality follows from the observation that $w^{\mathrm{T}}(\mathbf{X} - \operatorname{E}[\mathbf{X}])$ is a scalar.

Conversely, every symmetric positive semi-definite matrix is a covariance matrix. To see this, suppose $M$ is a $p \times p$ symmetric positive-semidefinite matrix. From the finite-dimensional case of the spectral theorem, it follows that $M$ has a nonnegative symmetric square root, which can be denoted by $\mathbf{M}^{1/2}$. Let $\mathbf{X}$ be any $p \times 1$ column vector-valued random variable whose covariance matrix is the $p \times p$ identity matrix. Then

$$\operatorname{var}(\mathbf{M}^{1/2}\mathbf{X}) = \mathbf{M}^{1/2} \operatorname{var}(\mathbf{X}) \mathbf{M}^{1/2} = \mathbf{M}.$$

$$= \operatorname{E}[b(X - \operatorname{E}[X])(X - \operatorname{E}[X])^{\top} b^{\top}]$$
$$= b\operatorname{E}[(X - \operatorname{E}[X])(X - \operatorname{E}[X])^{\top}]b^{\top}$$
$$= b\operatorname{Var}[X]b^{\top}$$

# PROB. & STAT. Contd.

**Sample mean is defined as:** $$\tilde{x} = \sum_{i=1}^{n} x_i P(x_i) = \frac{1}{n}\sum_{i=1}^{n} x_i$$ **where,**

**$P(x_i) = 1/n$.**

**Sample Variance is:** $$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \tilde{x})^2$$

**Higher order moments may also be computed:** $$E(x_i - \tilde{x})^3;\ E(x_i - \tilde{x})^4$$

**Covariance of a bivariate distribution:**

$$\sigma_{xy} = E[(x-\mu_x)(y-\mu_y)] = \frac{1}{n}\sum_{i=1}^{n}(x-\tilde{x})(y-\tilde{y})$$

**Second, third,… moments of the distribution p(x) are the expected values of: $x^2$, $x^3$,…**

**The $k^{th}$ central moment is defined as:**

$$E[(x-\mu_x)^k] = \sum_{i=1}^{n} (x-\mu_x)^k P(x_i)$$

**Thus, the second central moment (also called Variance) of a random variable x is defined as:**

$$\sigma_x^2 = E[\{x - E(x)\}^2] = E[(x-\mu_x)^2]$$

**S.D. of x is $\sigma_x$.**

$$\sigma_x^2 = E[\{x - E(x)\}^2] = E[(x-\mu_x)^2]$$

$$= E(x^2) - 2\mu_x^2 + \mu_x^2 = E(x^2) - \mu_x^2$$

*Thus*

$$E(x^2) = \sigma^2 + \mu^2$$

**If z is a new variable: z= ax + by; Then E(z) = E(ax + by)=aE(x) + bE(y).**

$$M_X(t) = \mathrm{E}[e^{tX}] = \int_{x=-\infty}^{\infty} e^{tx} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx =$$

$$\int_{x=-\infty}^{\infty} \frac{e^{-(x^2-2tx+t^2)/2} e^{t^2/2}}{\sqrt{2\pi}}\, dx = e^{t^2/2} \int_{x=-\infty}^{\infty} \frac{e^{-(x-t)^2/2}}{\sqrt{2\pi}}\, dx.$$

But this last integrand is a normal density with mean $t$ and variance 1, thus integrates to 1. Hence

$$M_X(t) = e^{t^2/2}.$$

Now we recall that

$$\mathrm{E}[X^k] = \left[\frac{d^k M_X(t)}{dt^k}\right]_{t=0},$$

so let's calculate successive derivatives:

$$\begin{aligned}
M_X'(t) &= te^{t^2/2} \\
M_X''(t) &= e^{t^2/2} + t^2 e^{t^2/2} = (1+t^2)e^{t^2/2} \\
M_X'''(t) &= 2te^{t^2/2} + (1+t^2)te^{t^2/2} = (3t+t^3)e^{t^2/2} \\
M_X^{(4)}(t) &= (3+3t^2)e^{t^2/2} + (3t^2+t^4)e^{t^2/2} = (3+6t^2+t^4)e^{t^2/2},
\end{aligned}$$

and it is fairly easy to continue this. Now simply evaluate all of these at $t = 0$ to get

$$\begin{aligned}
\mathrm{E}[X] &= 0 \\
\mathrm{E}[X^2] &= 1 \\
\mathrm{E}[X^3] &= 0 \\
\mathrm{E}[X^4] &= 3.
\end{aligned}$$

# MAXIMUM LIKELIHOOD ESTIMATE (MLE)

**The ML estimate (MLE) of a parameter is that value which, when substituted into the probability distribution (or density), produces that distribution for which the probability of obtaining the entire observed set of samples is maximized.**

**Problem: Find the maximum likelihood estimate for μ in a normal distribution.**

**Normal Density:**

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]$$

**Assuming all random samples to be independent:**

$$p(x_1,,,,x_n) = p(x_1).....p(x_n) = \prod_{i=1}^{n} p(x_i)$$

$$= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\frac{x-\mu}{\sigma})^2]$$

**Taking derivative (w.r.t. μ ) of the LOG of the above:**

$$\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu).2 = \frac{1}{\sigma^2}[\sum_{i=1}^{n} x_i - n\mu]$$

**Setting this term = 0, we get:**

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

**Also read about MAP estimate – Baye's is an example.**