What is Statistics?

Definition of Statistics

- **Statistics** 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**.

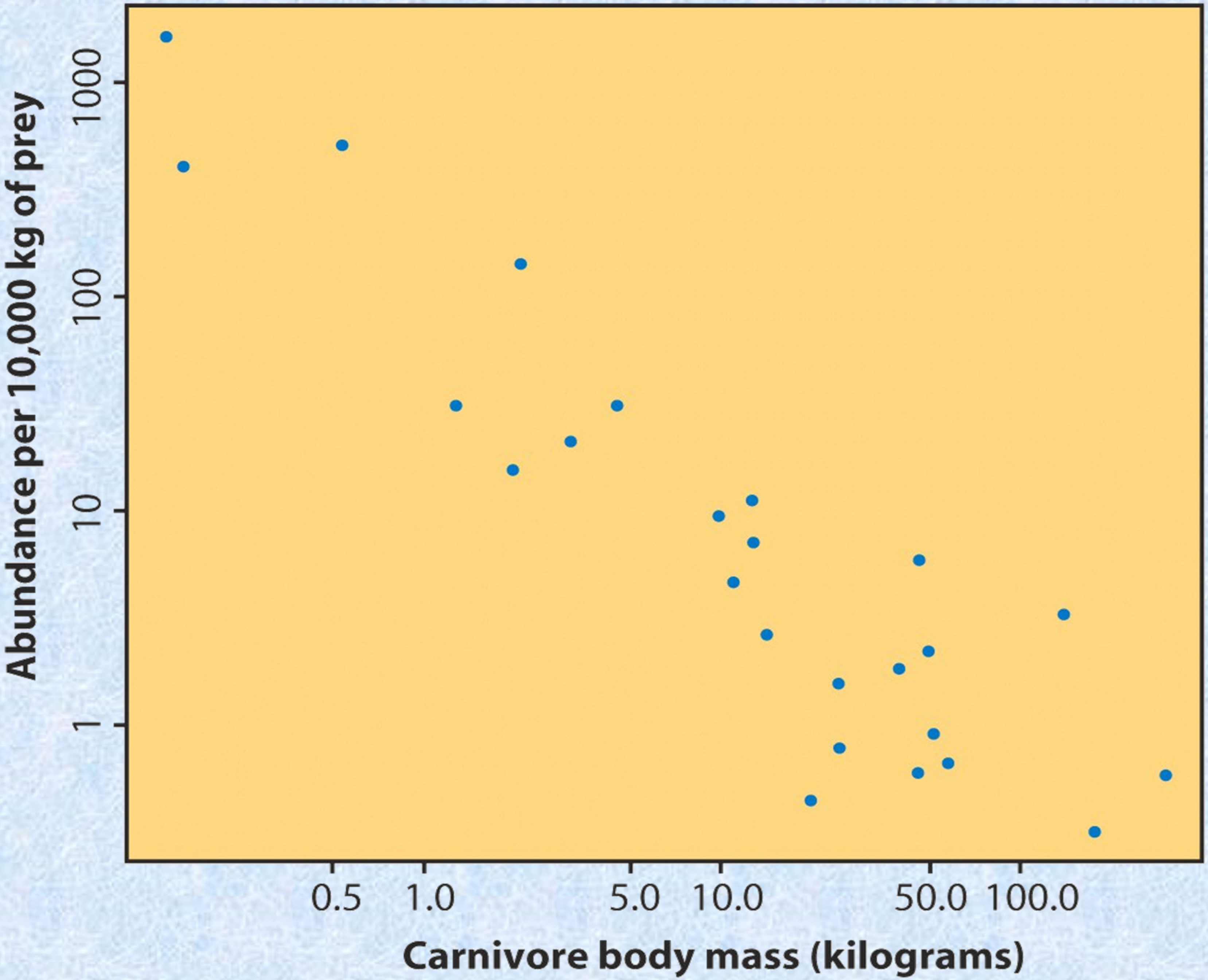
- Independent variable and dependent variable.

- 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 strength and direction of the ***linear association*** between two quantitative variables, usually labeled X and Y.

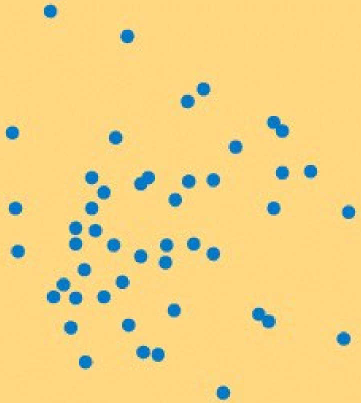
$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Facts about correlation

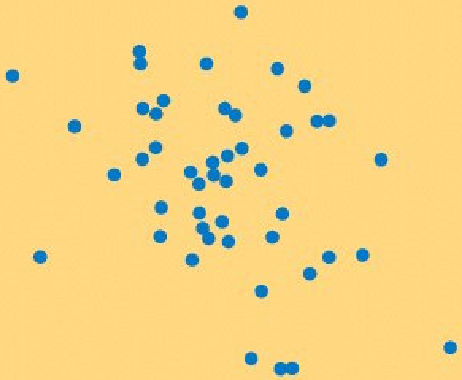
- What kind of variables do we use?
  - 1. No distinction between explanatory and response variables.
  - 2. Both variables should be quantitative
- Numerical properties
  - 1. $-1 \le r \le 1$
  - 2. r>0: positive association between variables
  - 3. r<0: negative association between variables
  - 4.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

$$\begin{array}{lcr} \multicolumn{3}{l}{\triangleleft -\mathit{negative\ relationship}- \triangleright \triangleleft -\mathit{positive\ relationship}- \triangleright} \\ -1 & 0 & 1 \\ \triangleleft ----\mathit{stronger} & & \mathit{stronger}---- \triangleright \end{array}$$

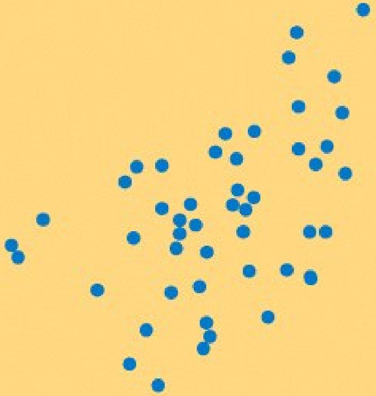
  - 6. Effected by a few outliers →not resistant.
  - 7. It doesn’t describe curved relationships
  - 8. Not easy to guess the value of r from the appearance of a scatter plot



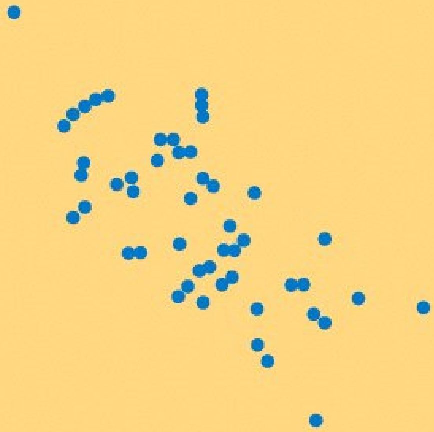
Correlation $r = 0$



Correlation $r = -0.3$

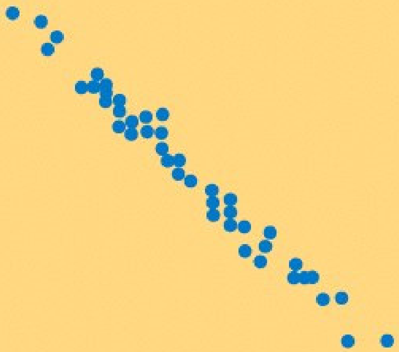


Correlation $r = 0.5$



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.**

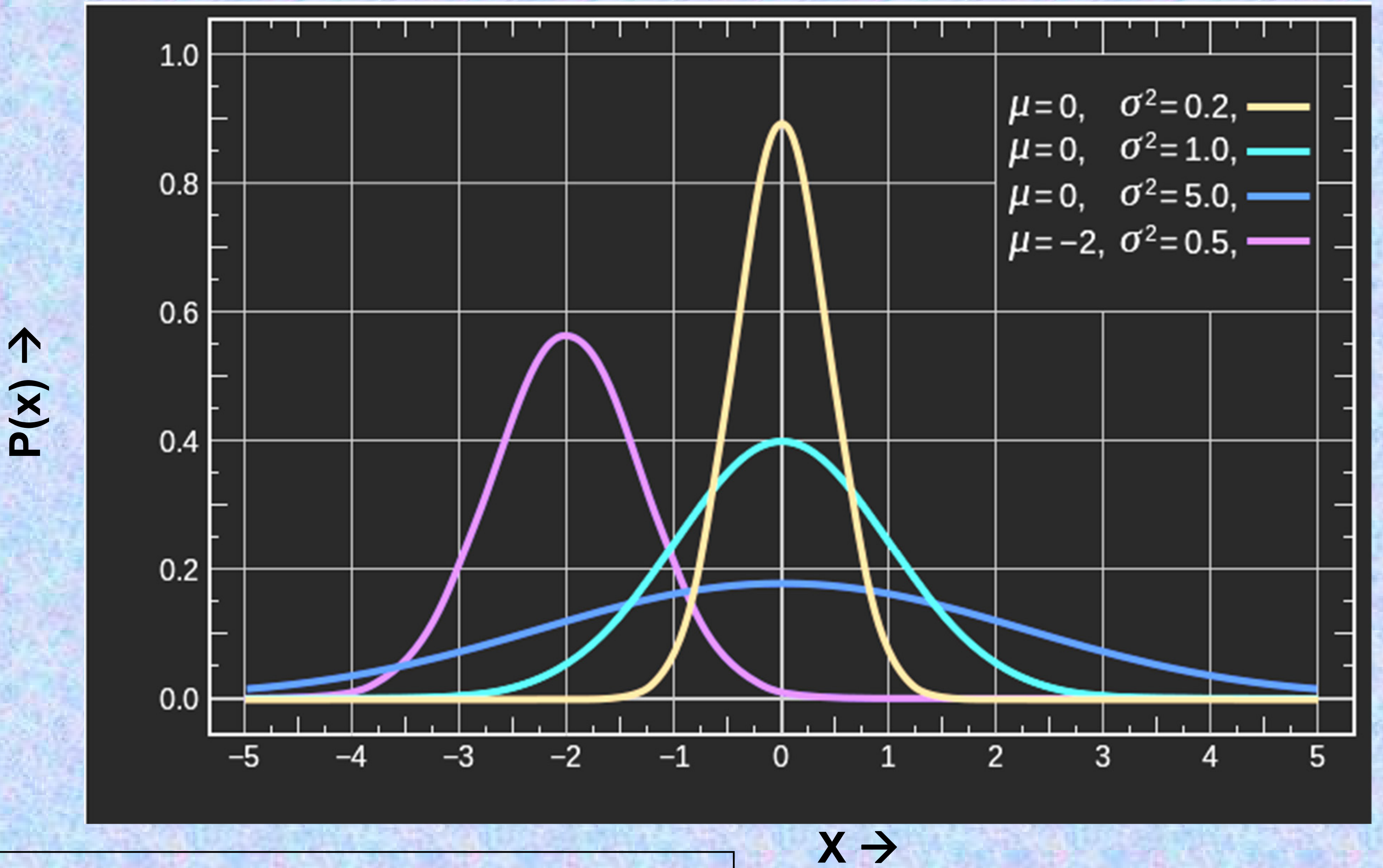
Its also called a BELL function/curve.

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 μ is called the mean or expectation (or median or mode) of the distribution.

The parameter σ is the standard deviation;
and variance is thus σ^2.



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

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

The normal distribution $p(x)$, with any mean μ and any positive deviation σ, has the following properties:

- **It is symmetric around the mean (μ) 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:

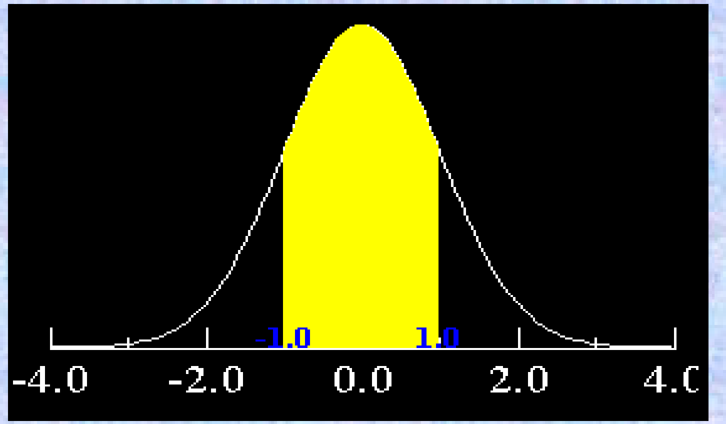
- **Its first derivative $p'(x)$ is: $-x.p(x)$.**
- **Its second derivative $p''(x)$ is: $(x^2 - 1).p(x)$**
- **More generally, its n-th derivative :**

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

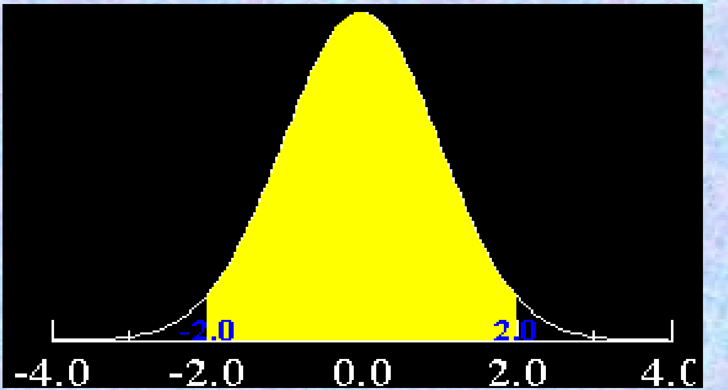
where, H_n is the Hermite polynomial of order n.

The 68 – 95 - 99.7% Rule:

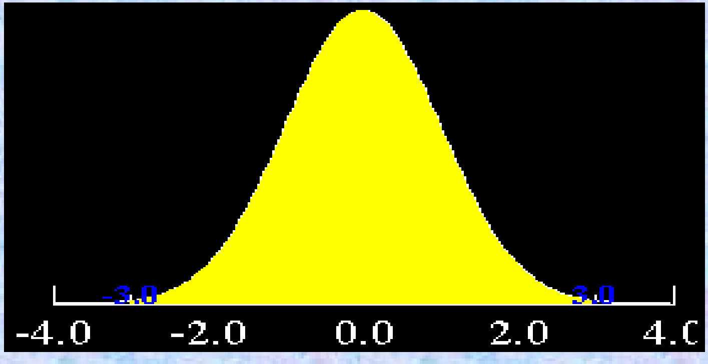
All normal density curves satisfy the following property which is often referred to as the Empirical Rule:



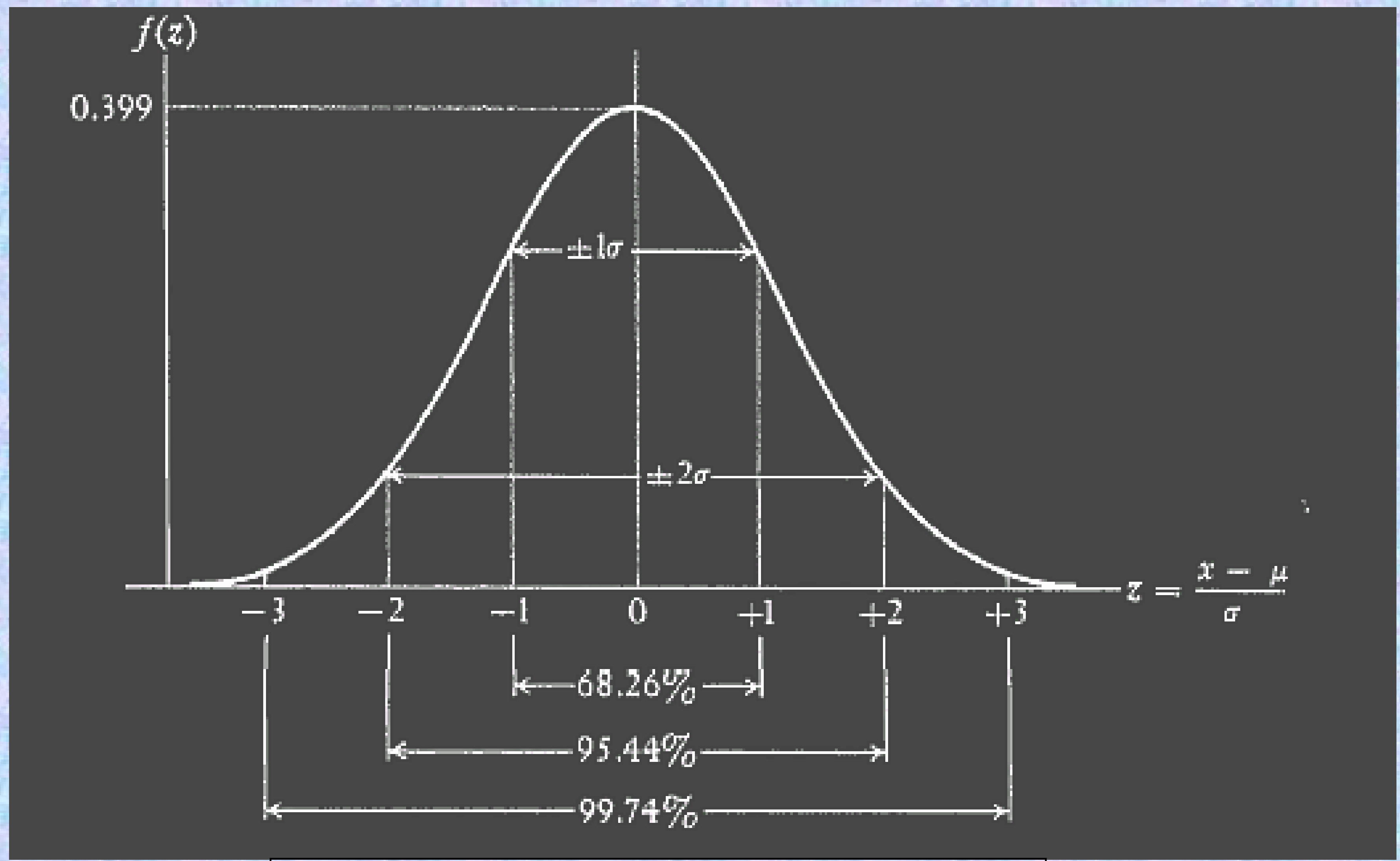
- 68% of the observations fall within 1 standard deviation of the mean, that is, between $(\mu-\sigma)\text{ and }(\mu+\sigma)$



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



- 99.7% of the observations fall within 3 standard deviations of the mean, that is, between $(\mu-3\sigma)\text{ and }(\mu+3\sigma)$

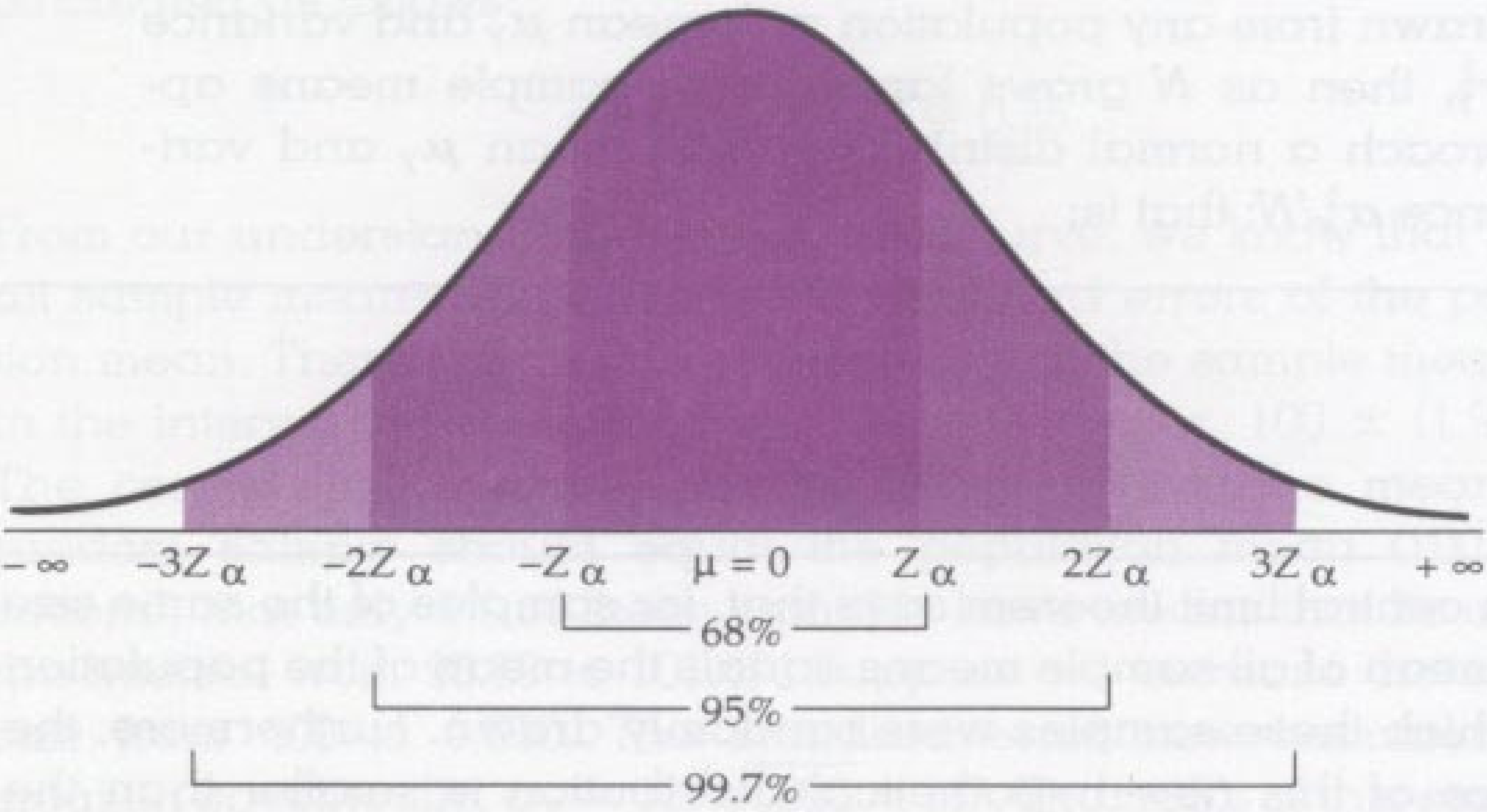


$$\boxed{p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]}$$

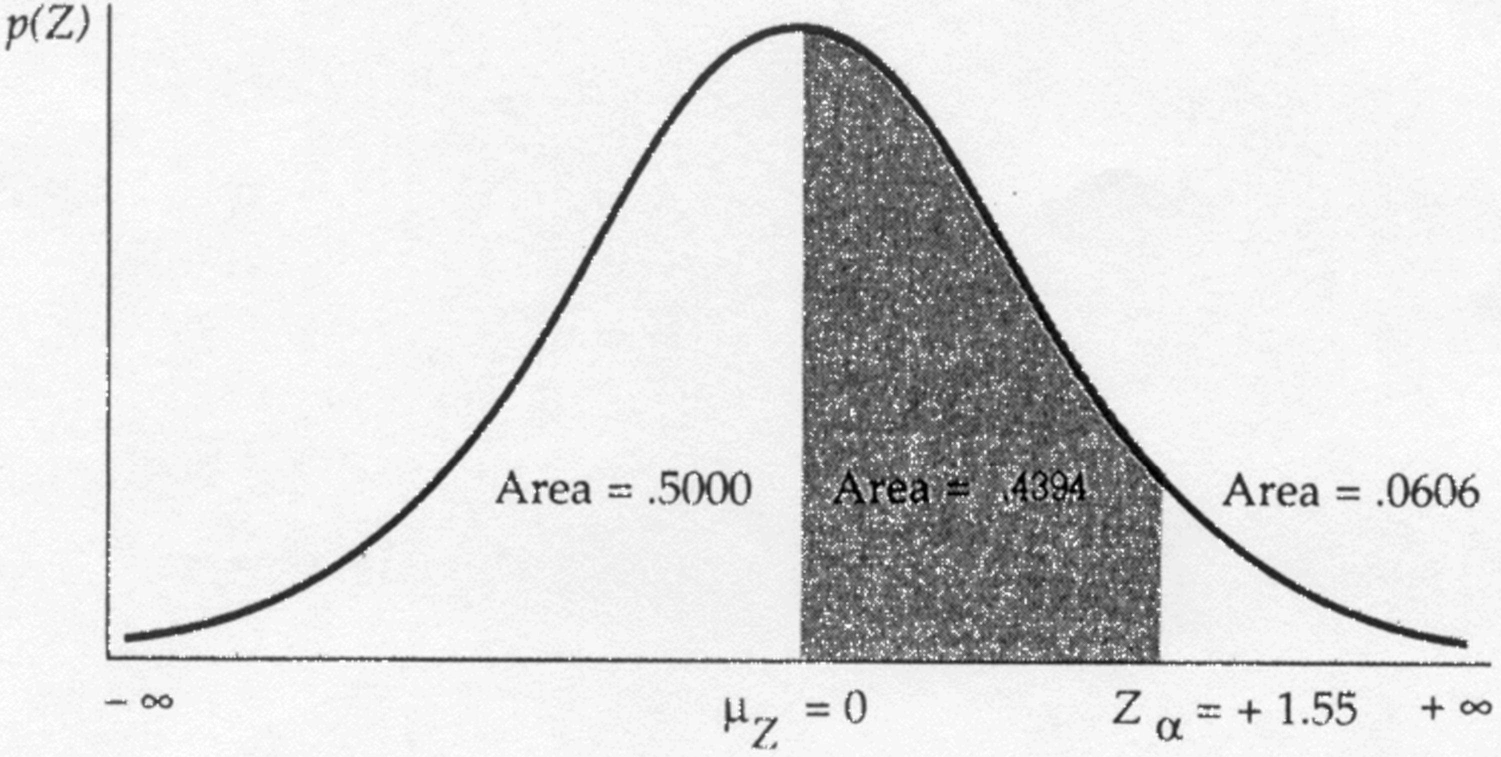
A normal distribution:

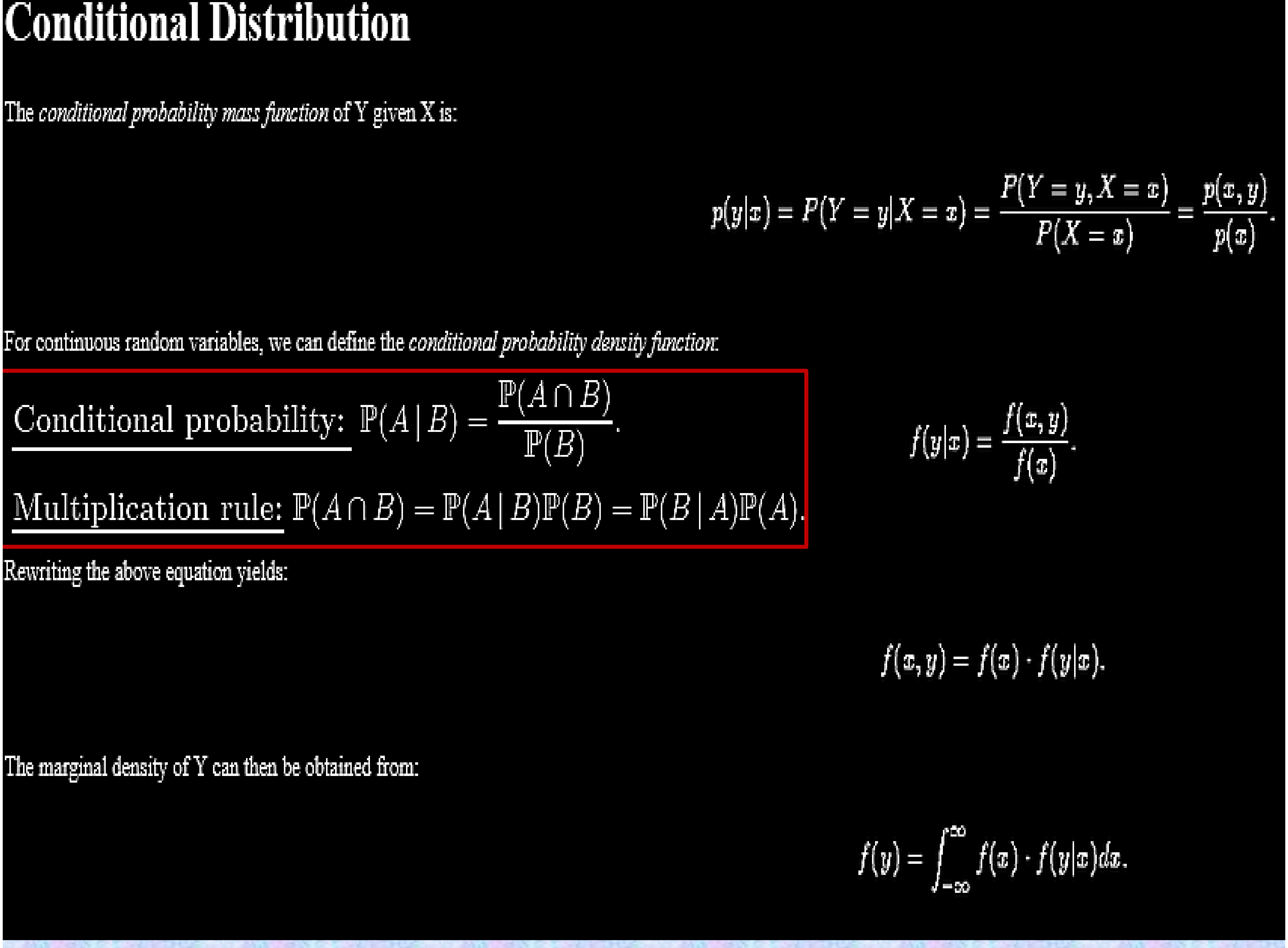
1. is **symmetrical** (both halves are *identical*);
2. is **asymptotic** (its *tails never touch* the underlying x-axis; the curve reaches to – ∞ and + ∞ and thus must be truncated);
3. has **fixed** and **known** *areas under the curve* (these fixed areas are marked off by units along the x-axis called **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*:

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

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

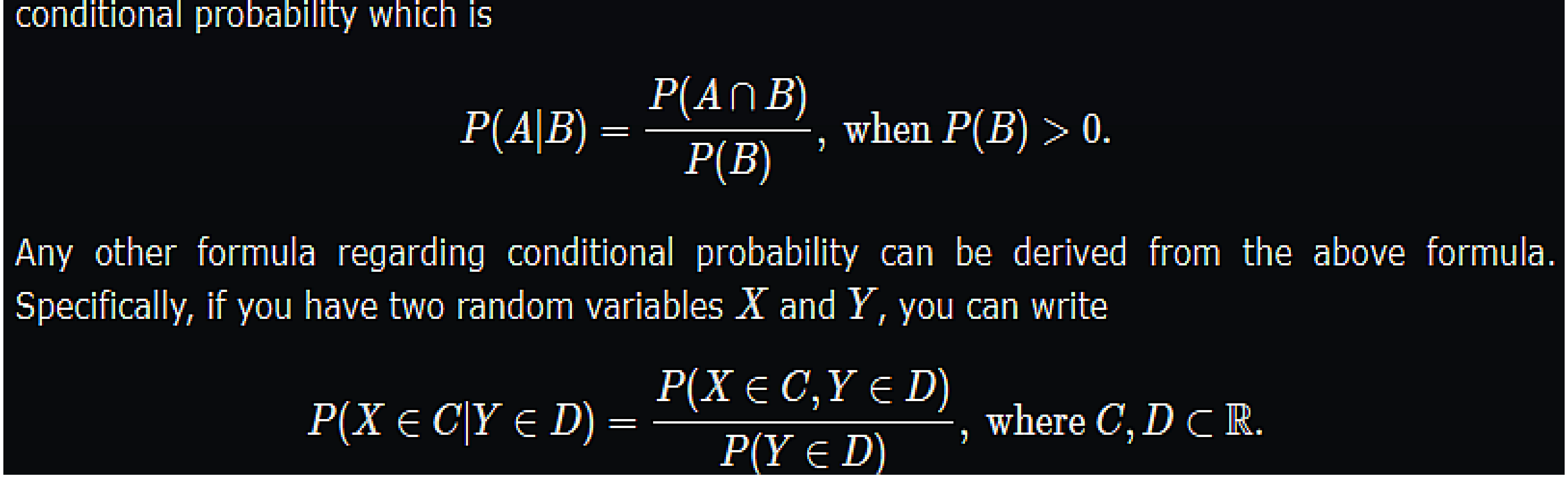
$$f(y|x) = \frac{f(x, y)}{f(x)}.$$

Rewriting the above equation yields:

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

The marginal density of Y can then be obtained from:

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

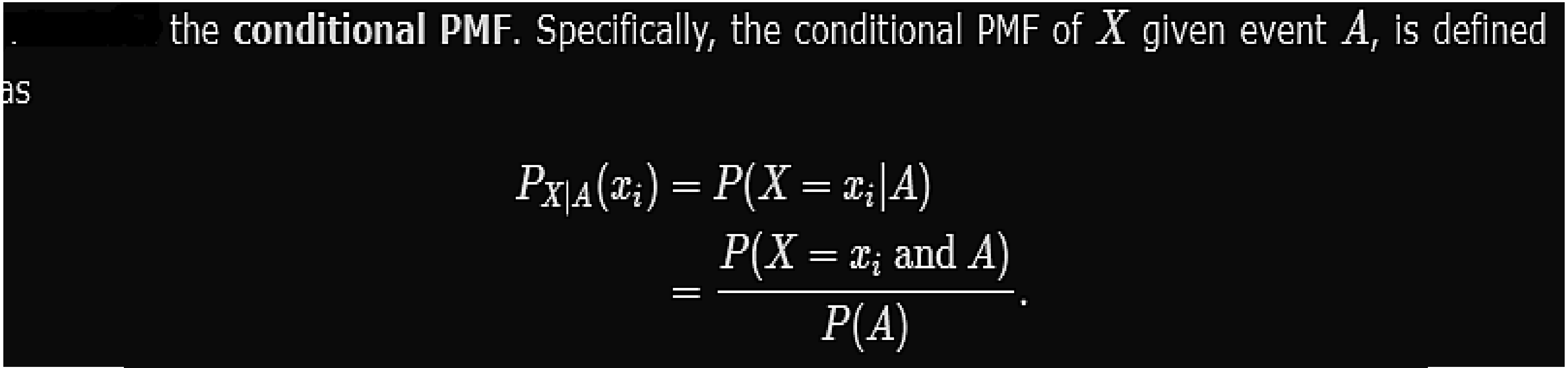


conditional probability which is

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

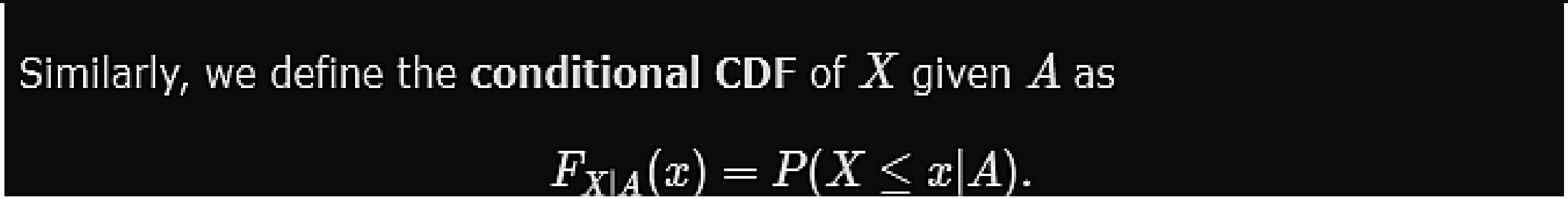
Any other formula regarding conditional probability can be derived from the above formula. Specifically, if you have two random variables X and Y, you can write

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



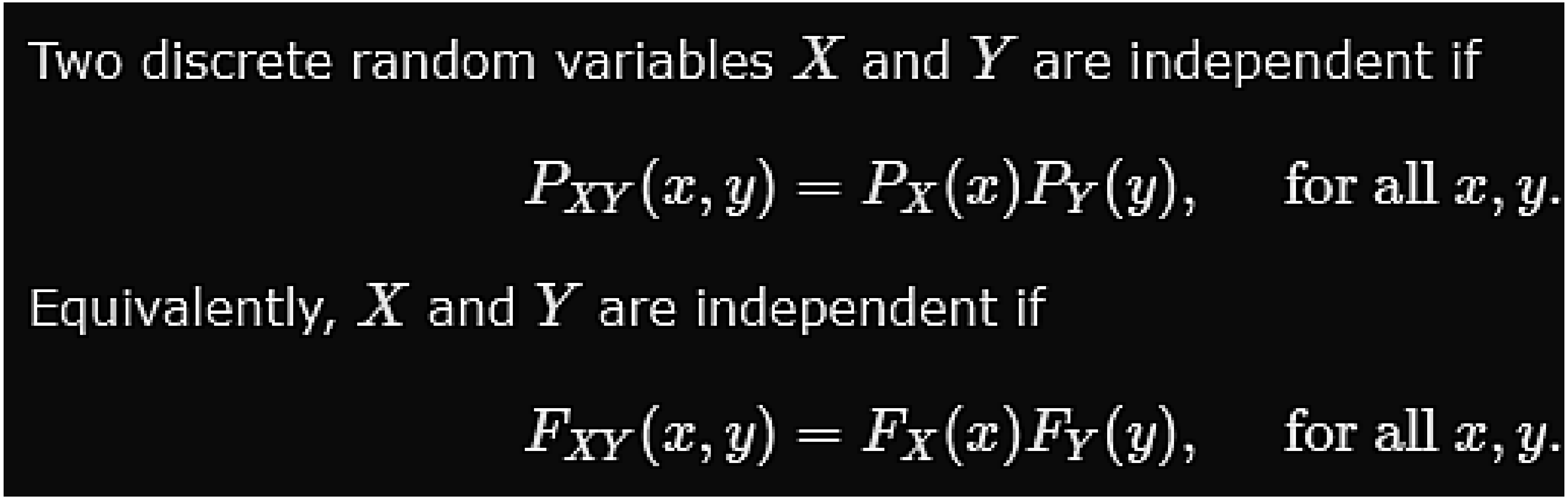
[redacted] the **conditional PMF**. Specifically, the conditional PMF of X given event A, is defined as

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



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

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

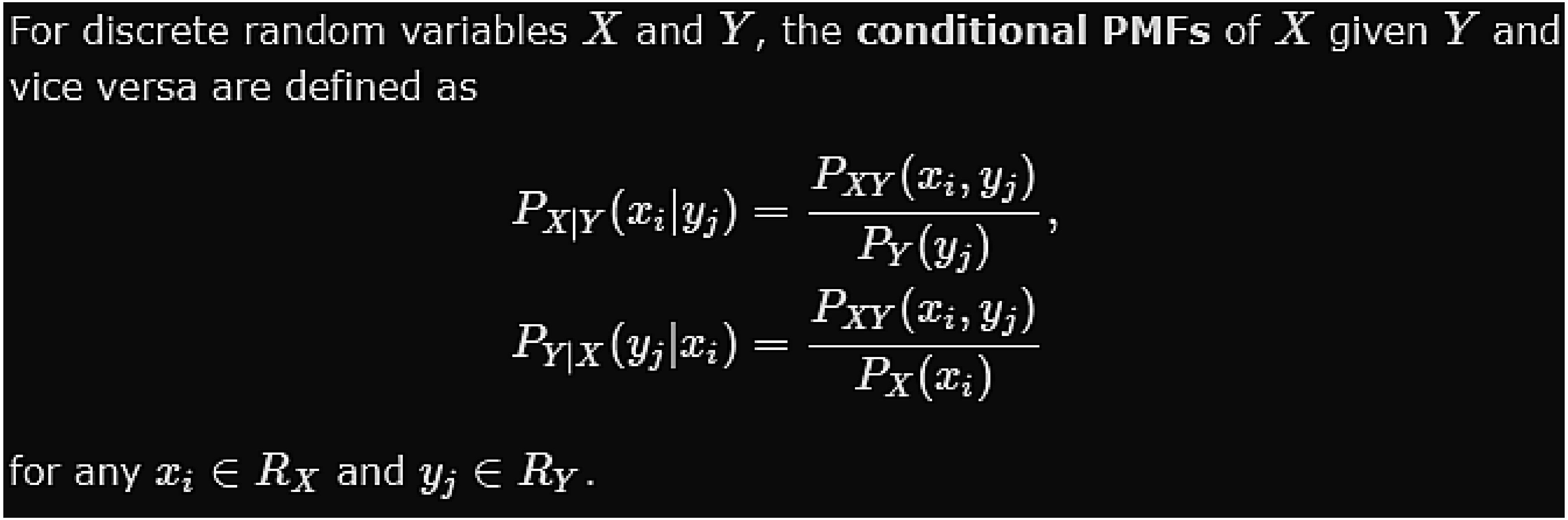


Two discrete random variables X and Y are independent if

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

Equivalently, X and Y are independent if

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

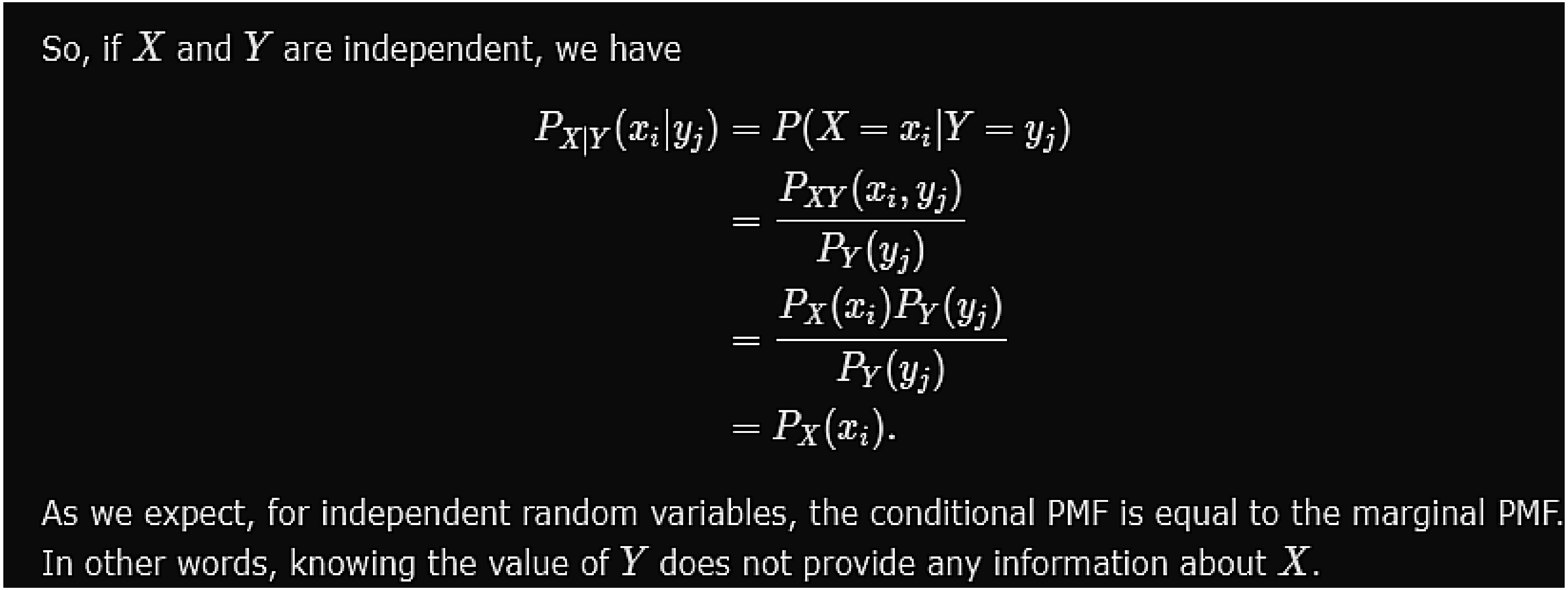


For discrete random variables X and Y, the **conditional PMFs** of X given Y and vice versa are defined as

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

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

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



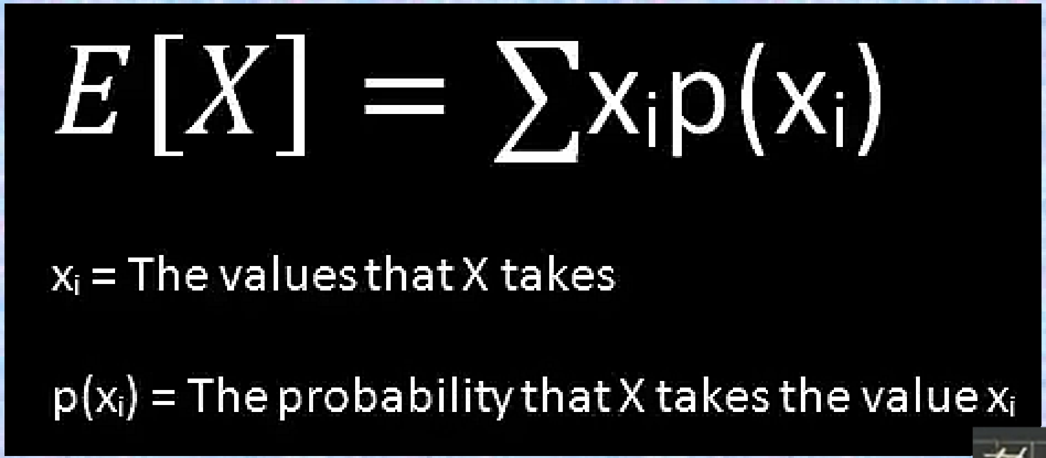
So, if X and Y are independent, we have

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

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

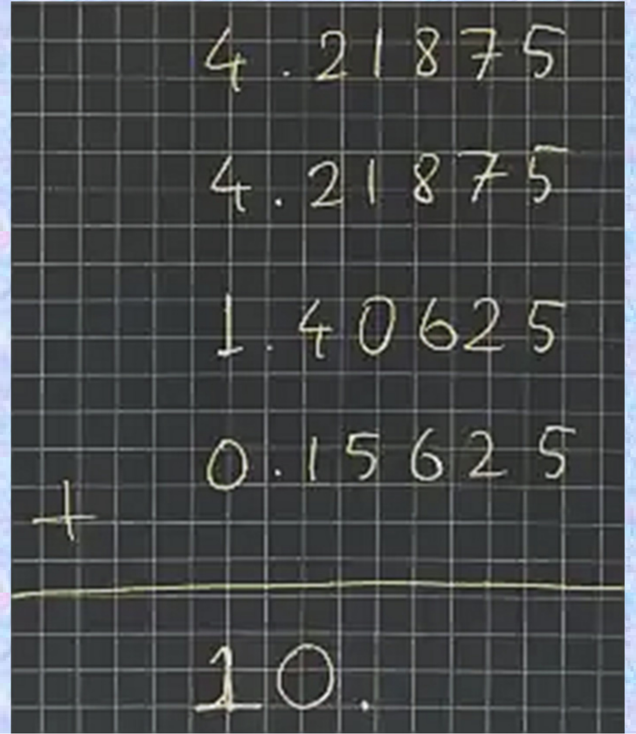
The expected value of a random variable is the weighted average of all possible values of the variable. The weight here means the probability of the random variable taking a specific value.



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

x_i = The values that X takes

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



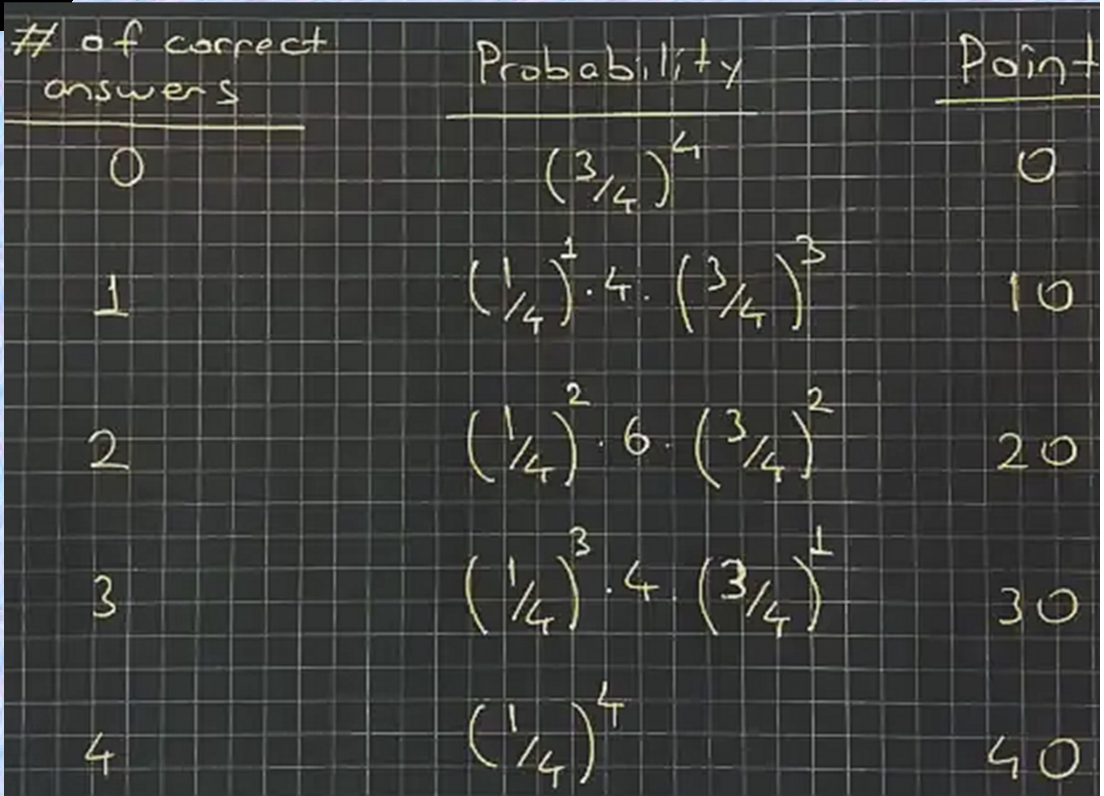
4.21875

4.21875

1.40625

+ 0.15625

10.



# 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

$$E[x] = \int_{5}^{10} x \, 0.2 \, dx = 0.2 \left. \frac{x^2}{2} \right|_{5}^{10}$$

$$= 0.1 \left. x^2 \right|_{5}^{10} = 0.1 (100 - 25)$$

Probability Density Function (PDF) of X

I his

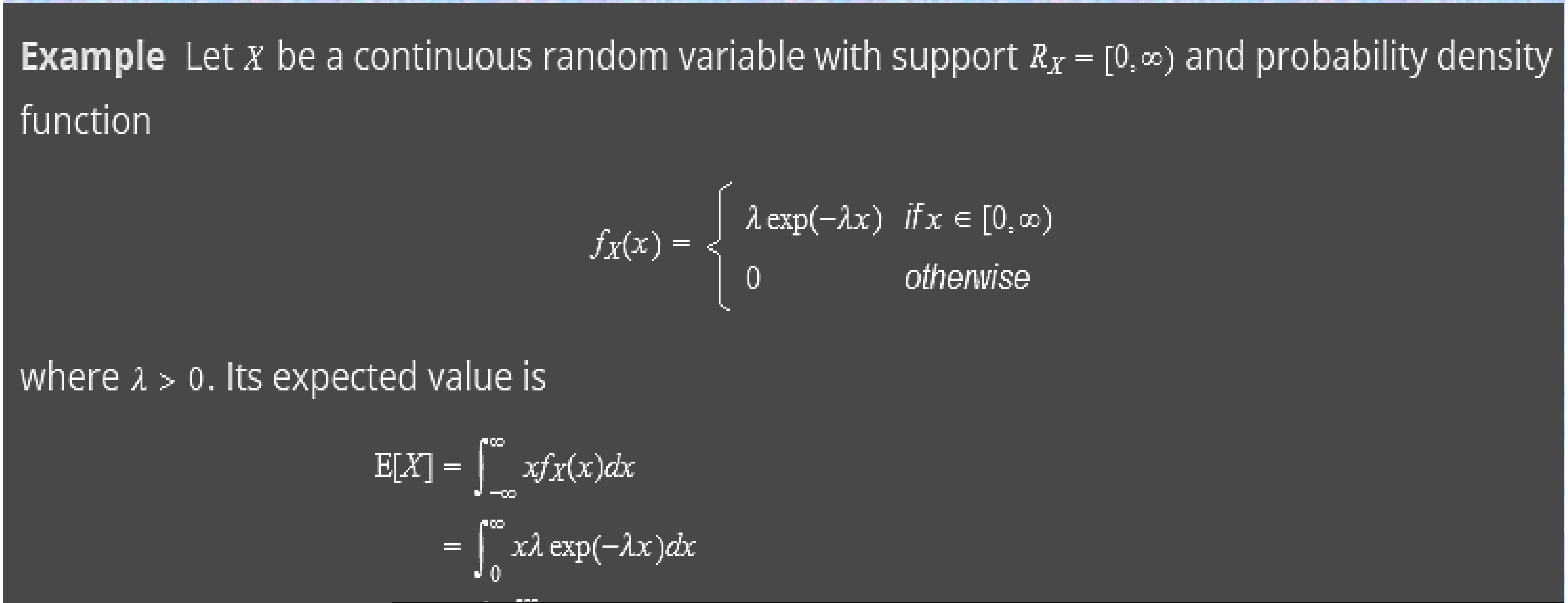
l

$$E[x] = \int_{0}^{8} x \, 0.03125 x \, dx$$

$$= 0.03125 \left. \frac{x^3}{3} \right|_{0}^{8}$$

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

$$= 5.33$$

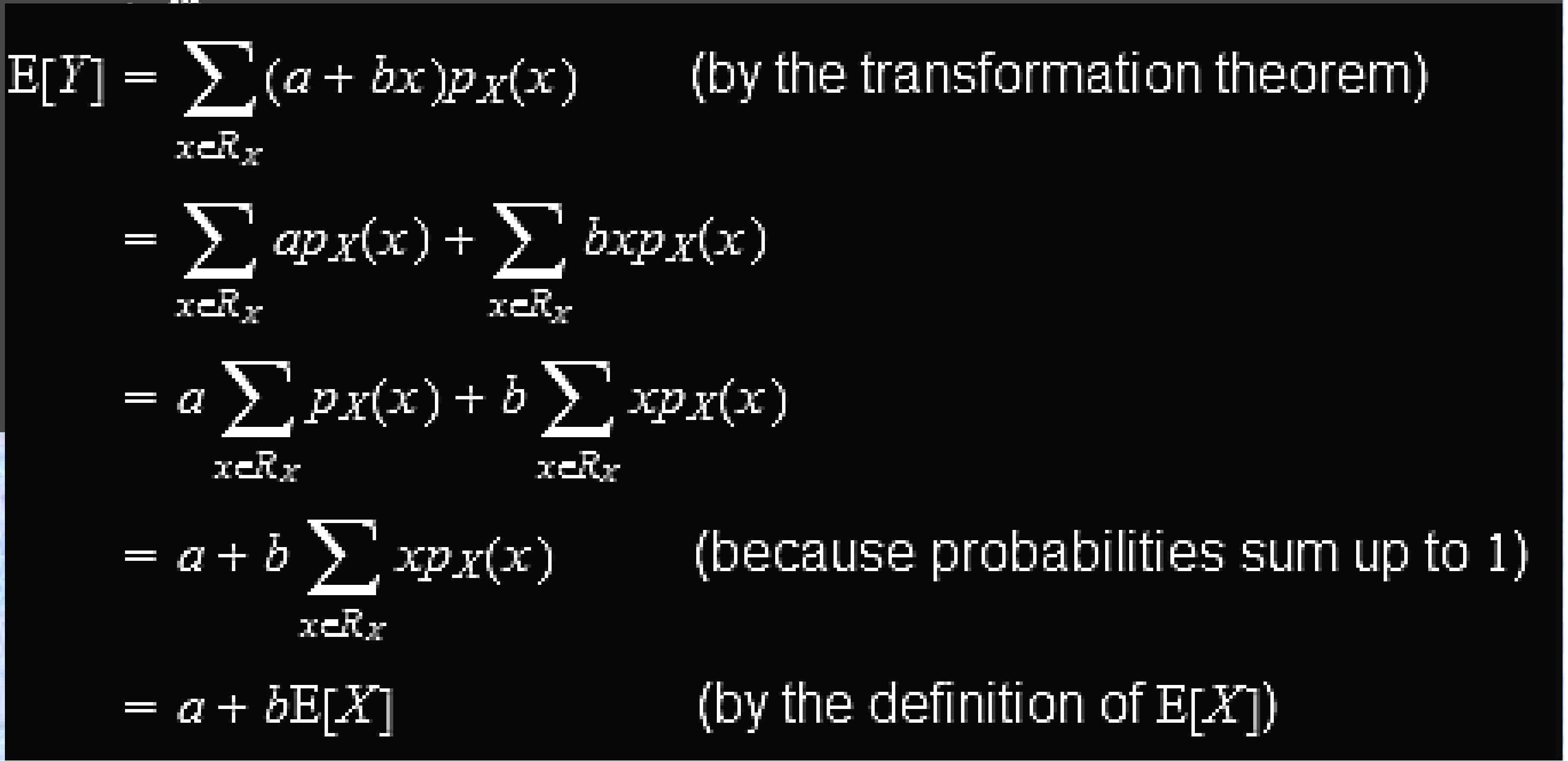


Example Let X be a continuous random variable with support $R_X = [0, \infty)$ and probability density function

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

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 continuous 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 discrete 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 independent random variables, and g, h be functions. Then

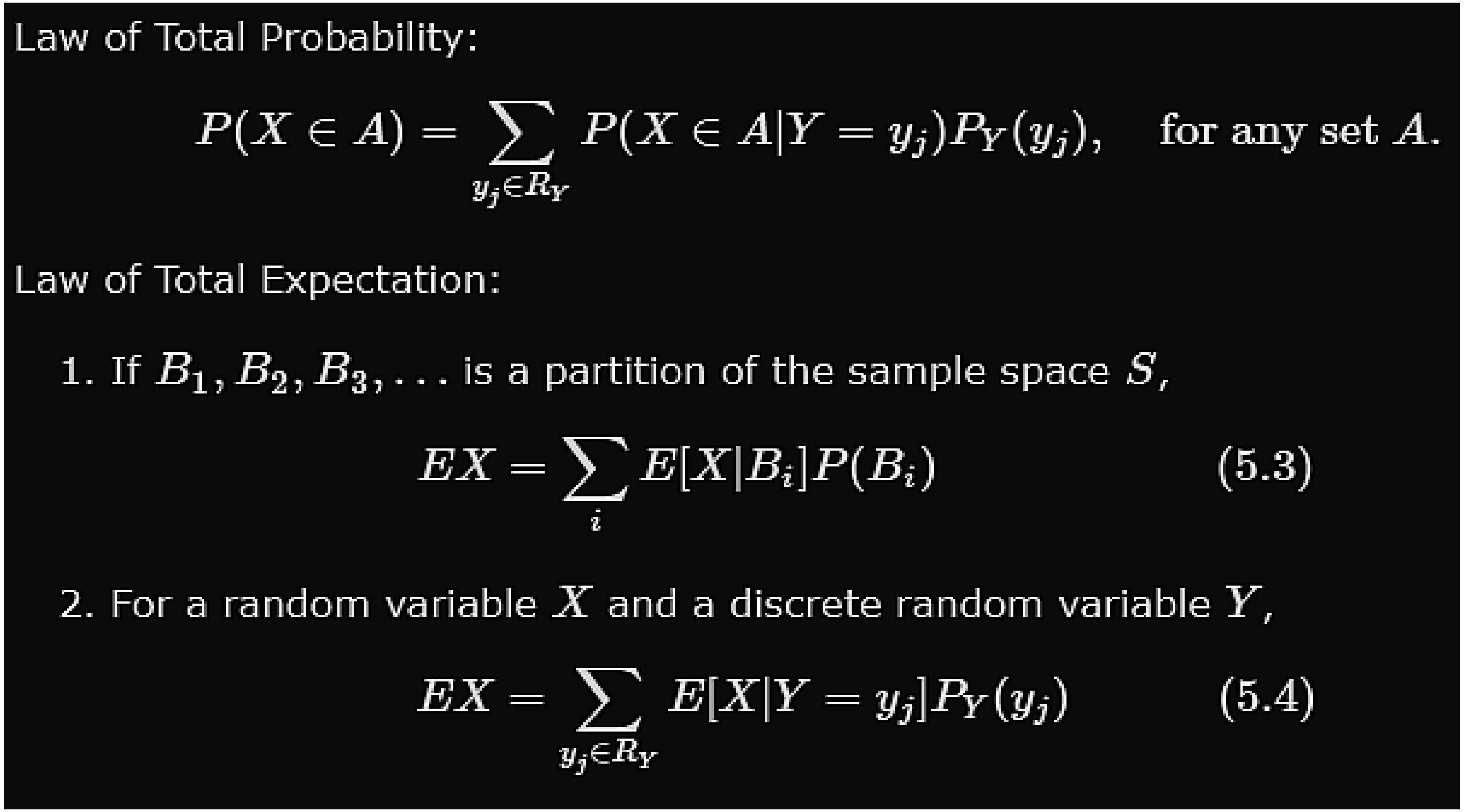
$$\begin{aligned}\mathbb{E}(XY) &= \mathbb{E}(X)\mathbb{E}(Y)\\ \mathbb{E}\Big(g(X)h(Y)\Big) &= \mathbb{E}\Big(g(X)\Big)\mathbb{E}\Big(h(Y)\Big).\end{aligned}$$

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*:

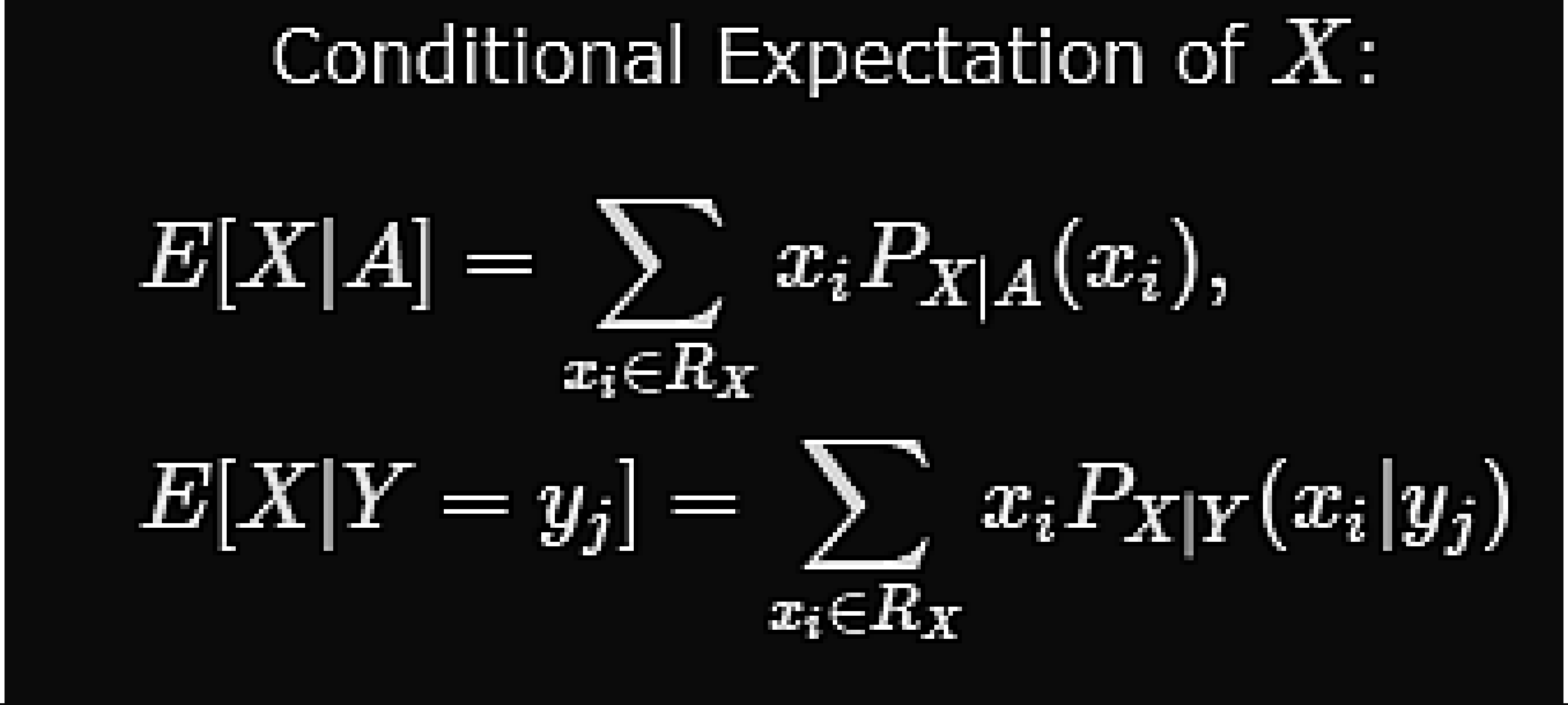
$$f(y|x) = \frac{f(x,y)}{f(x)}.$$

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*:

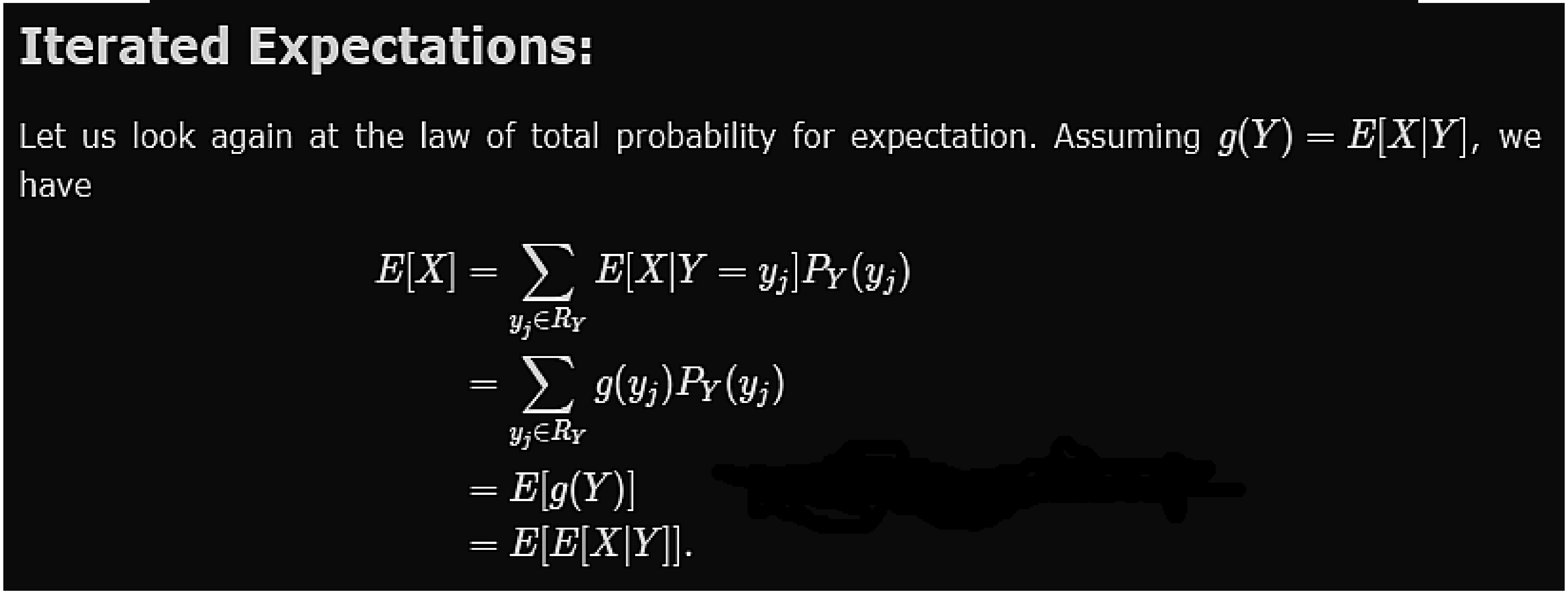
$$E[Y] = \begin{cases} \int_{-\infty}^{\infty} E[Y|X = x]f(x)dx & \text{if } X \text{ is continuous} \\ \sum_x E[Y|X = x]p(x) & \text{if } X \text{ is discrete.} \end{cases} \tag{39}$$



Conditional Expectation of X:

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

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

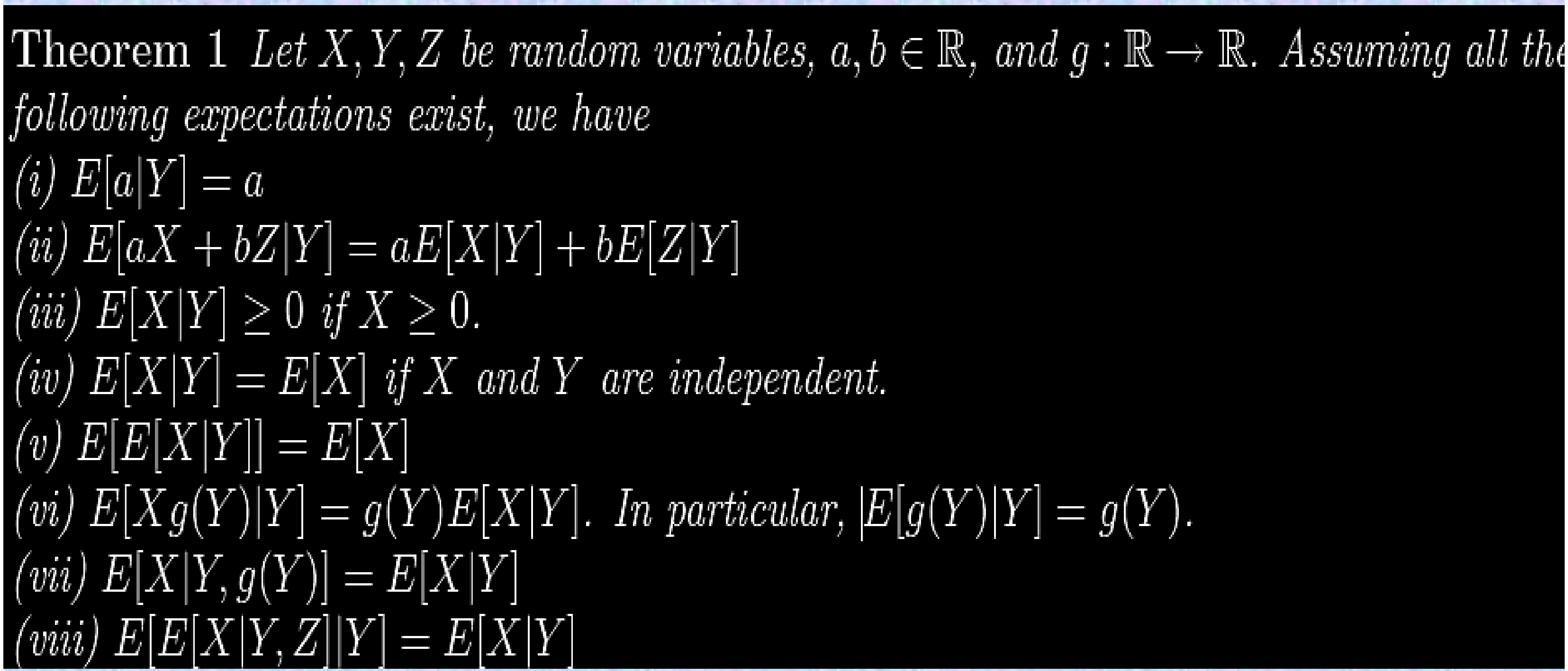


Iterated Expectations:

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

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

[redacted]



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:

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

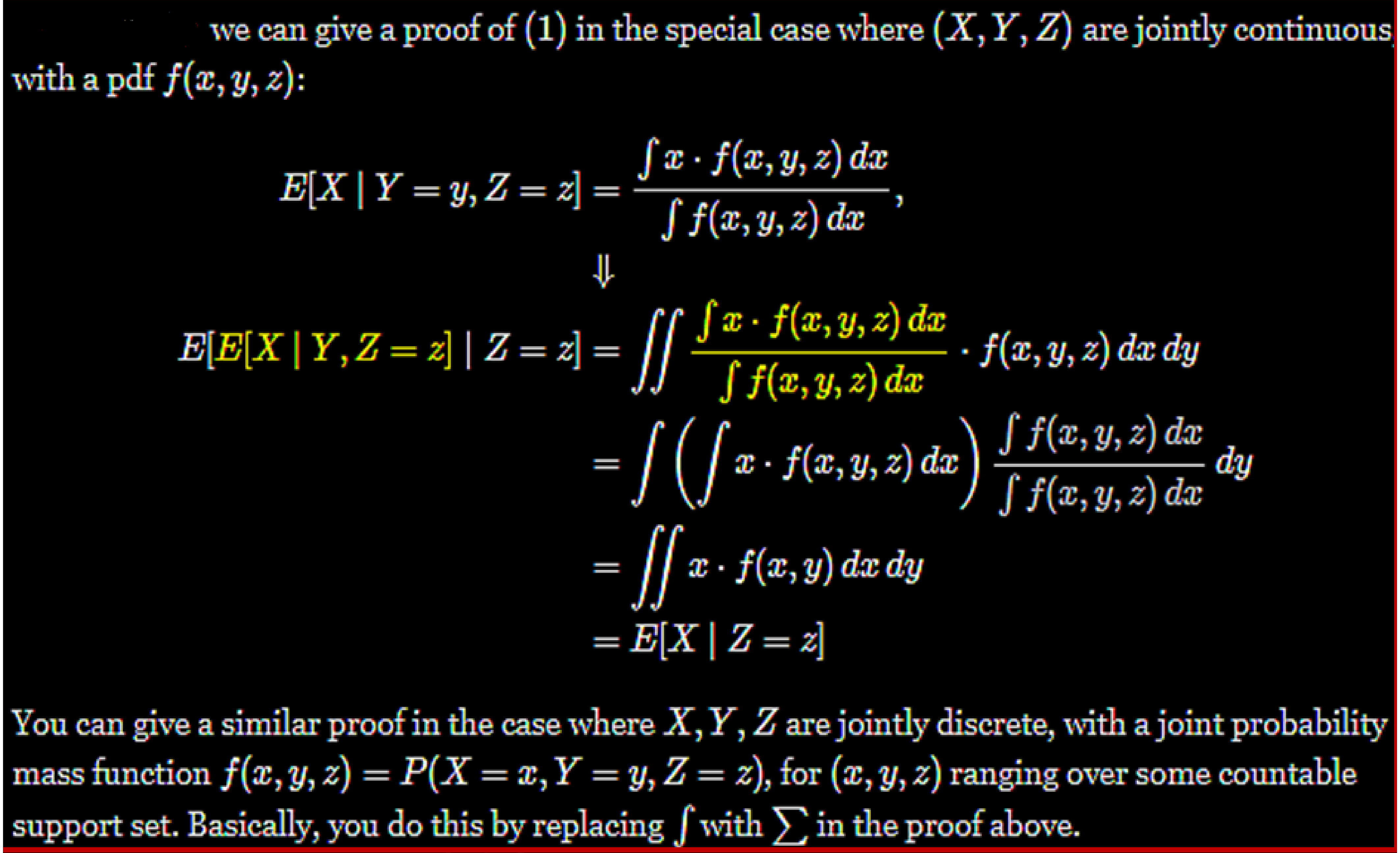
Laws of Total Expectation and Variance

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 any 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.*



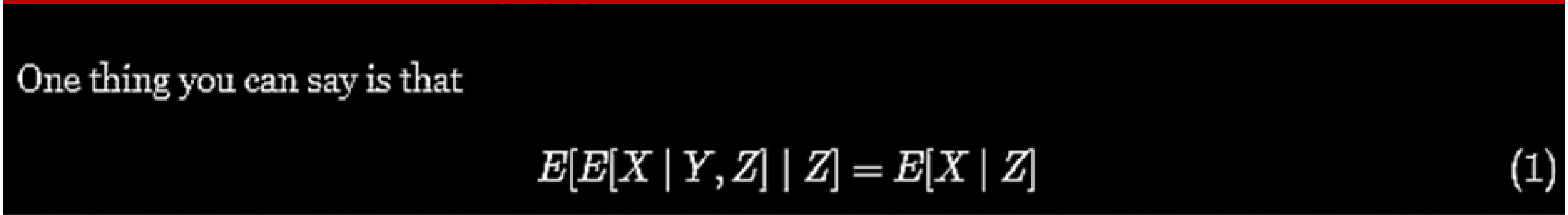
we can give a proof of (1) in the special case where (X, Y, Z) are jointly continuous with a pdf $f(x, y, z)$:

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

$$\Downarrow$$

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

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:

$$\begin{aligned}\text{RHS} &= \mathbb{E}_Y\Big[\mathbb{E}(g(X)\,|\,Y)\Big] = \mathbb{E}_Y\left[\sum_x g(x)\mathbb{P}(X=x\,|\,Y)\right]\\ &= \sum_y\left[\sum_x g(x)\mathbb{P}(X=x\,|\,Y=y)\right]\mathbb{P}(Y=y)\end{aligned}$$

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

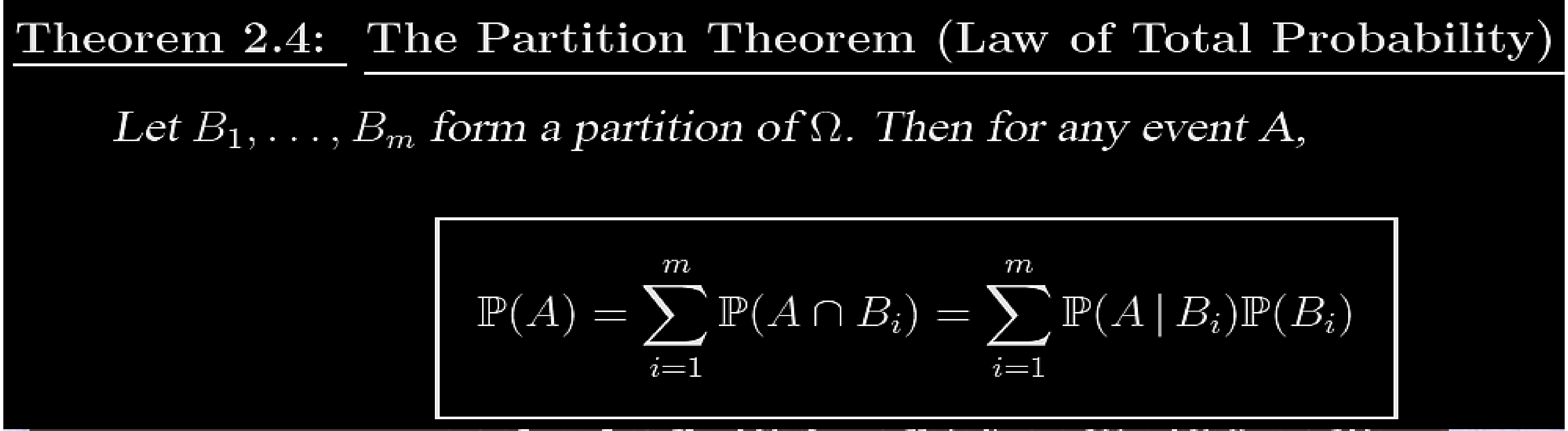
$$\mathbb{E}_Y[\text{Var}(X\,|\,Y)] + \text{Var}_Y[\mathbb{E}(X\,|\,Y)]$$

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

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

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

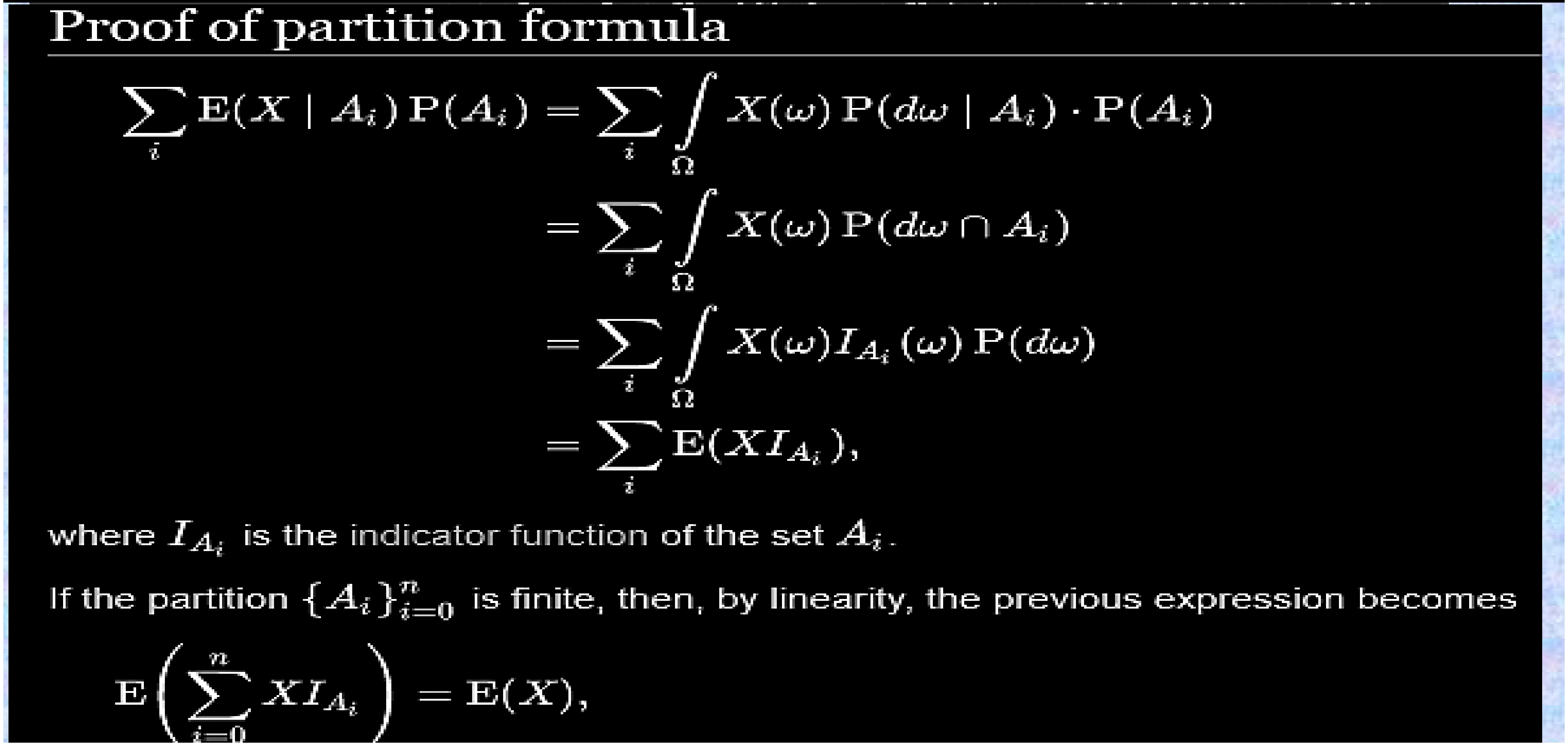
$$= \text{Var}(X) = \text{LHS}. \qquad \square$$



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

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

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



Proof of partition formula

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

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

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

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,

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

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

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

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

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:

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

μ - Mean; σ - S.D.; ρ_{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 kth 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.

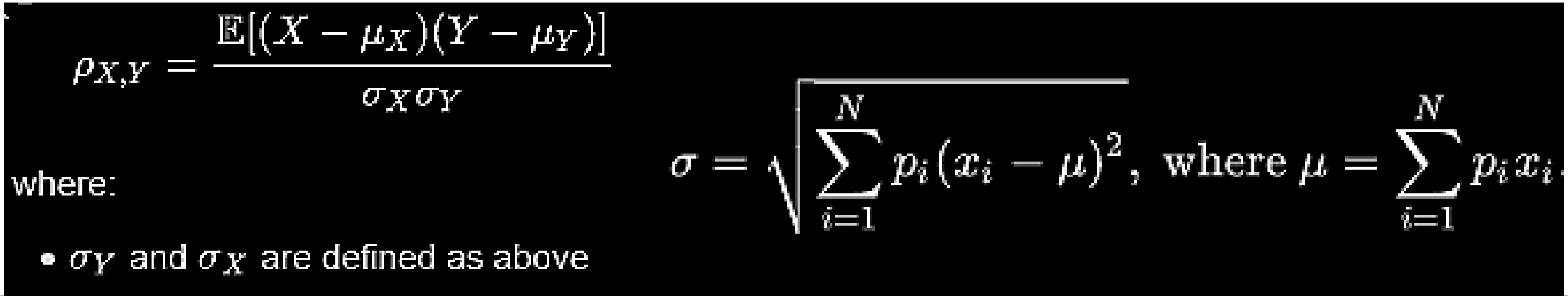
Correlation between x and y: $\rho_{xy} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y} = E[(\dfrac{x-\mu_x}{\sigma_x})(\dfrac{y-\mu_y}{\sigma_y})]$

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

For Z = ax + by ;

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

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

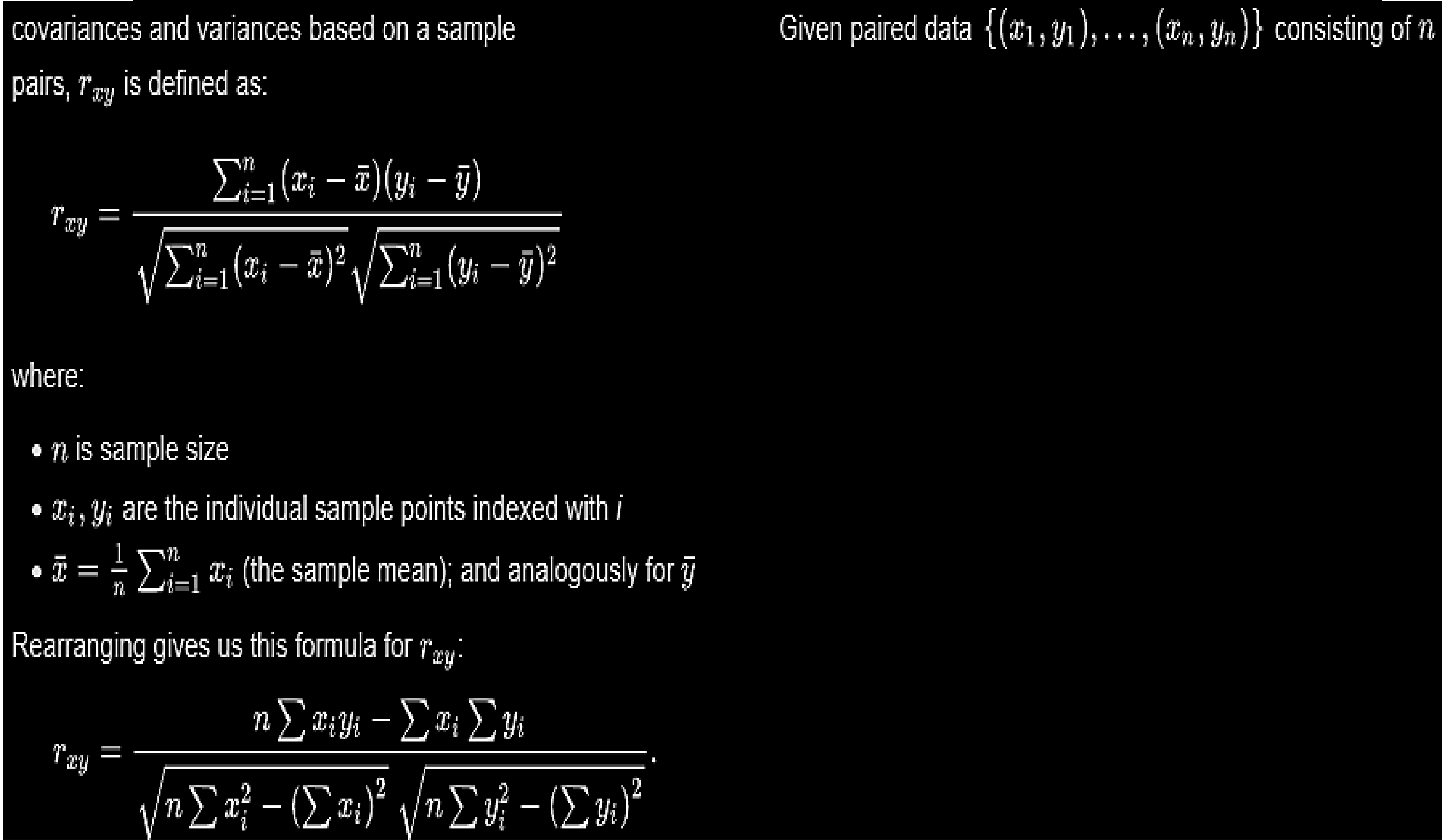


$$\rho_{X,Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y}$$

where:

- σ_Y and σ_X are defined as above

$$\sigma = \sqrt{\sum_{i=1}^{N} p_i(x_i-\mu)^2}, \text{ where } \mu = \sum_{i=1}^{N} p_i x_i$$



covariances and variances based on a sample

Given paired data $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ consisting of n pairs, r_{xy} is defined as:

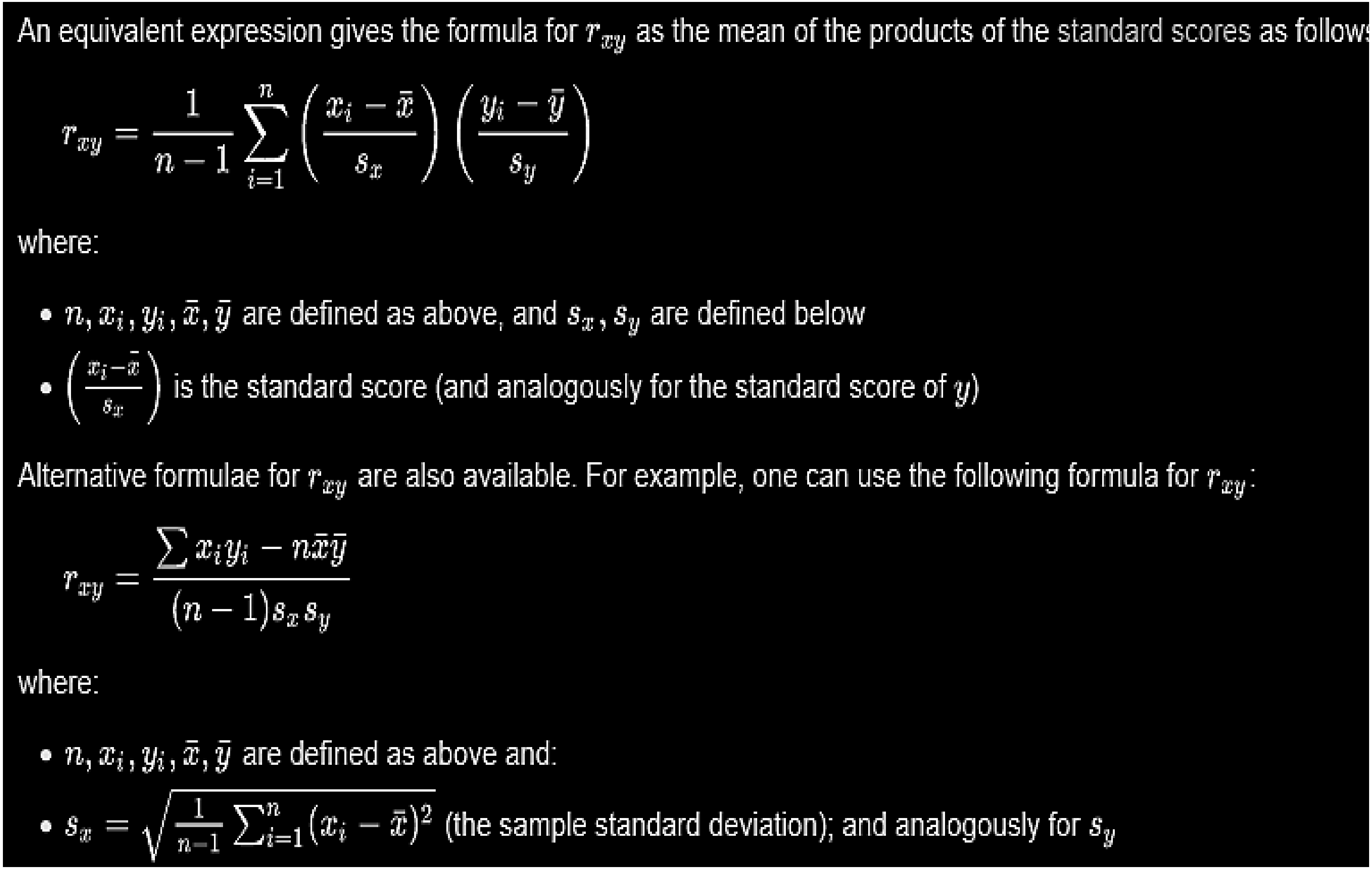
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

where:

- n is sample size
- x_i, y_i are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

Rearranging gives us this formula for r_{xy}:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\ \sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}.$$



An equivalent expression gives the formula for r_{xy} as the mean of the products of the standard scores as follow

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

where:

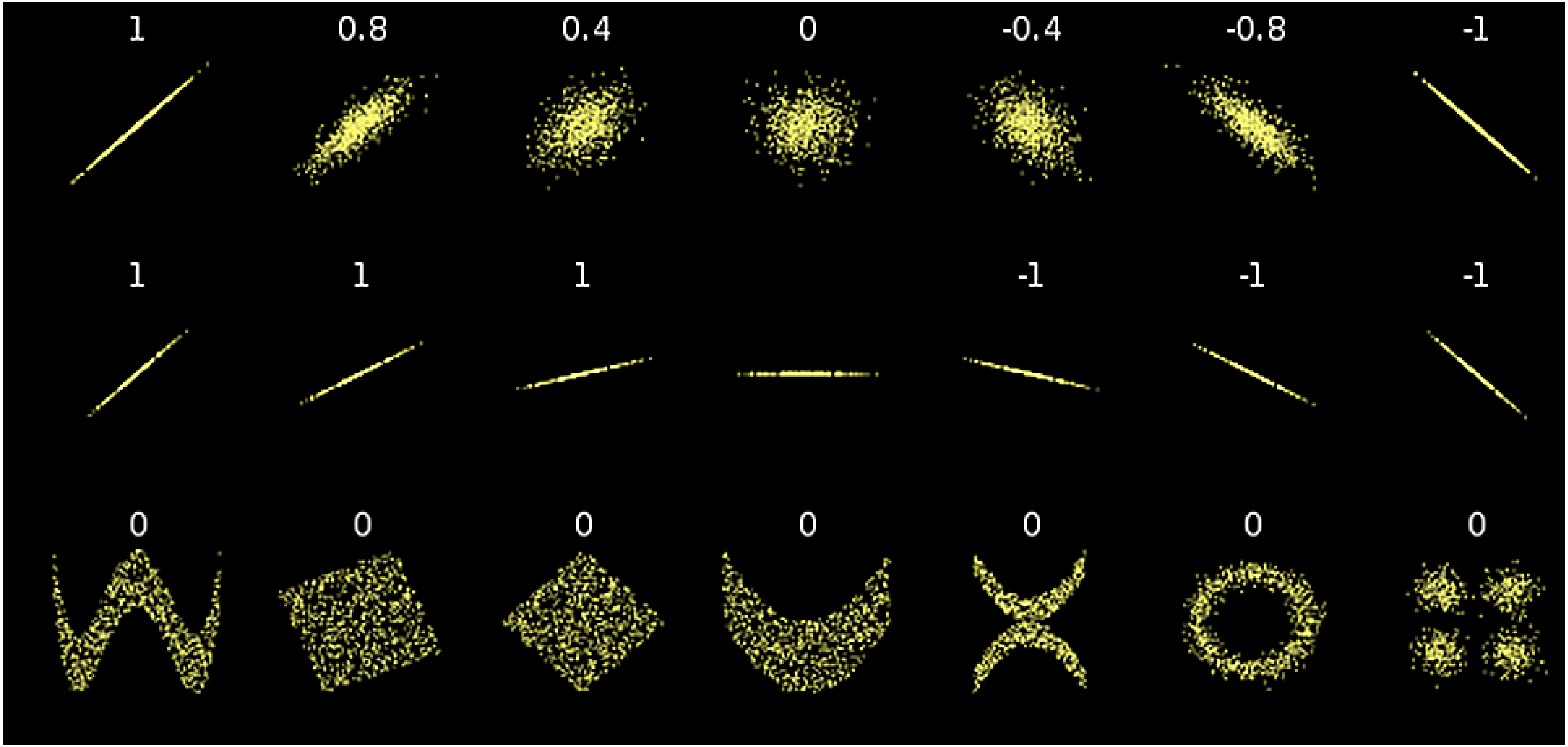
- $n, x_i, y_i, \bar{x}, \bar{y}$ are defined as above, and s_x, s_y are defined below
- $\left(\frac{x_i - \bar{x}}{s_x} \right)$ is the standard score (and analogously for the standard score of y)

Alternative formulae for r_{xy} are also available. For example, one can use the following formula for r_{xy}:

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1) s_x s_y}$$

where:

- $n, x_i, y_i, \bar{x}, \bar{y}$ are defined as above and:
- $s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$ (the sample standard deviation); and analogously for s_y

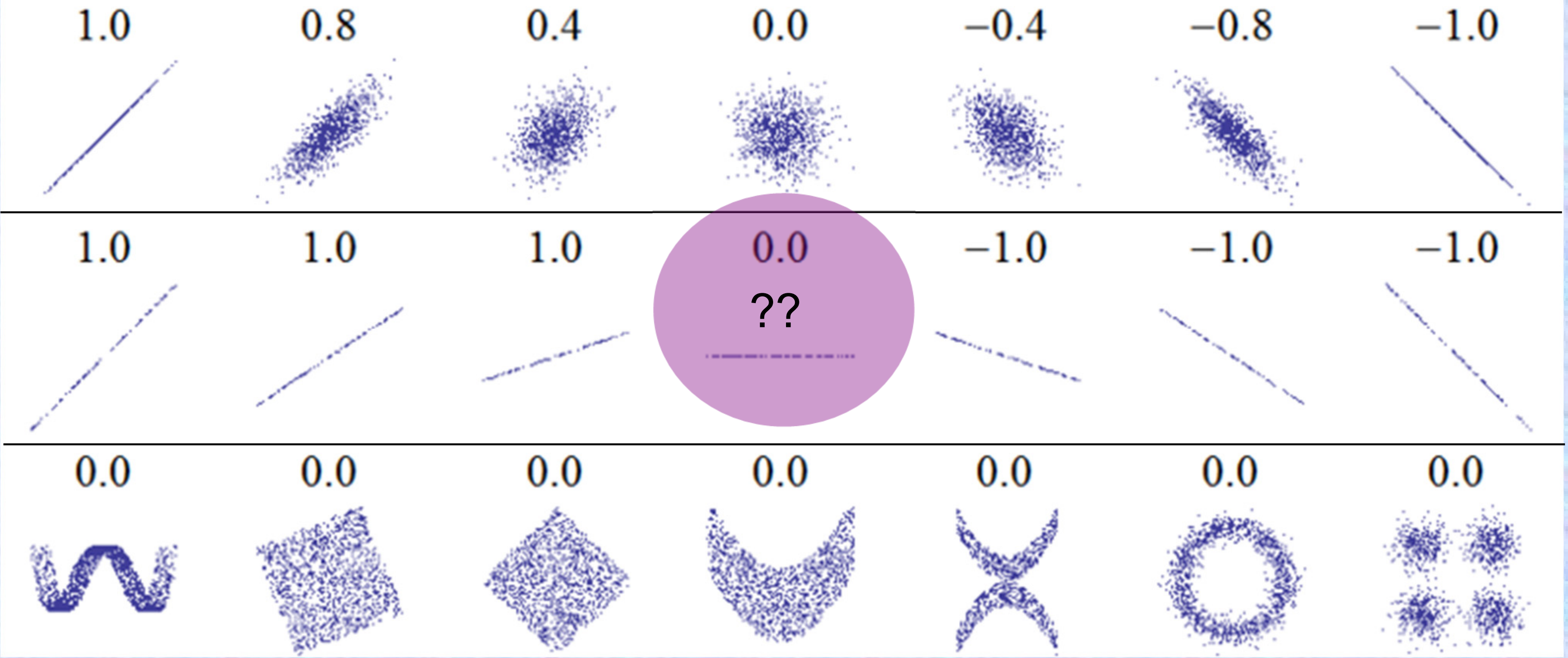


Several sets of (x, y) points, with the correlation coefficient of x and y for each set.

The correlation reflects the strength and direction of a linear relationship (top row),

but not the slope of that relationship (middle),

nor many aspects of nonlinear relationships (bottom).



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

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

The correlation coefficient can also be viewed as the cosine of the angle between the two vectors ($\Re^D$) of samples drawn from the two random variables - i.e. between the two observed vectors in N-dimensional space (for N observations of each variable) - http://www.hawaii.edu/powerkills/UC.HTM

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.

One defines also the correlation

$$\mathrm{Corr}[XY] = \frac{\mathrm{Cov}[XY]}{\sigma[X]\sigma[Y]} .$$

Here is a key connection between linear algebra and probability theory:

If X, Y are two random variables of zero mean, then the covariance $\mathrm{Cov}[XY] = \mathrm{E}[X \cdot Y]$ is the **dot product** of X and Y. The standard deviation of X is the length

f the angle between the two vectors.

Positive

egative correlation means an obtuse angle.

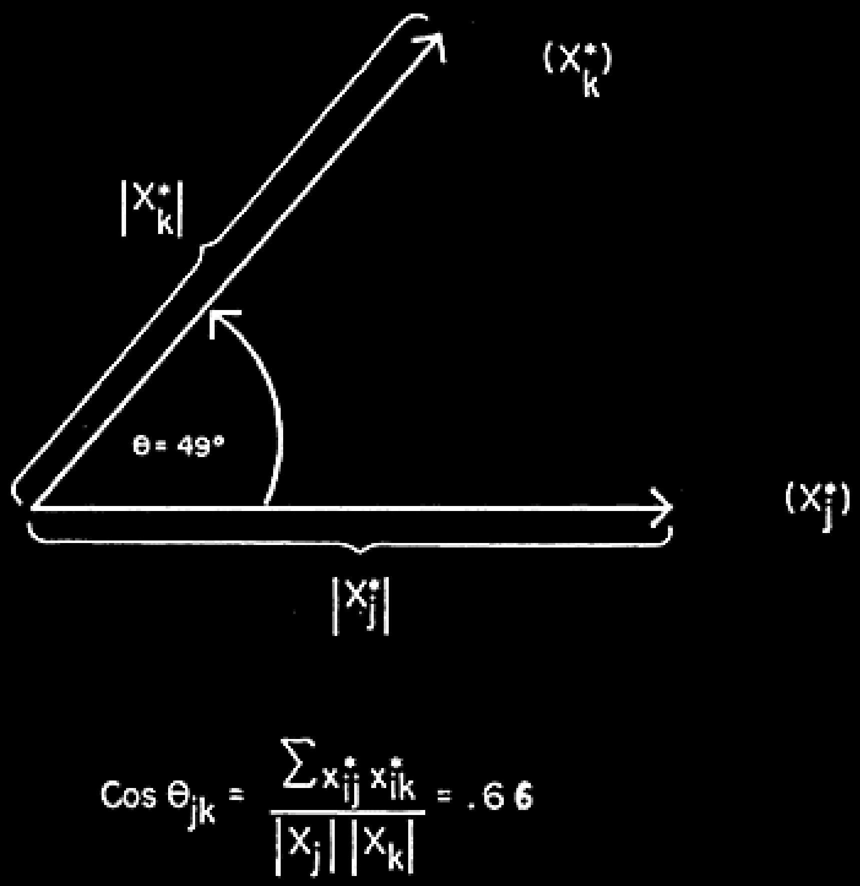
If correlation c

geometric significance of independence?

Two ra

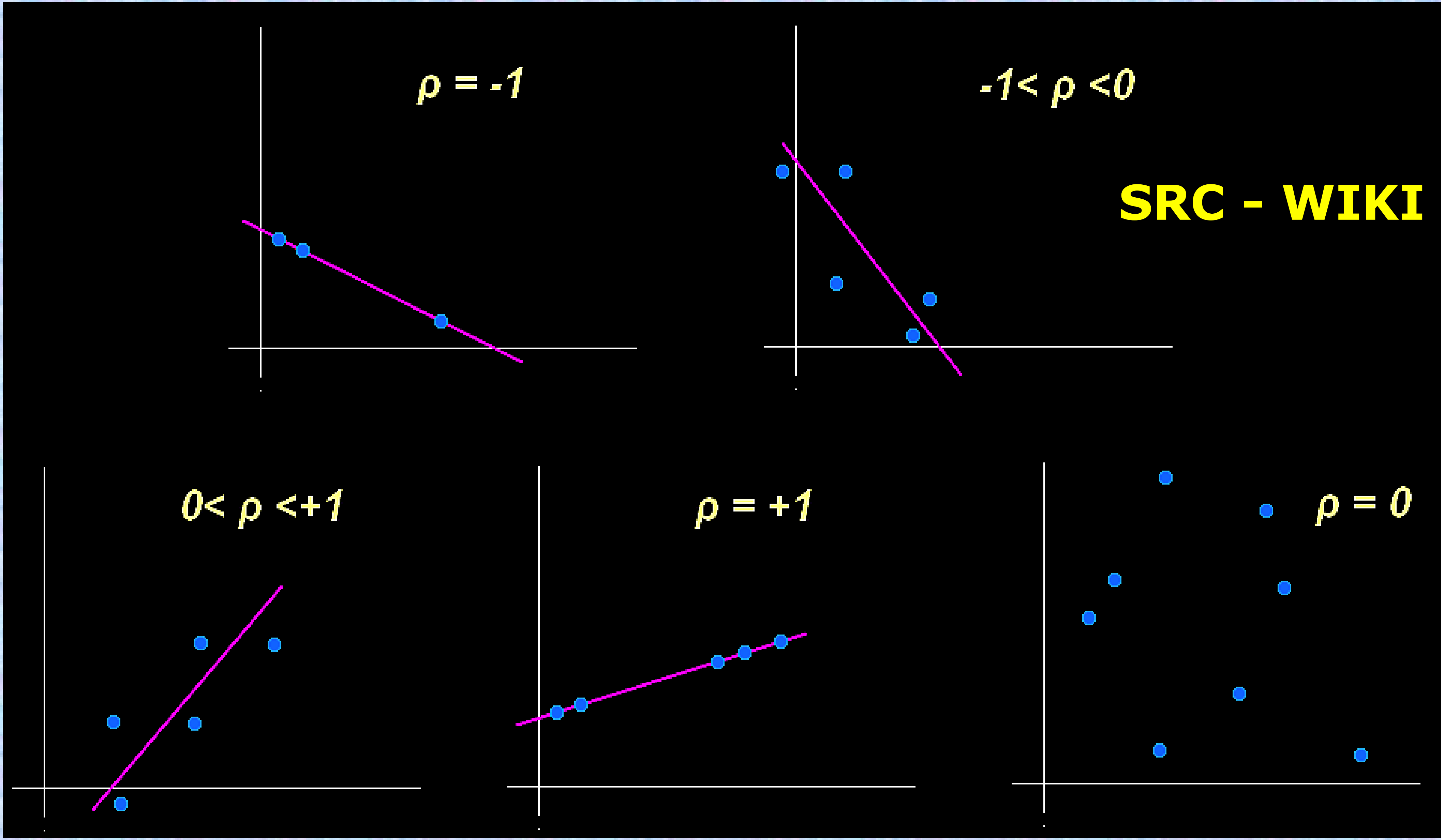
f and only if for any functions f, g the random

ated.



https://people

ath19b_2011/handouts/lecture12.pdf



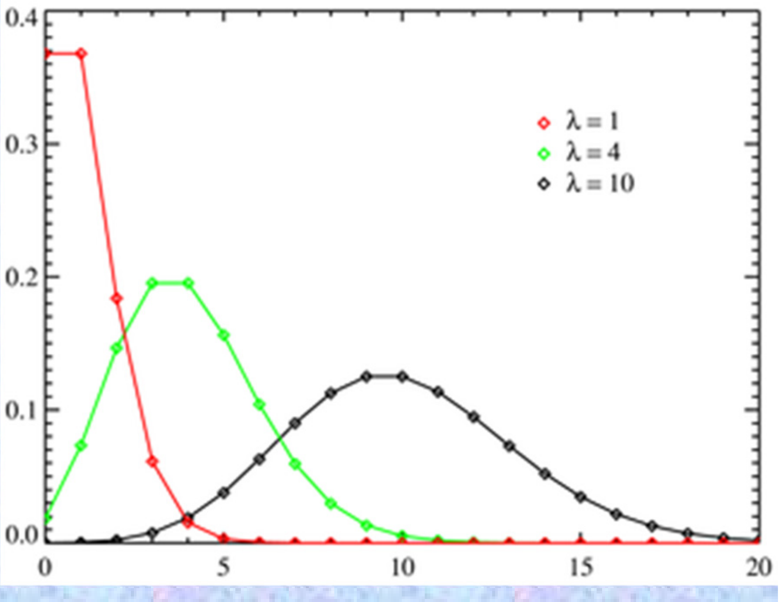
SRC - WIKI

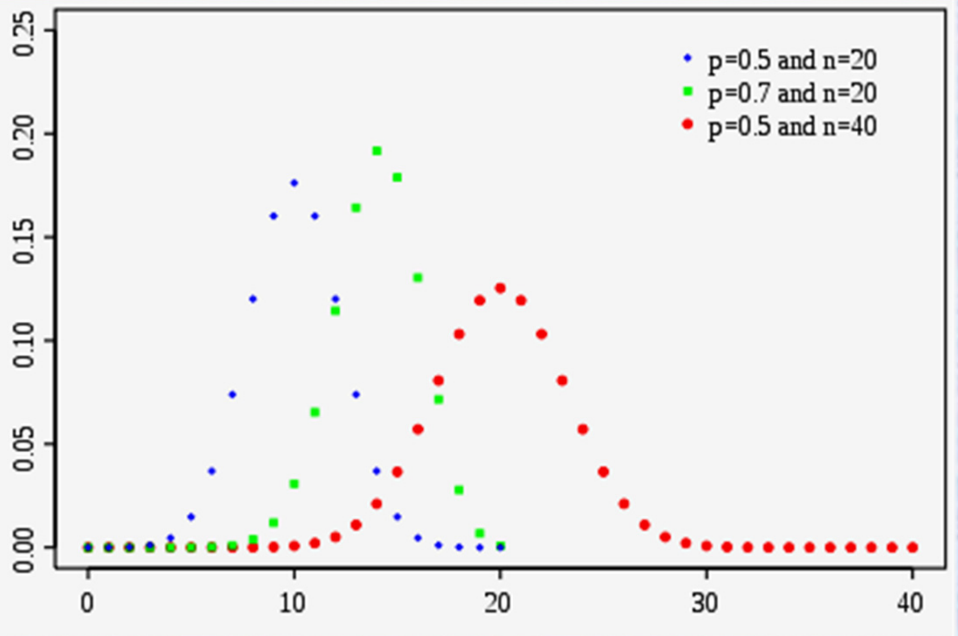
$$\rho_{X,Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \qquad \rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Other PDFs:

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

Poisson →





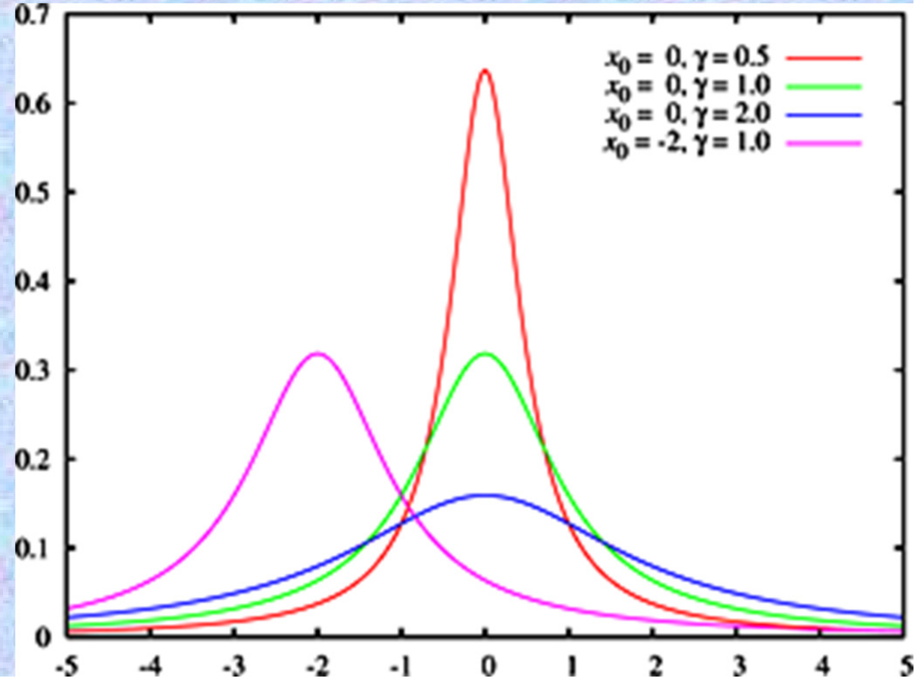
$$\Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

← Binomial

Cauchy

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

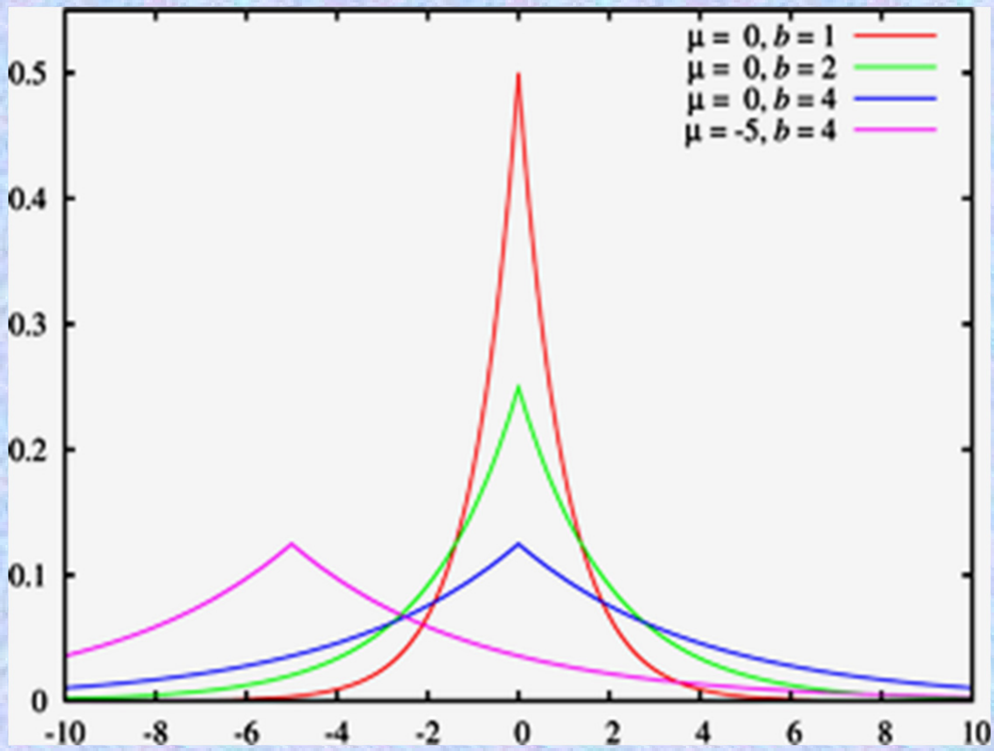
$$= \frac{1}{\pi} \left[\frac{\gamma}{(x - x_0)^2 + \gamma^2}\right]$$



LAPLACE:

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

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

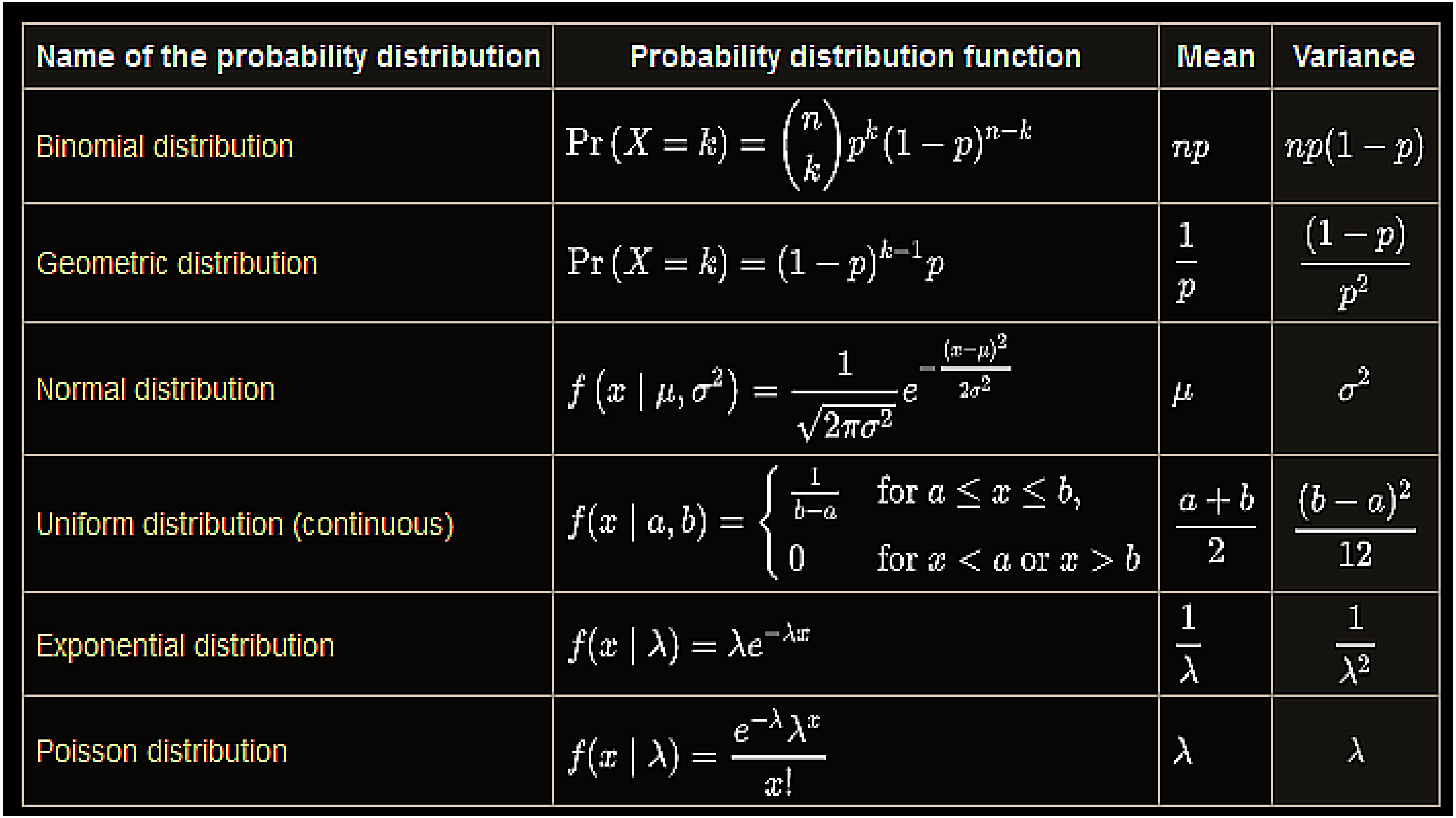


Double Exponential Density:

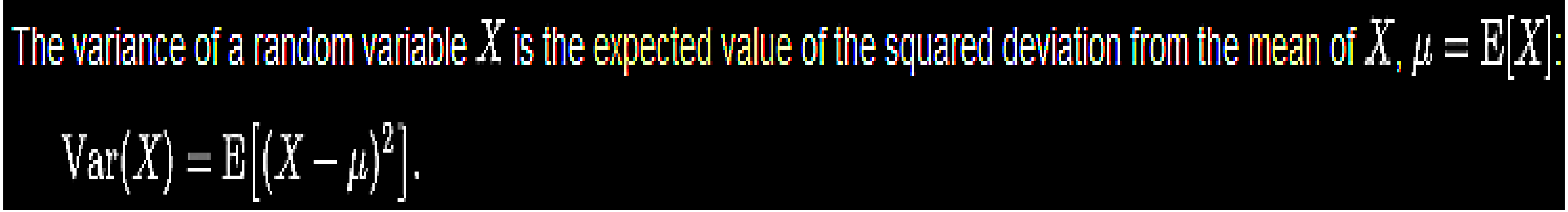
$$P(x)=\frac{1}{2b}e^{-\left|x-a/b\right|};$$

Read about:

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

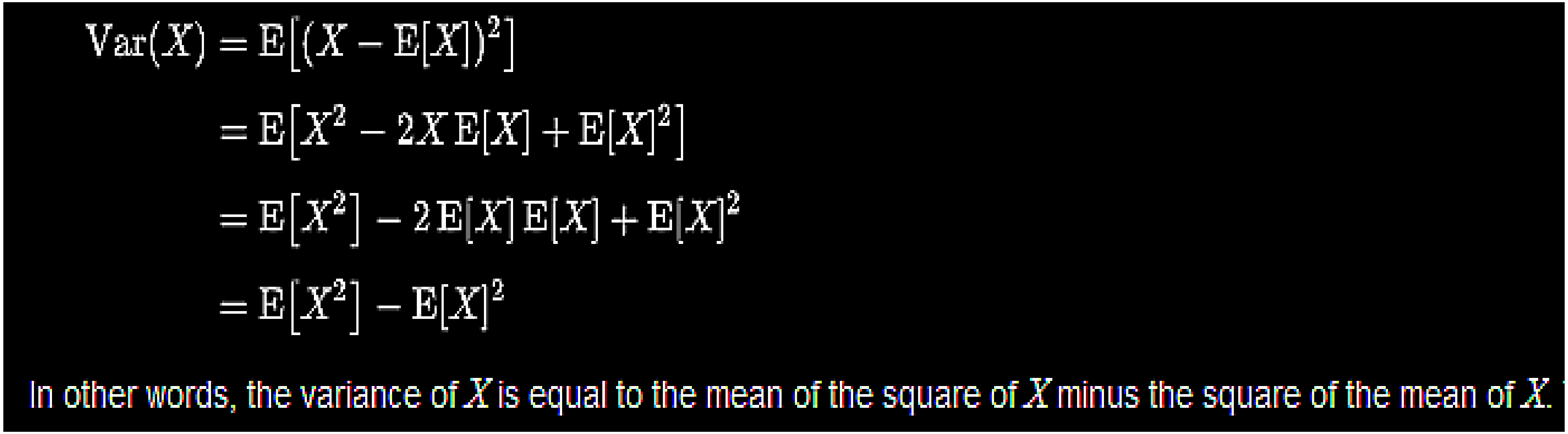


Name of the probability distribution	Probability distribution function	Mean	Variance
Binomial distribution	$\Pr(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
Geometric distribution	$\Pr(X=k)=(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{(1-p)}{p^2}$
Normal distribution	$f(x\mid\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^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	$f(x\mid\lambda)=\frac{e^{-\lambda}\lambda^x}{x!}$	λ	λ



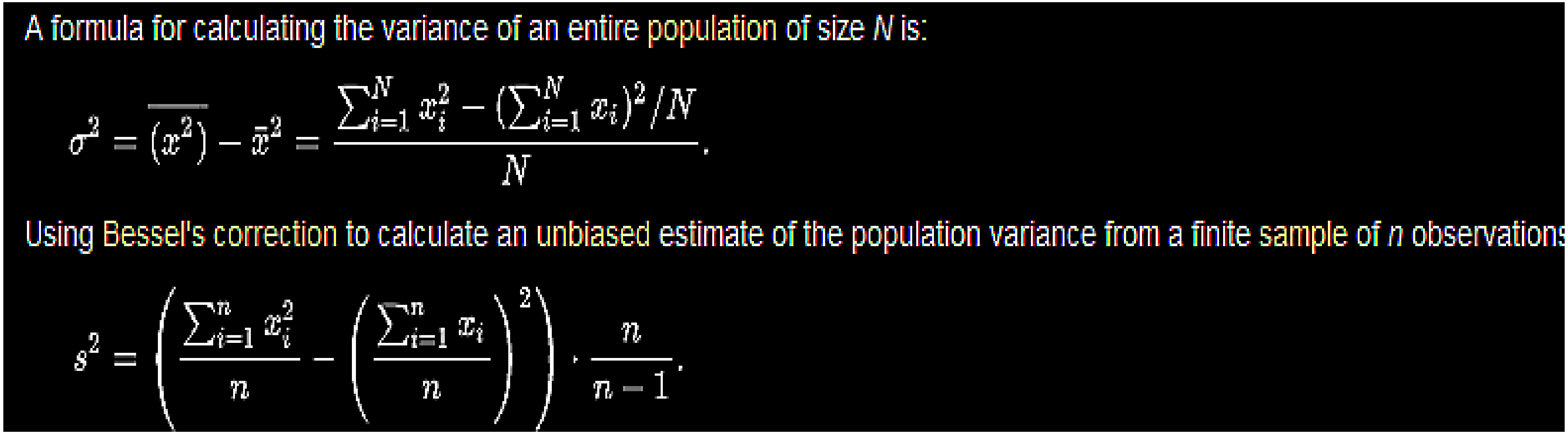
The variance of a random variable X is the expected value of the squared deviation from the mean of X, $\mu = \mathrm{E}[X]$:

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



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

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.

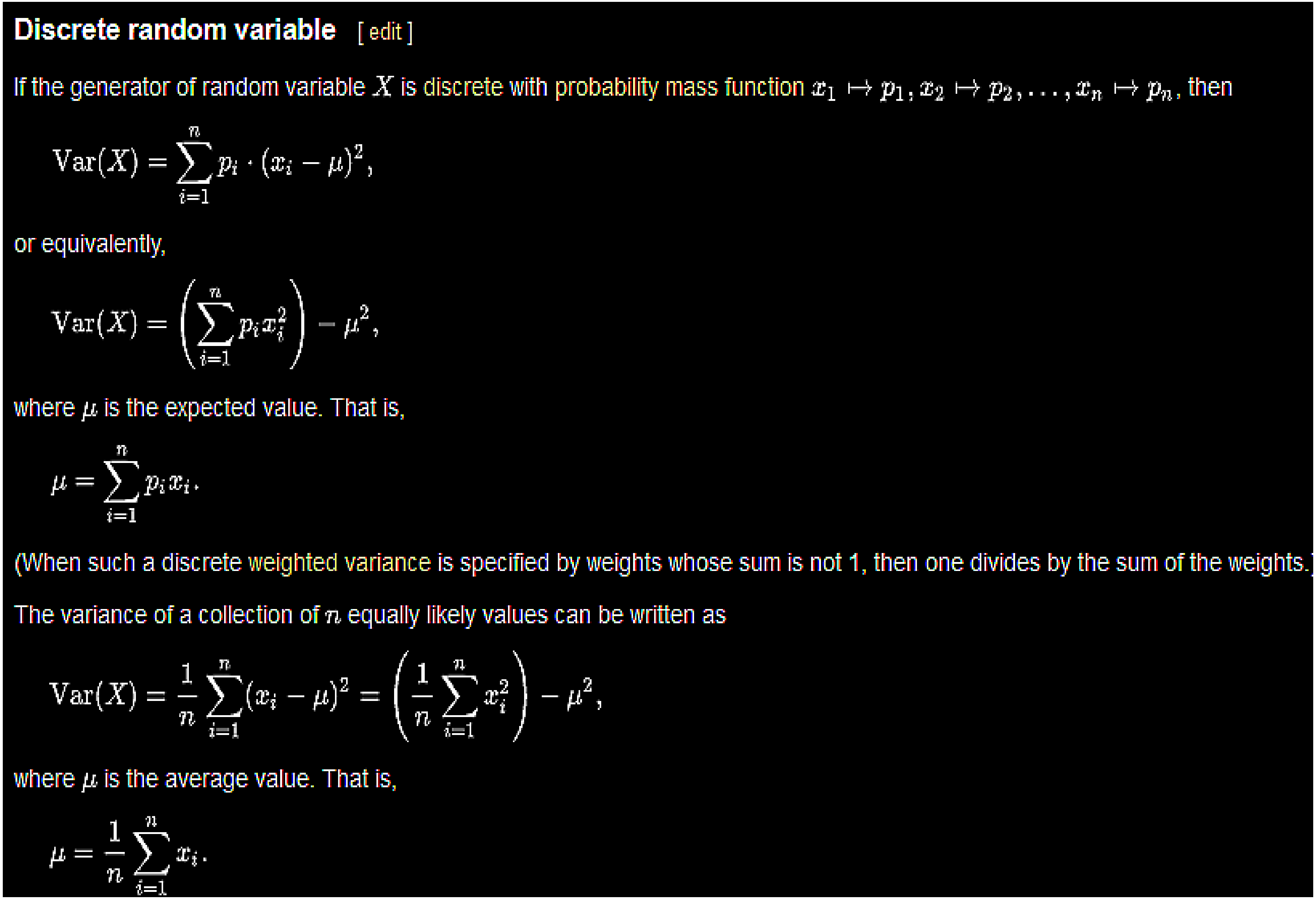


A formula for calculating the variance of an entire population of size N is:

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

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 μ is the expected value. That is,

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

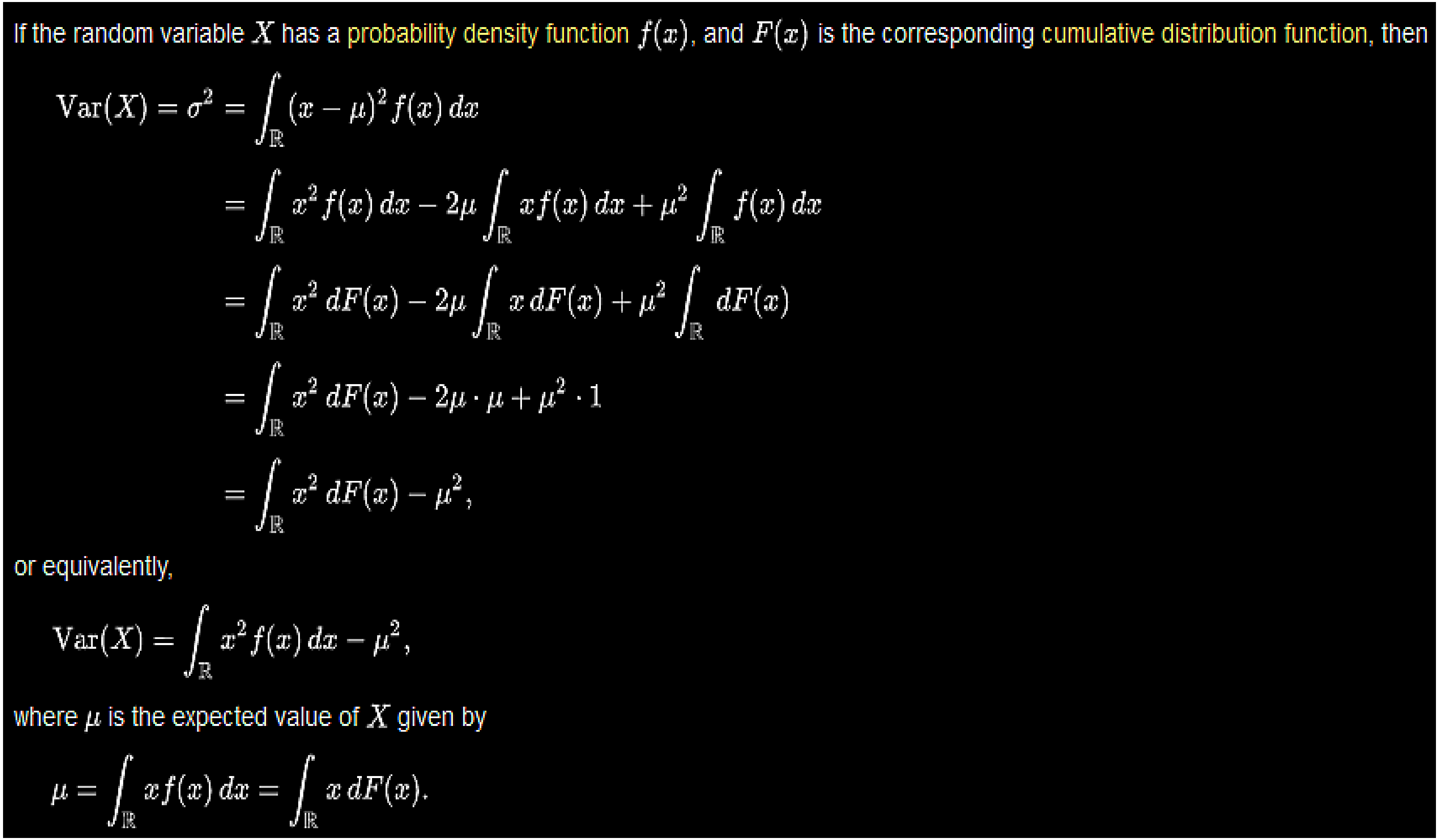
(When such a discrete weighted variance is specified by weights whose sum is not 1, then one divides by the sum of the weights.)

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 μ 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 μ is the expected value of X given by

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

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

om variable X is defined as

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

$$= E(X-\mu)^2 . = \sum_x (x-\mu$$

On simplification,

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

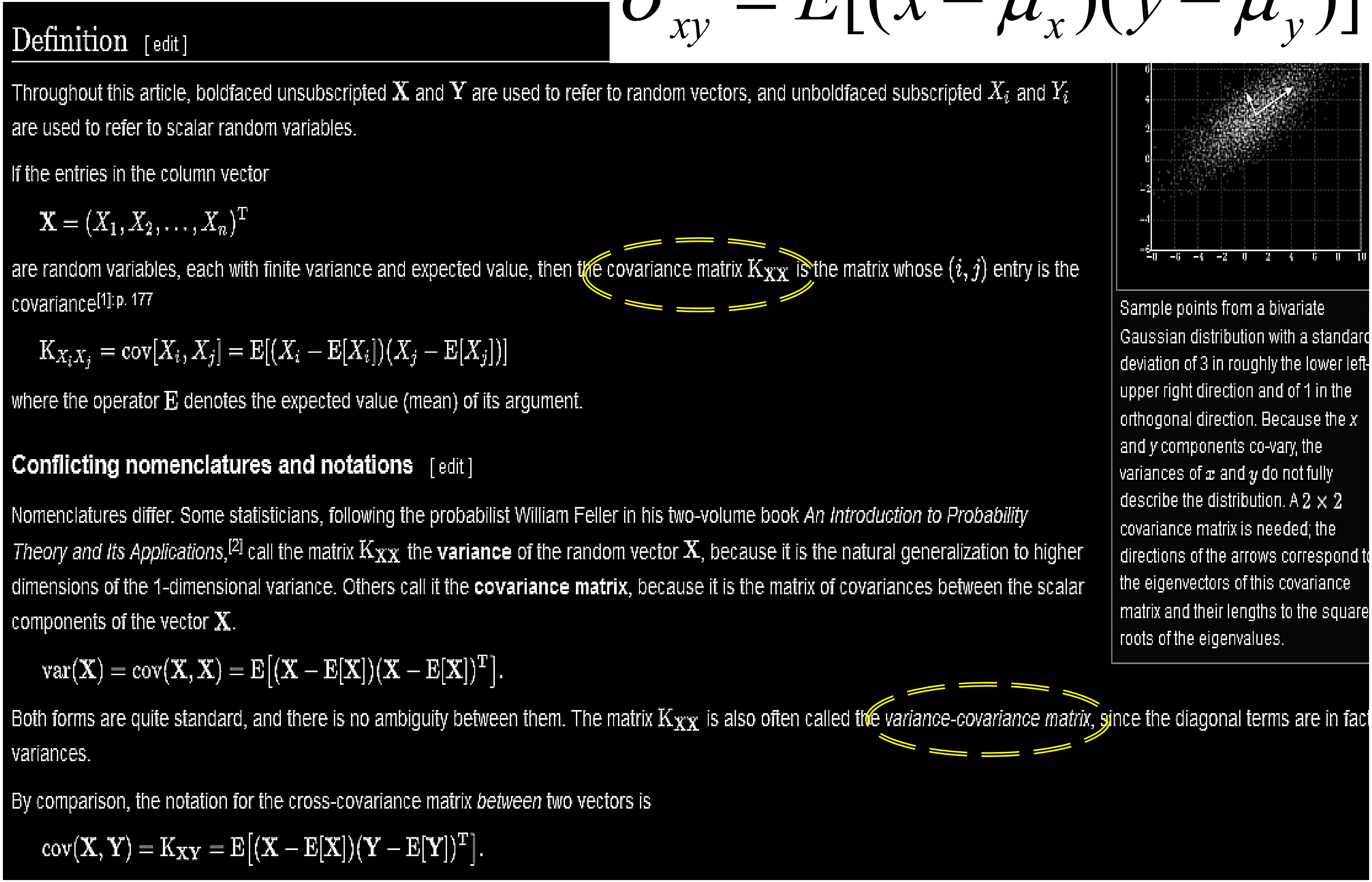
si

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

$$E^2(X) = E(X^2) - E^2(X).$$

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

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



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

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

where the operator E denotes the expected value (mean) of its argument.

Sample points from a bivariate Gaussian distribution with a standar 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 2×2 covariance matrix is needed; the directions of the arrows correspond t the eigenvectors of this covariance matrix and their lengths to the square roots of the eigenvalues.

Conflicting nomenclatures and notations [edit]

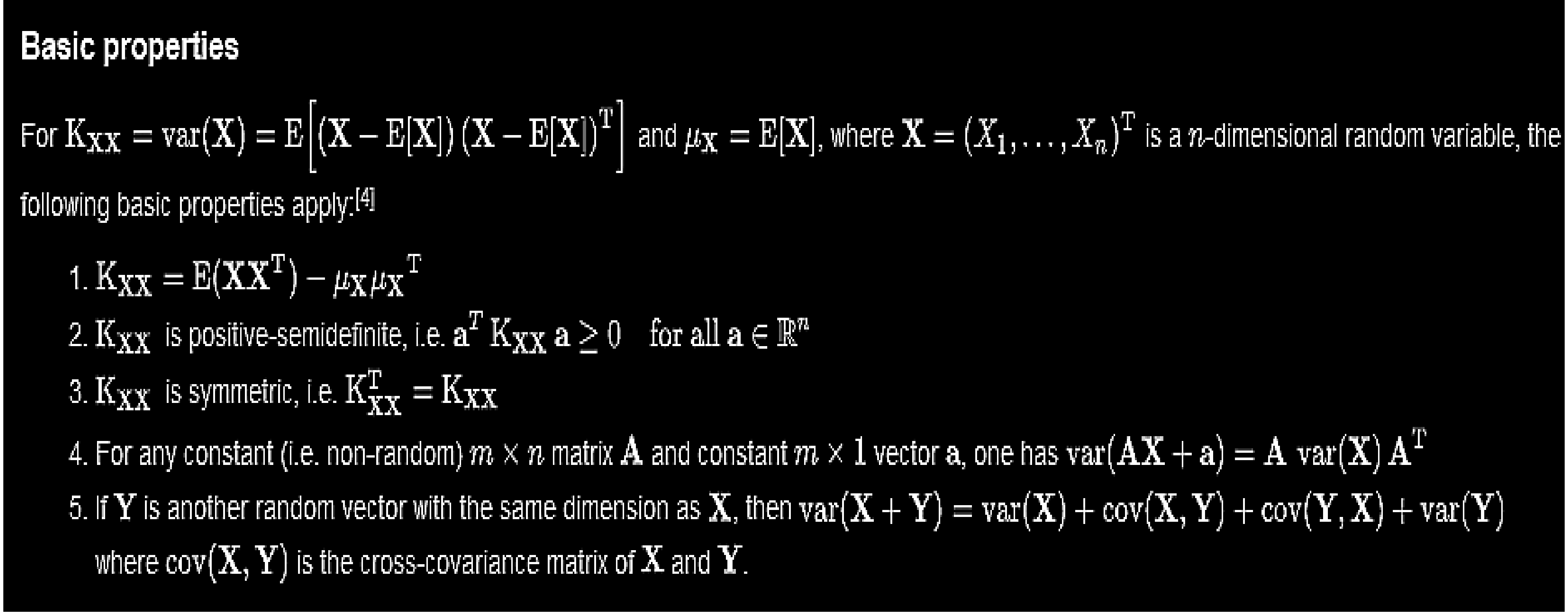
Nomenclatures differ. Some statisticians, following the probabilist William Feller in his two-volume book *An Introduction to Probability Theory and Its Applications*,[2] call the matrix $\mathrm{K}_{\mathbf{XX}}$ the **variance** of the random vector $\mathbf{X}$, because it is the natural generalization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

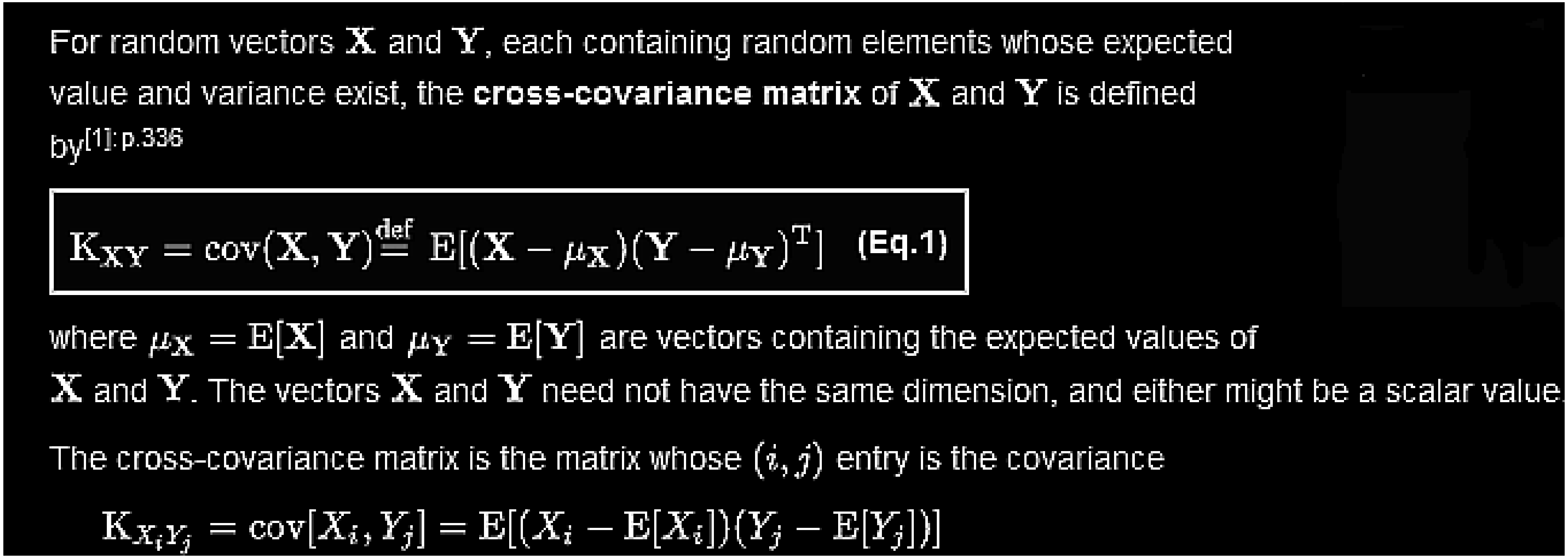
$$\operatorname{cov}(\mathbf{X},\mathbf{Y}) = \mathrm{K}_{\mathbf{XY}} = \mathrm{E}\left[(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\mathbf{Y} - \mathrm{E}[\mathbf{Y}])^{\mathrm{T}}\right].$$



Basic properties

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

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



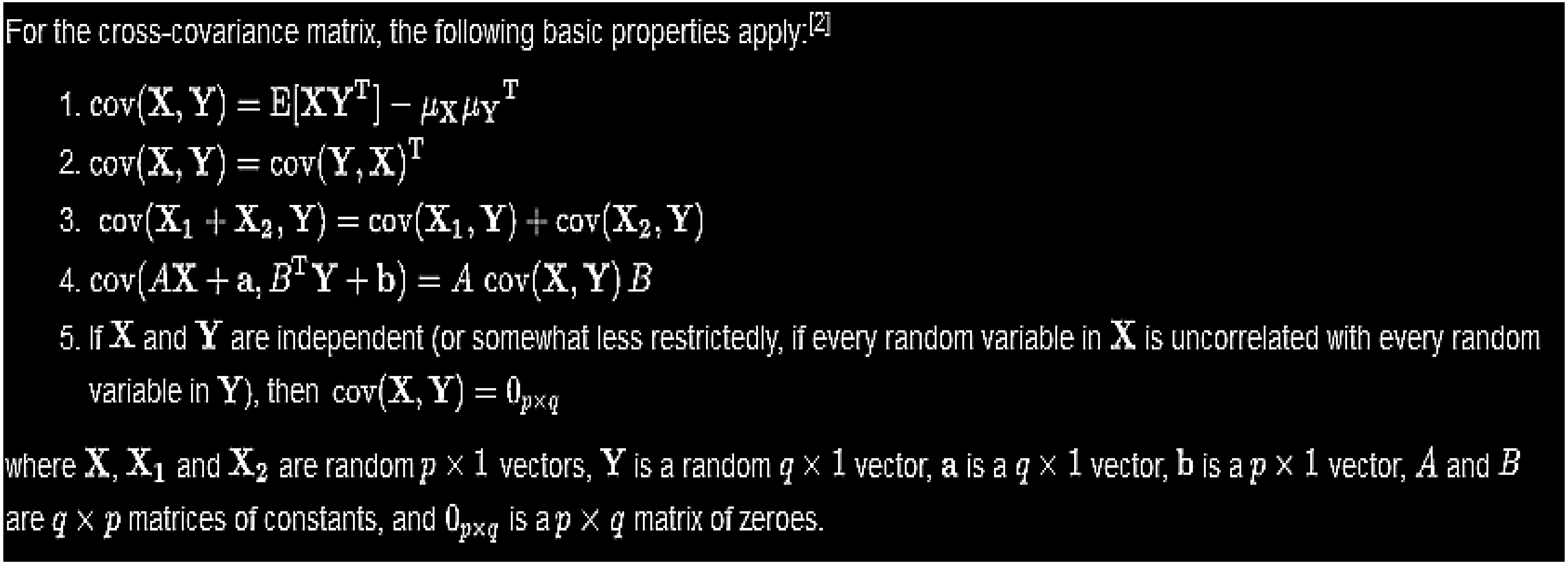
For random vectors $\mathbf{X}$ and $\mathbf{Y}$, each containing random elements whose expected value and variance exist, the **cross-covariance matrix** of $\mathbf{X}$ and $\mathbf{Y}$ is defined by[1]: p.336

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

where $\mu_{\mathbf{X}} = \operatorname{E}[\mathbf{X}]$ and $\mu_{\mathbf{Y}} = \operatorname{E}[\mathbf{Y}]$ 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

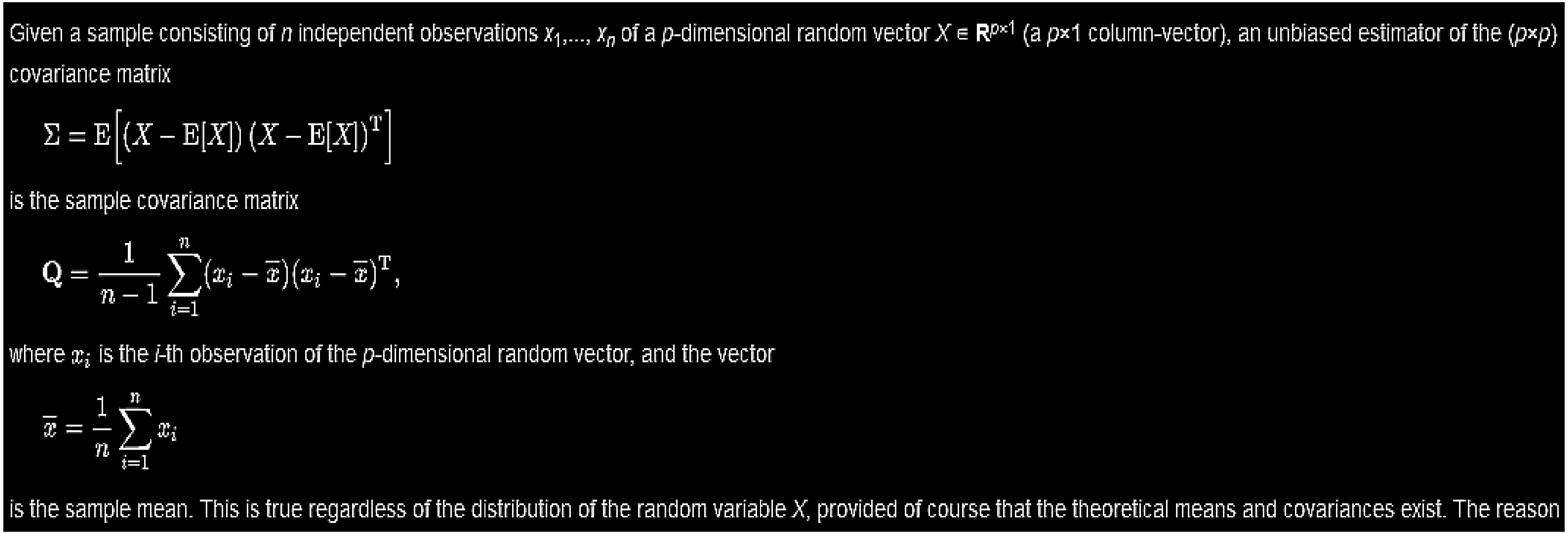
$$\operatorname{K}_{X_i Y_j} = \operatorname{cov}[X_i, Y_j] = \operatorname{E}[(X_i - \operatorname{E}[X_i])(Y_j - \operatorname{E}[Y_j])]$$



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*×1 column-vector), an unbiased estimator of the (*p*×*p*) covariance matrix

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

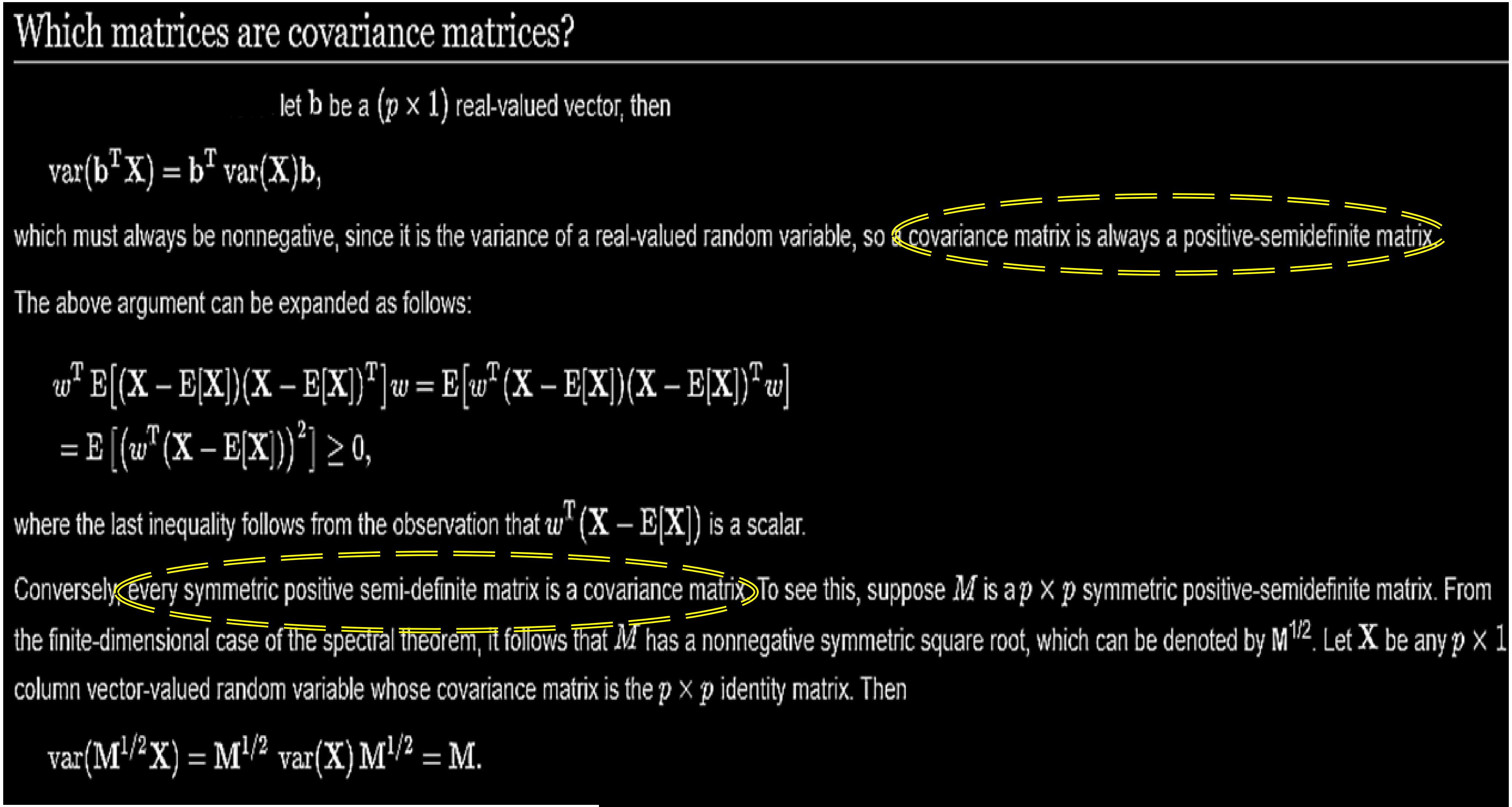
is the sample covariance matrix

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

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 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.

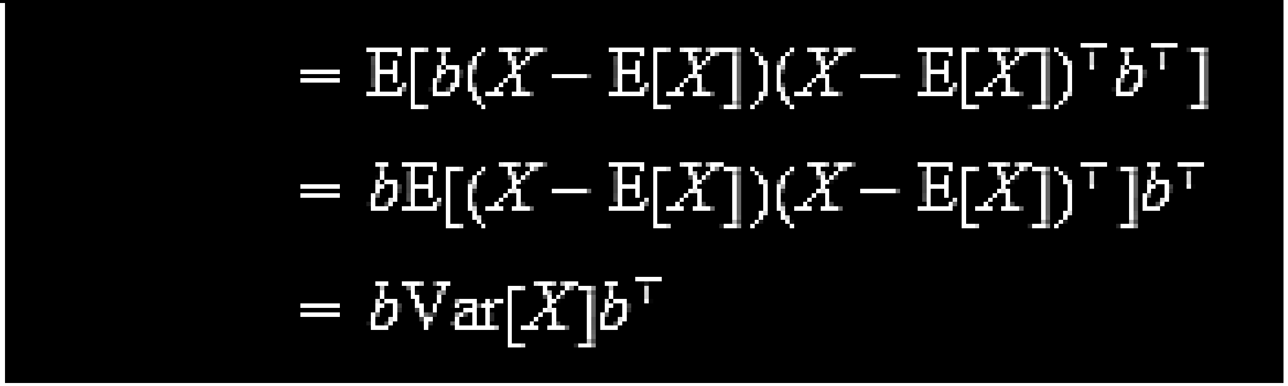
The above argument can be expanded as follows:

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

where the last inequality follows from the observation that $w^{\mathrm{T}}(\mathbf{X}-\mathrm{E}[\mathbf{X}])$ 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 $\mathrm{M}^{1/2}$. Let $\mathbf{X}$ be any $p \times 1$ column vector-valued random variable whose covariance matrix is the $p \times p$ identity matrix. Then

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



$$= \mathrm{E}[b(X-\mathrm{E}[X])(X-\mathrm{E}[X])^{\mathsf{T}}b^{\mathsf{T}}]$$
$$= b\mathrm{E}[(X-\mathrm{E}[X])(X-\mathrm{E}[X])^{\mathsf{T}}]b^{\mathsf{T}}$$
$$= b\mathrm{Var}[X]b^{\mathsf{T}}$$

PROB. & STAT. - *Revisited/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 σ_x.

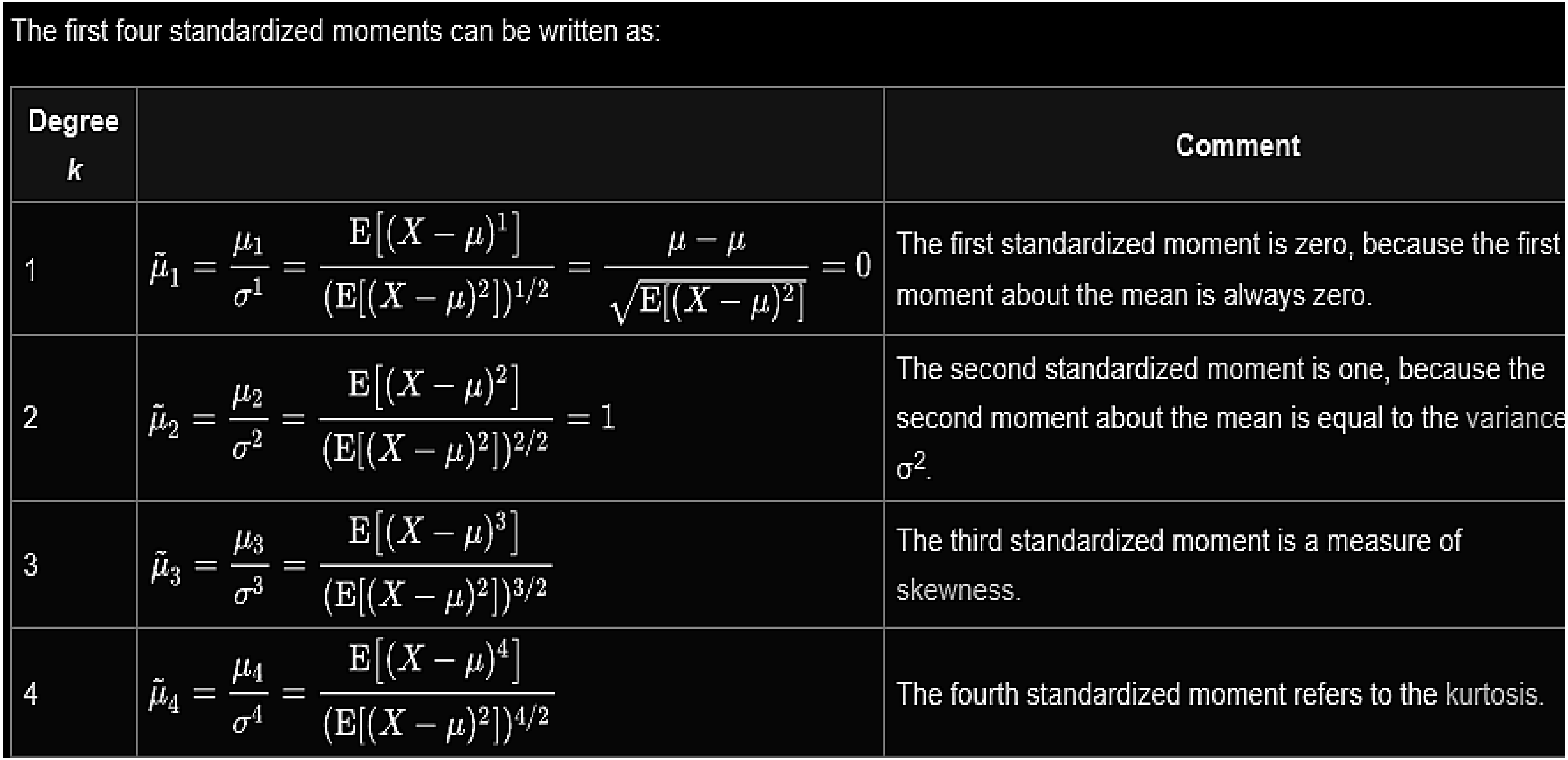
$$\sigma_x^2 = E[\{x-E(x)\}^2] = E[(x-\mu_x)^2]$$

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

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).



The first four standardized moments can be written as:

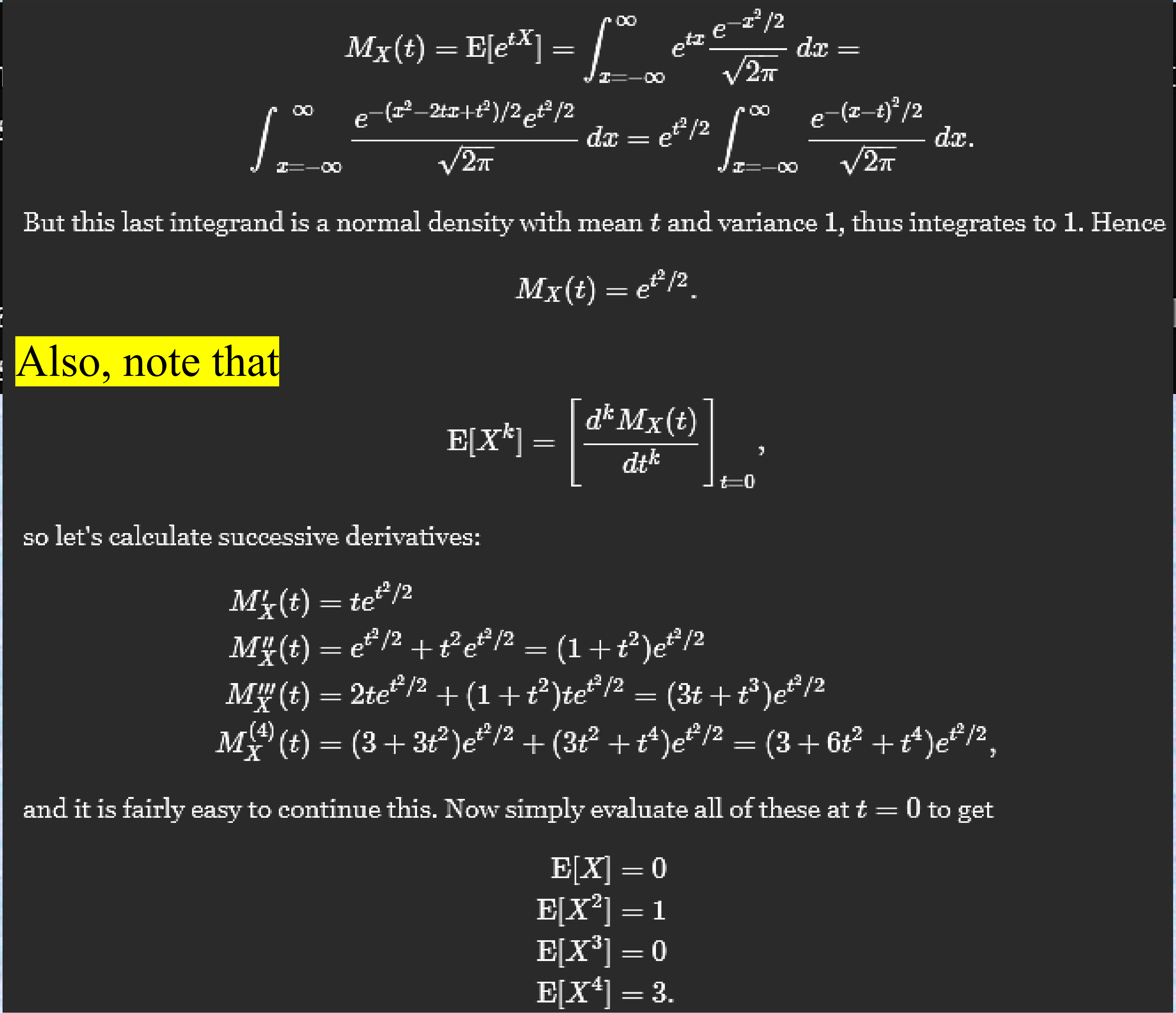
Degree *k*		**Comment**
1	$\tilde{\mu}_1 = \frac{\mu_1}{\sigma^1} = \frac{\mathrm{E}\left[(X-\mu)^1\right]}{(\mathrm{E}[(X-\mu)^2])^{1/2}} = \frac{\mu-\mu}{\sqrt{\mathrm{E}[(X-\mu)^2]}} = 0$	The first standardized moment is zero, because the first moment about the mean is always zero.
2	$\tilde{\mu}_2 = \frac{\mu_2}{\sigma^2} = \frac{\mathrm{E}\left[(X-\mu)^2\right]}{(\mathrm{E}[(X-\mu)^2])^{2/2}} = 1$	The second standardized moment is one, because the second moment about the mean is equal to the variance σ^2.
3	$\tilde{\mu}_3 = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{(\mathrm{E}[(X-\mu)^2])^{3/2}}$	The third standardized moment is a measure of skewness.
4	$\tilde{\mu}_4 = \frac{\mu_4}{\sigma^4} = \frac{\mathrm{E}\left[(X-\mu)^4\right]}{(\mathrm{E}[(X-\mu)^2])^{4/2}}$	The fourth standardized moment refers to the kurtosis.

The n ion

$M_X($

We sa hat

$M_X($



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

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

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

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

Also, note 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:

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

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

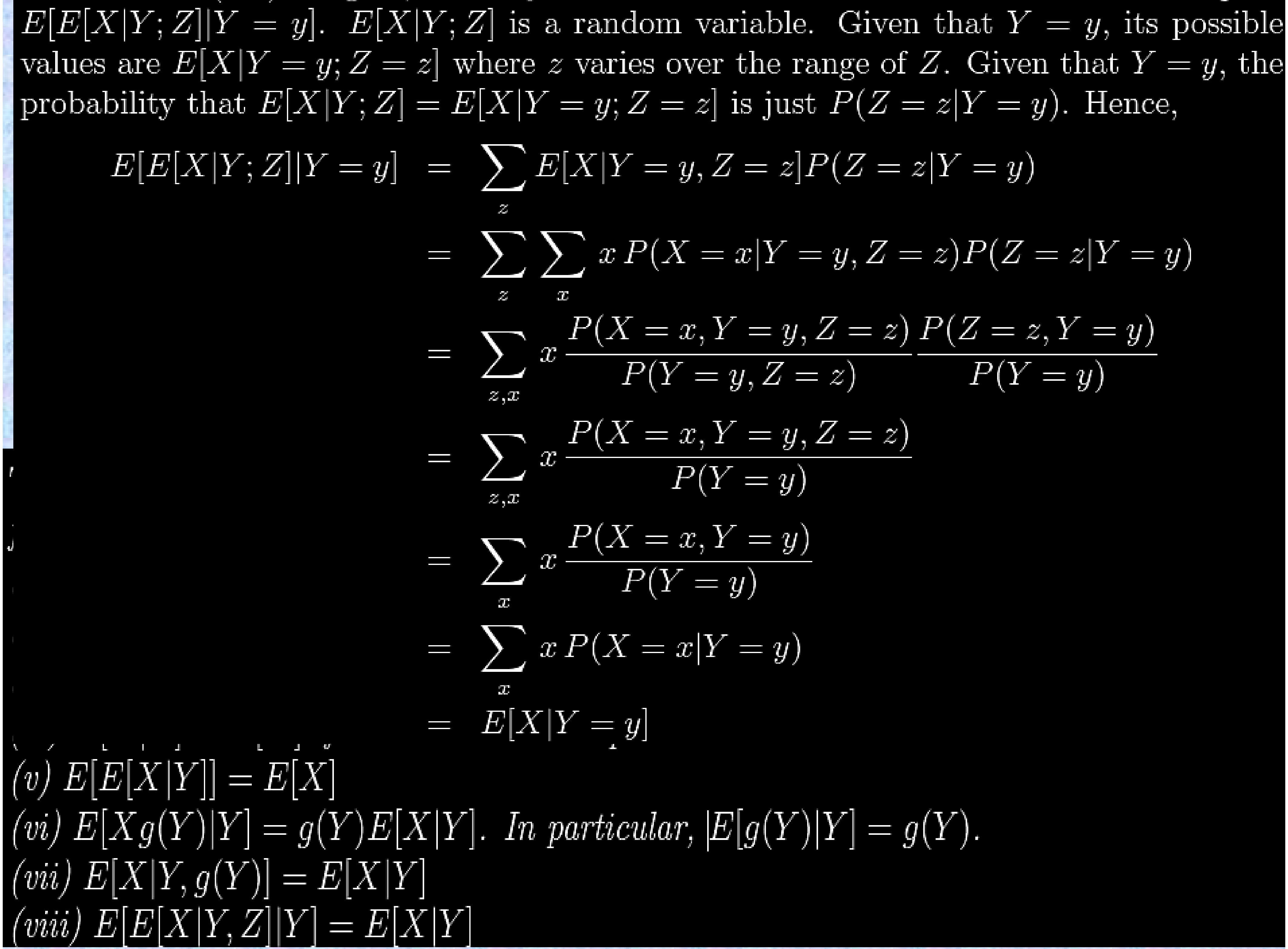
Taking derivative (w.r.t. μ)
of the LOG of the above:

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

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.



$E[E[X|Y;Z]|Y=y]$. $E[X|Y;Z]$ 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,

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

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

$E[E[X|Y;Z]|Y=y]$. $E[X|Y;Z]$ 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,

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

Sampling Distributions

http://grid.cs.gsu.edu/~skarmakar/math1070_slides.html

What are the main types of <u>sampling</u> and how is each done?

Simple Random Sampling: A simple random sample (**SRS**) of size *n* is produced by a scheme which ensures that each subgroup of the population of size *n* has an equal probability of being chosen as the sample.

Stratified Random Sampling: Divide the population into "strata". There can be any number of these. Then choose a simple random sample from each stratum. Combine those into the overall sample. That is a stratified random sample. (Example: Church A has 600 women and 400 women as members. One way to get a stratified random sample of size 30 is to take a SRS of 18 women from the 600 women and another SRS of 12 men from the 400 men.)

Multi-Stage Sampling: Sometimes the population is too large and scattered for it to be practical to make a list of the entire population from which to draw a SRS. For instance, when the a polling organization samples US voters, they do not do a SRS. Since voter lists are compiled by counties, they might first do a sample of the counties and then sample within the selected counties. This illustrates two stages.

< SRC: WIKI *>*

In statistics, a **simple random sample** is a subset of individuals (a sample) chosen from a larger set (a population). Each individual is chosen randomly and entirely by chance, such that each individual has the same probability of being chosen at any stage during the sampling process, and each subset of k individuals has the same probability of being chosen for the sample as any other subset of k individuals. This process and technique is known as simple random sampling, and should not be confused with systematic random sampling. A simple random sample is an unbiased surveying technique.

Systematic sampling (Sys-S) is a statistical method involving the selection of elements from an ordered sampling frame. The most common form of systematic sampling is an equi-probability method. In this approach, progression through the list is treated circularly, with a return to the top once the end of the list is passed. The sampling starts by selecting an element from the list at random and then every k-th element in the frame is selected, where k, the sampling interval (sometimes known as the *skip*): this is calculated as: $k = \mathrm{N/n}$ where n is the sample size, and N is the population size.

Systematic sampling (Sys-S) Example: Suppose a supermarket wants to study buying habits of their customers, then using systematic sampling they can choose every 10th or 15th customer entering the supermarket and conduct the study on this sample.

This is random sampling with a system. From the sampling frame, a starting point is chosen at random, and choices thereafter are at regular intervals. For example, suppose you want to sample 8 houses from a street of 120 houses. 120/8=15, so every 15th house is chosen after a random starting point between 1 and 15. If the random starting point is 11, then the houses selected are 11, 26, 41, 56, 71, 86, 101, and 116.

Sampling With Replacement and Sampling Without Replacement

Consider a population of potato sacks, each of which has either 12, 13, 14, 15, 16, 17, or 18 potatoes, and all the values are equally likely. Suppose that, in this population, there is exactly one sack with each number. So the whole population has seven sacks.

Sampling with replacement:

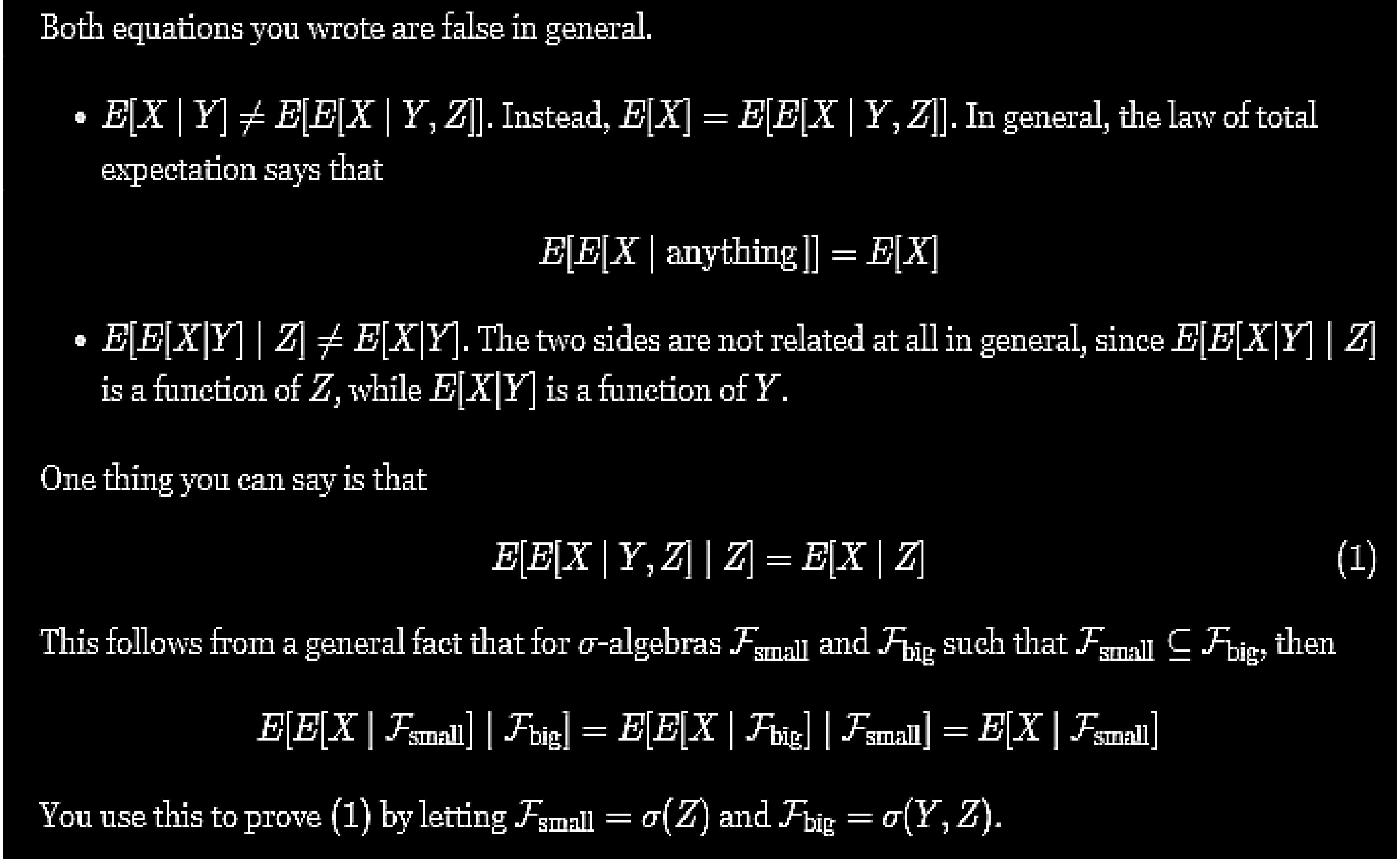
If I sample two with replacement, then I first pick one (say 14). I had a 1/7 probability of choosing that one. Then I replace it. Then I pick another. Every one of them still has 1/7 probability of being chosen. And there are exactly 49 different possibilities here.

Sampling without replacement:

If I sample two without replacement, then I first pick one (say 14). I had a 1/7 probability of choosing that one. Then I pick another. At this point, there are only six possibilities: 12, 13, 15, 16, 17, and 18. So there are only 42 different possibilities here (again assuming that we distinguish between the first and the second.)

Sampling distribution

- The sampling distribution of a statistic (not parameter) is the **distribution of values taken by the statistic** (not parameter) **in all possible samples** of the **same size** from the **same population**.



Both equations you wrote are false in general.

- $E[X \mid Y] \neq E[E[X \mid Y, Z]]$. Instead, $E[X] = E[E[X \mid Y, Z]]$. In general, the law of total expectation says that

$$E[E[X \mid \text{anything}\,]] = E[X]$$

- $E[E[X|Y] \mid Z] \neq E[X|Y]$. The two sides are not related at all in general, since $E[E[X|Y] \mid Z]$ is a function of Z, while $E[X|Y]$ is a function of Y.

One thing you can say is that

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

This follows from a general fact that for σ-algebras $\mathcal{F}_{\text{small}}$ and $\mathcal{F}_{\text{big}}$ such that $\mathcal{F}_{\text{small}} \subseteq \mathcal{F}_{\text{big}}$, then

$$E[E[X \mid \mathcal{F}_{\text{small}}] \mid \mathcal{F}_{\text{big}}] = E[E[X \mid \mathcal{F}_{\text{big}}] \mid \mathcal{F}_{\text{small}}] = E[X \mid \mathcal{F}_{\text{small}}]$$

You use this to prove (1) by letting $\mathcal{F}_{\text{small}} = \sigma(Z)$ and $\mathcal{F}_{\text{big}} = \sigma(Y, Z)$.

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

This follows from a general fact that for σ-algebras $\mathcal{F}_{\text{small}}$ and $\mathcal{F}_{\text{big}}$ such that $\mathcal{F}_{\text{small}} \subseteq \mathcal{F}_{\text{big}}$, then

$$E[E[X \mid \mathcal{F}_{\text{small}}] \mid \mathcal{F}_{\text{big}}] = E[E[X \mid \mathcal{F}_{\text{big}}] \mid \mathcal{F}_{\text{small}}] = E[X \mid \mathcal{F}_{\text{small}}]$$

You use this to prove (1) by letting $\mathcal{F}_{\text{small}} = \sigma(Z)$ and $\mathcal{F}_{\text{big}} = \sigma(Y, Z)$.

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

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

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.

Sampling Distribution Introduction

- In real life calculating parameters of populations is prohibitive because populations are very large.
- Rather than investigating the whole population, we take a sample, calculate a **statistic** related to the **parameter** of interest, and make an inference.
- The **sampling distribution** of the **statistic** is the tool that tells us how close is the statistic to the parameter.

Sample Statistics as Estimators of Population Parameters

- A **sample statistic** is a numerical measure of a summary characteristic of a sample.

A **population parameter** is a numerical measure of a summary characteristic of a population.

- An **estimator** of a population parameter is a sample statistic used to estimate or predict the population parameter.
- An **estimate** of a parameter is a *particular* numerical value of a sample statistic obtained through sampling.
- A **point estimate** is a single value used as an estimate of a population parameter.

Estimators

- The sample mean, $\overline{X}$, is the most common estimator of the population mean, μ.
- The sample variance, s^2, is the most common estimator of the population variance, σ^2.
- The sample standard deviation, s, is the most common estimator of the population standard deviation, σ.
- The sample proportion, $\hat{p}$, is the most common estimator of the population proportion, p.

Sampling Distribution of $\overline{X}$

- The **sampling distribution of $\overline{X}$** is the probability distribution of all possible values the random variable $\overline{X}$ may assume when a sample of size *n* is taken from a specified population.

Sampling Distribution of the Mean

- An example
  - A die is thrown infinitely many times. Let X represent the number of spots showing on any throw.
  - The probability distribution of X is

x	1	2	3	4	5	6
p(x)	1/6	1/6	1/6	1/6	1/6	1/6

E(X) = 1(1/6) + 2(1/6) + 3(1/6)+ ……………………….= 3.5

V(X) = $(1-3.5)^2(1/6)$ + $(2-3.5)^2(1/6)$ + ………….. …= 2.92

Throwing a dice twice – sampling distribution of sample mean

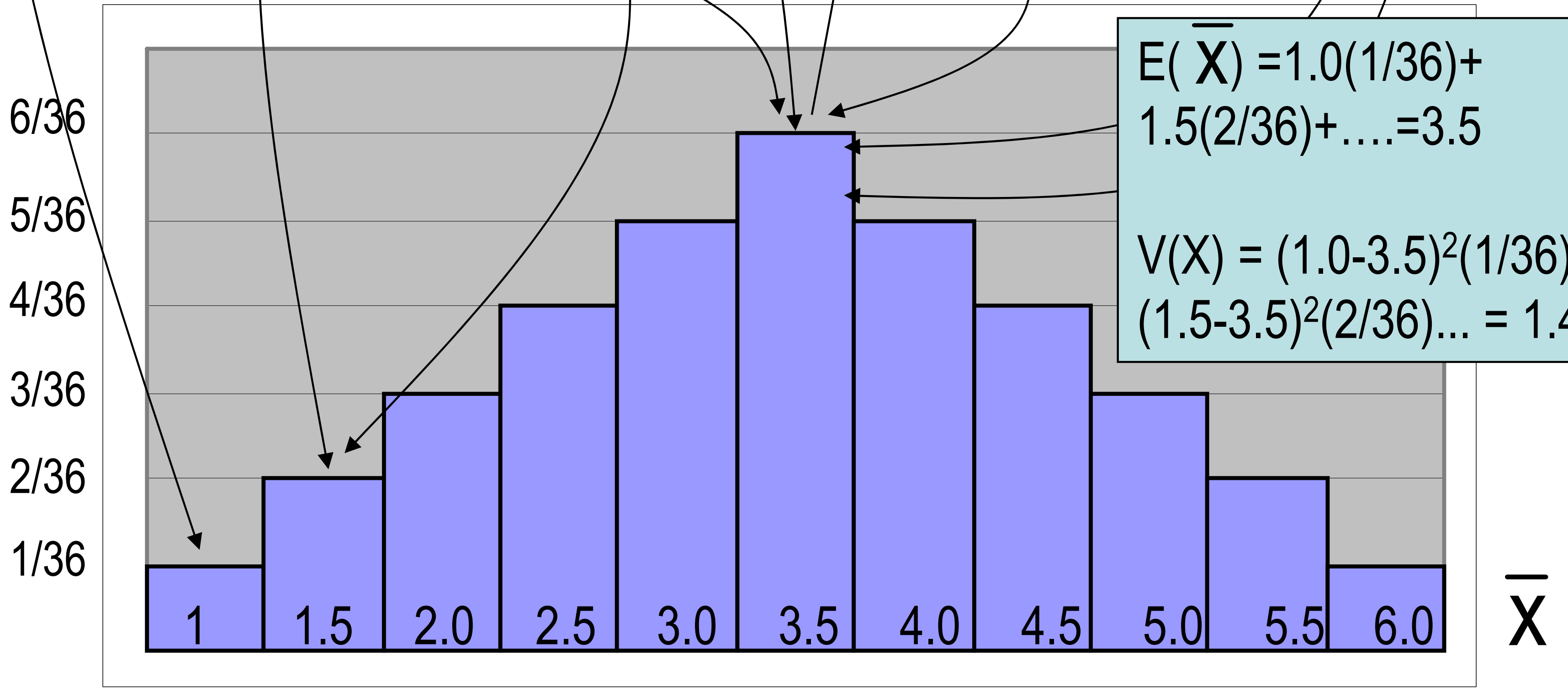
- Suppose we want to estimate μ from the mean $\bar{x}$ of a sample of size n = 2.
- What is the distribution of $\bar{x}$?

Throwing a die twice – sample mean

Sample		**Mean**	**Sample**		**Mean**	**Sample**		**Mean**
1	1,1	1	13	3,1	2	25	5,1	3
2	1,2	1.5	14	3,2	2.5	26	5,2	3.5
3	1,3	2	15	3,3	3	27	5,3	4
4	1,4	2.5	16	3,4	3.5	28	5,4	4.5
5	1,5	3	17	3,5	4	29	5,5	5
6	1,6	3.5	18	3,6	4.5	30	5,6	5.5
7	2,1	1.5	19	4,1	2.5	31	6,1	3.5
8	2,2	2	20	4,2	3	32	6,2	4
9	2,3	2.5	21	4,3	3.5	33	6,3	4.5
10	2,4	3	22	4,4	4	34	6,4	5
11	2,5	3.5	23	4,5	4.5	35	6,5	5.5
12	2,6	4	24	4,6	5	36	6,6	6

Sample		Mean	Sample		Mean	Sample		Mean
1	1,1	1	13	3,1	2	25	5,1	3
2	1,2	1.5	14	3,2	2.5	26	5,2	3.5
3	1,3	2	15	3,3	3	27	5,3	4
10	2,4	3	22	4,4	4	34	6,4	5
11	2,5	3.5	23	4,5	4.5	35	6,5	5.5
12	2,6	4	24	4,6	5	36	6,6	6

$$Note:\ \mu_{\bar{x}} = \mu_x \quad and \quad \sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{2}$$



E($\bar{X}$) =1.0(1/36)+ 1.5(2/36)+….=3.5

V(X) = $(1.0-3.5)^2$(1/36)+ $(1.5-3.5)^2$(2/36)... = 1.46

Sampling Distribution of the Mean

$n = 5$

$\mu_{\bar{x}} = 3.5$

$\sigma^2_{\bar{x}} = .5833\ (= \frac{\sigma^2_x}{5})$

6

$n = 10$

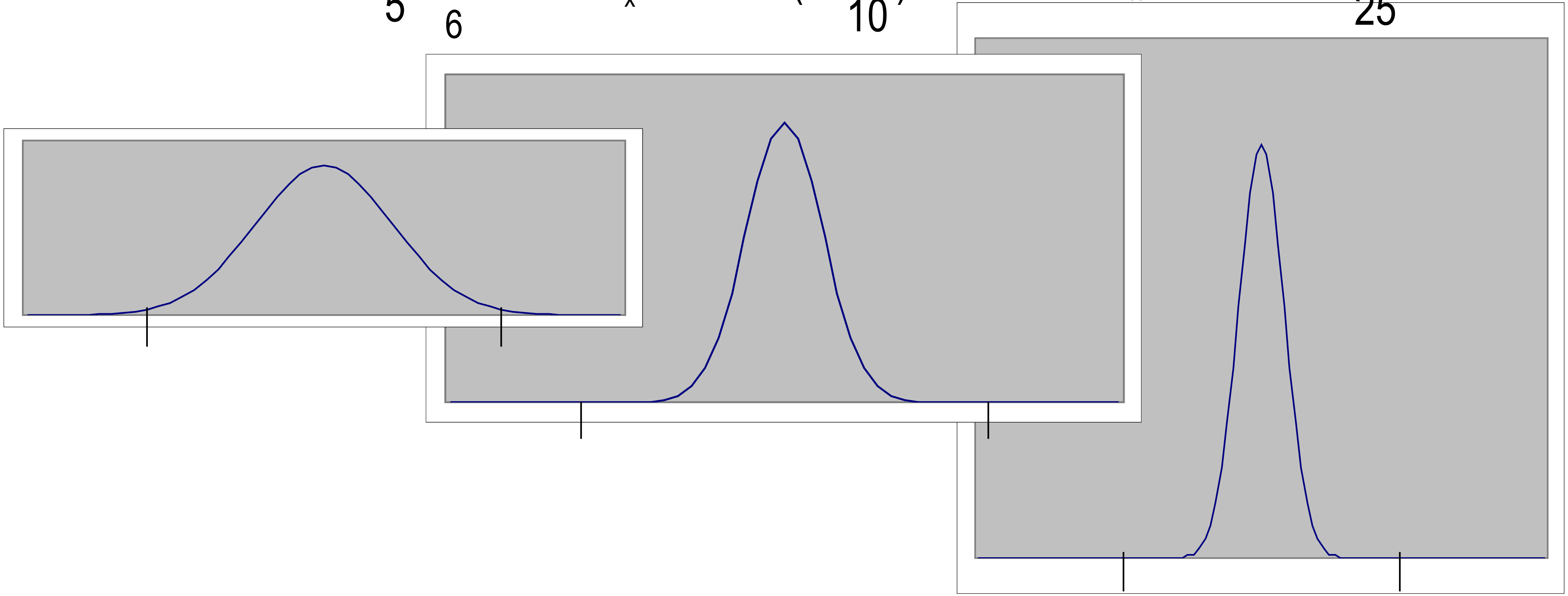
$\mu_{\bar{x}} = 3.5$

$\sigma^2_{\bar{x}} = .2917\ (= \frac{\sigma^2_x}{10})$

$n = 25$

$\mu_{\bar{x}} = 3.5$

$\sigma^2_{\bar{x}} = .1167\ (= \frac{\sigma^2_x}{25})$



Sampling Distribution of the Mean

$$n = 5$$

$$\mu_{\bar{x}} = 3.5$$

$$\sigma^2_{\bar{x}} = .5833\ (= \frac{\sigma^2_x}{5})$$

$$n = 10$$

$$\mu_{\bar{x}} = 3.5$$

$$\sigma^2_{\bar{x}} = .2917\ (= \frac{\sigma^2_x}{10})$$

$$n = 25$$

$$\mu_{\bar{x}} = 3.5$$

$$\sigma^2_{\bar{x}} = .1167\ (= \frac{\sigma^2_x}{25})$$

Notice that $\sigma^2_{\bar{x}}$ is smaller than σ^2_x. The larger the sample size the smaller $\sigma^2_{\bar{x}}$. Therefore, $\bar{X}$ tends to fall closer to μ, as the sample size increases.

Relationships between Population Parameters and the Sampling Distribution of the Sample Mean

The **expected value of the sample mean** is equal to the population mean:

$$E(\overline{X}) = \mu_{\overline{X}} = \mu_X$$

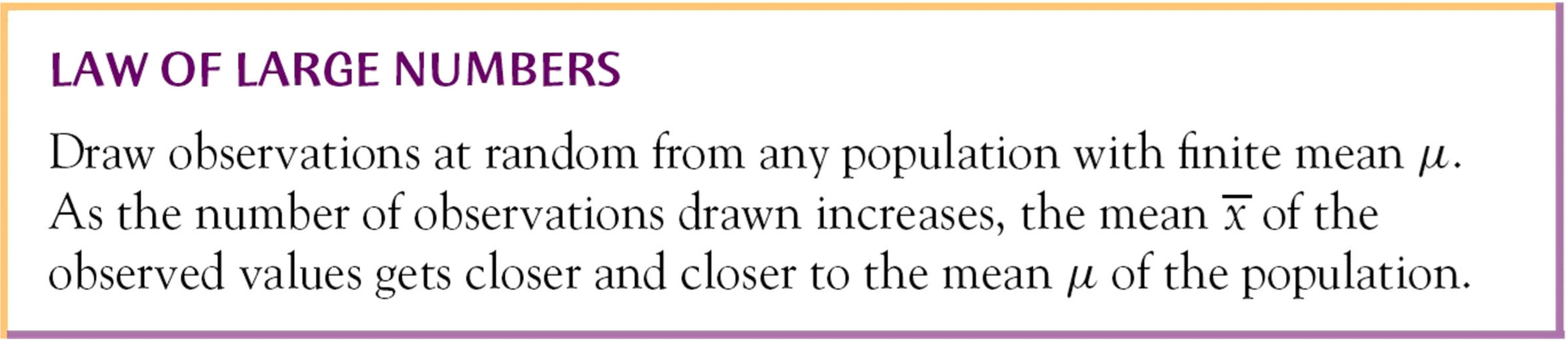
The **variance of the sample mean** is equal to the population variance divided by the sample size:

$$V(\overline{X}) = \sigma^2_{\overline{X}} = \frac{\sigma^2_X}{n}$$

The **standard deviation of the sample mean, known as the standard error of the mean,** is equal to the population standard deviation divided by the square root of the sample size:

$$\text{s.e.} = SD(\overline{X}) = \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$$

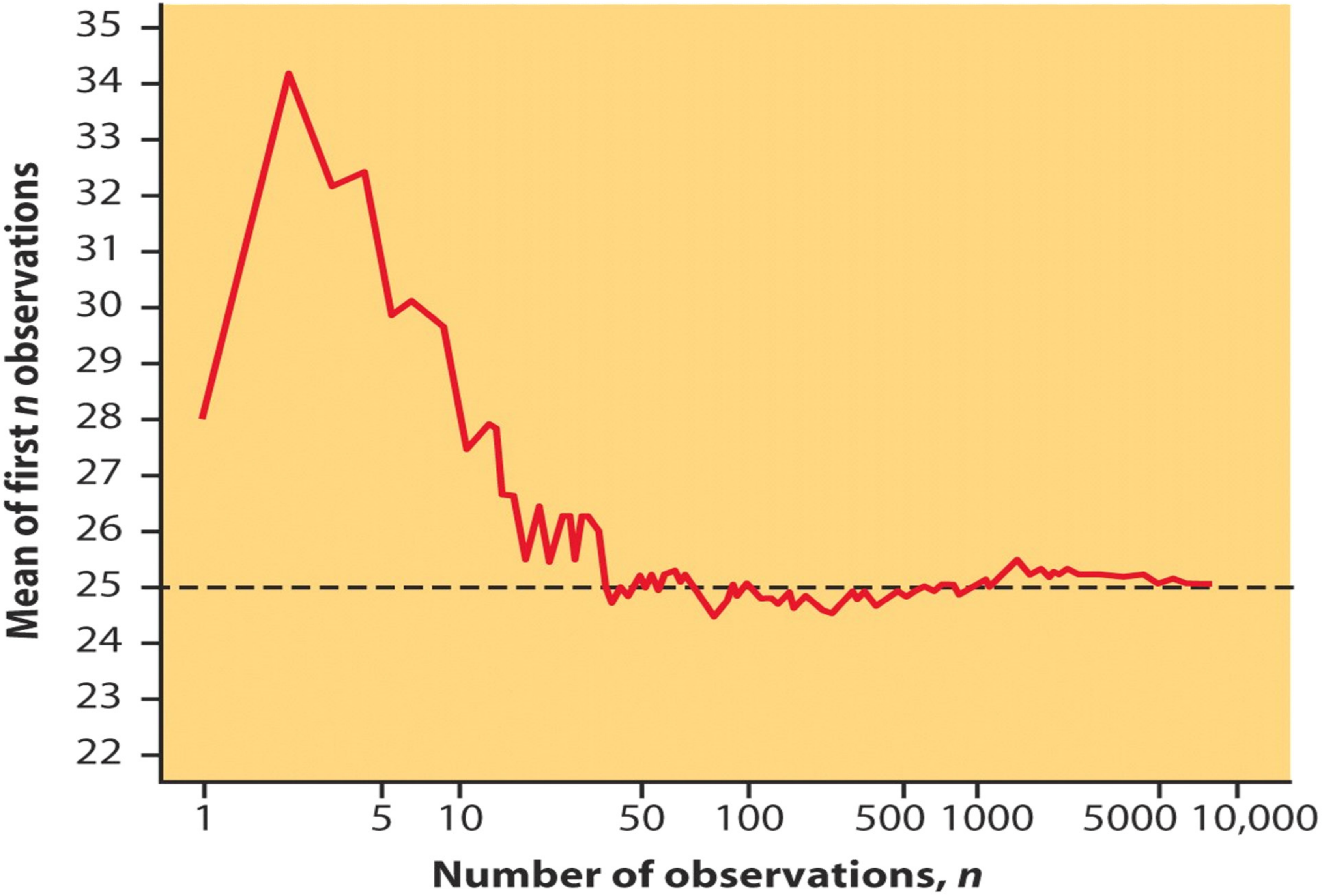
Law of Large Number



LAW OF LARGE NUMBERS

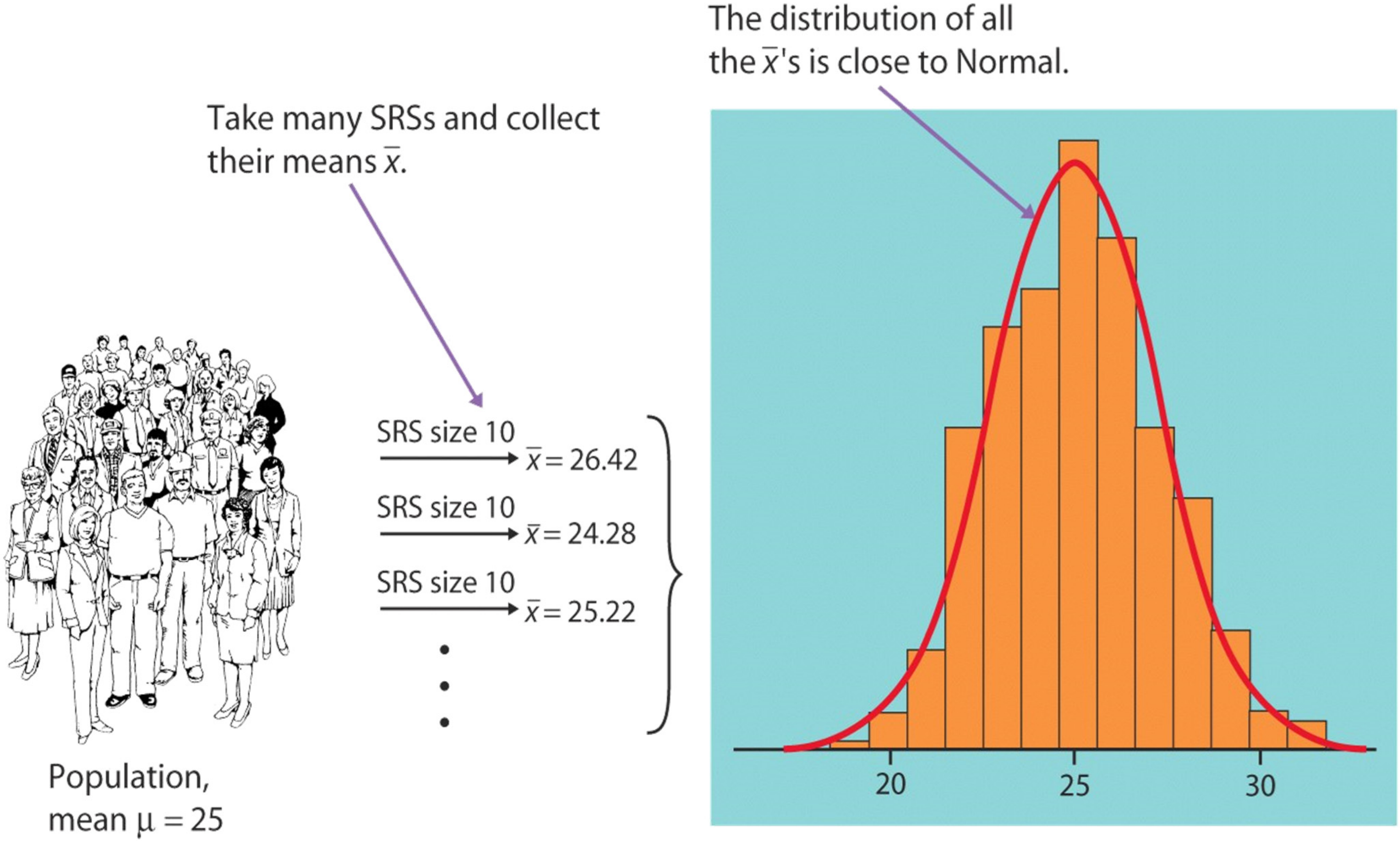
Draw observations at random from any population with finite mean μ. As the number of observations drawn increases, the mean $\overline{x}$ of the observed values gets closer and closer to the mean μ of the population.

How sample means approach the population mean (μ=25).



Example

- what would happen in many samples?

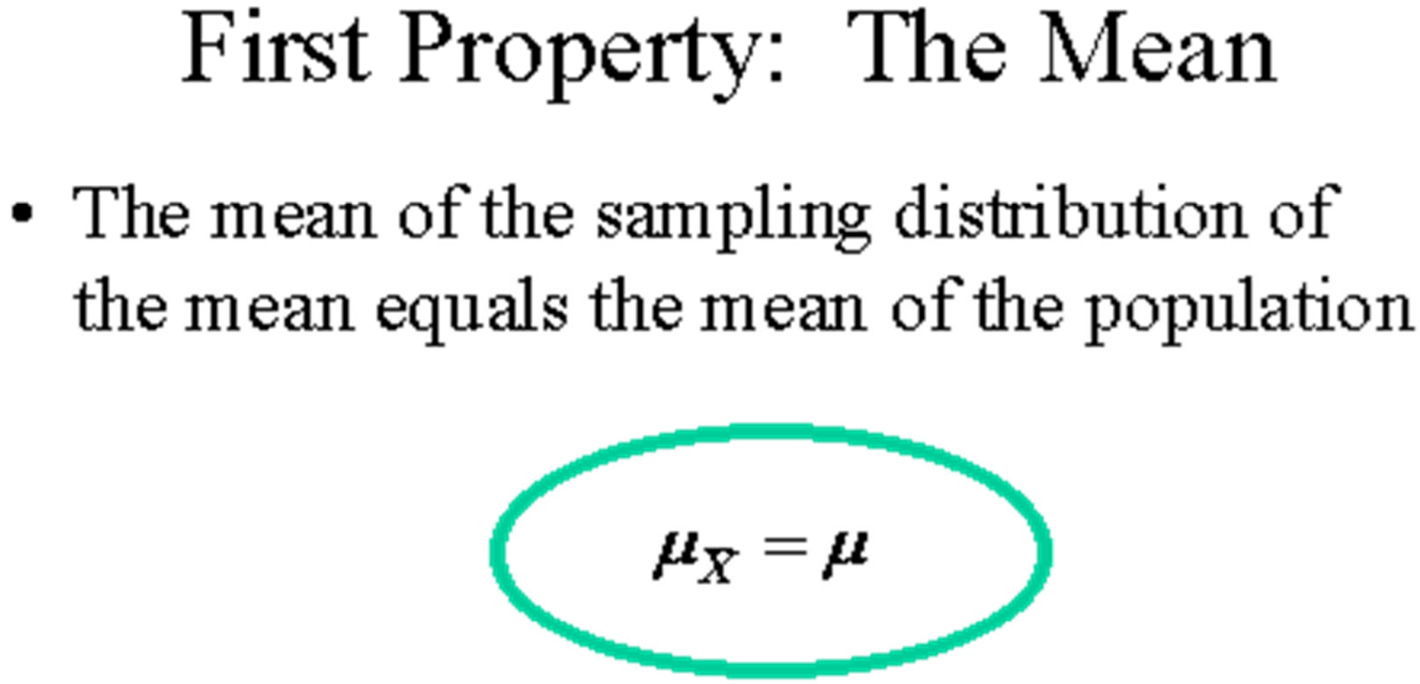


Recall Some Features of the Sampling Distribution

- It will approximate a normal curve even if the population you started with does NOT look normal

- Sampling distribution serves as a bridge between the sample and the population

Mean of a sample mean $\bar{x}$

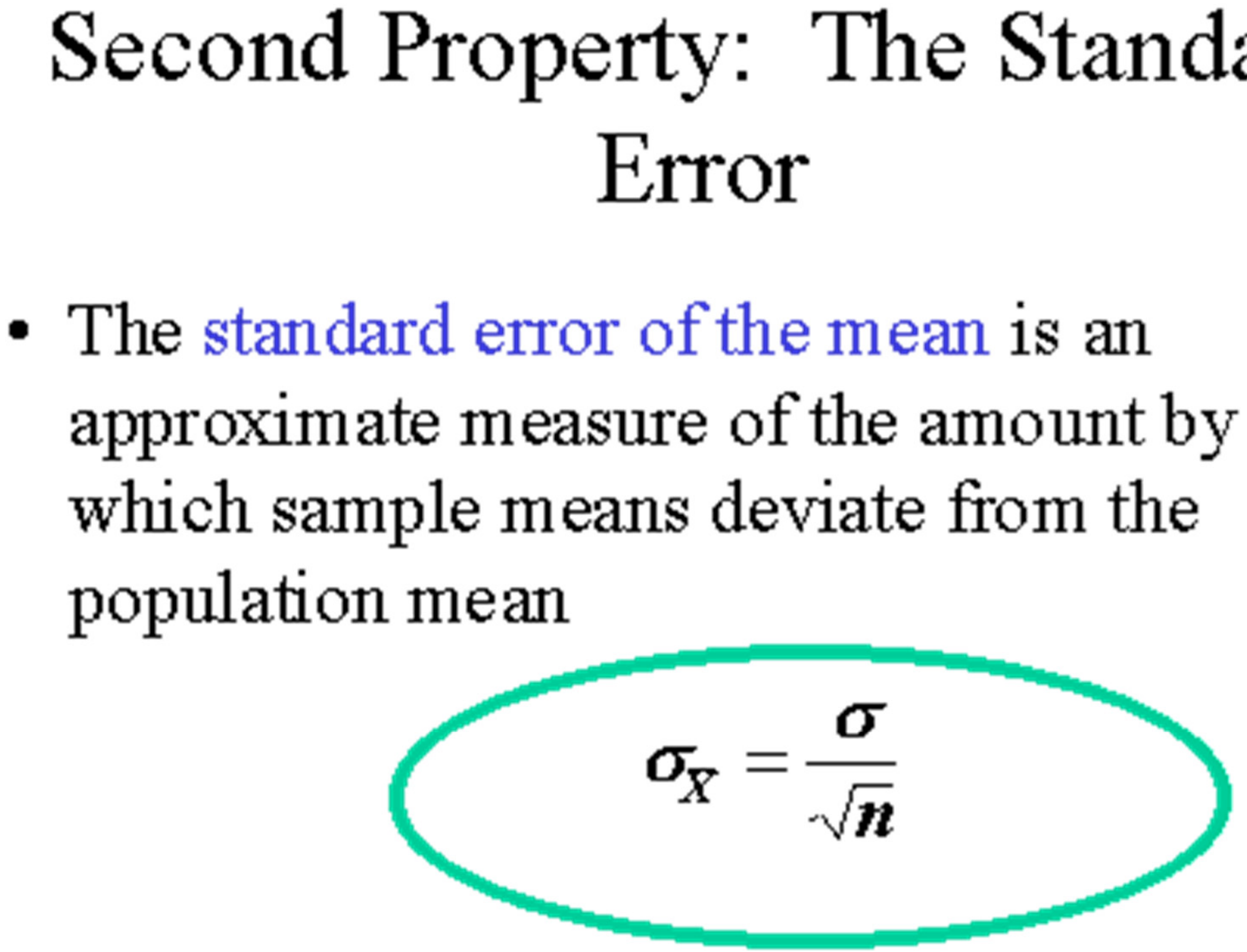


First Property: The Mean

- The mean of the sampling distribution of the mean equals the mean of the population

$$\mu_X = \mu$$

Standard Deviation of a sample mean $\bar{x}$



Second Property: The Standard Error

- The standard error of the mean is an approximate measure of the amount by which sample means deviate from the population mean

$$\sigma_{X} = \frac{\sigma}{\sqrt{n}}$$

Third Property: Sample Size and the Standard Deviation

- The larger the sample size, the smaller the standard deviation of the mean $\bar{x}$

Or

- As n increases, the standard deviation of the mean decreases

Example

- Population standard deviation = 100

$$\boldsymbol{For\ n} = 10,\ \sigma_X = \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}} = \frac{\mathbf{100}}{\sqrt{\mathbf{10}}} = \mathbf{31.62}$$

$$\boldsymbol{For\ n} = 100,\ \sigma_X = \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}} = \frac{\mathbf{100}}{\sqrt{\mathbf{100}}} = \mathbf{10.00}$$

$$\boldsymbol{For\ n} = 1000,\ \sigma_X = \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}} = \frac{\mathbf{100}}{\sqrt{\mathbf{1000}}} = \mathbf{3.16}$$

Sampling distribution of a sample mean $\overline{x}$

- Definition: For a random variable x and a given sample size n, the distribution of the variable $\overline{x}$, that is the distribution of all possible sample means, is called the sampling distribution of the sample mean.

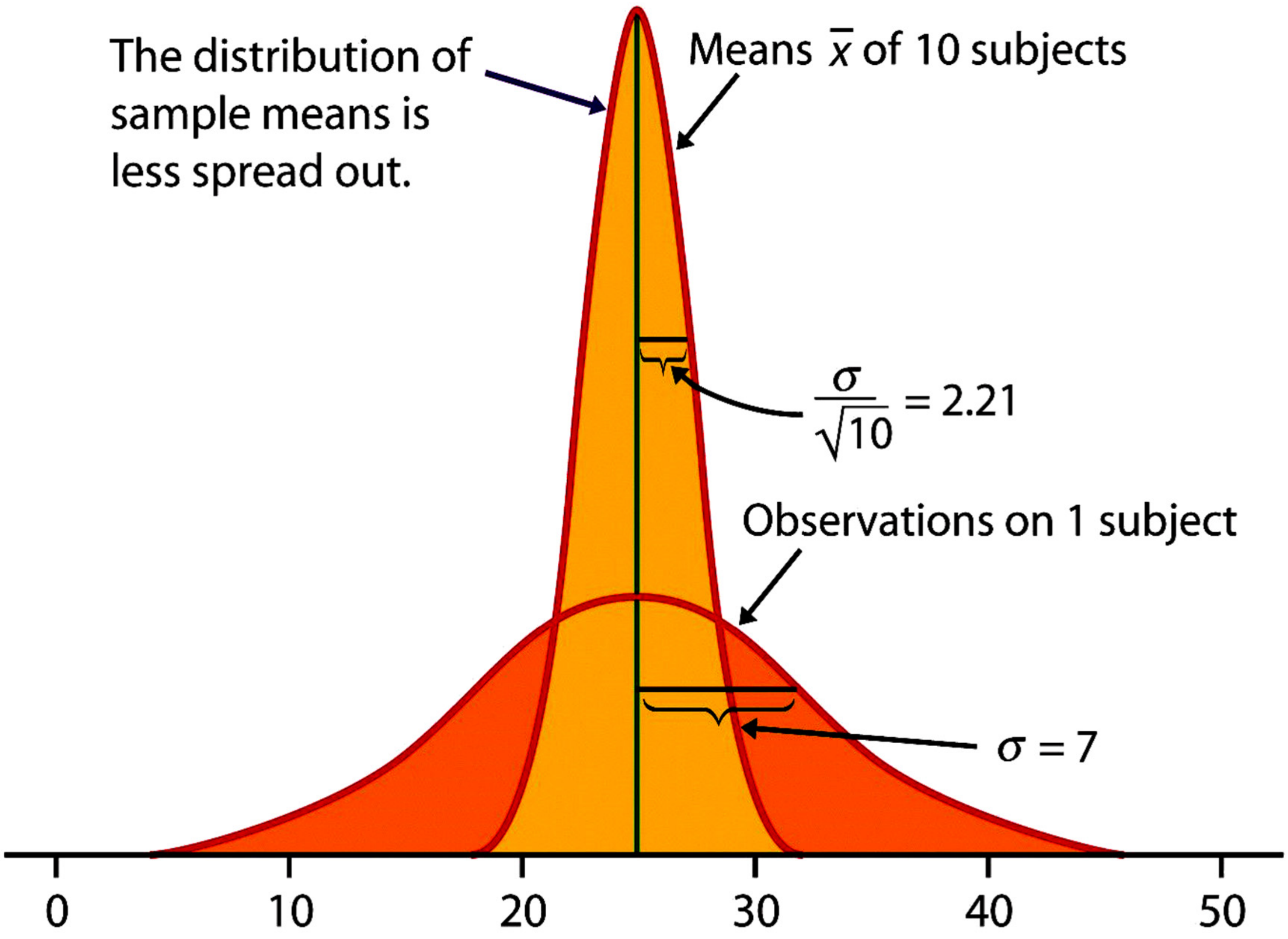
Sampling distribution of the sample mean

- Case 1. Population follows Normal distribution
  - Draw an SRS of size **n** from any population.
  - Repeat sampling.
  - Population follows a Normal distribution with mean **μ** and standard deviation **σ**.
  - Sampling distribution of $\overline{x}$ follows normal distribution as follows: **N(μ, σ/√n).**

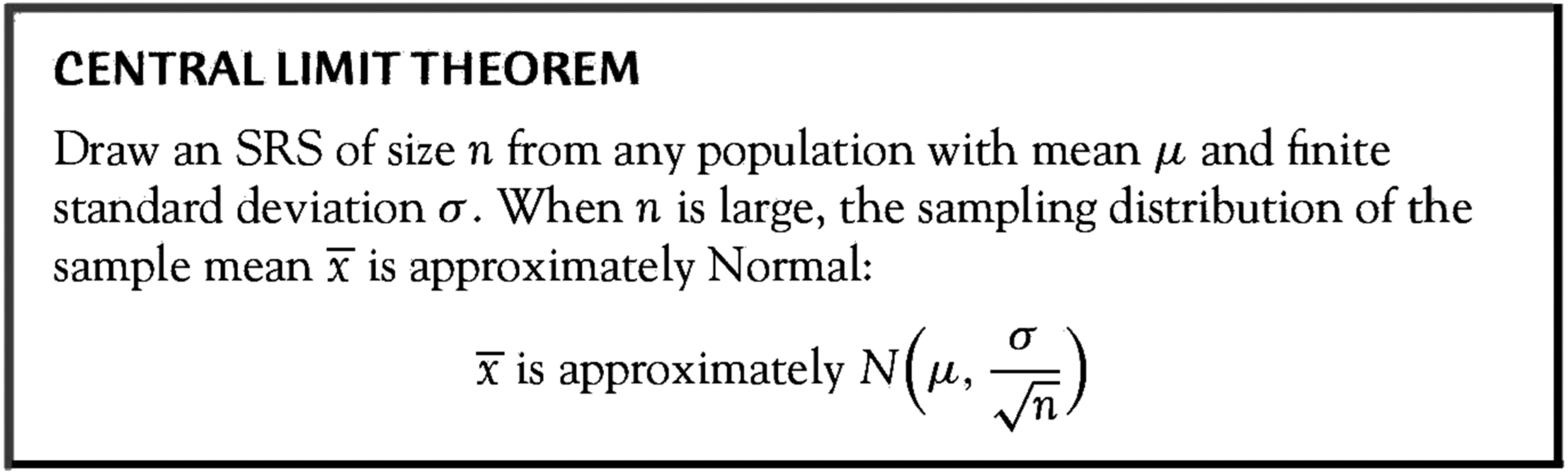
$\sigma / \sqrt{n}$

Example

(The population distribution follow a Normal distribution, then so does the sample mean)



The central limit theorem



CENTRAL LIMIT THEOREM

Draw an SRS of size n from any population with mean μ and finite standard deviation σ. When n is large, the sampling distribution of the sample mean $\overline{x}$ is approximately Normal:

$$\overline{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

This theorem tells us:

1. Small samples: Shape of sampling distribution is less normal
2. Large sample: Shape of sampling distribution is more normal.

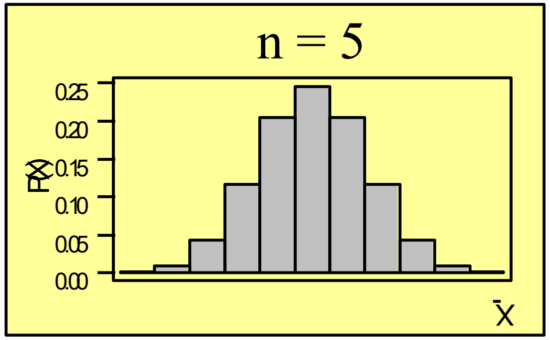
Sampling distribution of the sample mean

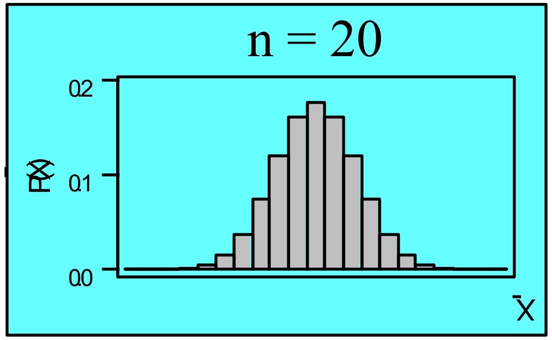
- Case 2. Population follows any distribution (CLT: Central limit theorem)
  - Draw an SRS of size **n** from any population.
  - Repeat sampling.
  - Population follows *a distribution* with mean **μ** and standard deviation **σ**.
  - When **n is large** (n>=30), sampling dist of $\bar{x}$ follows approximately Normal distribution as follows **N(μ, σ/√n).**

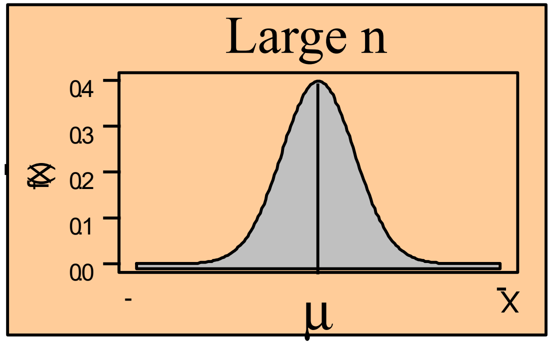
The Central Limit Theorem

When sampling from a population with mean μ and finite standard deviation σ, the sampling distribution of the sample mean will tend to be a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ as the sample size becomes large (n >30).

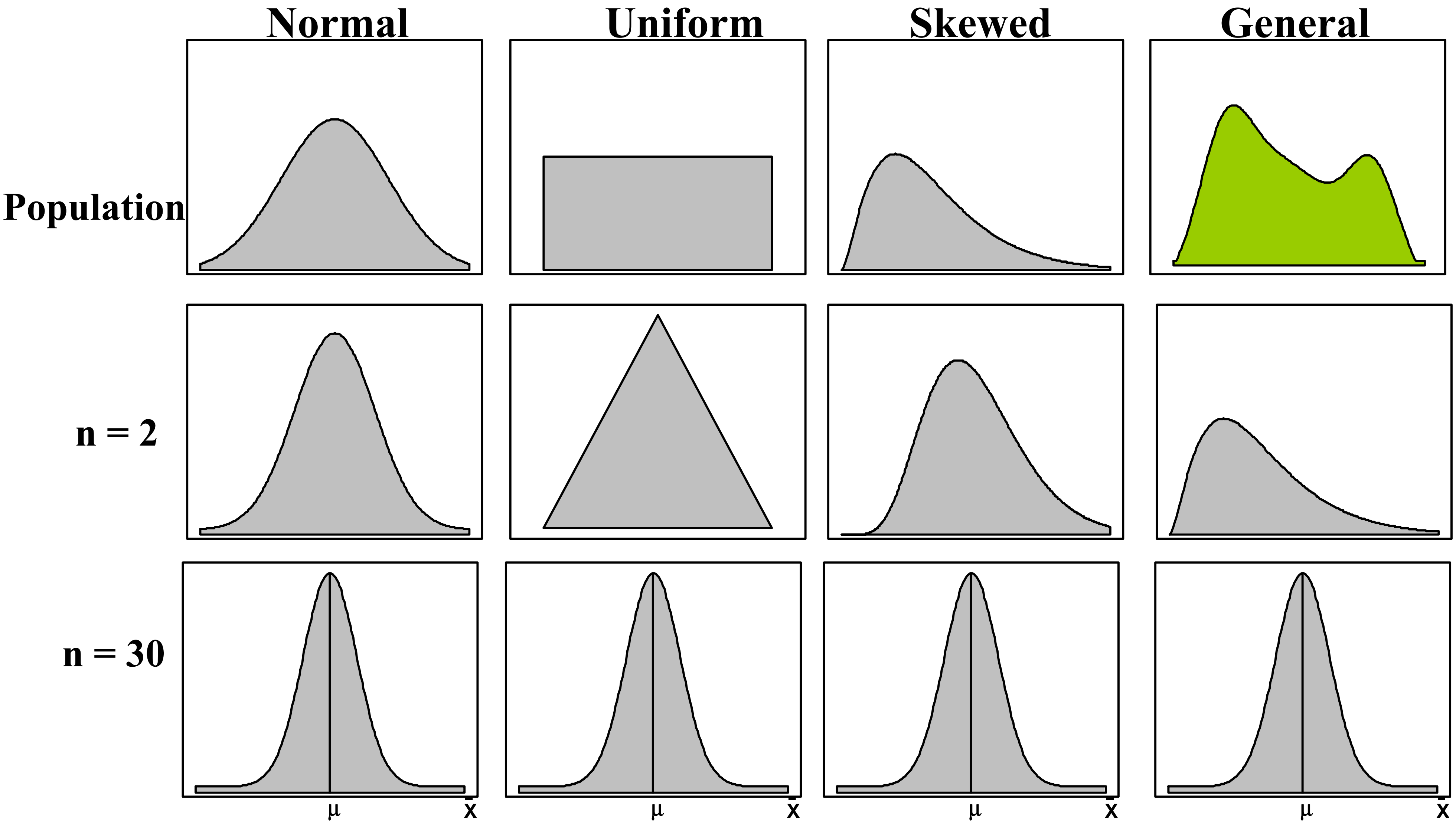
For “large enough” n: $\bar{X} \sim N(\mu, \sigma^2/n)$







The Central Limit Theorem Applies to Sampling Distributions from Any Population



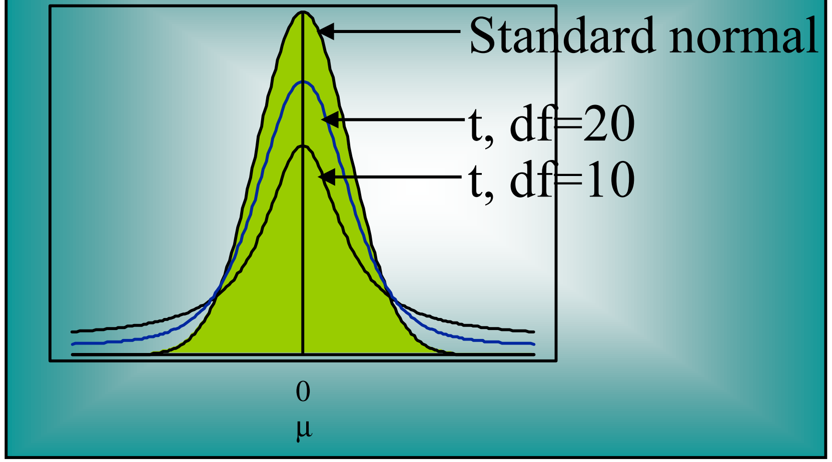
Student's *t* Distribution

If the population standard deviation, σ, is ***unknown***, replace σ with the sample standard deviation, *s*. If the population is normal, the resulting statistic:

$$t = \frac{\overline{X} - \mu}{s / \sqrt{n}}$$

has a ***t distribution* with *(n - 1) degrees of freedom***.

- The *t* is a family of bell-shaped and symmetric distributions, one for each number of degree of freedom.
- The expected value of *t* is 0.
- The variance of *t* is greater than 1, but approaches 1 as the number of degrees of freedom increases.
- The *t* distribution approaches a standard normal as the number of degrees of freedom increases.
- When the sample size is small (<30) we use t distribution.



Sampling Distributions

Finite Population Correction Factor

If the sample size is more than 5% of the population size and the sampling is done without replacement, then a correction needs to be made to the standard error of the means.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \bullet \sqrt{\frac{N-n}{N-1}}$$

Sampling Distribution of $\bar{x}$

Standard Deviation of $\bar{x}$

Finite Population | Infinite Population

$$\sigma_{\bar{x}} = \left(\frac{\sigma}{\sqrt{n}}\right)\sqrt{\frac{N-n}{N-1}}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- A finite population is treated as being infinite if $n/N \leq .05$.
- $\sqrt{(N-n)/(N-1)}$ is the finite correction factor.
- $\sigma_{\bar{x}}$ is referred to as the <u>standard error of the mean</u>.

The Sampling Distribution of the Sample Proportion, $\hat{p}$

The **sample proportion** is the percentage of successes in *n* binomial trials. It is the number of successes, *X*, divided by the number of trials, *n*.

Sample proportion: $\hat{p} = \dfrac{X}{n}$

As the sample size, *n,* increases, the sampling distribution of $\hat{p}$ approaches a **normal distribution** with mean *p* and standard deviation $\sqrt{\dfrac{p(1-p)}{n}}$