Uncertainty and Probability

MCSE 6309: Machine Learning

Dina Machuve (PhD)

Nelson Mandela African Institution of Science and Technology

July 2018

Acknowledgment

1. **Neil D. Lawrence** (2015), Machine and Intelligence Course, University of Sheffield
2. **Kevin P. Murphy** (2012), Machine Learning, A Probabilistic Perspective, Massachusetts Institute of Technology

What is Machine Learning?

$$\textit{data} \quad + \quad \textit{model} \quad \underrightarrow{\textit{compute}} \quad \textit{prediction}$$

- **data:** observations, could be actively or passively acquired (meta-data).

What is Machine Learning?

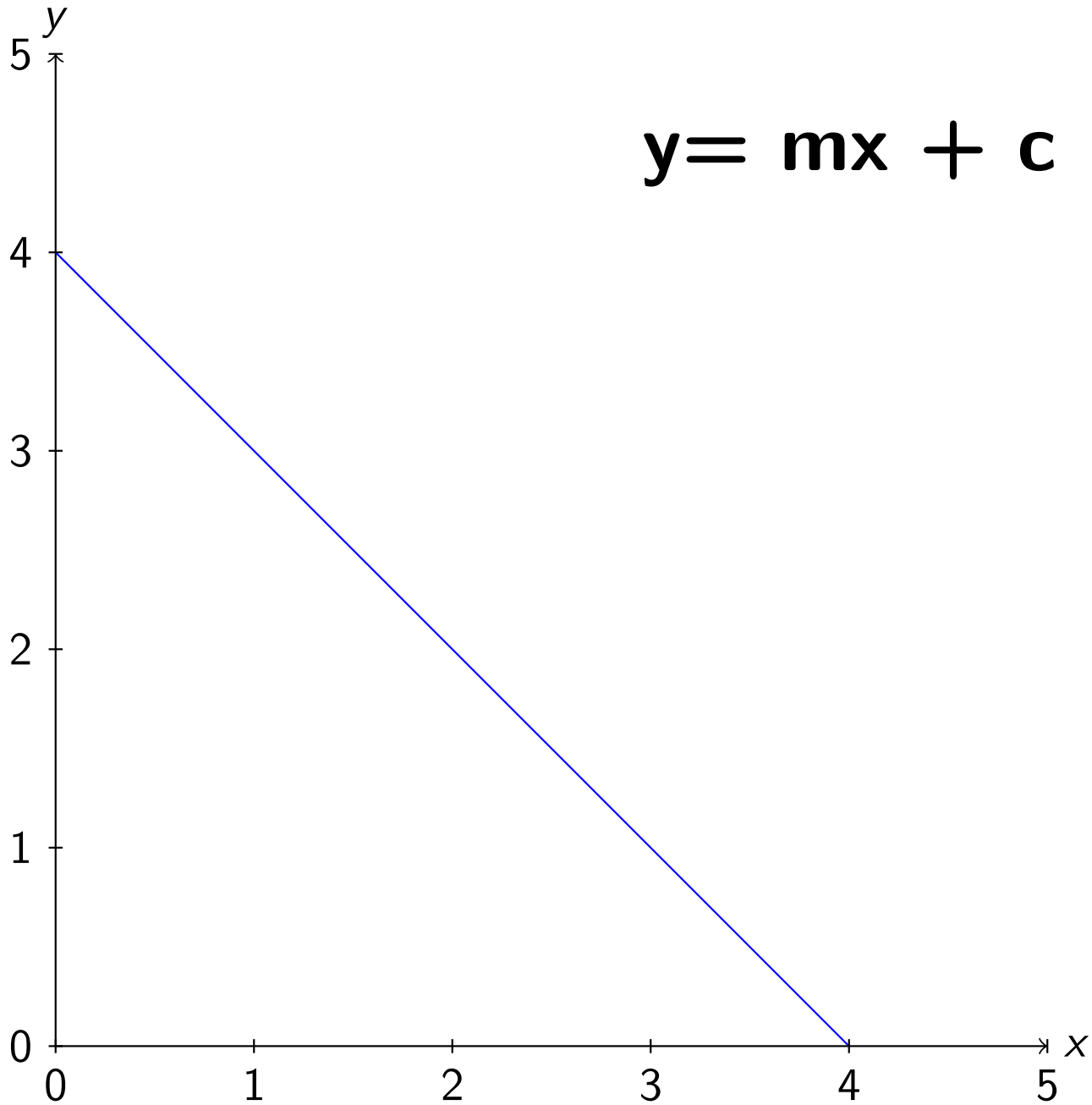
$$\textit{data} \quad + \quad \textit{model} \quad \underrightarrow{\textit{compute}} \quad \textit{prediction}$$

- **data:** observations, could be actively or passively acquired (meta-data).
- **model:** assumptions, based on previous experience (other data! transfer learning etc), or beliefs about the regularities of the universe. Inductive bias.

What is Machine Learning?

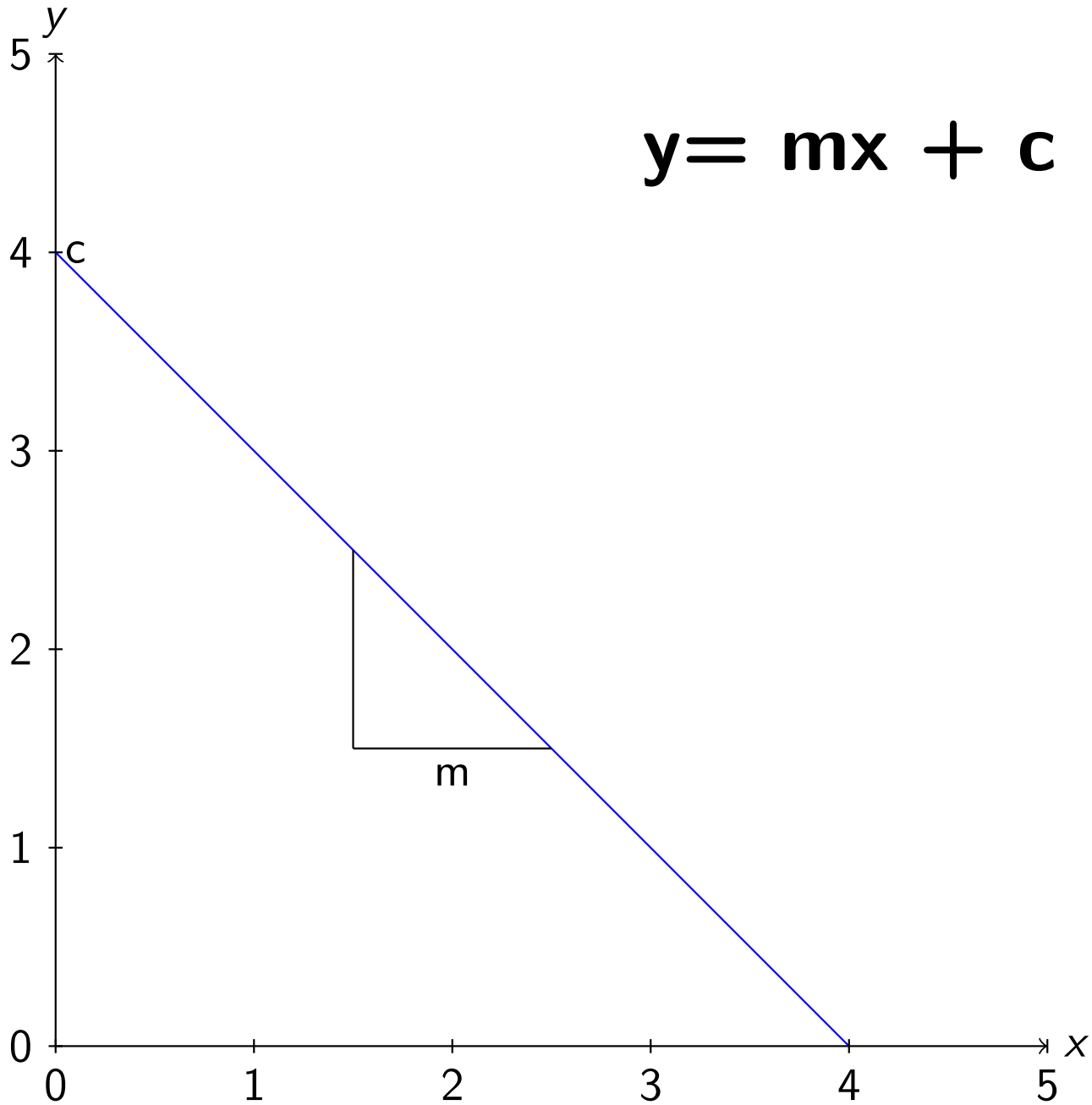
$$\textit{data} \quad + \quad \textit{model} \quad \underrightarrow{\textit{compute}} \quad \textit{prediction}$$

- **data:** observations, could be actively or passively acquired (meta-data).
- **model:** assumptions, based on previous experience (other data! transfer learning etc), or beliefs about the regularities of the universe. Inductive bias.
- **prediction:** an action to be taken or a categorization or a quality score.

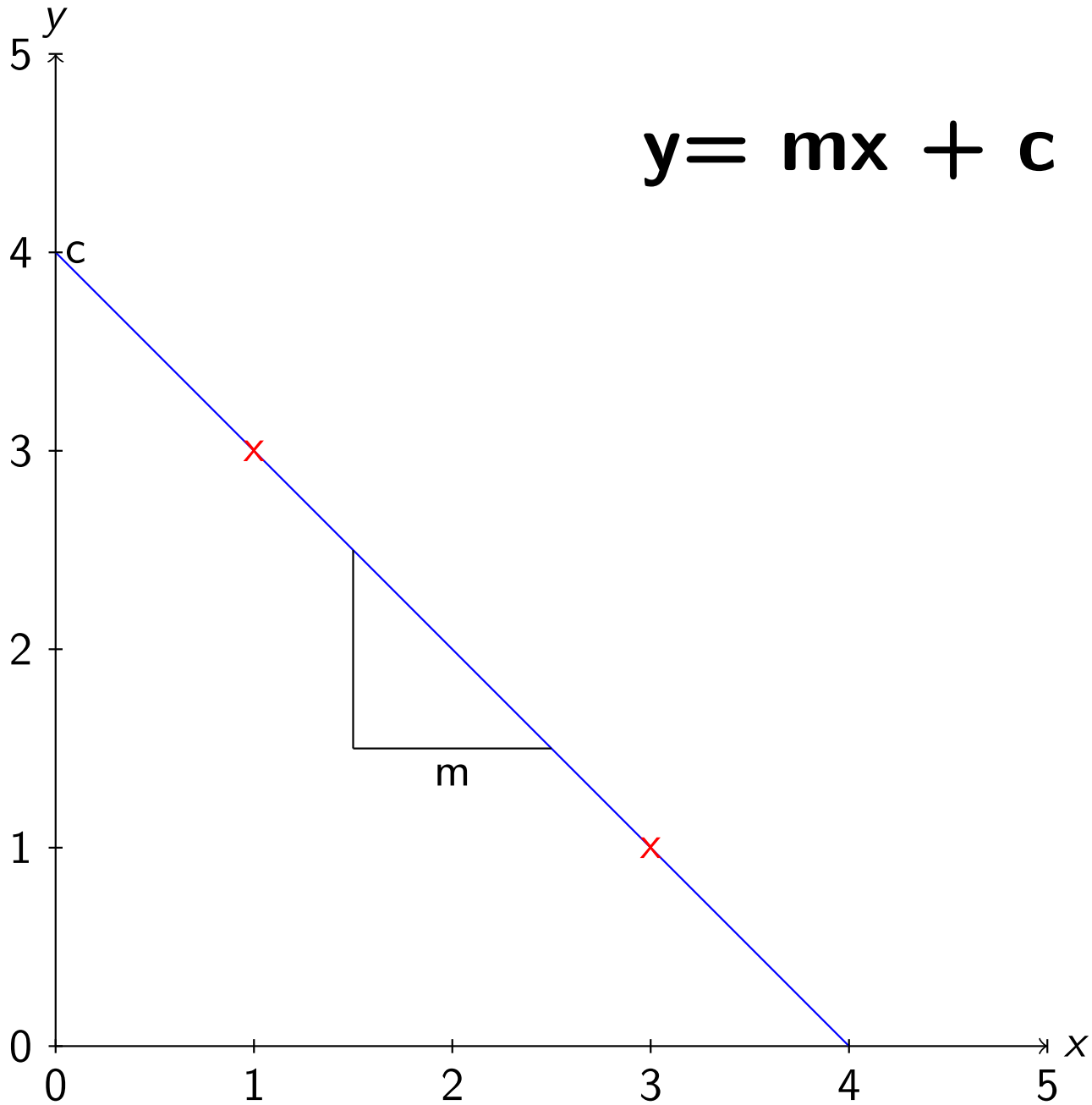


y= mx + c

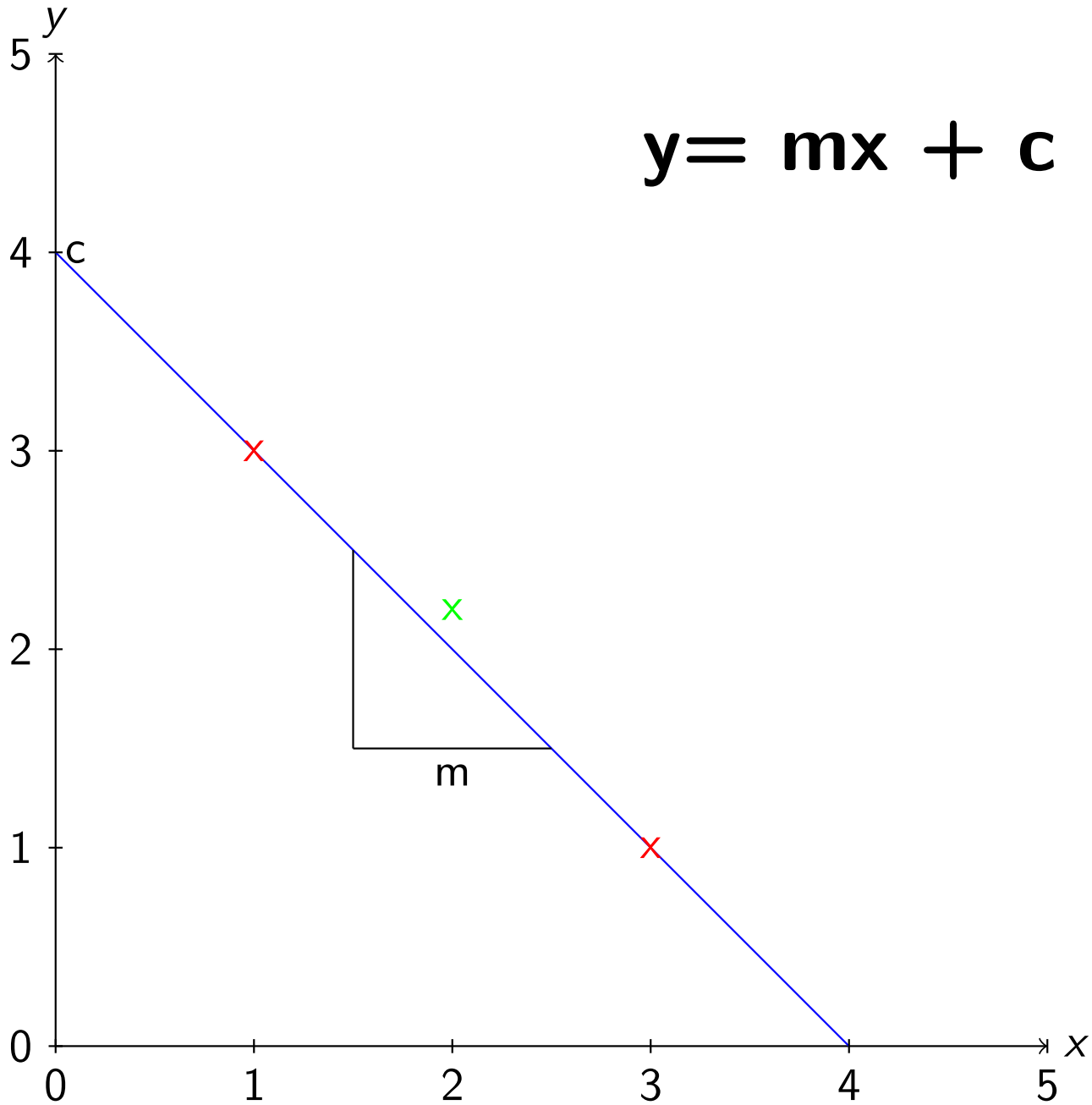
y= mx + c

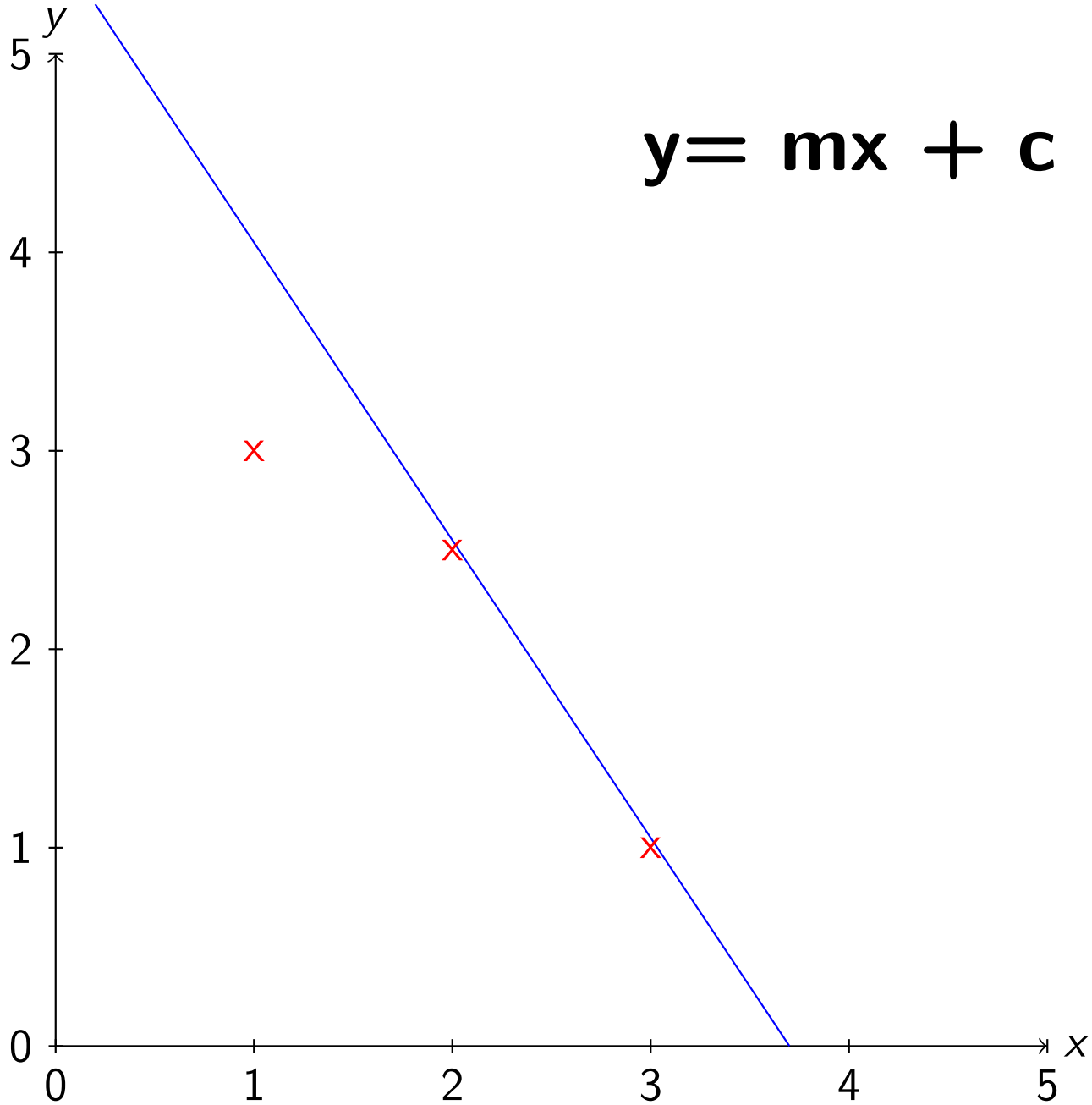


y= mx + c

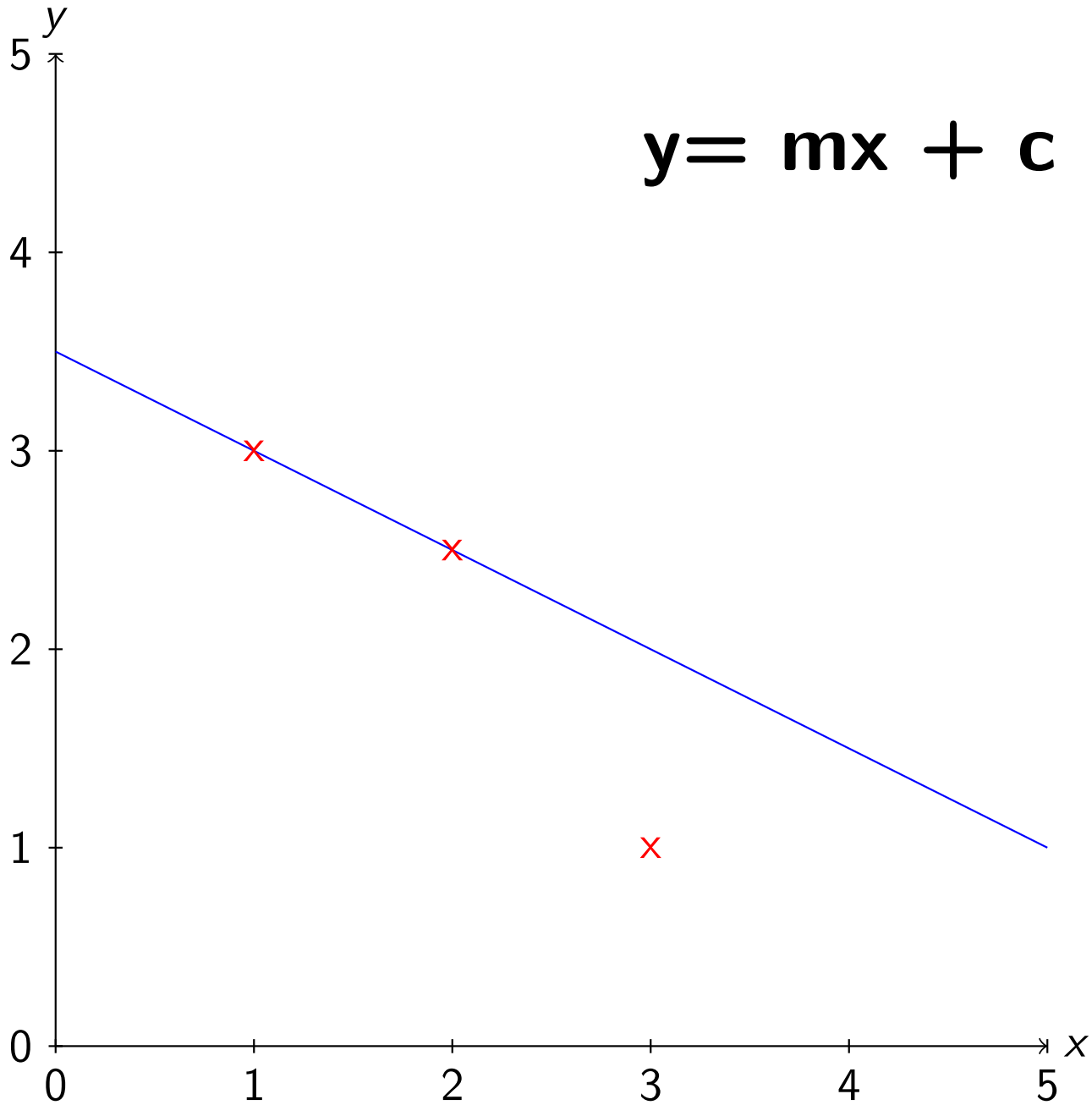


y= mx + c

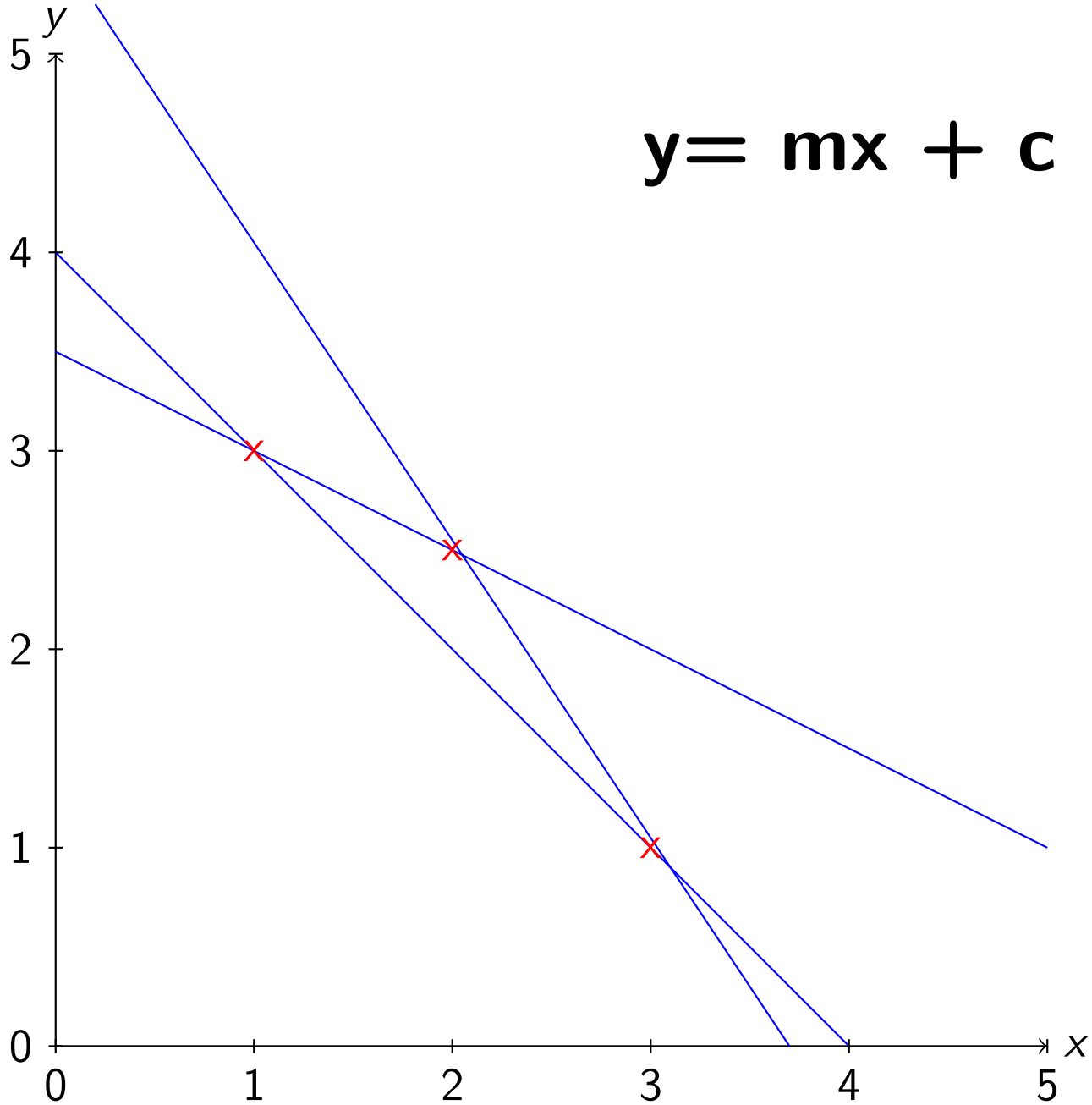




y= mx + c



y= mx + c



y= mx + c

$y = mx + c$

$$
\begin{aligned}
point\,1 : x = 1, y = 3 \\
3 = m + c \\
point\,2 : x = 3, y = 1 \\
1 = 3m + c \\
point\,3 : x = 2, y = 2.5 \\
2.5 = 2m + c
\end{aligned}
$$

other, we say that its choice is an effect without a cause. It is then, says Leibnitz, the blind chance of the Epicureans. The contrary opinion is an illusion of the mind, which, losing sight of the evasive reasons of the choice of the will in indifferent things, believes that choice is determined of itself and without motives.

We ought then to regard the present state of the universe as the effect of its anterior state and as the cause of the one which is to follow. Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it—an intelligence sufficiently vast to submit these data to analysis—it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes. The human mind offers, in the perfection which it has been able to give to astronomy, a feeble idea of this intelligence. Its discoveries in mechanics and geometry, added to that of universal gravity, have enabled it to comprehend in the same analytical expressions the past and future states of the system of the world. Applying the same method to some other objects of its knowledge, it has succeeded in referring to general laws observed phenomena and in foreseeing those which given circumstances ought to produce. All these efforts

height: "The day will come when, by study pursued through several ages, the things now concealed will appear with evidence; and posterity will be astonished that truths so clear had escaped us." Clairaut then undertook to submit to analysis the perturbations which the comet had experienced by the action of the two great planets, Jupiter and Saturn; after immense calculations he fixed its next passage at the perihelion toward the beginning of April, 1759, which was actually verified by observation. The regularity which astronomy shows us in the movements of the comets doubtless exists also in all phenomena.

The curve described by a simple molecule of air or vapor is regulated in a manner just as certain as the planetary orbits; the only difference between them is that which comes from our ignorance.

Probability is relative, in part to this ignorance, in part to our knowledge. We know that of three or a greater number of events a single one ought to occur; but nothing induces us to believe that one of them will occur rather than the others. In this state of indecision it is impossible for us to announce their occurrence with

$$y = mx + c + \epsilon$$

$$\begin{aligned}
point\,1 : x = 1, y = 3 \\
3 = m + c + \epsilon_1 \\
point\,2 : x = 3, y = 1 \\
1 = 3m + c + \epsilon_2 \\
point\,3 : x = 2, y = 2.5 \\
2.5 = 2m + c + \epsilon_3
\end{aligned}$$

Probability Review I

- We are interested in trials which result in two random variables, **X** and **Y**, each of which has an 'outcome' denoted by x or y.
- We summarize the notation and terminology for these distributions in the following table:

Terminology	**Notation**	**Description**
Joint Probability	$P(X = x, Y = y)$	The probability that $X = x$ and $Y = y$
Marginal Probability	$P(X = x)$	The probability that $X = x$ regardless of Y
Conditional Probability	$P(X = x \mid Y = y)$	The probability that $X = x$ given that $Y = y$

Table: The different basic probability distributions

Different Distributions

Terminology	**Notation**	**Description**
Joint Probability	$\lim_{N\to\infty} \frac{n_{X=3,Y=4}}{N}$	$P(X=3, Y=4)$
Marginal Probability	$\lim_{N\to\infty} \frac{n_{X=5}}{N}$	$P(X=5)$
Conditional Probability	$\lim_{N\to\infty} \frac{n_{X=3,Y=4}}{n_{Y=4}}$	$P(X=3\|Y=4)$

Table: Definition of probability distributions

Notational Details

- Typically we should write out $P(X = x, Y = y)$.
- In practice, we often use $P(x, y)$.
- This looks very much like we might write a multivariate function, e.g. $f(x, y) = \frac{x}{y}$
  - For a multivariate function though, $f(x, y) \neq f(y, x)$.
  - However $P(x, y) = P(y, x)$ because $P(X = x, Y = y) = P(Y = y, X = x)$
- We now quickly review the 'rules of probability'.

Normalization

All distributions are normalized. This is clear from the fact that $\sum_x n_x = N$, which gives

$$\sum_x P(x) = \frac{\sum_x n_x}{N} = \frac{N}{N} = 1$$

A similar result can be derived for the marginal and conditional distributions.

The Sum Rule

Ignoring the limit in our definitions:

- The marginal probability $P(y)$ is $\frac{n_y}{N}$ (ignoring the limit).
- The joint distribution $P(x, y)$ is $\frac{n_{x,y}}{N}$.
- $n_y = \sum_x n_{x,y}$ so

$$\frac{n_y}{N} = \sum_x \frac{n_{x,y}}{N}$$

in other words

$$P(y) = \sum_x P(x, y)$$

This is known as the sum rule of probability.

The Product Rule

- $P(x|y)$ is

$$\frac{n_{x,y}}{n_y}$$

.

- $P(x, y)$ is

$$\frac{n_{x,y}}{N} = \frac{n_{x,y}}{n_y}\frac{n_y}{N}$$

or in other words

$$P(x, y) = P(x|y)P(y)$$

This is known as the product rule of probability.

Bayes' Rule

From the product rule,

$$P(y, x) = P(x, y) = P(x|y)P(y),$$

so

$$P(y|x)P(x) = P(x|y)P(y)$$

which leads to Bayes' rule,

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Bayes' Theorem Example

- There are two barrels in front of you. Barrel One contains 20 apples and 4 oranges. Barrel Two other contains 4 apples and 8 oranges. You choose a barrel randomly and select a fruit. It is an apple. What is the probability that the barrel was Barrel One?

Bayes' Theorem Example: Answer I

We are given that:

$$P(F = A|B = 1) = 20/24$$
$$P(F = A|B = 2) = 4/12$$
$$P(B = 1) = 0.5$$
$$P(B = 2) = 0.5$$

Bayes Theorem Example: Answer II

- We use the sum rule to compute:

$$\begin{aligned} P(F = A) &= P(F = A|B = 1)P(B = 1) \\ &+P(F = A|B = 2)P(B = 2) \\ &= \frac{20}{24}x0.5 + \frac{4}{12}x0.5 = \frac{7}{12} \end{aligned}$$

Bayes Theorem Example: Answer II

- We use the sum rule to compute:

$$\begin{aligned} P(F = A) &= P(F = A|B = 1)P(B = 1) \\ &+P(F = A|B = 2)P(B = 2) \\ &= \frac{20}{24}x0.5 + \frac{4}{12}x0.5 = \frac{7}{12} \end{aligned}$$

- And Bayes theorem tells us that:

$$\begin{aligned} P(B = 1|F = A) &= \frac{P(F = A|B = 1)P(B = 1)}{P(F = A)} \\ &= \frac{\frac{20}{24}x0.5}{\frac{7}{12}} \\ &= \frac{5}{7} \end{aligned}$$

Reading & Exercises

Review the example on Bayes Theorem!

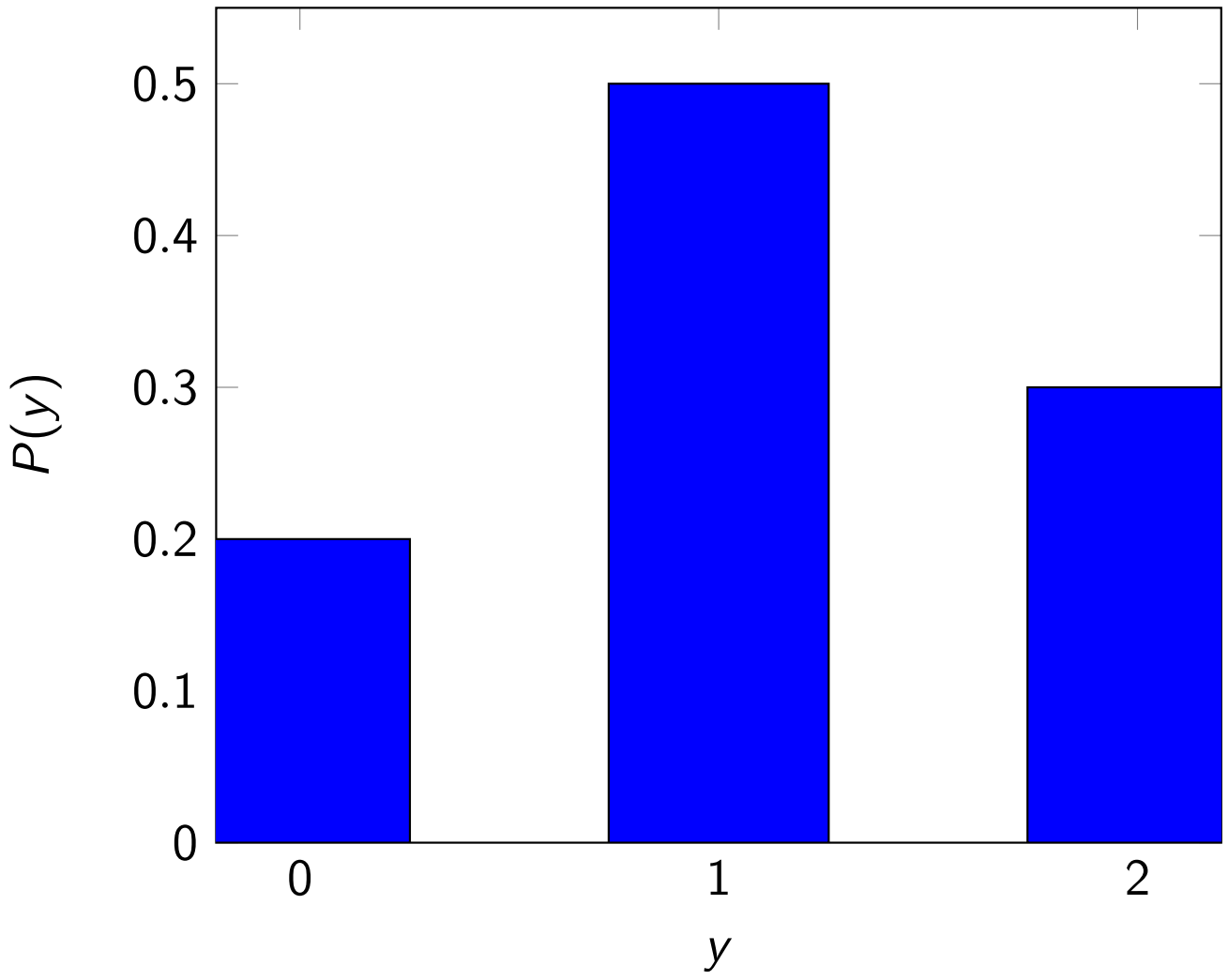
- Read and *understand* Bishop on probability distributions: page 12–20 (Section 1.2).
- Complete Exercise 1.3 in Bishop (page 58).

Distribution Representation

We can represent probabilities as tables

y	0	1	2
P(y)	0.2	0.5	0.3

Histogram



Histogram representation of the simple distribution.

Expectations of Distributions

- Writing down the entire distribution is tedious.
- Can summarize through expectations.

$$\langle f(y) \rangle_{P(y)} = \sum_y f(y) p(y)$$

- Consider:

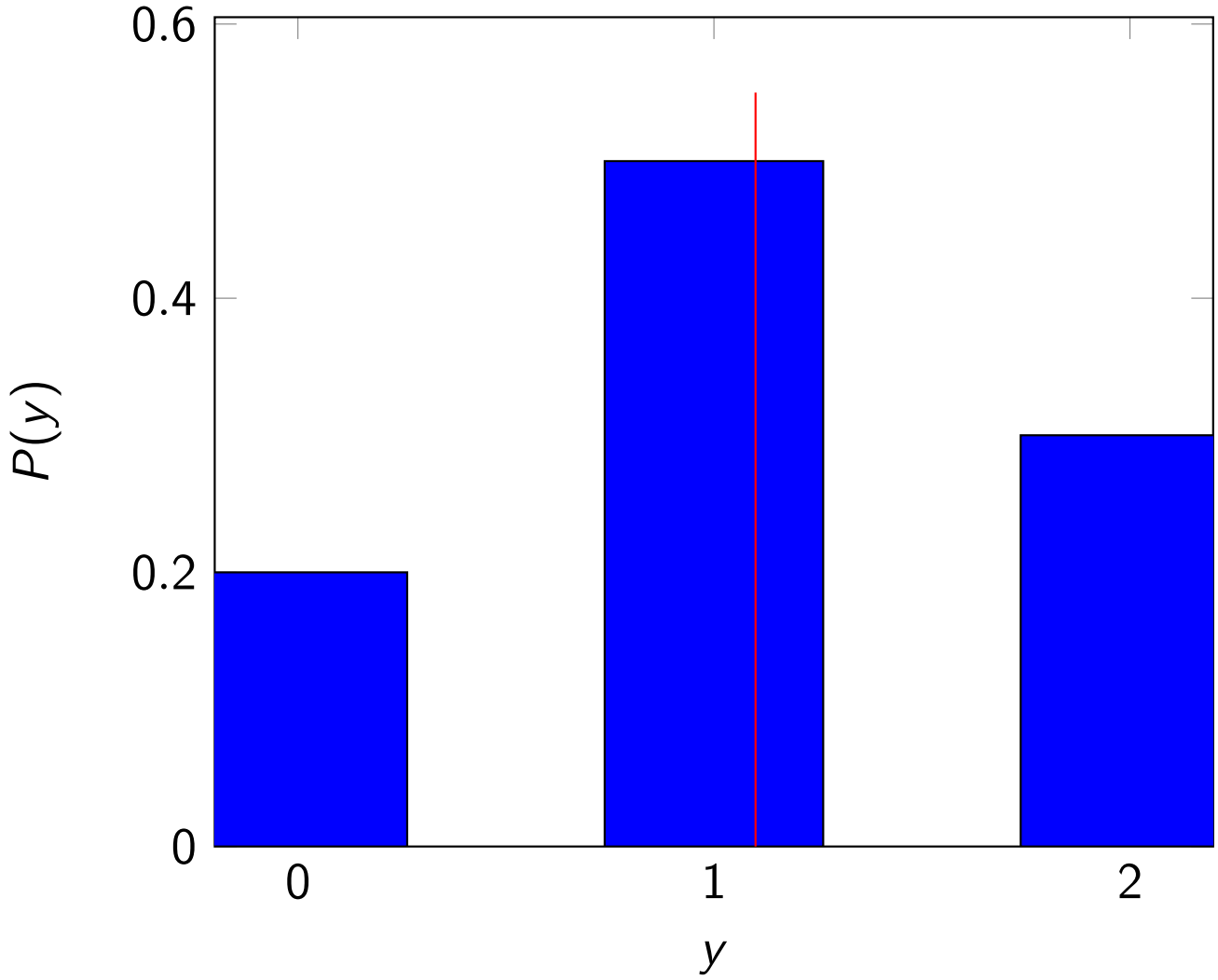
y	0	1	2
P(y)	0.2	0.5	0.3

- We have

$$\langle y \rangle_{P(y)} = 0.2x0 + 0.5x1 + 0.3x2 = 1.1$$

- This is the *first moment* or mean of the distribution.

Histogram



Histogram representation of the simple distribution including the expectation of y (red line), the mean of the distribution.

Variance and Standard Deviation

- Mean gives us the centre of the distribution.
- Consider:

y	0	1	2
y^2	0	1	4
$P(y)$	0.2	0.5	0.3

- Second moment is

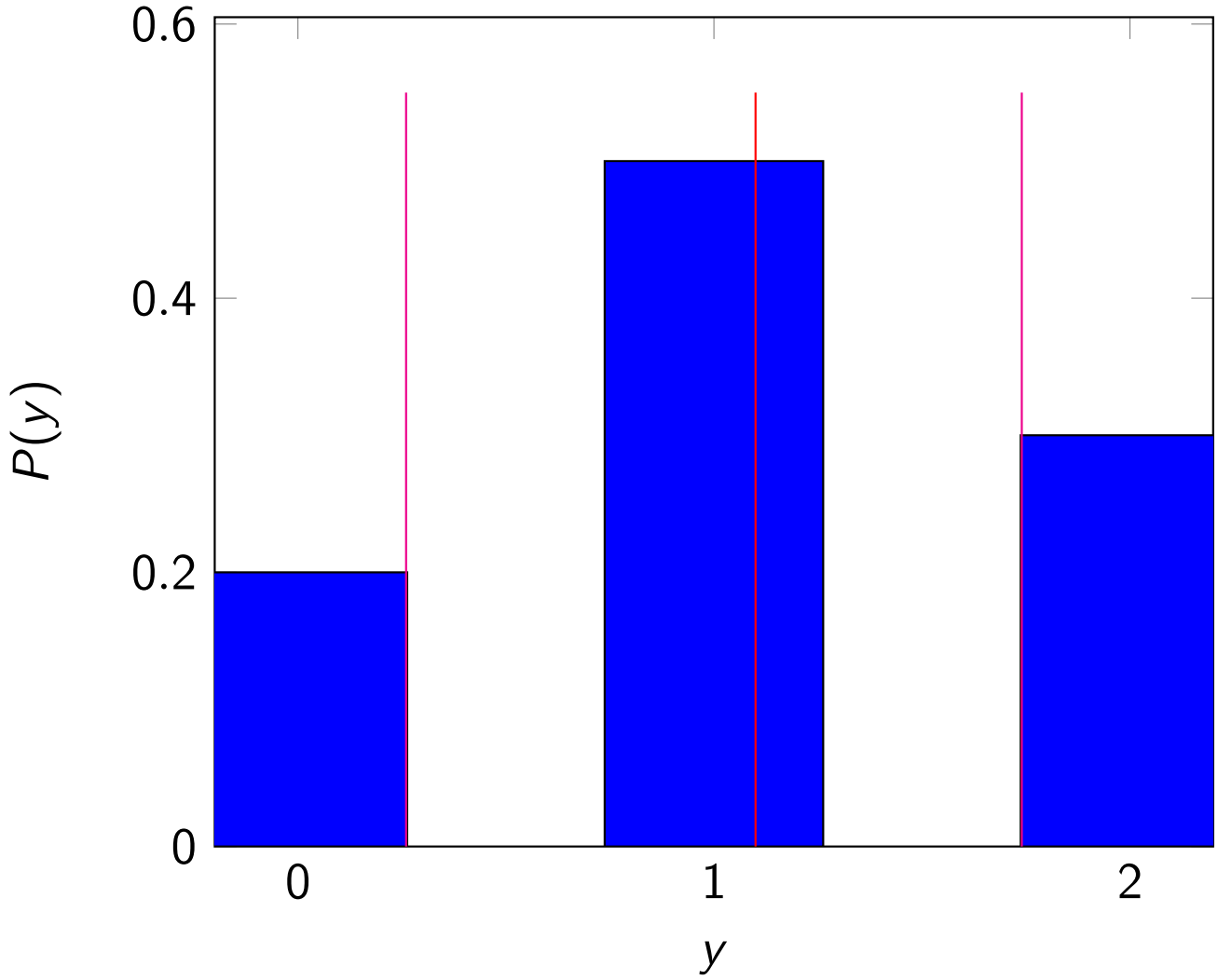
$$\langle y^2 \rangle_{P(y)} = 0.2x0 + 0.5x1 + 0.3x4 = 1.7$$

- Variance is

$$\langle y^2 \rangle - \langle y \rangle^2 = 1.7 - 1.1 \times 1.1 = 0.49$$

- Standard deviation is square root of variance.
- Standard deviation gives us the 'width' of the distribution.

Histogram



Histogram representation of the simple distribution including lines at one standard deviation from the mean of the distribution (magenta lines).

Expectation Computation Example

- Consider the following distribution.

y	1	2	3	4
$P(y)$	0.3	0.2	0.1	0.4

- What is the mean of the distribution?

Expectation Computation Example

- Consider the following distribution.

y	1	2	3	4
$P(y)$	0.3	0.2	0.1	0.4

- What is the mean of the distribution?
- What is the standard deviation of the distribution?

Expectation Computation Example

- Consider the following distribution.

y	1	2	3	4
$P(y)$	0.3	0.2	0.1	0.4

- What is the mean of the distribution?
- What is the standard deviation of the distribution?
- Are the mean and standard deviation representative of the distribution form?

Expectation Computation Example

- Consider the following distribution.

y	1	2	3	4
$P(y)$	0.3	0.2	0.1	0.4

- What is the mean of the distribution?
- What is the standard deviation of the distribution?
- Are the mean and standard deviation representative of the distribution form?
- What is the expected value of $-log \quad P(y)$?

Expectations Example: Answer

- We are given that:

y	1	2	3	4
$P(y)$	0.3	0.2	0.1	0.4
y^2	1	4	9	16
$-log(P(y))$	1.204	1.609	2.302	0.916

- Mean:

$$1x0.3 + 2x0.2 + 3x0.1 + 4x0.4 = 2.6$$

- Second moment:

$$1x0.3 + 4x0.2 + 9x0.1 + 16x0.4 = 8.4$$

- Variance:

$$8.4 - 2.6x2.6 = 1.64$$

- Standard deviation:

$$\sqrt{1.64} = 1.2806$$

- Expectation

$$-log(P(y)) : 0.3x1.204 + 0.2x1.609 + 0.1x2.302 + 0.4x0.916 = 1.280$$

Sample Based Approximation Example

- You are given the following values samples of heights of students,

i	1	2	3	4	5	6
y_i	1.76	1.73	1.79	1.81	1.85	1.80

- What is the sample mean?

Sample Based Approximation Example

- You are given the following values samples of heights of students,

i	1	2	3	4	5	6
y_i	1.76	1.73	1.79	1.81	1.85	1.80

- What is the sample mean?
- What is the sample variance?

Sample Based Approximation Example

- You are given the following values samples of heights of students,

i	1	2	3	4	5	6
y_i	1.76	1.73	1.79	1.81	1.85	1.80

- What is the sample mean?
- What is the sample variance?
- Can you compute sample approximation expected value of $-logP(y)$?
- Actually these 'data' were sampled from a Gaussian with mean 1.7 and standard deviation 0.15. Are your estimates close to the real values? If not why not?

Sample Based Approximation Example: Answer

- We can compute:

i	1	2	3	4	5	6
y_i	1.76	1.73	1.79	1.81	1.85	1.80
y_i^2	3.0976	2.9929	3.0204	3.2761	3.4225	3.2400

- Mean:

$$\frac{1.76 + 1.73 + 1.79 + 1.81 + 1.85 + 1.80}{6} = 1.79$$

- Second moment:

$$\frac{3.0976 + 2.9929 + 3.2041 + 3.2761 + 3.4225 + 3.24}{6} = 3.2055$$

- Variance:

$$3.20551.79x1.79 = 1.43x10^{-3}$$

- Standard deviation: 0.0379
- No, you can't compute it. You don't have access to $P(y)$ directly.

Reading

Read and *understand* Rogers and Girolami on:

1. Section 2.2 (pg 41–53).
2. Section 2.4 (pg 55–58).
3. Section 2.5.1 (pg 58–60).
4. Section 2.5.3 (pg 61–62).

If you are unfamiliar with probabilities you should complete the following exercises:

1. Bishop Exercise 1.7
2. Bishop Exercise 1.8
3. Bishop Exercise 1.9

For other material in Bishop read:

1. Probability densities: Section 1.2.1 (Pages 17–19).
2. Expectations and Covariances: Section 1.2.2 (Pages 19–20).
3. The Gaussian density: Section 1.2.4 (Pages 24–28)