FIT2086 Lecture 2
Probability and Probability Distributions

Daniel F. Schmidt

Faculty of Information Technology, Monash University

July 30, 2018

Outline

Outline

Some important notation – refresher

- We will use several bits of set notation in this lecture
  - We use $\{a, b, c\}$ to denote a set with elements a, b and c
  - We use $x \in \mathcal{X}$ to denote that x is an element of the set $\mathcal{X}$
    - Example: $3 \in \{1, 2, 3, 4, 5\}$
  - We use $A \subseteq \mathcal{X}$ to denote that A is a subset of the set $\mathcal{X}$
    - Example: $\{2, 3, 4\} \subseteq \{1, 2, 3, 4, 5\}$
- Some important sets:
  - $\mathbb{Z}$ is the set of all integers;
  - $\mathbb{Z}_+$ is the set of non-negative integers;
  - $\mathbb{R}$ is the set of all real numbers;
  - $\mathbb{R}_+$ is the set of non-negative numbers.

Random Variables (1)

- A random variable (RV) is a variable that takes on a value from a set of possible values with specified probabilities
  - We can let $\mathcal{X}$ denote the possible set of values
  - For now, let's just consider cases where $\mathcal{X}$ is discrete
- We often use capital letters to denote a random variable
- Example: let X be a random variable over $\mathcal{X} = \{1, 2, 3\}$ with:

$$X = \begin{cases} 1 & \text{with probability } 1/2 \\ 2 & \text{with probability } 1/4 \\ 3 & \text{with probability } 1/4 \end{cases},$$

Random Variables (2)

- A *realisation* of a random variable is a particular value from $\mathcal{X}$ drawn at random
- Consider our example distribution over $\mathcal{X} = \{1, 2, 3\}$ with:

$$X = \begin{cases} 1 & \text{with probability } 1/2 \\ 2 & \text{with probability } 1/4 \\ 3 & \text{with probability } 1/4 \end{cases},$$

- Twenty-two sample realisations are:

$$3, 3, 1, 3, 2, 1, 1, 1, 2, 3, 3, 2, 1, 3, 3, 2, 1, 2, 1, 2, 1, 1$$

- There are nine 1s, six 2s and seven 3s
  - We would expect 1s to appear more frequently the more realisations we take

Random Variables (2)

- A *realisation* of a random variable is a particular value from $\mathcal{X}$ drawn at random
- Consider our example distribution over $\mathcal{X} = \{1, 2, 3\}$ with:

$$X = \begin{cases} 1 & \text{with probability } 1/2 \\ 2 & \text{with probability } 1/4 \\ 3 & \text{with probability } 1/4 \end{cases},$$

- Twenty-two sample realisations are:

$$3, 3, 1, 3, 2, 1, 1, 1, 2, 3, 3, 2, 1, 3, 3, 2, 1, 2, 1, 2, 1, 1$$

- There are nine 1s, six 2s and seven 3s
  - We would expect 1s to appear more frequently the more realisations we take

Probability Distributions (1)

- We use the language of probability distributions to describe random variables
- The notation
$$\mathbb{P}(X = x),\ x \in \mathcal{X}$$
describes the probability that the RV X takes on the value x from $\mathcal{X}$.
- We can use this notation to describe the example random variable X from the previous slides
$$\mathbb{P}(X = 1) = 1/2,\ \ \mathbb{P}(X = 2) = 1/4,\ \ \mathbb{P}(X = 3) = 1/4$$

Probability Distributions (2)

- Review of facts regarding probability distributions
- Fact 1: A probability distribution satisfies:

$$\mathbb{P}(X = x) \in [0, 1] \text{ for all } x \in \mathcal{X}$$

and

$$\sum_{x \in \mathcal{X}} \mathbb{P}(X = x) = 1$$

Probability Distributions (3)

- Fact 2: The probability of $(X \in A_1 \text{ OR } X \in A_2)$, with $A_1, A_2 \subset \mathcal{X}$ and $A_1 \cap A_2 = \emptyset$ is

$$\mathbb{P}(X \in A_1 \cup A_2) = \mathbb{P}(X \in A_1) + \mathbb{P}(X \in A_2),$$

with "$\cap$" set intersection, "$\cup$" set union and $\emptyset$ the empty set.

- Example: If X follows the probability distribution

$$\mathbb{P}(X = 1) = 1/2, \quad \mathbb{P}(X = 2) = 1/4, \quad \mathbb{P}(X = 3) = 1/4$$

then $\mathbb{P}(X \geq 2)$ is

$$\begin{aligned} \mathbb{P}(X \in \{2\} \cup \{3\}) &= \mathbb{P}(X = 2) + \mathbb{P}(X = 3) \\ &= 1/4 + 1/4 \\ &= 1/2 \end{aligned}$$

Probability Distributions (3)

- Fact 2: The probability of $(X \in A_1 \text{ OR } X \in A_2)$, with $A_1, A_2 \subset \mathcal{X}$ and $A_1 \cap A_2 = \emptyset$ is

$$\mathbb{P}(X \in A_1 \cup A_2) = \mathbb{P}(X \in A_1) + \mathbb{P}(X \in A_2),$$

with "$\cap$" set intersection, "$\cup$" set union and $\emptyset$ the empty set.

- Example: If X follows the probability distribution

$$\mathbb{P}(X = 1) = 1/2, \;\; \mathbb{P}(X = 2) = 1/4, \;\; \mathbb{P}(X = 3) = 1/4$$

then $\mathbb{P}(X \geq 2)$ is

$$\begin{aligned} \mathbb{P}(X \in \{2\} \cup \{3\}) &= \mathbb{P}(X = 2) + \mathbb{P}(X = 3) \\ &= 1/4 + 1/4 \\ &= 1/2 \end{aligned}$$

Probability Distributions of Two RVs (1)

- Now let us consider the case of two RVs $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$
  - $\mathcal{X}$ and $\mathcal{Y}$ are the sets of values X and Y can take, respectively
  - $\mathcal{X} \times \mathcal{Y}$ is the set of values the pair can assume

- Example: If $\mathcal{X} = \{1, 2, 3\}$ and $\mathcal{Y} = \{1, 2\}$, then

$$\mathcal{X} \times \mathcal{Y} = \{\{1,1\}, \{2,1\}, \{3,1\}, \{1,2\}, \{2,2\}, \{3,2\}\}$$

- Example: An example distribution over $\mathcal{X} \times \mathcal{Y}$:

	$X = 1$	$X = 2$	$X = 3$
$Y = 1$	0.05	0.15	0.1
$Y = 2$	0.25	0.15	0.3

Probability Distributions of Two RVs (1)

- Now let us consider the case of two RVs $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$
  - $\mathcal{X}$ and $\mathcal{Y}$ are the sets of values X and Y can take, respectively
  - $\mathcal{X} \times \mathcal{Y}$ is the set of values the pair can assume
- Example: If $\mathcal{X} = \{1, 2, 3\}$ and $\mathcal{Y} = \{1, 2\}$, then

$$\mathcal{X} \times \mathcal{Y} = \{\{1,1\},\{2,1\},\{3,1\},\{1,2\},\{2,2\},\{3,2\}\}$$

- Example: An example distribution over $\mathcal{X} \times \mathcal{Y}$:

	$X = 1$	$X = 2$	$X = 3$
$Y = 1$	0.05	0.15	0.1
$Y = 2$	0.25	0.15	0.3

Probability Distributions of Two RVs (2)

- We can define a probability distribution over (X, Y) as before:

$$\mathbb{P}(X = x, Y = y) \in [0, 1] \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y}$$

which satisfies

$$\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y) = 1$$

- $\mathbb{P}(X = x, Y = y)$ is the *joint* probability of $X = x$ and $Y = y$
    - That is, the probability of $X = x$ AND $Y = y$

- Example: The example distribution from previous slide

$$\begin{aligned} \mathbb{P}(X = 1, Y = 1) &= 0.05 \\ \mathbb{P}(X = 1, Y = 2) &= 0.25 \\ \mathbb{P}(X = 2, Y = 1) &= 0.15 \end{aligned}$$

and so on.

Probability Distributions of Two RVs (2)

- We can define a probability distribution over (X, Y) as before:

$$\mathbb{P}(X = x, Y = y) \in [0, 1] \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y}$$

which satisfies

$$\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y) = 1$$

- $\mathbb{P}(X = x, Y = y)$ is the *joint* probability of $X = x$ and $Y = y$
  - That is, the probability of $X = x$ AND $Y = y$

- Example: The example distribution from previous slide

$$\begin{aligned}
\mathbb{P}(X = 1, Y = 1) &= 0.05 \\
\mathbb{P}(X = 1, Y = 2) &= 0.25 \\
\mathbb{P}(X = 2, Y = 1) &= 0.15
\end{aligned}$$

and so on.

The Sum Rule (1)

The Sum Rule

The sum rule is given by:

$$\mathbb{P}(X = x) = \sum_{y \in \mathcal{Y}} P(X = x, Y = y)$$

The probability $\mathbb{P}(X = x)$ is called the *marginal* probability.

- The marginal probability $\mathbb{P}(X = x)$ is the probability of seeing $X = x$ irrespective of what value Y takes on

The Sum Rule (2)

- Example:

	$X=1$	$X=2$	$X=3$
$Y=1$	0.05	0.15	0.1
$Y=2$	0.25	0.15	0.3

- Then

$$\begin{aligned}\mathbb{P}(Y=1) &= 0.05+0.15+0.1=0.3\\ \mathbb{P}(Y=2) &= 0.25+0.15+0.3=0.7\end{aligned}$$

so that the probability of seeing a $Y=2$ is significantly higher than the probability of seeing a $Y=1$, irrespective of the value of X.

Conditional Probability (1)

Conditional Probability

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

The probability $\mathbb{P}(X = x \,|\, Y = y)$ is called the probability of $X = x$, conditional on $Y = y$.

- The conditional probability $\mathbb{P}(X = x \,|\, Y = y)$ is the (joint) probability of seeing $X = x$ and $Y = y$, divided by the (marginal) probability that we have observed $Y = y$.

Conditional Probability (2)

- Example:

	$X=1$	$X=2$	$X=3$
$Y=1$	0.05	0.15	0.1
$Y=2$	0.25	0.15	0.3

- Then

$$\begin{aligned}\mathbb{P}(X=1\,|\,Y=1) &= \mathbb{P}(X=1,Y=1)/\mathbb{P}(Y=1)\\ &= 0.05/0.3 \approx 0.1667\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(X=1\,|\,Y=2) &= \mathbb{P}(X=1,Y=2)/\mathbb{P}(Y=2)\\ &= 0.25/0.7 \approx 0.3571\end{aligned}$$

so that seeing $X=1$ is twice as likely when $Y=2$ as compared to the case that $Y=1$.

Independent Random Variables (4)

- Independent random variables are very important
- X and Y are considered independent if

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$$

  for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$.
- This implies that

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

  $\Rightarrow$ Knowing about Y tells us nothing new about X
- An even more special class are independent and identically distributed (i.i.d.) random variables
  - $X_1 \in \mathcal{X}$, $X_2 \in \mathcal{X}$ are i.i.d. if they are independent and

$$\mathbb{P}(X_1 = x) = \mathbb{P}(X_2 = x) \text{ for all } x \in \mathcal{X}$$

Independent Random Variables (4)

- Independent random variables are very important
- X and Y are considered independent if

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$$

  for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$.

- This implies that

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

  $\Rightarrow$ Knowing about Y tells us nothing new about X

- An even more special class are independent and identically distributed (i.i.d.) random variables
  - $X_1 \in \mathcal{X}$, $X_2 \in \mathcal{X}$ are i.i.d. if they are independent and $\mathbb{P}(X_1 = x) = \mathbb{P}(X_2 = x)$ for all $x \in \mathcal{X}$

Continuous Random Variables (1)

- So far we have considered only discrete random variables
- The ideas extend to the case that the values X can take on form a continuum, that is, $\mathcal{X} \subseteq \mathbb{R}$
- X now follows a probability density function (pdf) $p(x)$.
- A pdf satisfies:

$$p(x) \geq 0 \text{ for all } x \in \mathcal{X}$$

and

$$\int_{\mathcal{X}} p(x) dx = 1$$

Continuous Random Variables (2)

- The probability that X lies in an interval (a, b) is

$$\mathbb{P}(a < X < b) = \int_a^b p(x)dx.$$

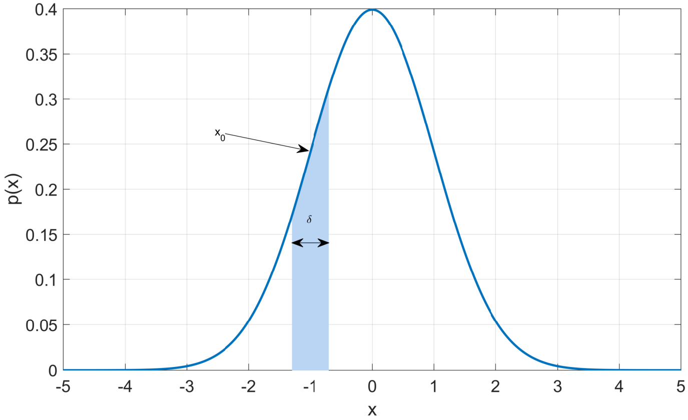
- More generally, the probability $X \in A$, where $A \subset \mathcal{X}$ is

$$\mathbb{P}(X \in A) = \int_A p(x)dx.$$

- This implies that $\mathbb{P}(X = x) = 0$
  $\Rightarrow$ One of the most confusing aspects of continous RVs

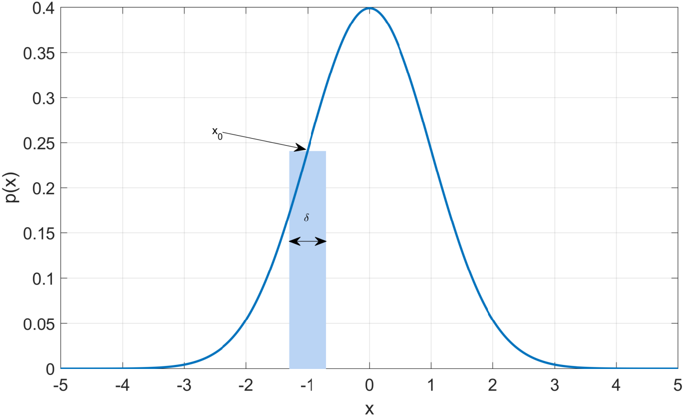
Continuous Random Variables (3)

- Example: Probability of $(x_0 - \delta/2 < X < x_0 + \delta/2)$



Continuous Random Variables (4)

- If δ is small enough then $\int_{x_0-\delta/2}^{x_0+\delta/2} p(x)dx \approx p(x_0)\delta$
  $\Rightarrow$ Take $\delta \to 0$ and it is clear why $\mathbb{P}(X = x) = 0$.



Cumulative Distribution Functions (1)

- The cumulative distribution function (cdf) of a continuous RV is:
$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} p(x')dx'$$
that is, the probability that X is less than some value x
- Let's introduce some shorthand notation for discrete RVs:
$$\mathbb{P}(X = x) \equiv p(x)$$
- Then, if X is a discrete RV over the integers (or a subset)
$$\mathbb{P}(X \leq x) = \sum_{x' \leq x} p(x')$$
- It follows that
$$\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x)$$

Cumulative Distribution Functions (2)

- The inverse cdf is

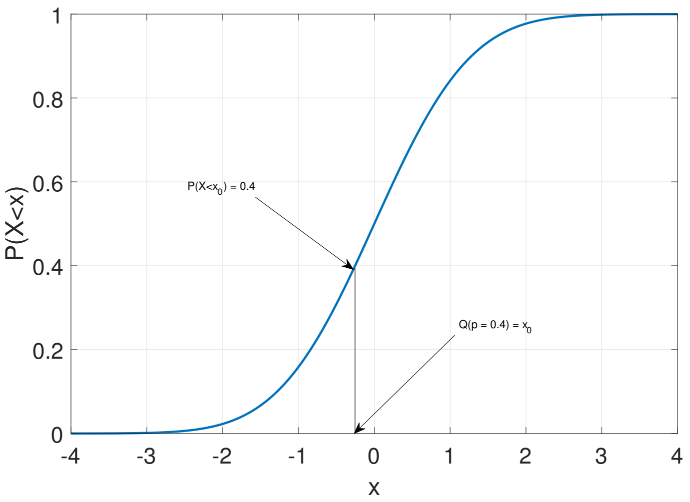
$$Q(p) = \{x \in \mathcal{X} : \mathbb{P}(X \leq x) = p\}$$

  which is sometimes called the quantile function.

- In words, the quantile function says: find the the value x such that the probability that $X \leq x$ is p

- For example:
  - $Q(p = 1/2)$ is the median;
  - $Q(p = 1/4)$ is the first quartile; and
  - $Q(p = 3/4)$ is the third quartile.

Cumulative Distribution Functions (3)



Expected Values (1)

- Given a distribution, we can define the expected value of the RV:

$$\mathbb{E}\left[X\right] = \sum_{x \in \mathcal{X}} x\, p(x)$$

  recalling that $p(x) \equiv \mathbb{P}(X = x)$.

- The expected value is the average value over $\mathcal{X}$, weighted by the probability of each particular $x \in \mathcal{X}$ appearing.
- For continuous RVs, replace the sum with an integral:

$$\mathbb{E}\left[X\right] = \int x\, p(x) dx$$

- Example: $\mathbb{P}(X = 1) = 0.5$, $\mathbb{P}(X = 2) = 0.4$, $\mathbb{P}(X = 3) = 0.1$:

$$\mathbb{E}\left[X\right] = 1 \cdot 0.5 + 2 \cdot 0.4 + 3 \cdot 0.1 = 1.6$$

Expected Values (1)

- Given a distribution, we can define the expected value of the RV:
$$\mathbb{E}\left[X\right] = \sum_{x \in \mathcal{X}} x\, p(x)$$
recalling that $p(x) \equiv \mathbb{P}(X = x)$.
- The expected value is the average value over $\mathcal{X}$, weighted by the probability of each particular $x \in \mathcal{X}$ appearing.
- For continuous RVs, replace the sum with an integral:
$$\mathbb{E}\left[X\right] = \int x\, p(x) dx$$
- Example: $\mathbb{P}(X = 1) = 0.5$, $\mathbb{P}(X = 2) = 0.4$, $\mathbb{P}(X = 3) = 0.1$:
$$\mathbb{E}\left[X\right] = 1 \cdot 0.5 + 2 \cdot 0.4 + 3 \cdot 0.1 = 1.6$$

Expected Values (2)

- More generally:

$$\mathbb{E}\left[f(X)\right] = \sum_{x \in \mathcal{X}} f(x)p(x)$$

where $f(x)$ is any function of x.

- Example: $\mathbb{P}(X = 1) = 0.5$, $\mathbb{P}(X = 2) = 0.4$, $\mathbb{P}(X = 3) = 0.1$:

$$\mathbb{E}\left[\log X\right] = \log 1 \cdot 0.5 + \log 2 \cdot 0.4 + \log 3 \cdot 0.1 = 0.3871$$

where $\log x$ is the natural logarithm (sometimes called $\ln$).

Variance (1)

- This lets us define important properties such as the variance:

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \sum_{x \in \mathcal{X}} (x - \mathbb{E}[X])^2 p(x)\end{aligned}$$

$\Rightarrow$ The expected squared deviation around the mean

- The larger $\mathbb{V}[X]$ the more variation around the mean
- The standard deviation is equal to $\sqrt{\mathbb{V}[X]}$.

- Example: $\mathbb{P}(X = 1) = 0.5$, $\mathbb{P}(X = 2) = 0.4$, $\mathbb{P}(X = 3) = 0.1$; recall that in this case, $\mathbb{E}[X] = 1.6$, so:

$$\mathbb{V}[X] = (1-1.6)^2 \cdot 0.5 + (2-1.6)^2 \cdot 0.4 + (3-1.6)^2 \cdot 0.1 = 0.44$$

Variance (1)

- This lets us define important properties such as the variance:

$$\begin{aligned} \mathbb{V}[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \sum_{x \in \mathcal{X}} (x - \mathbb{E}[X])^2 p(x) \end{aligned}$$

$\Rightarrow$ The expected squared deviation around the mean

- The larger $\mathbb{V}[X]$ the more variation around the mean
- The standard deviation is equal to $\sqrt{\mathbb{V}[X]}$.

- Example: $\mathbb{P}(X = 1) = 0.5$, $\mathbb{P}(X = 2) = 0.4$, $\mathbb{P}(X = 3) = 0.1$; recall that in this case, $\mathbb{E}[X] = 1.6$, so:

$$\mathbb{V}[X] = (1-1.6)^2 \cdot 0.5 + (2-1.6)^2 \cdot 0.4 + (3-1.6)^2 \cdot 0.1 = 0.44$$

Variance (2)

- A useful alternative expression for variance is:

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

  where the third step follows from properties of sums/integrals

- Variance is sum of expected squared value of X, minus square of expected value of X
  $\Rightarrow$ Use this to find variance for our example on previous slide

Covariance/Correlation (1)

- For two variables X and Y we can define the covariance:

$$\begin{aligned} \operatorname{cov}(X,Y) &= \mathbb{E}\left[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\right] \\ &= \mathbb{E}[XY]-\mathbb{E}[X]\,\mathbb{E}[Y] \end{aligned}$$

and from this, we can define the correlation:

$$\operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\mathbb{V}[X]\;\mathbb{V}[Y]}}$$

$\Rightarrow$ Compare to the sample correlation formula in Lecture 1.

Covariance/Correlation (2)

- Positive covariance/correlation:
  $\Rightarrow$ if X greater than $\mathbb{E}[X]$ then likely Y is *greater* than $\mathbb{E}[Y]$
- Negative covariance/correlation:
  $\Rightarrow$ if X greater than $\mathbb{E}[X]$ then likely Y is *less* than $\mathbb{E}[Y]$
- Covariance between $(-\infty, \infty)$,
  - Depends on scale (unit of measurement) of variables X and Y
- Correlation between $[-1, 1]$,
  - Independent of scale of variables
- If X, Y independent, $\mathrm{cov}(X, Y) = \mathrm{corr}(X, Y) = 0$
  $\Rightarrow$ Converse is **not** true!

Covariance/Correlation (2)

- Positive covariance/correlation:
  $\Rightarrow$ if X greater than $\mathbb{E}\left[X\right]$ then likely Y is *greater* than $\mathbb{E}\left[Y\right]$
- Negative covariance/correlation:
  $\Rightarrow$ if X greater than $\mathbb{E}\left[X\right]$ then likely Y is *less* than $\mathbb{E}\left[Y\right]$
- Covariance between $(-\infty, \infty)$,
  - Depends on scale (unit of measurement) of variables X and Y
- Correlation between $[-1, 1]$,
  - Independent of scale of variables
- If X, Y independent, $\operatorname{cov}\left(X, Y\right) = \operatorname{corr}\left(X, Y\right) = 0$
  $\Rightarrow$ Converse is **not** true!

Expectations and Independent RVs

- In general, expectation of a function of two RVs is

$$\mathbb{E}\left[f(X,Y)\right] = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} f(x,y)p(x,y)$$

- Fact 1: Due to linearity of expectation, we have

$$\mathbb{E}\left[f(X)+g(Y)\right] = \mathbb{E}\left[f(X)\right] + \mathbb{E}\left[g(Y)\right]$$

for all RVs X and Y, and

- Fact 2: For independent RVs, we have

$$\mathbb{E}\left[f(X)f(Y)\right] = \mathbb{E}\left[f(X)\right]\mathbb{E}\left[g(Y)\right]$$

implying that

$$\mathbb{V}\left[X+Y\right] = \mathbb{V}\left[X\right] + \mathbb{V}\left[Y\right]$$

for X and Y independent.

Expectations and Independent RVs

- In general, expectation of a function of two RVs is

$$\mathbb{E}\left[f(X,Y)\right] = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} f(x,y)p(x,y)$$

- Fact 1: Due to linearity of expectation, we have

$$\mathbb{E}\left[f(X)+g(Y)\right] = \mathbb{E}\left[f(X)\right] + \mathbb{E}\left[g(Y)\right]$$

for all RVs X and Y, and

- Fact 2: For independent RVs, we have

$$\mathbb{E}\left[f(X)f(Y)\right] = \mathbb{E}\left[f(X)\right]\mathbb{E}\left[g(Y)\right]$$

implying that

$$\mathbb{V}\left[X+Y\right] = \mathbb{V}\left[X\right] + \mathbb{V}\left[Y\right]$$

for X and Y independent.

Existence of Expected Values

- Expected values do not always exist
- If $\mathcal{X}$ is *finite*, then $\mathbb{E}\left[X\right]$ always exists
- However, in general, $\mathcal{X}$ will not be finite
- $\mathcal{X}$ is usually the set of integers $\mathbb{Z}$ or real numbers $\mathbb{R}$
  $\Rightarrow$ In this case, expectations are not guaranteed to exist
- In contrast, the quantiles (such as median) *always* exist

Outline

Probability Distributions as Models

- We can use probability distributions as models of reality
- For example, imagine we knew the distribution of heights of people in Australia, say $p(h)$
- We could then make predictions/statements about heights
  - For example, the proportion of people taller than $1.7m$

$$\mathbb{P}(H > 1.7) = \int_{1.7}^{\infty} p(h)dh$$

- Then a company clothing could use this information to determine how much product of different sizes to stock

Parametric Probability Distributions (1)

- So far we have built probability distributions by directly specifying the probabilities for each element $x \in \mathcal{X}$
- This is fine if $\mathcal{X}$ is a small finite set
- But if $\mathcal{X}$ is large, or infinite (for example, all the integers), this approach no longer works
- Instead it is usual to use parametric probability distributions
- We will look at several important distributions:
    - The Gaussian distribution;
    - The Bernoulli distribution;
    - The binomial distribution;
    - The Poisson distribution.

Parametric Probability Distributions (1)

- So far we have built probability distributions by directly specifying the probabilities for each element $x \in \mathcal{X}$
- This is fine if $\mathcal{X}$ is a small finite set
- But if $\mathcal{X}$ is large, or infinite (for example, all the integers), this approach no longer works
- Instead it is usual to use parametric probability distributions
- We will look at several important distributions:
  - The Gaussian distribution;
  - The Bernoulli distribution;
  - The binomial distribution;
  - The Poisson distribution.

Parametric Probability Distributions (2)

- We specify the probability density function by

$$p(x \,|\, \boldsymbol{\theta}), \;\; x \in \mathcal{X}, \; \boldsymbol{\theta} \in \Theta$$

or, for discrete RVs we use the shorthand notation:

$$\mathbb{P}(X = x \,|\, \boldsymbol{\theta}) \equiv p(x \,|\, \boldsymbol{\theta}), \;\; x \in \mathcal{X}, \; \boldsymbol{\theta} \in \Theta$$

where
  - $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ are the parameters that control distribution of the probabilities;
  - Θ is the set of valid parameters for the model.

$\Rightarrow$ by changing $\boldsymbol{\theta}$ we can change the distribution.

- Usually, the number of parameters $k \ll |\mathcal{X}|$

Parametric Probability Distributions (3)

- The properties of the RV are determined by $p(x \,|\, \boldsymbol{\theta})$
- For example, the mean is

$$\mathbb{E}\left[X\right] = f(\boldsymbol{\theta}),$$

  where $f(\cdot)$ is a function that depends on $\boldsymbol{\theta}$ and $p(x \,|\, \boldsymbol{\theta})$.

- The same applies to the variance, cdf, quantiles, etc.
- The parameterisation will not be unique
  $\Rightarrow$ there are often several common parameterisations for the same distribution

Gaussian Distribution (1)

- Let's begin with the case that $\mathcal{X} = \mathbb{R}$
  $\Rightarrow$ that is, we want a distribution over all the real numbers
- Probably the most important distribution for real numbers is the Gaussian (normal) distribution
  $\Rightarrow$ named after Carl Friedrich Gauss (1777-1855)
- The pdf for a Gaussian distribution is given by

$$p(x \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right)$$

  where
  - μ is the mean of the distribution;
  - σ^2 is the variance of the distribution;

  so that $\boldsymbol{\theta} = (\mu, \sigma^2)$ for the Gaussian distribution.

Gaussian Distribution (1)

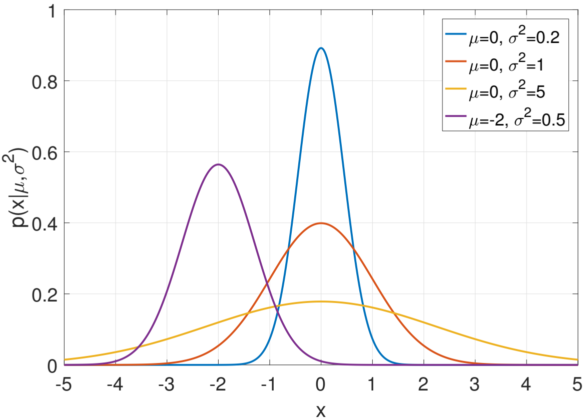
- Let's begin with the case that $\mathcal{X} = \mathbb{R}$
  $\Rightarrow$ that is, we want a distribution over all the real numbers
- Probably the most important distribution for real numbers is the Gaussian (normal) distribution
  $\Rightarrow$ named after Carl Friedrich Gauss (1777-1855)
- The pdf for a Gaussian distribution is given by

$$p(x \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right)$$

  where
  - μ is the mean of the distribution;
  - σ^2 is the variance of the distribution;

  so that $\boldsymbol{\theta} = (\mu, \sigma^2)$ for the Gaussian distribution.

Gaussian Distribution (2)



Probability density functions for several normal (Gaussian) distributions. The orange curve is the *standard normal distribution*. Note that the normal distribution is symmetric and tails off to zero as $|x| \to \infty$.

Gaussian Distribution (3)

- If X follows a Gaussian distribution, we write that

$$X \sim N(\mu, \sigma^2)$$

where "$\sim$" is read as "is distributed per a"

- An important property of Gaussian RVs is self-similarity
- Every Gaussian distribution is a translated and scaled version of the standard normal distribution $N(0,1)$
- If $Z \sim N(0,1)$, then

$$X = \sigma Z + \mu$$

is distributed as per $N(\mu, \sigma^2)$

Gaussian Distribution (4)

- If $X \sim N(\mu, \sigma^2)$, then

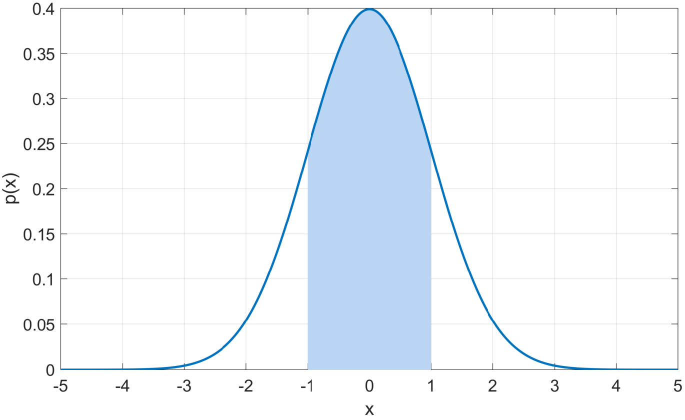
$$\begin{aligned} \mathbb{E}[X] &= \mu, \\ \mathbb{V}[X] &= \sigma^2. \end{aligned}$$

- The Gaussian distribution is symmetric around μ, so that:
  - its mode is μ;
  - its median is μ.

- The cdf for the Gaussian has no closed form
  - Most packages have algorithms to evaluate it numerically
  - Some well-known rules regarding the cdf are ...

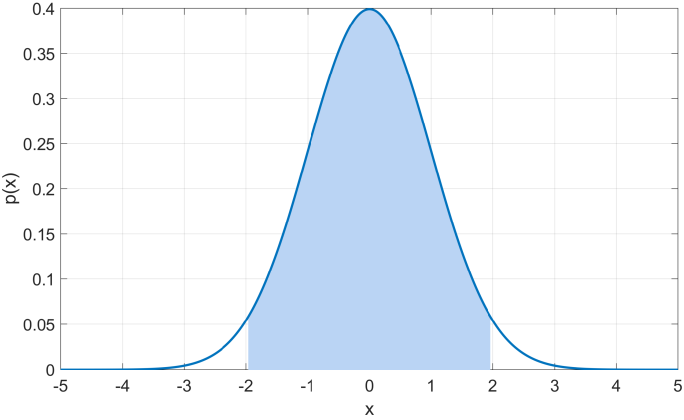
Gaussian Distribution (5)

- For any $N(\mu, \sigma^2)$:
  - 68.27% of probability falls within $(\mu - \sigma, \mu + \sigma)$



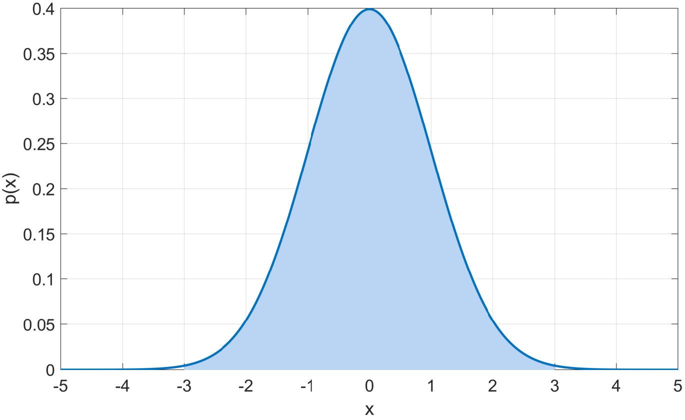
Gaussian Distribution (6)

- For any $N(\mu, \sigma^2)$:
  - 95.45% of probability falls within $(\mu - 2\sigma, \mu + 2\sigma)$



Gaussian Distribution (7)

- For any $N(\mu, \sigma^2)$:
  - 99.73% of probability falls within $(\mu - 3\sigma, \mu + 3\sigma)$



Bernoulli Distribution (1)

- Let's consider the case of discrete, binary RVs, i.e., $\mathcal{X} = \{0, 1\}$
- The Bernoulli distribution models these variables

$$\mathbb{P}(X = 1 \,|\, \theta) = \theta, \;\; \theta \in [0, 1]$$

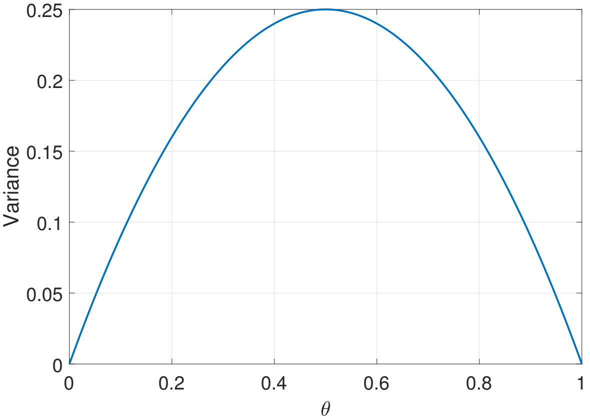
so that the parametric probability distribution follows:

$$p(x \,|\, \theta) = \theta^x (1 - \theta)^{(1-x)}$$

- The parameter θ is the probability of observing a "success"
- If X follows a Bernoulli distribution, we write $X \sim \mathrm{Be}(\theta)$
- It is easy to see that

$$\begin{aligned} \mathbb{E}[X] &= \theta \\ \mathbb{V}[X] &= \theta(1 - \theta) \end{aligned}$$

Bernoulli Distribution (2)



Variance of a Bernoulli random variable as a function of θ. The variance is maximum when $\theta = 1/2$ and smallest for $\theta = 0$ and $\theta = 1$.

Binomial Distribution (1)

- Now consider n binary RVs $\mathbf{X} = (X_1, \ldots, X_n)$.
  - Example realisation: $\mathbf{x} = (0, 1, 1, 1, 0, 1, 0, 0, 1, 1)$
- The sum

$$m(\mathbf{x}) \equiv m = \sum_{j=1}^{n} x_j$$

  counts the number of "successes"
  $\Rightarrow$ in our example, $m = 6$
- Given n, the count is a RV, say M, over the sample space $\{0, 1, 2, \ldots, n\}$

Binomial Distribution (2)

- The binomial distribution describes the probability that M takes a particular value m

$$p(m \,|\, \theta) = \binom{n}{m} \prod_{i=1}^{n} p(x_i \,|\, \theta) = \binom{n}{m} \theta^m (1-\theta)^{(n-m)}$$

where

$$\binom{n}{m} = \frac{n!}{(n-m)!m!}$$

is the number of ways of choosing m objects out of n identical objects (the binomial coefficient)
$\Rightarrow m! = 1 \times 2 \times 3 \times \ldots \times m$ is the factorial function

- This captures the fact that, for $1 \leq m \leq (n-1)$ there are multiple sequences with m successes out of n trials

Binomial Distribution (3)

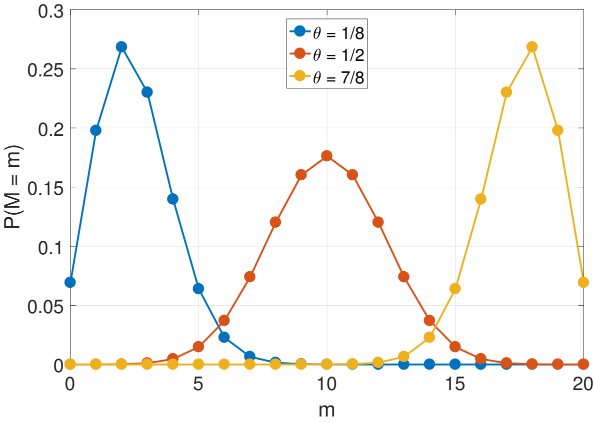
- Example: The following six sequences have $m = 2$ successes out of $n = 4$ trials:

$$
\begin{gathered}
1,1,0,0 \\
1,0,1,0 \\
1,0,0,1 \\
0,1,1,0 \\
0,1,0,1 \\
0,0,1,1
\end{gathered}
$$

so that

$$p(m = 2 \mid \theta) = \binom{n=4}{m=2} \theta^2 (1-\theta)^{(4-2)}$$

Binomial Distribution (4)



Binomial distribution for $n = 20$ and $\theta = 1/8$, $\theta = 1/12$, $\theta = 7/8$. The distribution is defined only on the integers – the connecting lines are only guides for the eye. Note that $\theta = 7/8$ is a mirror of $\theta = 1/8$.

Binomial Distribution (5)

- If M follows a binomial distribution, we write

$$M \sim \mathrm{Bin}(\theta, n)$$

- As m is a sum of independent Bernoulli RVs we have

$$\begin{aligned} \mathbb{E}\left[M\right] &= n\theta \\ \mathbb{V}\left[M\right] &= n\theta(1-\theta) \end{aligned}$$

Poisson Distribution (1)

- What if our data is non-negative integers; for example:
  - number of telephone calls made in an hour
  - number of people kicked to death by horses in a year
  - Sample space is then $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$
- One suitable distribution is the Poisson distribution
  - Named after Simeon Poisson (1781–1840)
- Has the form

$$p(k \mid \lambda) = \frac{\lambda^k \exp(-\lambda)}{k!}$$

where λ is often called the *rate*.

Poisson Distribution (1)

- What if our data is non-negative integers; for example:
  - number of telephone calls made in an hour
  - number of people kicked to death by horses in a year
  - Sample space is then $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$
- One suitable distribution is the Poisson distribution
  - Named after Simeon Poisson (1781–1840)
- Has the form

$$p(k \mid \lambda) = \frac{\lambda^k \exp(-\lambda)}{k!}$$

where λ is often called the *rate*.

Poisson Distribution (2)

- If X is distributed per a Poisson distribution we write

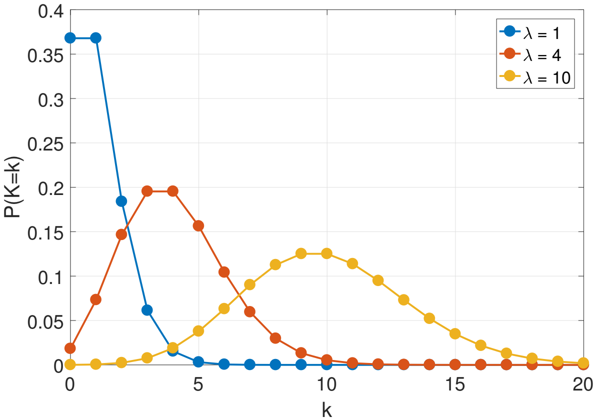
$$X \sim \mathrm{Pois}(\lambda)$$

- The Poisson distribution has

$$\begin{aligned} \mathbb{E}\left[X\right] &= \lambda \\ \mathbb{V}\left[X\right] &= \lambda \end{aligned}$$

- The Poisson distribution is an example of a distribution in which the variance grows with the mean

Poisson Distribution (3)



Poisson distribution for $\lambda = 1$, $\lambda = 4$ and $\lambda = 10$. The distribution is defined only on the integers – the connecting lines are only guides for the eye.

Poisson Distribution (4)

- The Poisson distribution models the number of events in an interval of time
- When is the Poisson appropriate?
    - The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently.
    - The rate at which events occur is constant. The rate cannot be higher in some intervals and lower in other intervals.
    - Two events cannot occur at exactly the same instant.
    - The probability of an event in a small interval is proportional to the length of the interval.

(taken from Wikipedia)

Chebyshev's Inequality (1)

- Named after P. Chebyshev (1821-1894)
- If X is a RV with mean μ and variance σ^2, then for any $k > 0$

$$\mathbb{P}(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

- This inequality allows us to compute (bounds on) probabilities even when only the mean and variance are known

Chebyshev's Inequality (2)

- Chebyshev's bound if only $\mathbb{E}[X] = 0$, $\mathbb{V}[X] = 1$ is known:
  - $\mathbb{P}(|X| \geq 1) \leq 1$;
  - $\mathbb{P}(|X| \geq 2) \leq 0.25$;
  - $\mathbb{P}(|X| \geq 3) \leq 0.1112$;
- Compare to the situation that we know $X \sim N(0, 1)$:
  - $\mathbb{P}(|X| \geq 1) = 0.3173$;
  - $\mathbb{P}(|X| \geq 2) = 0.0455$;
  - $\mathbb{P}(|X| \geq 3) = 0.0027$.

$\Rightarrow$ Chebyshev's bounds very general but not always accurate.

Weak Law of Large Numbers (1)

- An important application of Chebyshev's inequality is to prove the weak law of large numbers.
- Let $X_1, \ldots, X_n$ be RVs with $\mathbb{E}[X_i] = \mu$; then for any $\varepsilon > 0$

$$\mathbb{P}\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right\} \to 0 \text{ as } n \to \infty.$$

- Informally, you can think of this result as saying that the (sample) mean of a realisation of random variables converges to the expected value as the number of realisations grows larger and larger.

Reading/Terms to Revise

- Reading for this week: Chapters 4 and 5 of Ross.
- Terms you should know:
  - Random variable;
  - Conditional Probability;
  - Probability density function;
  - Expectations;
  - Variance and co-variance;
  - Normal, Bernoulli, binomial and Poisson distributions
  - Law of large numbers