CS430/910: Foundations of Data Analytics

Statistics | Dr Greg Watson

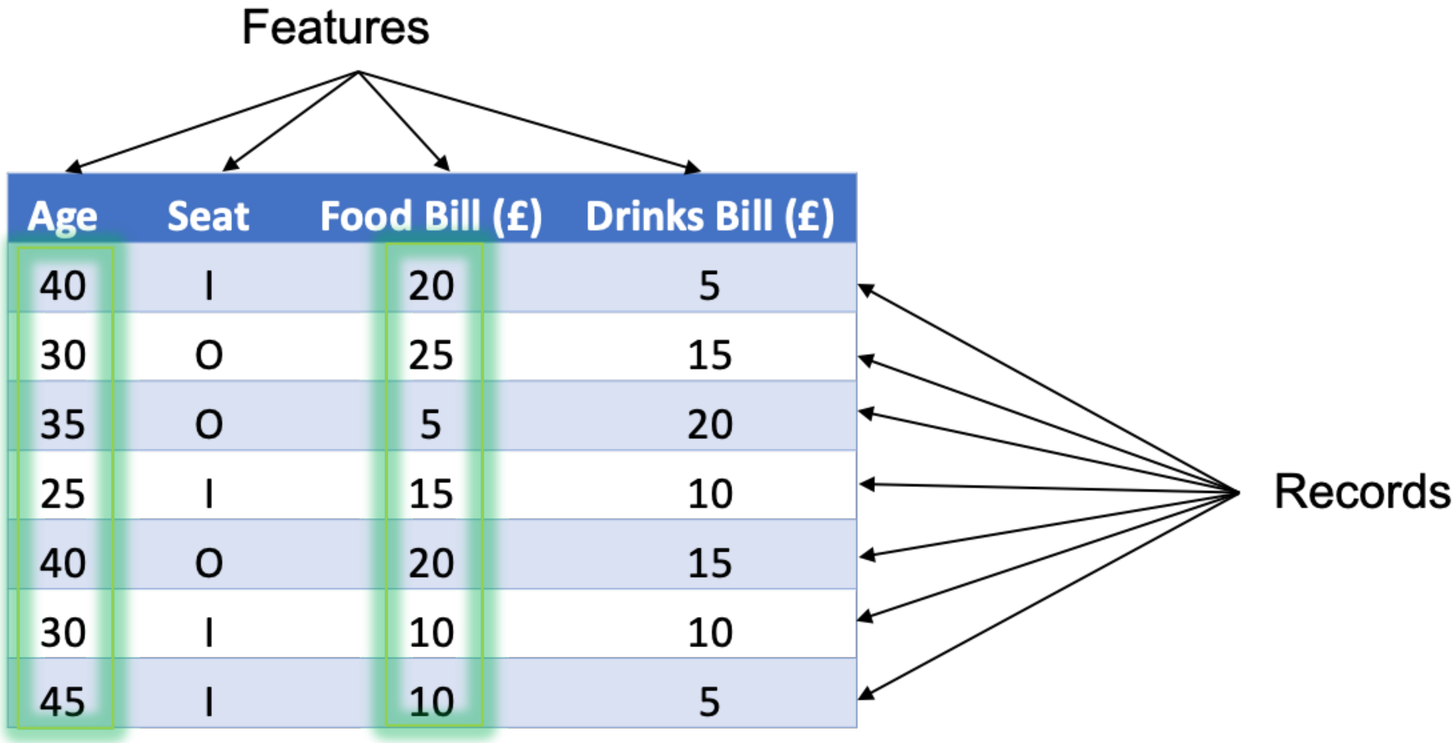
Part A: Random Variables and Probability

The Statistical Lens

- It is helpful to use the tools of statistics to view data:
  - Each attribute can be viewed as describing a random variable.
  - Look at statistical properties of this random variable.
  - Can also look at combinations of attributes.
  - We study the **empirical** distribution given by the data.

Data Representation

- Radical change of terminology: features to be called “Random Variables”.
- Random Variable ≈ possible outcome of something happening.



Random Variables and Sample Spaces

Definition: A *Random Variable* is a function $X : \Omega \to \mathbb{R}$.

- Ω is a set corresponding to a random experiment:
  - …also called the “sample space”.
  - …contains all possible outcomes/samples.
- A “random experiment” is an experiment which randomly results in a single outcome out of a set (that set is Ω).
- Example: Coin tossing (once):
  - Experiment = Tossing the coin.
  - Random? Yes, because the outcome could be either heads or tails.
  - Sample space: $\Omega = \{H, T\}$.

Sample Spaces

More Examples

- Coins tossing (either two coins at once, or one coin twice)
  - Experiment: the tossing of the coin
  - Random? Yes, unless they are rigged…
  - Sample space… it depends:
    - If we care about order, $\Omega = \{HH, HT, TH, TT\}$
    - If we don't care about order, $\Omega = \{HH, HT, TT\}$ (if I get HT, I don't care which coin was H and which coin was T)

Sample Spaces

More Examples

- Rolling two dice:
  - Experiment: the rolling
  - Random? Yes, unless they are rigged…
  - Sample space (say order matters), $\Omega = \{11,12,21,...,66\}$ (36 elements in the set)

Sample Spaces and σ-Field

Definition: A Sample Space Ω is a set whose elements are called elementary events (as potential outcomes of a random experiment).

- The set could be finite (rolling a dice), countably infinite (picking a natural number), or uncountable (picking a real number)

  On their own, sample spaces are not very interesting.

Definition: A collection of subsets $\mathcal{F}$ of Ω is called a σ-Field if:

$$\Omega \in \mathcal{F}$$

$$A \in \mathcal{F} \implies A^c \in \mathcal{F}$$

$$A_n \in \mathcal{F}, \forall n \in \mathbb{N} \implies \bigcup_{n \in \mathbb{N}} A_n \in \mathcal{F}$$

An element of $\mathcal{F}$ is called an elementary event if it consists of a single point from Ω; otherwise it is called an event.

Ω and $\mathcal{F}$

- Ω - set of possible elementary events of a single (one-shot) random experiment.
- $\mathcal{F}$ - set of events observable to an outsider (who cannot see the outcome, but is allowed to ask)
- Say that Alice rolls a dice once, i.e. $\Omega = \{1,2,3,4,5,6\}$
- Bob starts questioning (Did that happen? What about that?, …)
  - Generous Alice: $\mathcal{F} = 2^{\Omega}$: Bob will always get an answer (yes/no)
  - Secretive Alice: $\mathcal{F} = \{\Omega, \emptyset\}$: Bob can only find out whether something at all happened!
  - Odd Alice: $\mathcal{F} = \{\Omega, \emptyset, \{1,3,5\}, \{2,4,6\}\}$: Bob can only find out whether an odd or even number showed up. Alice will not answer “Did 5 happen?” or “Did 2 or 3 happen?”

Probability

Definition: Given a sample space Ω and a corresponding σ-Field $\mathscr{F}$, the pair $(\Omega, \mathscr{F})$ is called a *Measurable Space*.

- Definition (Probability): Given a measurable space $(\Omega, \mathscr{F})$, a probability (measure) is a countably additive function $P : \mathscr{F} \to [0,1]$ such that:

$$P(\Omega) = 1$$

$$P\left(\bigcup_n A_n\right) = \sum_n P(A_n), \forall A_i \cap A_j = \varnothing$$

Probability

- P essentially assigns a probability (measure) to each set in $\mathcal{F}$.
- Could there be many probability measures on the same $(\Omega, \mathcal{F})$?

Probability Spaces $(\Omega, \mathscr{F}, P)$

- Coin tossing: $\Omega = \{H, T\}$

Case 1:

$\mathscr{F} = \{\{H, T\}, \varnothing, \{H\}, \{T\}\}$

$P(H) = 0.75, P(T) = 0.25$

Case 2:

$\mathscr{F} = \{\{H, T\}, \varnothing, \{H\}, \{T\}\}$

$P(H) = 1, P(T) = 0$

Case 3:

$\mathscr{F} = \{\{H, T\}, \varnothing\}$

$P(\{H, T\}) = 1$

- Case 2: Rigged coin!
- Case 3: Alice doesn't talk to Bob.
- Important note: In all cases, the rules on $\mathscr{F}$ and P are obeyed!

Probability Spaces $(\Omega, \mathscr{F}, P)$

- Rolling two dice: $\Omega = \{11,12,21,13,...,66\}$

Case 1:

$\mathscr{F} = 2^{\Omega}$

$P(ij) = \frac{1}{36}$

Case 2:

$\mathscr{F} = 2^{\Omega}$

$P(3j) = \frac{1}{6}$

$P(ij) = 0, \forall i \neq 3$

Case 3:

$\mathscr{F} = \{\Omega, \varnothing, \{ii : \forall i\}, \{ij : \forall i \neq j\}\}$

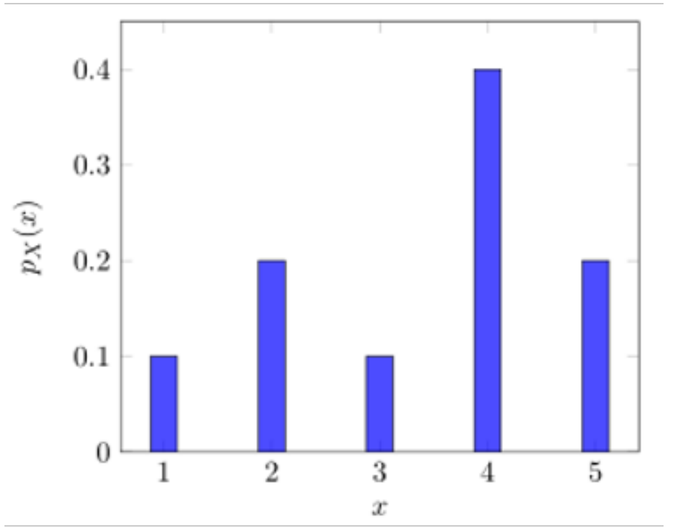
$P(\{ii : \forall i\}) = 0.8$

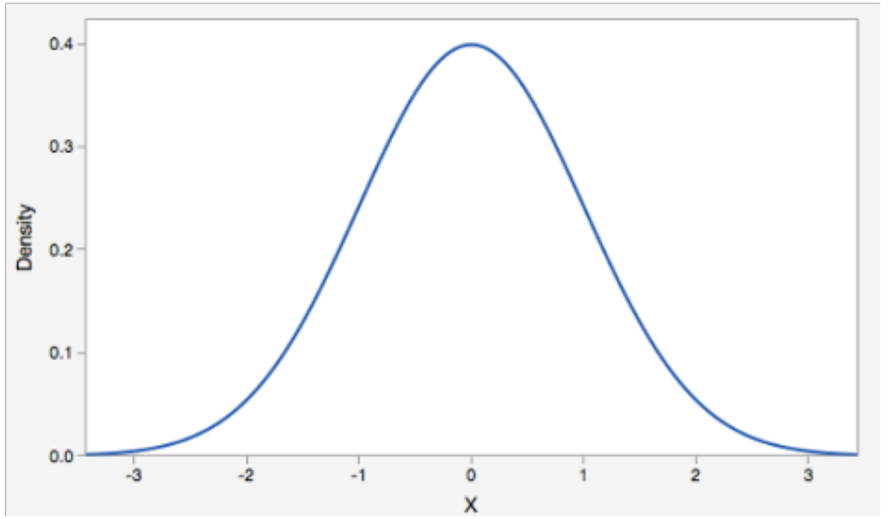
$P(\{ij : \forall i \neq j\}) = 0.2$

- Case 2: First dice is rigged (only shows 3).
- Case 3: Alice only answers to a small number of questions.

Two Types of Random Variables

- We know that a random variable is a measurable function: $X : \Omega \rightarrow \mathbb{R}$.
- A random variable can either be discrete (countable number of values), or continuous.
- Example of discrete random variable: X is the number of students attending some CS430/910 lecture.
- Example of a continuous random variable: X is the time that someone wakes up tomorrow (assuming infinite precision).





Illustrative Examples of Random Variables

- The number of winners in a football pool next week.
- The number of hurricanes that will hit the United States next year.
- The number of claims that will be submitted to an insurance company next year.

} Discrete Random Variables taking a discrete number of values!

- The amount of rainfall that the city of London will receive next year.
- The lifetime of a newly bought battery.
- The duration of your next phone call.

} Continuous Random Variables taking a continuum of values!

Discrete Random Variables

- A random variable X is said to be discrete if its set of possible values is finite or countably infinite.
- The set of possible values of X is called the range of X and is denoted by I.
- The probabilities associated with these possible values are determined by the probability measure P on the sample space of the chance experiment.
- The *Probability Mass Function* (PMF) of a discrete random variable X is defined by $P(X = x)$ for $x \in I$, where the notation $P(X = x)$ is shorthand for:

$$P(X = x) = P(\{\omega : X(\omega) = x\})$$

Discrete Random Variables

- A random variable X is said to be discrete if its set of possible values is finite or countably infinite.
- The set of possible values of X is called the range of X and is denoted by I.
- The probabilities associated with these possible values are determined by the probability measure P on the sample space of t
- The *Probability Mass Function* (PMF) of a discre by $P(X = x)$ for $x \in I$, where the notation $P(X$

$P(X = x)$ is the probability mass assigned by P to the set of all outcomes ω for which $X(\omega) = x$.

$$P(X = x) = P(\{\omega : X(\omega) = x\})$$

Discrete Random Variables

Example

- Imagine rolling a fair dice twice.
- Sample space (Ω) consists of 36 equally likely outcomes (i, j).
- The random variable X is defined as the smallest of the two numbers.
  - X assigns the numerical value $min(i, j)$ to the outcome (i, j) of the sample space.
  - Therefore, the range of X is $\{1,2,3,4,5,6\}$.

Tijms, H., 2017. Probability: a lively introduction. Cambridge University Press.

Discrete Random Variables

Example

- X takes on the value 1 if one of the 11 outcomes $(1,1), (1,2), \ldots, (1,6), (2,1), (3,1), \ldots, (6,1)$ occurs.
- Hence, $P(X = 1) = \frac{11}{36}$.
- In the same way, $P(X = 2) = \frac{9}{36}$, $P(X = 3) = \frac{7}{36}$, $P(X = 4) = \frac{5}{36}$, $P(X = 5) = \frac{3}{36}$, $P(X = 6) = \frac{1}{36}$.

Tijms, H., 2017. Probability: a lively introduction. Cambridge University Press.

Expected Value

Of a Discrete Random Variable

- Imagine a casino where the player has a 70% probability of losing £1, a 25% chance of winning £2 and a 5% chance of winning £3.
- In approximately $0.7n$ repetitions of the game, the player loses £1.
- In approximately $0.25n$ repetitions of the game, the player gains £2.
- In approximately $0.05n$ repetitions of the game, the player gains £3.
- Therefore, total win:

$$(0.7n) \times (-1) + (0.25n) \times 2 + (0.05n) \times 3 = -(0.05)n$$

Expected Value

Of a Discrete Random Variable

- Imagine a casino where the player has a 70% probability of losing £1, a 25% chance of winning £2 and a 5% chance of winning 3%.
- In approximately $0.7n$ repetitions of the game, the player loses £1.
- In approximately $0.25n$ repetitions of the game, the player gains £2.
- In approximately $0.05n$ repetitions of the game, the player gains £3.
- Therefore, total win:

Average “win” is actually a loss!

$$(0.7n) \times (-1) + (0.25n) \times 2 + (0.05n) \times 3 = -(0.05)n$$

Expected Value

Of a Discrete Random Variable

- Therefore, total win:

$$(0.7n) \times (-1) + (0.25n) \times 2 + (0.05n) \times 3 = -(0.05)n$$

- We can define a random variable X as the win achieved in a single repetition of the game.

- In this case, -0.05 is the expected value of X.

$$-1 \times P(X = -1) + 2 \times P(X = 2) + 3 \times P(X = 3) = -0.05$$

Expected Value

Of a Discrete Random Variable

- General definition:

$$E(X) = \sum_{x \in I} xP(X = x)$$

- Also known as the *mean* or *average value* of X.
- Two useful properties:

1. $E(X + Y) = E(X) + E(Y)$
2. $E(aX + b) = aE(X) + b$ (for any constants a and b)

Variance

Of a Discrete Random Variable

- The variance of a random variable X is denoted by:

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

where $\mu = E[X]$.

- The variance is a measure of the spread of the possible values of X.
- The standard deviation is denoted as follows:

$$\sigma(X) = \sqrt{var(X)}$$

Variance

Of a Discrete Random Variable

- Question: Why would we use standard deviation rather than variance?
- Answer: Standard deviation has the same units (e.g. pounds) as $E(X)$, whereas variance has been squared.
- Two useful properties:

1. $var(aX + b) = a^2 var(X)$
2. $\sigma(aX + b) = |a|\sigma(X)$

Continuous Random Variable

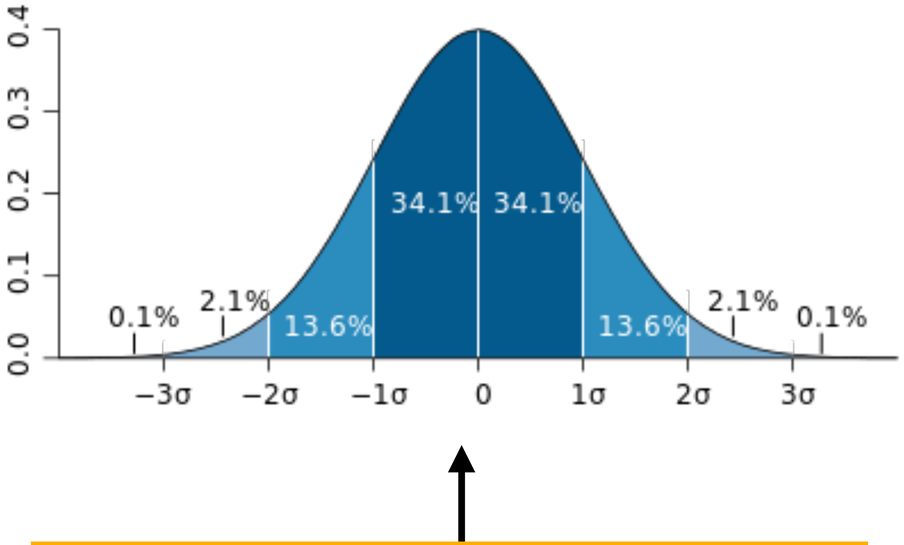
- Definition: A random variable is continuous if $P(X = x) = 0 \forall x$
- A Cumulative Distribution Function (usually denoted by $F(x)$), describes the probability that X takes on a value less than or equal to x.
  - $F(x) = P(X \leq x)$
- A Continuous Random Variable does not assign probabilities to points, but to intervals:

$$P(a \leq X \leq b) = F(b) - F(a)$$

Probability Distributions

Probability Density Function (PDF)

- Given a continuous random variable, X:
  - A *Probability Density Function* (PDF) specifies the probability of a random variable falling within a range of values.
  - $P(\alpha \leq X \leq \beta)$
    - Can be calculated using integration.
  - An analogue of the Probability Mass Function (PMF) of discrete random variables.



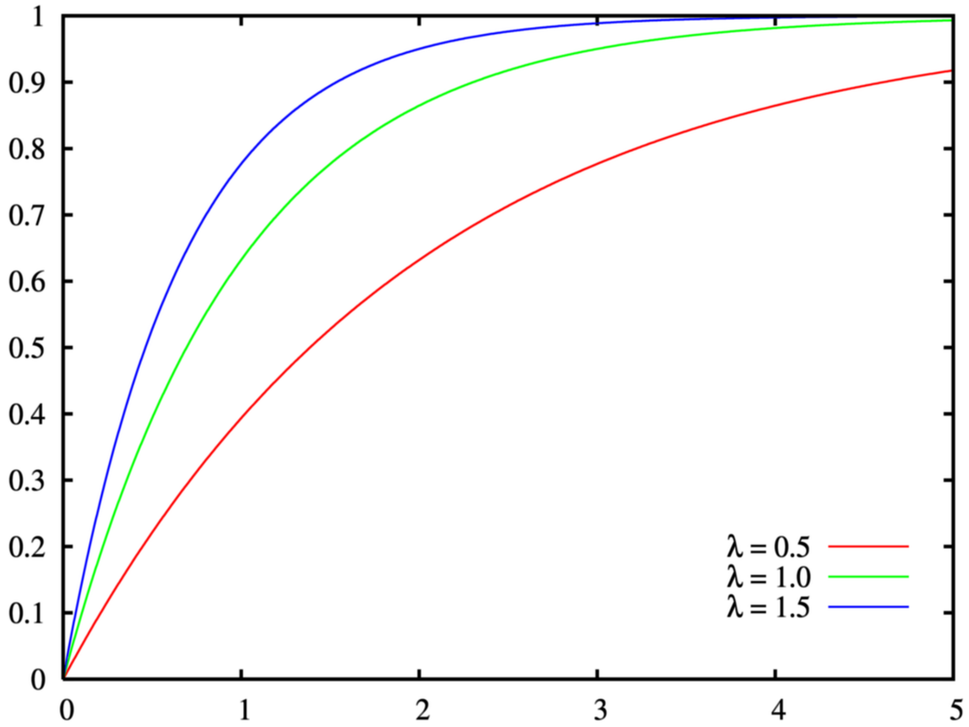
Probability Density Function of a Normal Distribution.

Figure by: M. W. Toews, CC BY 2.5 <https://creativecommons.org/licenses/by/2.5>, via Wikimedia Commons

Probability Distributions

Cumulative Distribution Function (CDF)

- Given a continuous random variable, X:
  - A Cumulative Distribution Function (usually denoted by $F(x)$), describes the probability that X takes on a value less than or equal to x.
  - $F(x) = P(X \leq x)$
- $F(x)$ is an increasing function ($F(x_1) \leq F(x_2)$ if $x_1 \leq x_2$).
- $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$



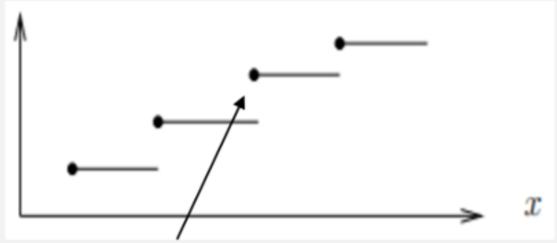
Probability Distributions

Cumulative Distribution Function (CDF)

- Both Discrete and Continuous Random Variables have CDFs.

Discrete CDF

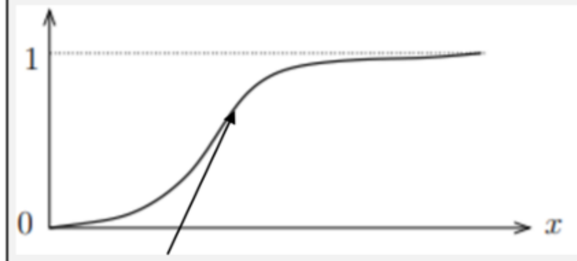
$$F(x) := P(X \leq x) = \sum_{y \leq x} P(y)$$



Probability accumulates in discrete steps

Continuous CDF

$$F(x) := P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$



Probability accumulates continuously

Images from Fewster: https://www.stat.auckland.ac.nz/~fewster/210/notes.php

Probability Distributions

Complementary Cumulative Distribution Function (CCDF)

- We can also calculate the opposite of the value of the Cumulative Distribution Function.
- The *Complementary Cumulative Distribution Function* (CCDF) is defined as:
  - $\bar{F}(x) = P(X > x) = 1 - P(X \leq x) = 1 - F(x)$

Acknowledgements

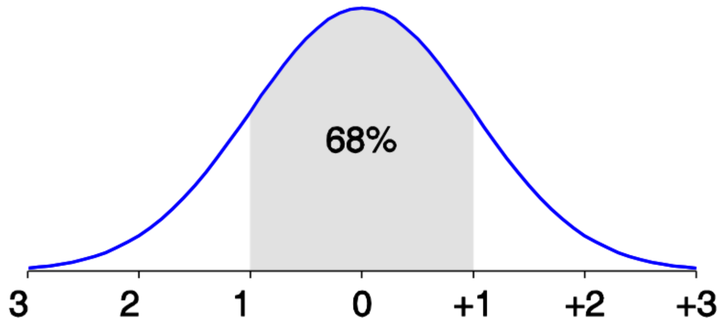
- Florin Ciucu [Warwick, CS430/CS910]
- Tijms, H., 2017. Probability: a lively introduction. Cambridge University Press.
- zedStatistics: https://www.youtube.com/watch?v=2kg1O0j1J9c
- jbStatistics: https://www.youtube.com/watch?v=zq9Oz82iHf0
- Han, J. and Kamber, M., 2006. *Data Mining: Concepts and Techniques*. Elsevier.
- Han, J., Pei, J. and Kamber, M., 2012. *Data Mining: Concepts and Techniques*. Elsevier.

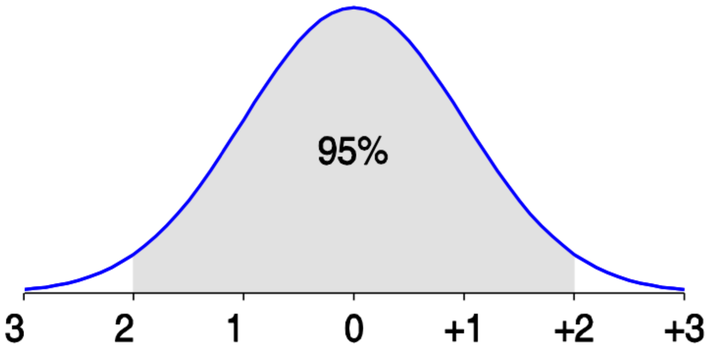
Part B: Statistical Distributions & Correlation

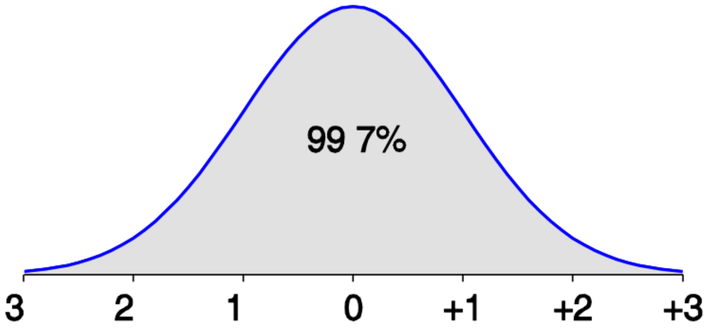
Statistical Distributions in Data

Normal Distribution

- Many familiar distributions model observed data.
- Normal distribution: characterised by mean μ and variance σ^2.
- From $\mu - \sigma$ to $\mu + \sigma$: contains about 68% of the measurements (μ: mean, σ: standard deviation).
- From $\mu - 2\sigma$ to $\mu + 2\sigma$: contains about 95% of data.
- From $\mu - 3\sigma$ to $\mu + 3\sigma$: contains about 99.7% of data.





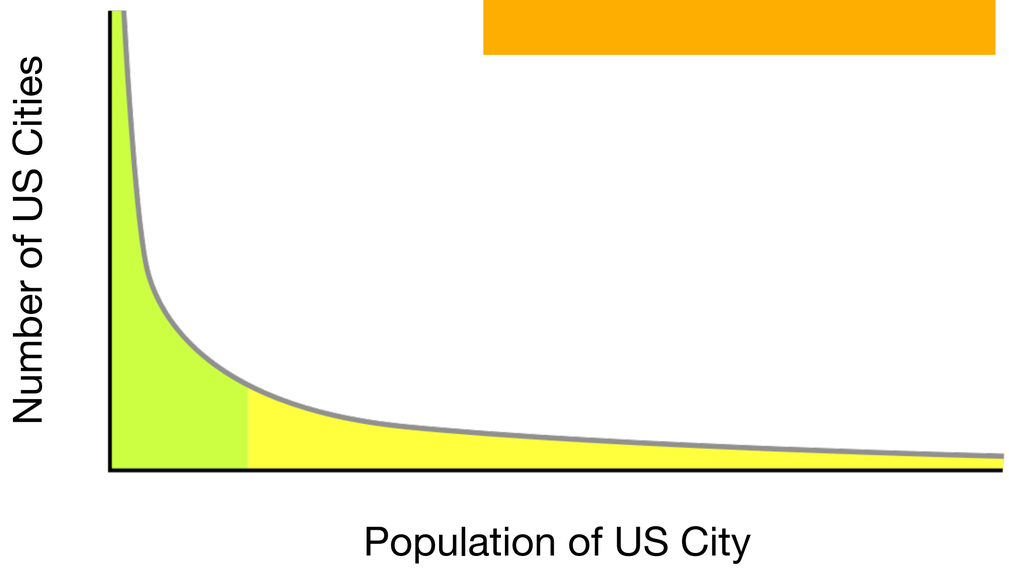


Statistical Distributions in Data

Pareto Distribution

- Not all things that we can measure peak around a typical value…
- Consider the number of people living in US cities:
  - (1) New York (~9 million)
  - (2) Los Angeles (~4 million)
  - …
  - (10) San Jose, California (~ 1 million)
  - (11) Austin, Texas (~960,000)
  - …
  - (50) Arlington, Texas (~400,000)

Large number of US cities have small population.



Population data from the United States Census Bureau: http://census.gov

Statistical Distributions in Data

Pareto Distribution

- Arises in many cases:
    1. Number of people living in cities
    2. Popularity of products from retailers
    3. Frequency of word use in written text
    4. Wealth distribution (99% vs 1%)
    5. Video popularity

Statistical Distributions in Data

Geometric Distribution

- Suppose that an event happens with probability p (for a small p) (Independently at each time step).
- **Discrete Case**: The *Geometric Distribution* is the probability distribution showing the “waiting time” between events in Poisson processes (events which occur at a constant average rate).

- $P(X = x) = (1 - p)^{x-1}p$, for $x > 0$
- $P[X \geq x] = (1 - p)^{x}$
- $E[X] = \frac{1}{p}$
- $Var[X] = \frac{(1 - p)}{p^2}$

Statistical Distributions in Data

Geometric Distribution

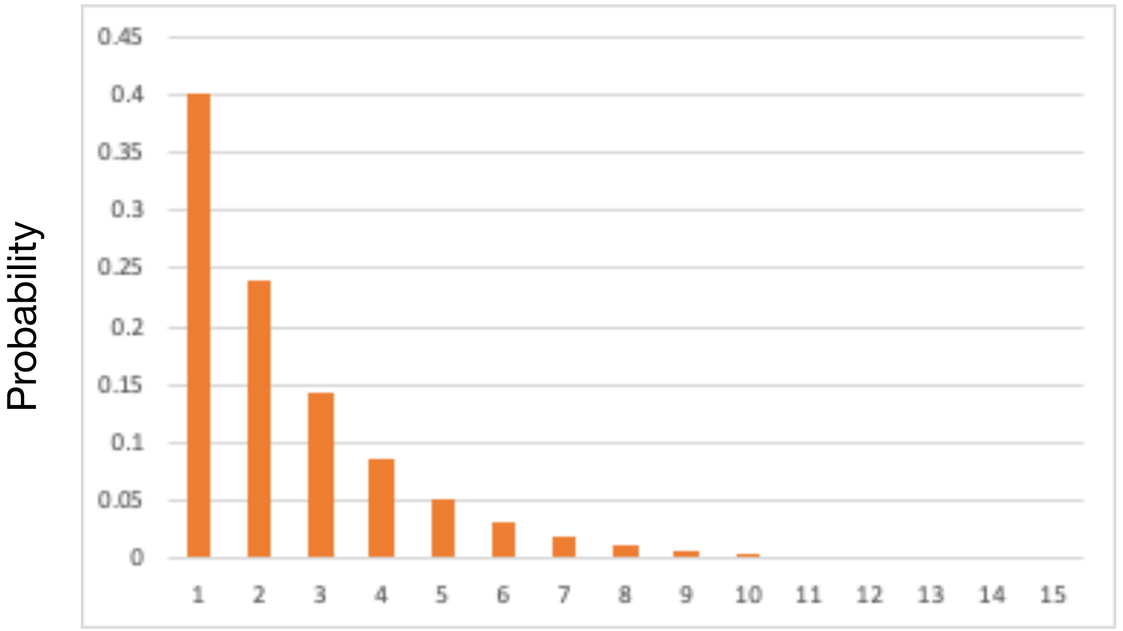
- Consider: 40%* of staff at the University of Warwick are academic staff. If we randomly select a staff member, what is the probability that the 4th person selected is the first academic?
- $p = 0.4$ (probability of selecting academic)
- $1 - p = 0.6$ (probability of not selecting academic)
- $x = 4$
- $$\begin{aligned} P(X = x) &= (1-p)^{x-1}p \\ &= (0.6)^3 \times 0.4 \\ &= 0.0864 \end{aligned}$$
- $$\begin{aligned} P(X \geq x) &= (1-p)^{x} \\ &= (0.6)^4 \\ &= 0.1296 \end{aligned}$$

*40% may not be accurate!

Statistical Distributions in Data

Geometric Distribution

- $E[X] = \frac{1}{p} = \frac{1}{0.4} = 2.5$
- $Var[X] = \frac{(1-p)}{p^2} = \frac{0.6}{0.4^2} = 3.75$



Statistical Distributions in Data

Exponential Distribution

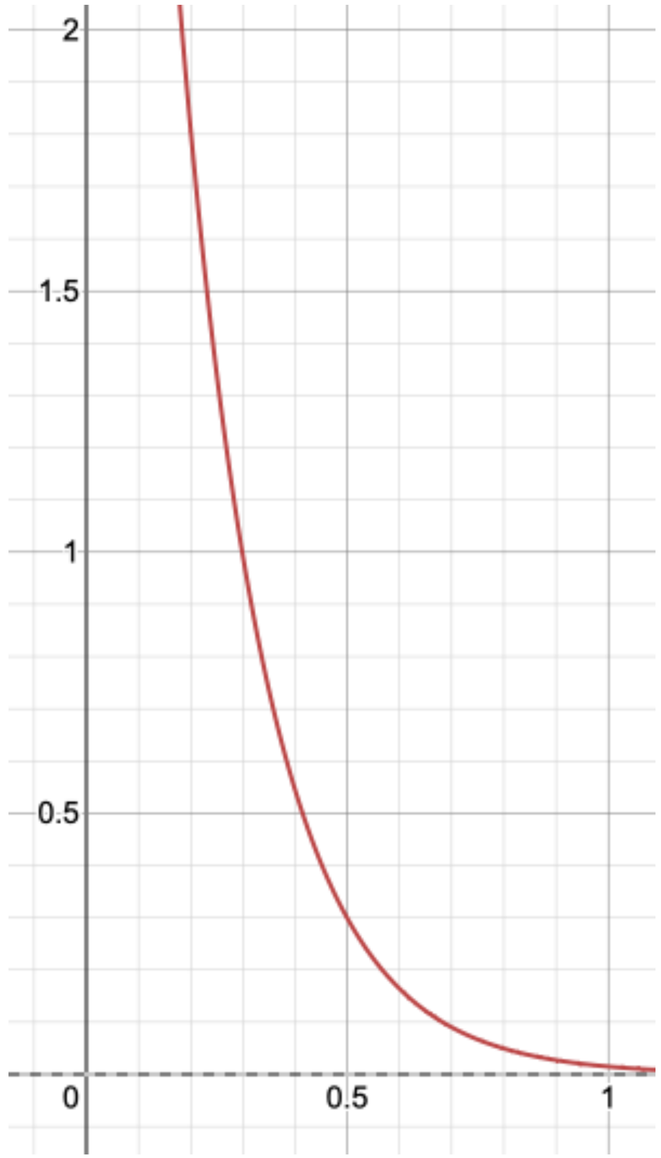
- Suppose that an event happens with probability p (for a small p) (Independently at each time step).
- **Continuous Case**: The *Exponential Distribution* is the probability distribution showing the “waiting time” between events in Poisson processes (events which occur at a constant average rate).

- PDF: $\lambda e^{-\lambda x}$, for parameter λ, $x \geq 0$
- $P[X \geq x] = e^{-\lambda x}$
- $E[X] = \frac{1}{\lambda}$
- $Var[X] = \frac{1}{\lambda^2}$

Statistical Distributions in Data

Exponential Distribution

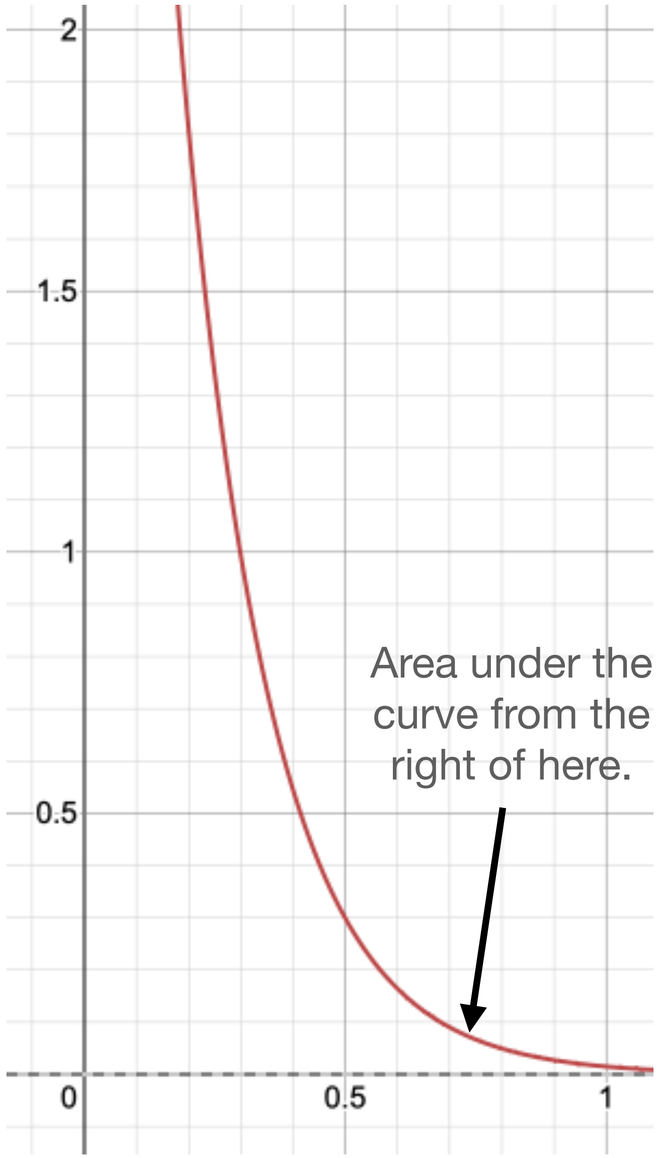
- Example - The rate of the number 11 bus arriving on campus is 6 per hour.
  - i.e. $\lambda = 6$
  - $\mu = \frac{1}{\lambda} = \frac{1}{6}$ (if 6 buses arrive per hour, then mean time between each bus is $\frac{1}{6}$ hours)!
  - Could graph the PDF using $\lambda e^{-\lambda x}$, i.e. $6e^{-6x}$.



Statistical Distributions in Data

Exponential Distribution

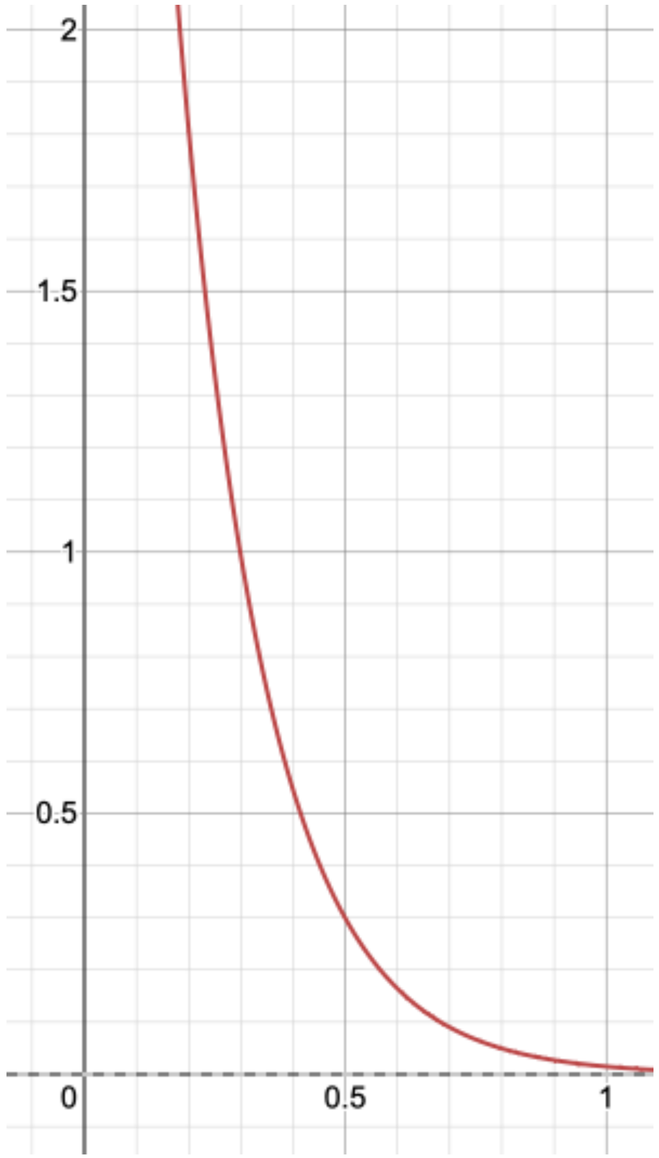
- Example - Find the probability that the next bus arrives after 45 minutes (0.75 hours).
  - $\lambda = 6$
  - $P[X \geq x] = e^{-\lambda x} = e^{-6 \times 0.75} = 0.0111089965$



Statistical Distributions in Data

Exponential Distribution

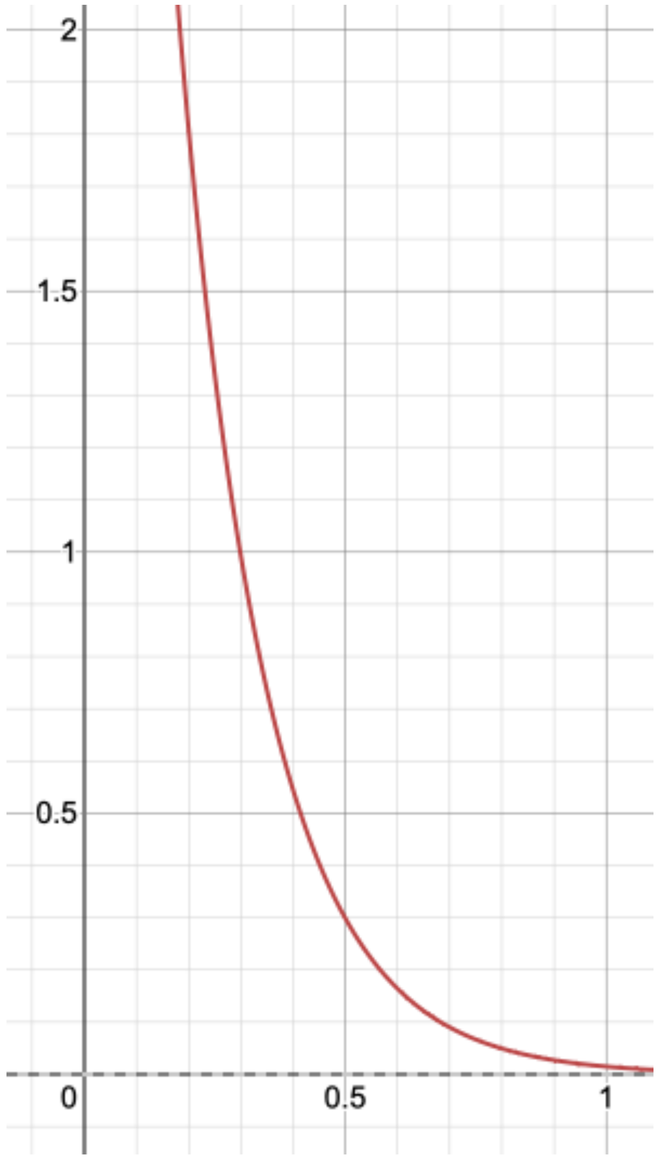
- Why is the PDF decreasing?
- Probability of bus arriving in first minute ($\frac{1}{60}$ of hour):
  - $P[X < x] = 1 - e^{-\lambda x} = 1 - e^{-6 \times \frac{1}{60}} = 0.095$
- Probability of bus arriving in second minute:
  - = Probability of bus **NOT** arriving in first minute × Probability of bus arriving in following minute
  - The probability of arriving in second minute is unaffected by the fact that a bus did not arrive in the first minute, due to *Memorylessness*.



Statistical Distributions in Data

Exponential Distribution

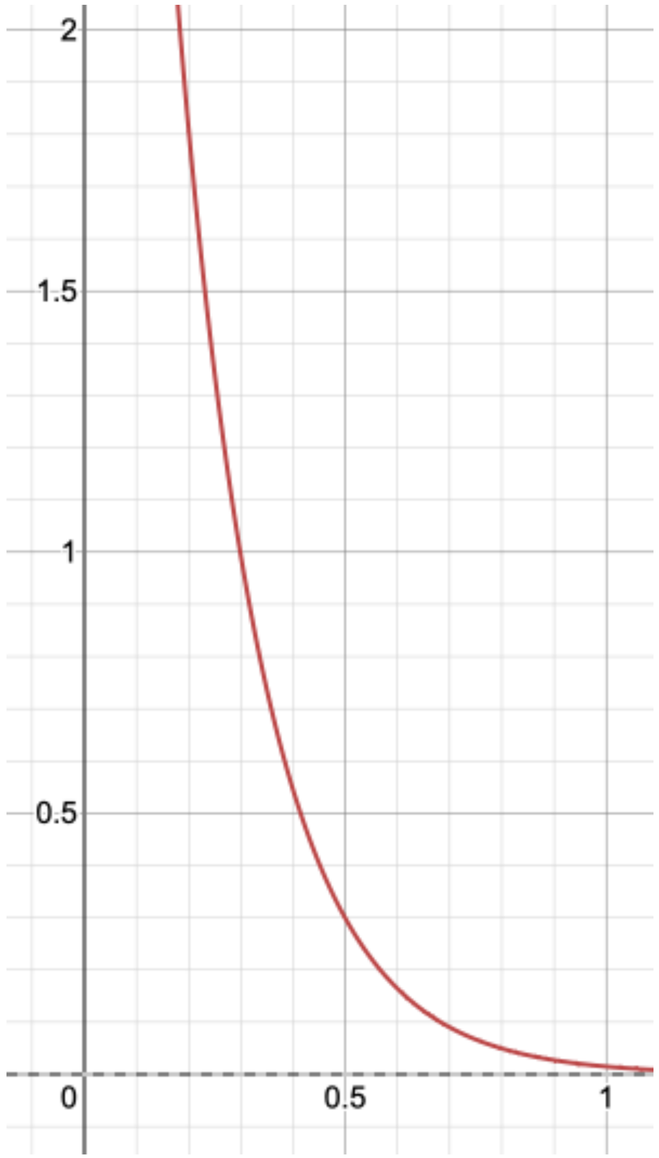
- Memorylessness:
  - $P[X > x + y \,|\, X > x] = P[X > y]$
  - Let x be 1 and y be 2:
    - $P[X > 3 \,|\, X > 1] = P[X > 2]$
    - We've currently been waiting for 1 minute. The probability of the bus arriving in 2+ minutes is unaffected by the fact that a bus has not arrived in the previous minute!



Statistical Distributions in Data

Exponential Distribution

- Probability of bus arriving in first minute ($\frac{1}{60}$ of hour):
  - $P[X < x] = 1 - e^{-\lambda x} = 1 - e^{-6 \times \frac{1}{60}} = 0.095$
- Probability of bus arriving in second minute:
  - = Probability of **NOT** arriving in first minute × probability of arriving in following minute
  - =(1-0.095) × 0.095
- Probability of bus arriving in third minute:
  - = Probability of **NOT** arriving in first or second minute × probability of arriving in following minute
  - $=(1-0.095)^2 \times 0.095$



Sample vs. Population

- Imagine we want to gather some data from all people in the world with a certain characteristic (gene, disease, etc.).
- Population = All people in the world with that characteristic.
- Can we really gather data from all of those people?
  - Probably not!
- Instead we gather data from a sample (a subset of the population).
  - This can introduce some problems (e.g. sampling error, selection bias,…).

Measures of Correlation

- Let's say we want to measure the correlation between two numeric variables.
  - Covariance of random variables X and Y:

$$\begin{aligned} Cov(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - YE[X] - XE[Y] + E[X]E[Y]] \\ &= E[XY] - E[Y]E[X] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

  - Notice: $Cov(X, X) = E[X^2] - E[X]^2 = Var(X)$
  - If X and Y are independent, then $E[XY] = E[X]E[Y]$: the covariance is 0
  - But if the covariance is 0, X and Y can still be related (dependent)…

Measures of Correlation

- But if the covariance is 0, X and Y can still be related (dependent)…
- Consider the following (table to the right):

$$E[X] = (-2 \times 0.25) + (-1 \times 0.25) + (1 \times 0.25) + (2 \times 0.25) = 0$$

$$E[XY] = ((-2 \times 4) \times 0.25) + ((-1 \times 1) \times 0.25) + ((1 \times 1) \times 0.25) + ((2 \times 4) \times 0.25) = 0$$

- Covariance is 0!

X =	x	-2	-1	1	2
Y =	x^2	4	1	1	4
	P(X=x)	0.25	0.25	0.25	0.25

Measures of Correlation

- But if the covariance is 0, X and Y can still be related (dependent)…
- Consider the following (table to the right):

$$E[X] = (-2 \times 0.25) + (-1 \times 0.25) + (1 \times 0.25) + (2 \times 0.25) = 0$$

$$E[XY] = ((-2 \times 4) \times 0.25) + ((-1 \times 1) \times 0.25) + ((1 \times 1) \times 0.25) + ((2 \times 4) \times 0.25) = 0$$

- Covariance is 0!

X =	x	-2	-1	1	2
Y =	x^2	4	1	1	4
	P(X=x)	0.25	0.25	0.25	0.25

0, yet obviously related (dependent).

Measures of Correlation

- For a population, we can calculate the covariance by:

$$Cov(A, B) = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n}$$

- For a sample, we can calculate the covariance by:

$$Cov(A, B) = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n - 1}$$

Measures of Correlation

Example

- Table to the right shows variables, X and Y, representing the intake of two academic departments over 5 years.
- $\bar{X} = \frac{240 + 290 + 310 + 300 + 280}{5} = \frac{1420}{5} = 284$
- $\bar{Y} = \frac{200 + 240 + 270 + 280 + 310}{5} = \frac{1300}{5} = 260$
- (Sample) $Cov(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{n - 1} = 725$
- As we have a positive covariance, we can say that variables rise together.

	X	Y
1	240	200
2	290	240
3	310	270
4	300	280
5	280	310

Measures of Correlation

Example

- Table to the right shows variables, X and Y, representing the intake of two academic departments over 5 years.
- $\bar{X} = \frac{240 + 290 + 310 + 300 + 280}{5} = \frac{1420}{5} = 284$
- $\bar{Y} = \frac{200 + 240 + 270 + 280 + 310}{5} = \frac{1300}{5} = 260$
- (Sample) $Cov(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{n - 1} = 725$
- As we have a positive covariance, we can say that variables rise together.

	X	Y
1	240	200
2	290	240
3	310	270

However, covariance is unbounded, and is heavily influenced by the scale of the variables X and Y.

Measuring Correlation

- Pearson product-moment correlation coefficient (PMCC):

  for a population:

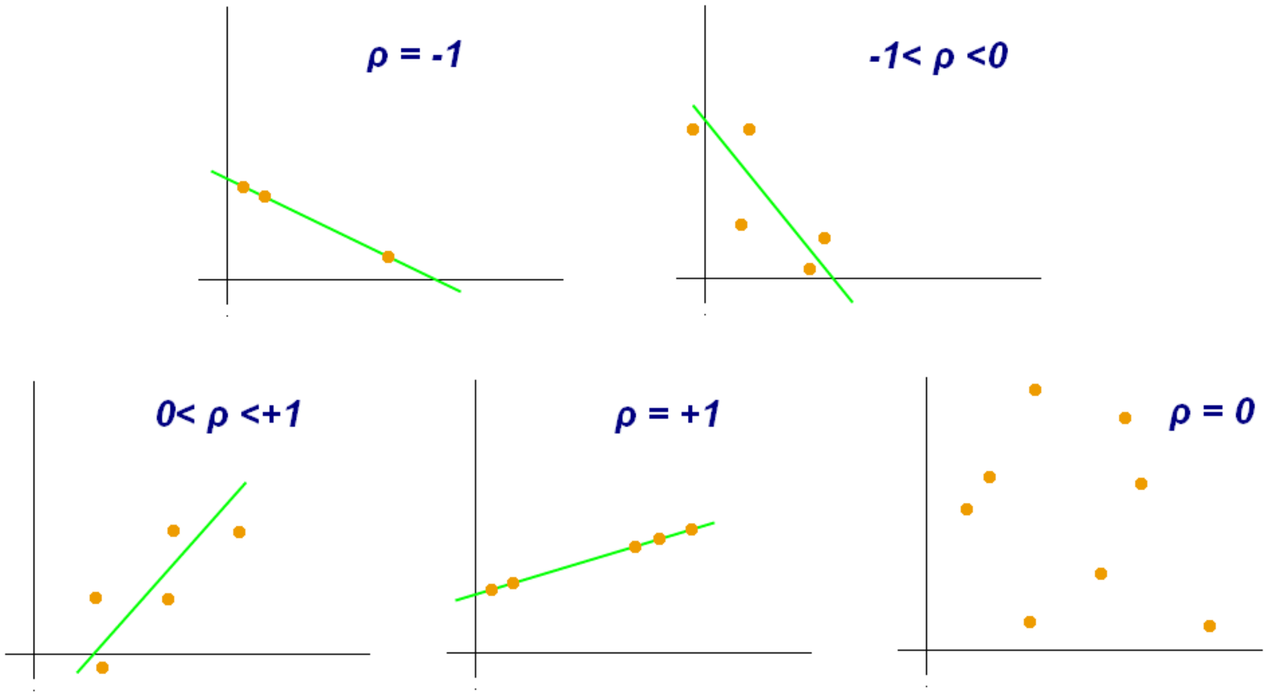
$$\rho_{X,Y} = \frac{E[XY] - E[X]E[Y]}{\sqrt{E[X^2] - (E[X])^2}\sqrt{E[Y^2] - (E[Y])^2}} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

  for a sample:

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

Measuring Correlation

- The Pearson product-moment correlation coefficient (PMCC) has a range of -1 to 1:
  - **-1**: Perfect negative linear relationship
  - **0**: No linear relationship
  - **1**: Perfect positive linear relationship



χ^2 (Chi-Square) Test

Correlation Analysis

- Pearson's product-moment coefficient is used for numerical variables. We need an approach for categorical (discrete) data. In this case, a correlation relationship can be discovered between variables A and B using a χ^2 test.

- Suppose A has c distinct values $(a_1, a_2, \ldots, a_c)$, and B has r distinct values $(b_1, b_2, \ldots, b_r)$. The records described by A and B can be shown as a Contingency Table, with the c values of A making up the columns, and the r values of B making up the rows.

- Let (A_i, B_j) denote the event that variable A takes on value a_i and variable B takes on value b_j.

χ^2 (Chi-Square) Test

Correlation Analysis

- Let (A_i, B_j) denote the event that variable A takes on value a_i and variable B takes on value b_j. Each and every possible (A_i, B_j) joint event has its own cell (or slot) in the table. The χ^2 (also known as the Pearson χ^2 statistic) is computed as:

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where o_{ij} is the observed frequency (i.e. actual count) of the joint event (A_i, B_j) and e_{ij} is the expected frequency of (A_i, B_j).

χ^2 (Chi-Square) Test

Correlation Analysis

- χ^2 tests the hypothesis that A and B are independent.

- The **Null Hypothesis** (H_0): A and B are independent.
- The **Alternative Hypothesis** (H_A): A and B are not independent.

χ^2 (Chi-Square) Test

Correlation Analysis

- e_{ij} can be computed as:

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n}$$

where n is the number of records, $count(A = a_i)$ is the number of records having value a_i for A, and $count(B = b_j)$ is the number of records having value b_j for B.

- NOTE: Cells that contribute the most to the χ^2 value are those whose actual count is very different from that expected.

χ^2 (Chi-Square) Test

Correlation Analysis

- The test is based on a significance level, with $(r-1)\times(c-1)$ degrees of freedom.
- We need to chose a significance level.
  Example: If the significance level $= 0.05$ and our degree of freedom = 1, then we can be 95% (1-0.05) sure that we are correctly rejecting our null hypothesis if the value of χ^2 exceeds 3.841.

	Significance Level										
Degrees of freedom (df)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.63	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.61	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.81	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59

χ^2 (Chi-Square) Test

Example

- Suppose that a group of 1,500 people was surveyed. The sex of each person, as well as their preference for fiction or non-fiction material.

	Male	Female	Total
Fiction	250	200	450
Non-Fiction	50	1000	1050
Total	300	1200	1500

Table from: Han, J., Pei, J. and Kamber, M., 2012. *Data Mining: Concepts and Techniques*. Elsevier.

χ^2 (Chi-Square) Test

Example

- We now need to calculate the expected frequency for each combination. Recall:

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n}$$

- So, for the pair $(male, fiction)$:

$$e_{ij} = \frac{count(male) \times count(fiction)}{n}$$

$$e_{11} = \frac{300 \times 450}{1500}$$

$$= 90$$

	Male	Female	Total
Fiction	250	200	450
Non-Fiction	50	1000	1050
Total	300	1200	1500

Table from: Han, J., Pei, J. and Kamber, M., 2012. *Data Mining: Concepts and Techniques*. Elsevier.

χ^2 (Chi-Square) Test

Example

- We now need to calculate the expected frequency for each combination. They have been added in brackets to the table.

- Now, for the χ^2 test. Recall:

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$\chi^2 = \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} = 507.93$$

	Male	Female	Total
Fiction	250 (90)	200 (360)	450
Non-Fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Table from: Han, J., Pei, J. and Kamber, M., 2012. *Data Mining: Concepts and Techniques*. Elsevier.

χ^2 (Chi-Square) Test

Example

$$\chi^2 = 507.93$$

- Choose a significance level of 0.05. DF is 1 $((2-1)(2-1) = (1)(1) = 1)$.
- 507.93 > 3.841. So we can reject the null hypothesis that sex and preferred reading are independent, and conclude that the two are strongly correlated for the given group of people.

	Significance Level										
Degrees of freedom (df)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.63	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.61	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.81	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59

	Male	**Female**	**Total**
Fiction	250 (90)	200 (360)	450
Non-Fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Bottom table from: Han, J., Pei, J. and Kamber, M., 2012. *Data Mining: Concepts and Techniques*. Elsevier.

Acknowledgements

- Florin Ciucu [Warwick, CS430/CS910]
- Tijms, H., 2017. Probability: a lively introduction. Cambridge University Press.
- zedStatistics: https://www.youtube.com/watch?v=2kg1O0j1J9c
- jbStatistics: https://www.youtube.com/watch?v=zq9Oz82iHf0
- Han, J. and Kamber, M., 2006. *Data Mining: Concepts and Techniques*. Elsevier.
- Han, J., Pei, J. and Kamber, M., 2012. *Data Mining: Concepts and Techniques*. Elsevier.