

Statistics for Data Analytics

Dr. John S. Butler

Lecturer

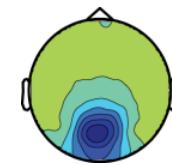
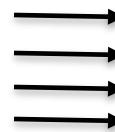
- John Butler
- e-mail: john.s.butler@tudublin.ie

My Background

- Numerical Analysis (Trinity College Dublin, PhD work)
 - Robust Numerical methods of Prandtl Boundary Layer Problems

- Self-motion Perception (Max Planck Institute for Biological Cybernetics)
 - Walking
 - Driving

- Unisensory and Multisensory processing
 - Developmental Disorders (Albert Einstein College of Medicine)
 - Autism Spectrum Disorder, Niemann Pick Type C
 - Movement Disorders (Trinity Centre for Bioengineering)
 - Parkinson's Disease
 - Dystonia



BrightSpace

The screenshot shows the BrightSpace Learning Management System (LMS) interface. At the top, there is a navigation bar with links for 'My Home', 'Progress', 'ePortfolio', 'Content', 'Assessment', 'Module Tools', 'Library', and 'Help'. Below the navigation bar is a banner with the module title 'Intro to Prob and Stat Inference MATH9901: 2022-23'. The main content area displays five course modules with their completion status:

Module	Topics Completed	Total Topics	Completion Percentage	Action
Introduction	6	34	18%	Continue
Data Types and Measures of Location	4	11	36%	Continue
Mathematical Probability	3	23	13%	Continue
Continuous Probability Distribution...	0	5	0%	Continue
None examinable material	0	1	0%	Continue

On the right side, there are two boxes: 'Announcements' (which states 'There are no announcements to display.') and 'Calendar' (which shows 'Monday, 19 September 2022'). The BrightSpace logo is visible in the top left corner of the page.

OneNote

OneNote

Home Insert Draw View Class Notebook

Table File Printout Picture Screenshot Link Equation Date Date & Time

Audio Recording Meeting Stickers Details

Probability and Statistical Inference 2021-2022

Counting Probability Discrete Distributions Review 1 Hypothesis Testing Regression Revision

Counting Example Question – Ice Hockey Problem Sheet 1a Problem Sheet 1b COVID Example

52
4 suits
13 cards types 1-10,J,Q,K

$\frac{52 \times 51 \times 50 \times 49}{4} \times 4!$ order

$52 \times 51 \times 50 \times 49 \times 4!$

$5! = 5 \times 4 \times 3 \times 2 \times 1$

$\binom{52}{4} = \frac{52 \times 51 \times 50 \times 49}{4 \times 3 \times 2 \times 1} = 270,725$

$\binom{52}{4} = \frac{52!}{(48! \cdot 4!)} = 270,725 = \frac{52 \times 51 \times 50 \times 49}{4 \times 3 \times 2 \times 1}$

3. A bank issues bank cards with PINs consisting of 4 digits, each one {0,1,2,...,9}. How many unique PINs are there if

- Any 4-digit code can be used.
 $10^4 = 10 \times 10 \times 10 \times 10 = 10,000$
- The digits must be different.

$10 \times 9 \times 8 \times 7 = 5040$

${}^4P_4 = \frac{10!}{6!} = 5,040 = \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{6 \times 5 \times 4 \times 3 \times 2 \times 1}$

4. In a lottery, each ticket has 5 one-digit numbers 0-9 which is not repeated on it.
i. You win if your ticket has the digits in any order. What are the total number of possible combinations?

10 1 8 7 6

OLLScoil TEICNEOLAÍOCHTA BHÁILE ÁTHA CLIATH

DUBLIN TECHNOLOGICAL UNIVERSITY DUBLIN

The Whole Module in one hours



Probability and Statistical Inference

Cheat Sheet

John S Butler (TU Dublin)

Course Twitter Account

Data Type

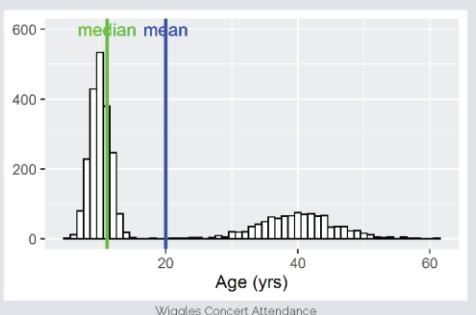
- Categorical
- Interval
- Ordinal
- Ratio

Measures of Location

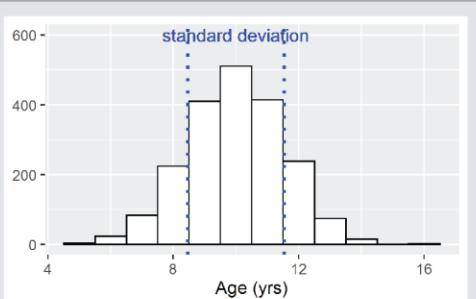
Different aspects of a distribution of data can be summarised by the measures of location:

- The First Moment: Mean, Mode or Median;
- The Second Moment: Variance, Standard Deviation;
- The Third Moment: Skewness.

First Moment: Middle

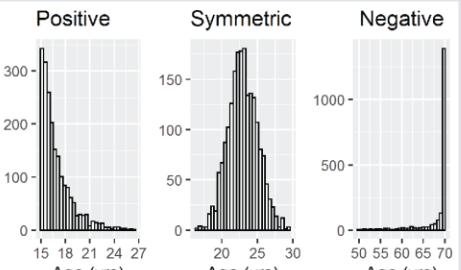


Second Moment: Spread



Measures of Location (cont.)

Third Moment: Symmetry



Left: One Direction Concert Attendance
Middle: Harry Styles Concert Attendance
Right: André Rieu Concert Attendance

Mathematical Probability

Definitions

Define some event A that can be the outcome of an experiment. $\Pr(A)$ is the probability of a given event A will happen. Rules:

- $\Pr(A)$ is between 0 and 1, $0 \leq \Pr(A) \leq 1$;
- $\Pr(A) = 1$, means it will definitely happen;
- $\Pr(A) = 0$, means it will definitely **not** happen;
- $\Pr(A) = 0.05$, is arbitrarily considered unlikely.

Sample Space and Events

The **Sample Space**, S , of an experiment is the universal set of all possible outcomes for that experiment, defined so, no two outcomes can occur simultaneously. For example:

- Throwing a die $S = \{1, 2, 3, 4, 5, 6\}$;
 - Tossing two coins $S = \{HH, TH, HT, TT\}$.
- An event, A , is a subset of the sample space S . For example:
- Throwing a die $S = \{3, 4, 6\}$;
 - Tossing two coins $S = \{TH, TT\}$.

Axioms of Probabilities

For an event A subset S associated a number $\Pr(A)$, the probability of A , which must have the following properties

- $\Pr(A \cap B) = 0$; $\Pr(A \cup B) = \Pr(A) + \Pr(B)$;
- Probability of the Null Event $\Pr(\emptyset) = 0$;
- The probability of the complement of A , $\Pr(\bar{A}) = 1 - \Pr(A)$;
- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

Conditional Probability

The Conditional Probability $\Pr(A|B)$ denotes the probability of the event A occurring given that the event B has occurred.

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

Example: The rain in Ireland

A normal probability would be what is the probability it is going to rain, $\Pr(\text{rain})$. A conditional probability would be what is the probability it is going to rain **given** that you are in Ireland, $\Pr(\text{rain}|Ireland)$,

$$\Pr(\text{rain}|Ireland) = \frac{\Pr(\text{rain} \cap \text{Ireland})}{\Pr(\text{Ireland})},$$

where the probability of rain is $\Pr(\text{rain}) = 0.3$, the probability of being in Ireland is $\Pr(\text{Ireland}) = 0.4$ and the probability of being in Ireland and it raining is $\Pr(\text{rain} \cap \text{Ireland}) = 0.2$.

$$\Pr(\text{rain}|Ireland) = \frac{0.2}{0.4} = 0.5,$$

You could be interested in the probability that you are in Ireland **given** that it is raining,

$$\Pr(\text{Ireland}|\text{rain}) = \frac{\Pr(\text{rain} \cap \text{Ireland})}{\Pr(\text{rain})} = \frac{0.2}{0.3} = 0.75.$$

Bayes Theorem

Bayes Theorem states

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}.$$

Example: Diagnostic test

The probability that an individual has a rare disease is $\Pr(\text{Disease}) = 0.01$. The probability that a diagnostic test results in a positive (+) test **given you have** the disease is $\Pr(+|\text{Disease}) = 0.95$. On the other hand, the probability that the diagnostic test results in a positive (+) test **given you do not have** the disease is $\Pr(+|\text{No Disease}) = 0.1$. This raises the important question if you are given a positive diagnosis, what is the probability you have the disease $\Pr(\text{Disease}|+)$? From Bayes Theorem we have:

$$\Pr(\text{Disease}|+) = \frac{\Pr(+|\text{Disease})\Pr(\text{Disease})}{\Pr(+)}$$

The probability of a positive test is,

$$\Pr(+) = \Pr(+|\text{Disease})\Pr(\text{Disease}) + \Pr(+|\text{No Disease})\Pr(\text{No Disease}),$$

$$\Pr(+) = 0.1085.$$

$$\Pr(\text{Disease}|+) = \frac{\Pr(+|\text{Disease})\Pr(\text{Disease})}{\Pr(+)} = \frac{0.95 \times 0.01}{0.1085} = 0.0875576.$$

This can also be done in a simple table format, by assume a population of 10,000

Group	+ Diagnosis	- Diagnosis	Total
Disease	95	5	100
No Disease	990	8,910	9,900
Total	1,085	8,915	10,000

From the table we can calculate the same answer,
 $\Pr(\text{Disease}|+) = \frac{95}{1085} = 0.0875576$.

Probability and Statistical Inference

Cheat Sheet

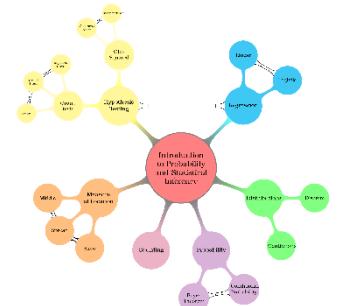
John S Butler (TU Dublin)

Course Twitter Account

Data Type

- Categorical
 - Interval
 - Ordinal
 - Ratio

Week 1

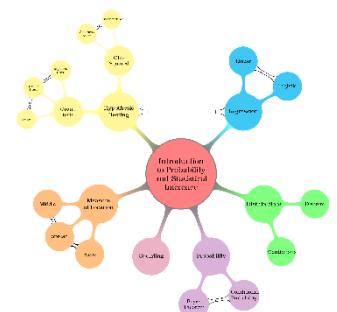


Measures of Location

Different aspects of a distribution of data can be summarised by the measures of location:

1. The First Moment: Mean, Mode or Median;
2. The Second Moment: Variance, Standard Deviation;
3. The Third Moment: Skewness.

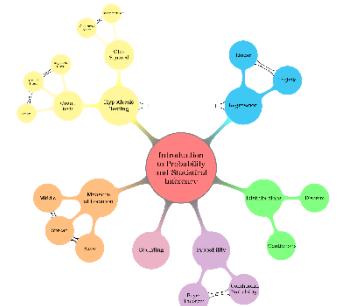
Week 1



What Concert is it

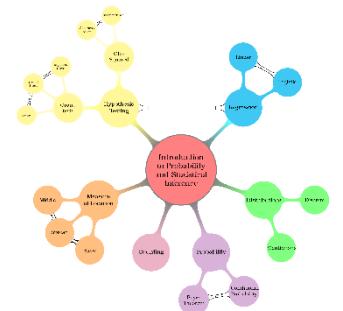
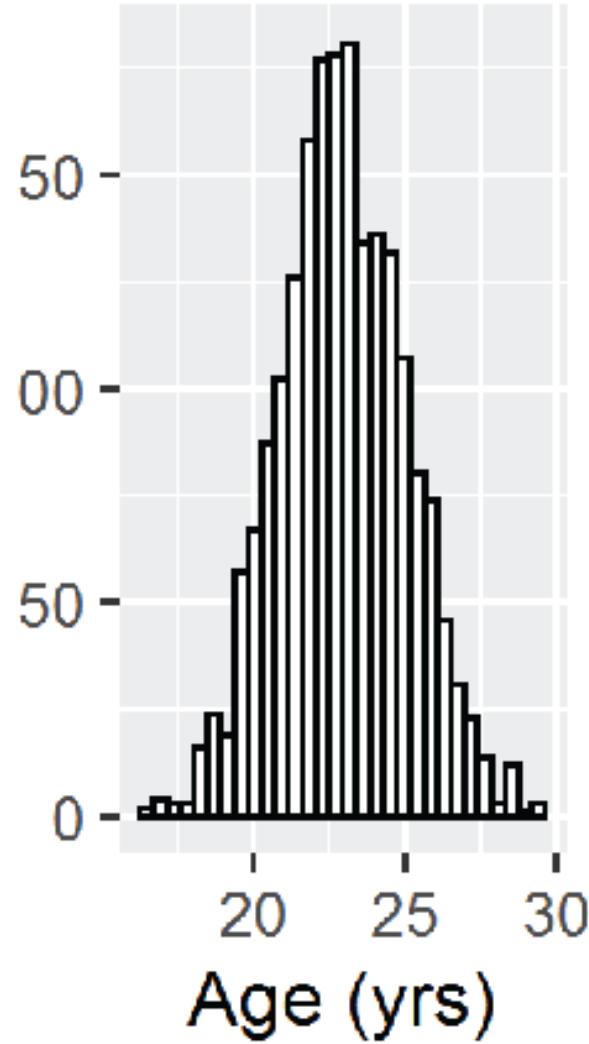
- The mean age is 23?
- Is it a Taylor Swift Concert?
- Is it a Wiggles Concert?

Week 1



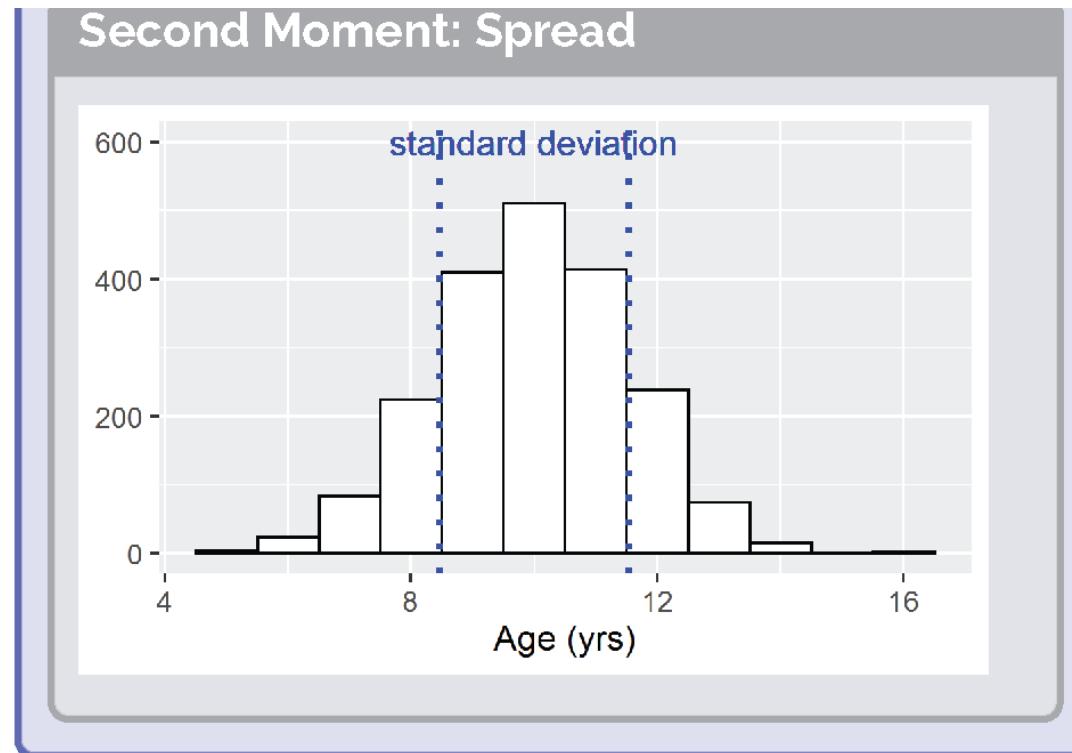
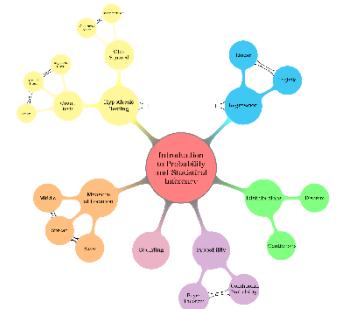
Mean (average) and Median

Taylor Swift Concert



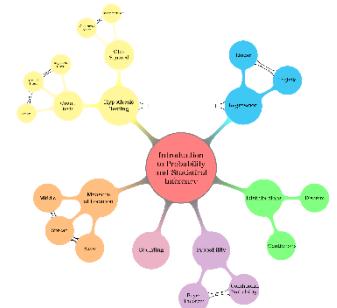
Spread

Week 1



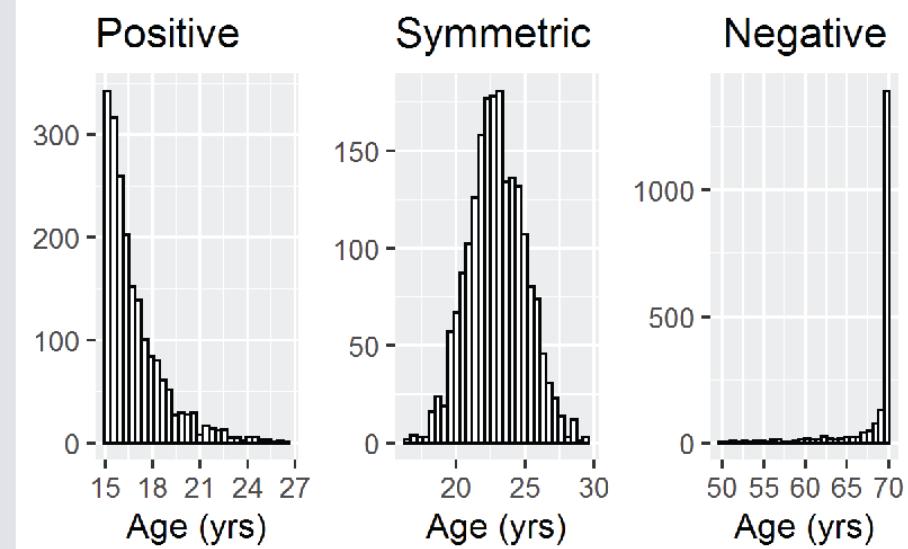
Skewness

Week 1



Measures of Location (cont.)

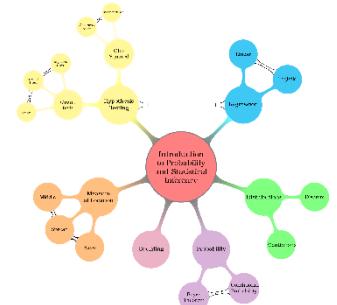
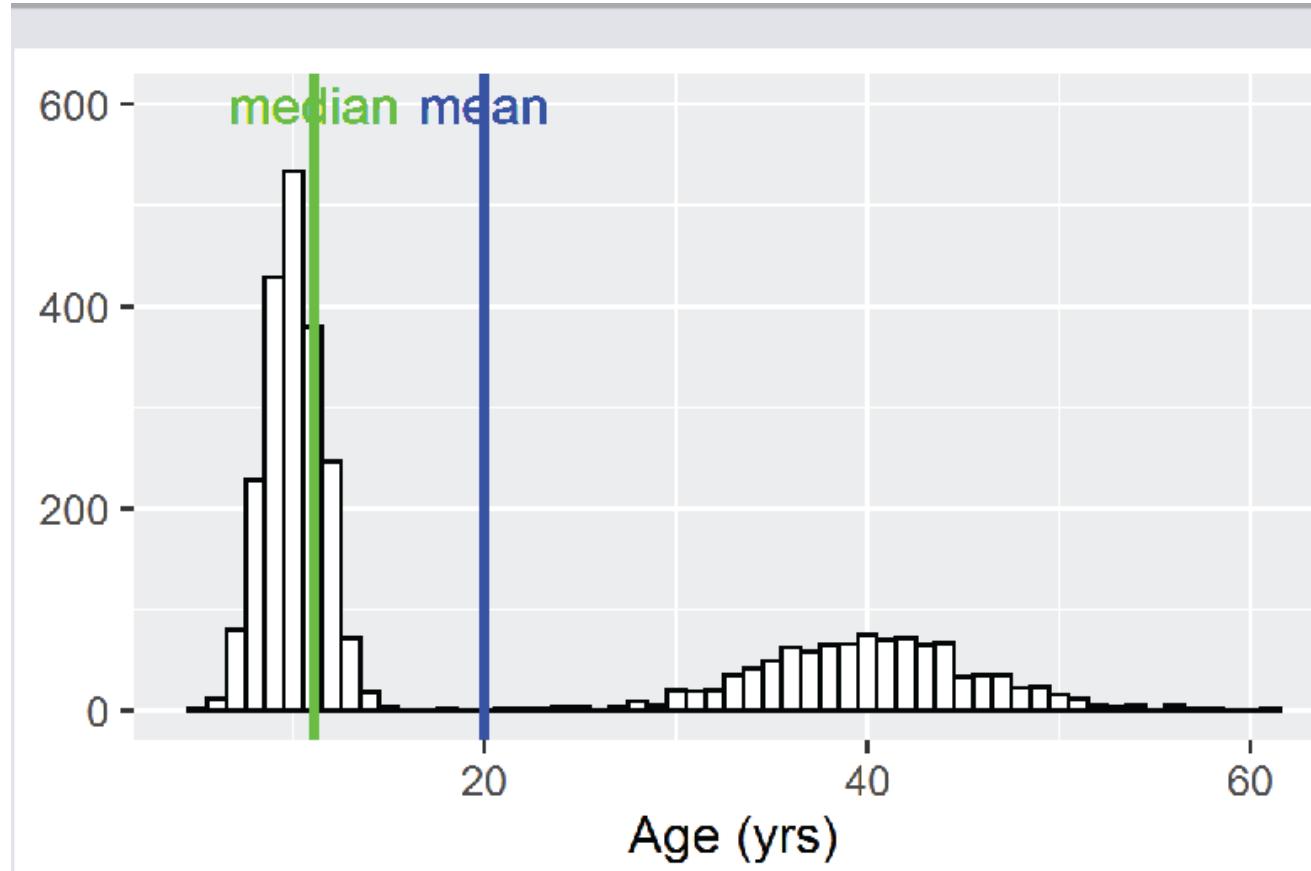
Third Moment: Symmetry



Left: One Direction Concert Attendance
Middle: Harry Styles Concert Attendance
Right: André Rieu Concert Attendance

The Wiggles Break everything

Wiggles Concert





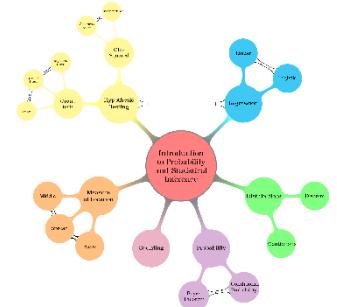
Counting

- Number of Royal Flushes
- Probability of a Royal Flush





Week 2



Mathematical Probability

Definitions

Define some event A that can be the outcome of an experiment.

$\Pr(A)$ is the probability of a given event A will happen.

Rules:

- $\Pr(A)$ is between 0 and 1, $0 \leq \Pr(A) \leq 1$;
- $\Pr(A) = 1$, means it will definitely happen;
- $\Pr(A) = 0$, means it will definitely **not** happen;
- $\Pr(A) = 0.05$, is arbitrarily considered unlikely.

Sample Space and Events

The **Sample Space**, S , of an experiment is the universal set of all possible outcomes for that experiment, defined so, no two outcomes can occur simultaneously. For example:

- Throwing a die $S = \{1, 2, 3, 4, 5, 6\}$;
- Tossing two coins $S = \{HH, TH, HT, TT\}$.

An event, A , is a subset of the sample space S . For example:

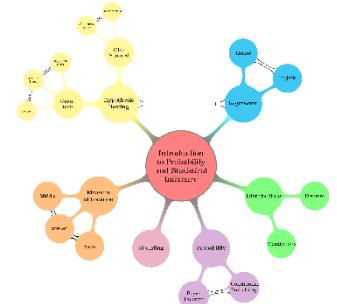
- Throwing a die $S = \{3, 4, 6\}$;
- Tossing two coins $S = \{TH, TT\}$.

Axioms of Probabilities

For an event A subset S associated a number $\Pr(A)$, the probability of A , which must have the following properties

- $\Pr(A \cap B) = 0$; $\Pr(A \cup B) = \Pr(A) + \Pr(B)$;
- Probability of the Null Event $\Pr(\emptyset) = 0$;
- The probability of the complement of A , $\Pr(\bar{A}) = 1 - \Pr(A)$;
- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

Week 2



Conditional Probability

The Conditional Probability $\Pr(A|B)$ denotes the probability of the event A occurring given that the event B has occurred.

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

Example: The rain in Ireland

A normal probability would be what is the probability it is going to rain, $\Pr(\text{rain})$. A conditional probability would be what is the probability it is going to rain **given** that you are in Ireland, $\Pr(\text{rain}|\text{Ireland})$,

$$\Pr(\text{rain}|\text{Ireland}) = \frac{\Pr(\text{rain} \cap \text{Ireland})}{\Pr(\text{Ireland})},$$

where the probability of rain is $\Pr(\text{rain}) = 0.3$, the probability of being in Ireland is $\Pr(\text{Ireland}) = 0.4$ and the probability of being in Ireland and it raining is $\Pr(\text{rain} \cap \text{Ireland}) = 0.2$.

$$\Pr(\text{rain}|\text{Ireland}) = \frac{0.2}{0.4} = 0.5,$$

You could be interested in the probability that you are in Ireland **given** that it is raining,

$$\Pr(\text{Ireland}|\text{rain}) = \frac{\Pr(\text{rain} \cap \text{Ireland})}{\Pr(\text{rain})} = \frac{0.2}{0.3} = 0.75.$$

Bayes Theorem

Bayes Theorem states

$$\Pr(A|B) = \frac{\Pr(B|A)P(A)}{\Pr(B)}.$$

Example: Diagnostic test

The probability that an individual has a rare disease is $\Pr(\text{Disease}) = 0.01$. The probability that a diagnostic test results in a positive (+) test *given you have the disease* is $\Pr(+|\text{Disease}) = 0.95$. On the other hand, the probability that the diagnostic test results in a positive (+) test *given you do not have the disease* is $\Pr(+|\text{No Disease}) = 0.1$. This raises the important question if you are given a positive diagnosis, what is the probability you have the disease $\Pr(\text{Disease}|+)$? From Bayes Theorem we have:

$$\Pr(\text{Disease}|+) = \frac{\Pr(+|\text{Disease}) \Pr(\text{Disease})}{\Pr(+)}$$

The probability of a positive test is,

$$\Pr(+) = \Pr(+|\text{Disease}) \Pr(\text{Disease}) + \Pr(+|\text{No Disease}) \Pr(\text{No Disease}),$$

$$\Pr(+) = 0.1085.$$

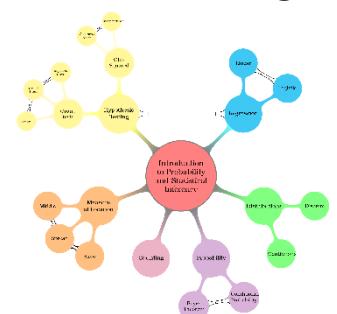
$$\Pr(\text{Disease}|+) = \frac{\Pr(+|\text{Disease}) \Pr(\text{Disease})}{\Pr(+)} = \frac{0.95 \times 0.01}{0.1085} = 0.0875576.$$

This can also be done in a simple table format, by assume a population of 10,000

Group	+ Diagnosis	- Diagnosis	Total
Disease	95	5	100
No Disease	990	8,910	9,900
Total	1,085	8,915	10,000

From the table we can calculate the same answer,
 $\Pr(\text{Disease}|+) = \frac{95}{1085} = 0.0875576$.

Week 2-3



Probability and Statistical Inference

Cheat Sheet

John S Butler (TU Dublin)

Course Twitter Account

Data Type

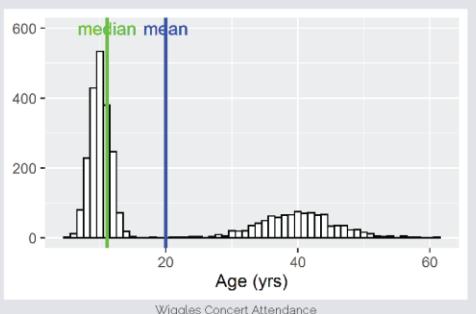
- Categorical
- Interval
- Ordinal
- Ratio

Measures of Location

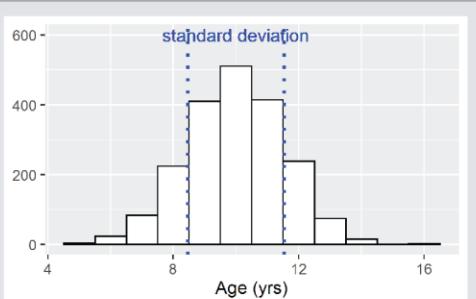
Different aspects of a distribution of data can be summarised by the measures of location:

- The First Moment: Mean, Mode or Median;
- The Second Moment: Variance, Standard Deviation;
- The Third Moment: Skewness.

First Moment: Middle

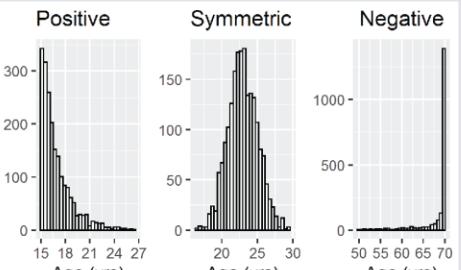


Second Moment: Spread



Measures of Location (cont.)

Third Moment: Symmetry



Left: One Direction Concert Attendance
Middle: Harry Styles Concert Attendance
Right: André Rieu Concert Attendance

Mathematical Probability

Definitions

Define some event A that can be the outcome of an experiment.

$\Pr(A)$ is the probability of a given event A will happen.

Rules:

- $\Pr(A)$ is between 0 and 1, $0 \leq \Pr(A) \leq 1$;
- $\Pr(A) = 1$, means it will definitely happen;
- $\Pr(A) = 0$, means it will definitely **not** happen;
- $\Pr(A) = 0.05$, is arbitrarily considered unlikely.

Sample Space and Events

The **Sample Space**, S , of an experiment is the universal set of all possible outcomes for that experiment, defined so, no two outcomes can occur simultaneously. For example:

- Throwing a die $S = \{1, 2, 3, 4, 5, 6\}$;
 - Tossing two coins $S = \{HH, TH, HT, TT\}$.
- An event, A , is a subset of the sample space S . For example:
- Throwing a die $S = \{3, 4, 6\}$;
 - Tossing two coins $S = \{TH, TT\}$.

Axioms of Probabilities

For an event A subset S associated a number $\Pr(A)$, the probability of A , which must have the following properties

- $\Pr(A \cap B) = 0$; $\Pr(A \cup B) = \Pr(A) + \Pr(B)$;
- Probability of the Null Event $\Pr(\emptyset) = 0$;
- The probability of the complement of A , $\Pr(\bar{A}) = 1 - \Pr(A)$;
- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

Conditional Probability

The Conditional Probability $\Pr(A|B)$ denotes the probability of the event A occurring given that the event B has occurred.

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

Example: The rain in Ireland

A normal probability would be what is the probability it is going to rain, $\Pr(\text{rain})$. A conditional probability would be what is the probability it is going to rain **given** that you are in Ireland, $\Pr(\text{rain}|Ireland)$,

$$\Pr(\text{rain}|Ireland) = \frac{\Pr(\text{rain} \cap \text{Ireland})}{\Pr(\text{Ireland})},$$

where the probability of rain is $\Pr(\text{rain}) = 0.3$, the probability of being in Ireland is $\Pr(\text{Ireland}) = 0.4$ and the probability of being in Ireland and it raining is $\Pr(\text{rain} \cap \text{Ireland}) = 0.2$.

$$\Pr(\text{rain}|Ireland) = \frac{0.2}{0.4} = 0.5,$$

You could be interested in the probability that you are in Ireland **given** that it is raining,

$$\Pr(\text{Ireland}|\text{rain}) = \frac{\Pr(\text{rain} \cap \text{Ireland})}{\Pr(\text{rain})} = \frac{0.2}{0.3} = 0.75.$$

Bayes Theorem

Bayes Theorem states

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}.$$

Example: Diagnostic test

The probability that an individual has a rare disease is $\Pr(\text{Disease}) = 0.01$. The probability that a diagnostic test results in a positive (+) test **given you have** the disease is $\Pr(+|\text{Disease}) = 0.95$. On the other hand, the probability that the diagnostic test results in a positive (+) test **given you do not have** the disease is $\Pr(+|\text{No Disease}) = 0.1$. This raises the important question if you are given a positive diagnosis, what is the probability you have the disease $\Pr(\text{Disease}|+)$? From Bayes Theorem we have:

$$\Pr(\text{Disease}|+) = \frac{\Pr(+|\text{Disease})\Pr(\text{Disease})}{\Pr(+)}$$

The probability of a positive test is,

$$\Pr(+) = \Pr(+|\text{Disease})\Pr(\text{Disease}) + \Pr(+|\text{No Disease})\Pr(\text{No Disease}),$$

$$\Pr(+) = 0.1085.$$

$$\Pr(\text{Disease}|+) = \frac{\Pr(+|\text{Disease})\Pr(\text{Disease})}{\Pr(+)} = \frac{0.95 \times 0.01}{0.1085} = 0.0875576.$$

This can also be done in a simple table format, by assume a population of 10,000

Group	+ Diagnosis	- Diagnosis	Total
Disease	95	5	100
No Disease	990	8,910	9,900
Total	1,085	8,915	10,000

From the table we can calculate the same answer,
 $\Pr(\text{Disease}|+) = \frac{95}{1085} = 0.0875576$.



Discrete Distribution

Probability Mass Functions

Event Number i	0	1	2	3	4
Event Value x_i	-1	0	1	2	3
Probability of Event $\Pr(x_i)$	0.3	0.1	0.3	0.1	0.2

The expected value of the distribution is:

$$\mu = E[X] = \sum_i x_i \Pr(x_i),$$

$$\sum_i x_i p(x_i) = -1 \times 0.4 + 0 \times 0.1 + 1 \times 0.3 + 0.1 \times 2 + 0.2 \times 3 = 0.7,$$

The variance of the distribution is:

$$Var[X] = \sum_i (x_i - \mu)^2 p(x_i) = \sum_i (x_i - 0.7)^2 p(x_i).$$

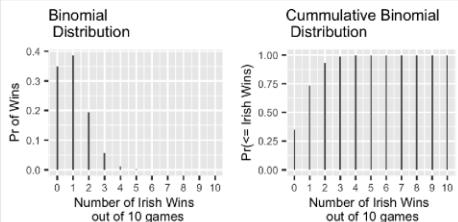
Binomial Distribution

The formula for the Binomial distribution is:

$$\Pr(k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, 2, \dots, n,$$

$$E[k] = np, \quad Var[k] = npq,$$

where n is the total of games, k is the number of 'wins', p is the probability of a 'win', $q = 1 - p$ probability of a 'loss'.



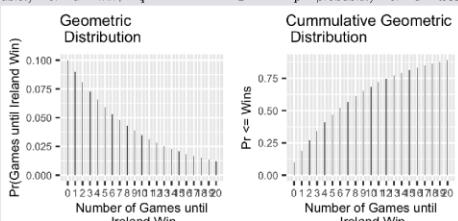
Geometric Distribution

The formula for the Geometric distribution is:

$$\Pr(k) = q^{k-1} p, \quad k = 1, 2, \dots$$

$$E[k] = \frac{1}{p}, \quad Var[k] = \frac{q}{p^2},$$

k is the number of events until one 'win', p is the probability of a 'win', $q = 1 - p$ probability of a 'loss'.



Discrete Distribution

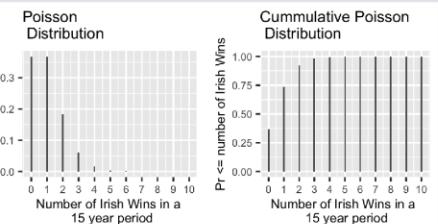
Poisson Distribution

The formula for the Poisson distribution is:

$$\Pr(k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots$$

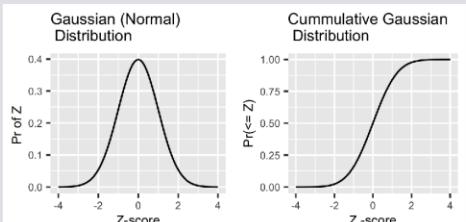
$$E[k] = \lambda, \quad Var[k] = \lambda,$$

where λ is the mean and standard deviation of the distribution and k is the number of 'wins' in a specified time or space.

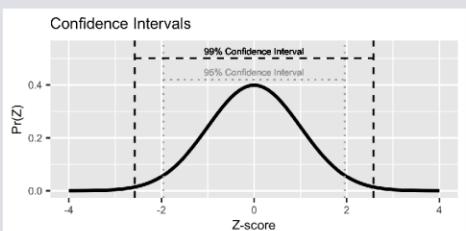


Continuous Distribution

Normal Distribution



Confidence Intervals



Hypothesis Testing

Five steps for Hypothesis testing:

- State the Null Hypothesis H_0 :
- State an Alternative Hypothesis H_α :
- Calculate a Test Statistic (see below);
- Calculate a p-value and/or set a rejection region;
- State your conclusions.

z-test

Continuous Data

The test statistic is given by

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1),$$

where \bar{x} is the observed mean, μ is the historical mean, σ is the standard deviation and n is the number of observations. $\mathcal{N}(0, 1)$ is the normal distribution with a mean of 0 and a standard deviation of 1.

Do supplements make you faster?

The effect of a food supplements on the response time in rats is of interest to a biologist. They have established that the normal response time of rats is $\mu = 1.2$ seconds. The $n = 100$ rats were given a new food supplement. The following summary statistics were recorded from the data: $\bar{x} = 1.05$ and $\sigma = 0.5$ seconds

- The rats in the study are the same as normal rats, $H_0 : \mu = 1.2$.
- The rats are different, $H_\alpha : \mu \neq 1.2$.
- Calculate a Test Statistic $Z = \frac{1.05 - 1.2}{0.5/\sqrt{100}} = -3$
- Reject the Null hypothesis H_0 if $Z < -1.96$ and $Z > 1.96$
- The data suggests that rats are faster with the new food.

Proportional Data

The test statistic is given by

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \sim \mathcal{N}(0, 1),$$

where \hat{p} is the observed proportion, p_0 is the historical proportion, q_0 is the complement $q_0 = 1 - p_0$, and n is the number of observations.

Discrete Distribution

Probability Mass Functions

Event Number i	0	1	2	3	4
Event Value x_i	-1	0	1	2	3
Probability of Event $\Pr(x_i)$	0.3	0.1	0.3	0.1	0.2

The expected value of the distribution is:

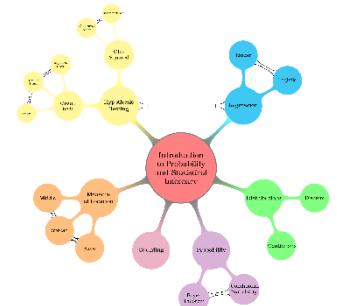
$$\mu = E[X] = \sum_i x_i \Pr(x_i),$$

$$\sum_i x_i p(x_i) = -1 \times 0.4 + 0 \times 0.1 + 1 \times 0.3 + 0.1 \times 2 + 0.2 \times 3 = 0.7,$$

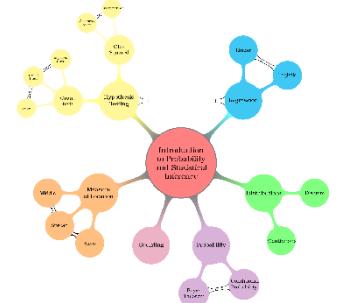
The variance of the distribution is:

$$Var[X] = \sum_i (x_i - \mu)^2 p(x_i) = \sum_i (x_i - 0.7)^2 p(x_i).$$

Week 3



Week 3



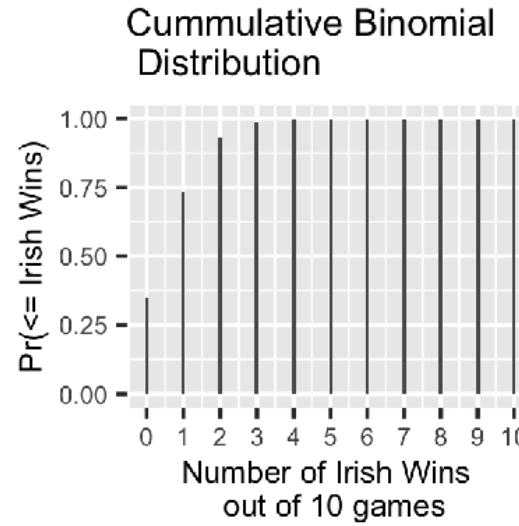
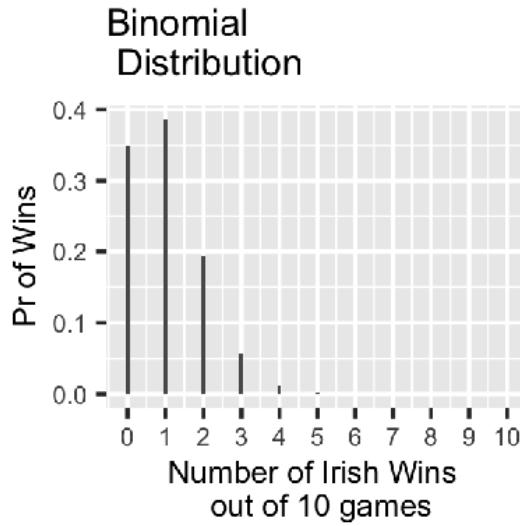
Binomial Distribution

The formula for the Binomial distribution is:

$$\Pr(k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, 2, \dots, n,$$

$$E[k] = np, \quad \text{Var}[k] = npq,$$

where n is the total of games, k is the number of "wins", p is the probability of a "win", $q = 1 - p$ probability of a "loss".



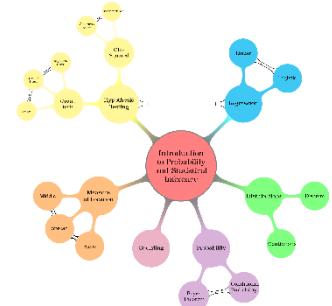
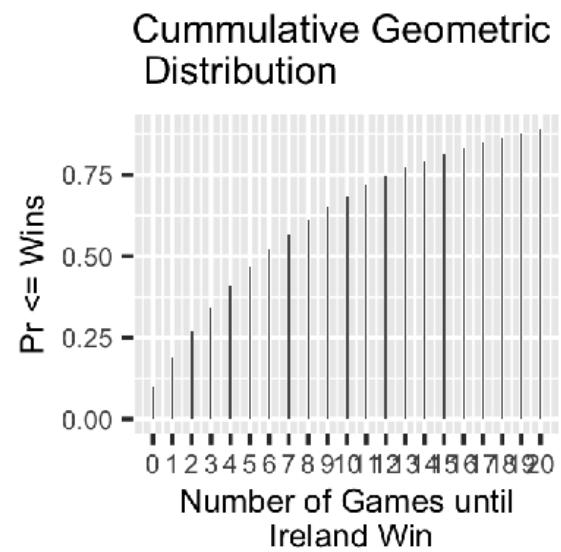
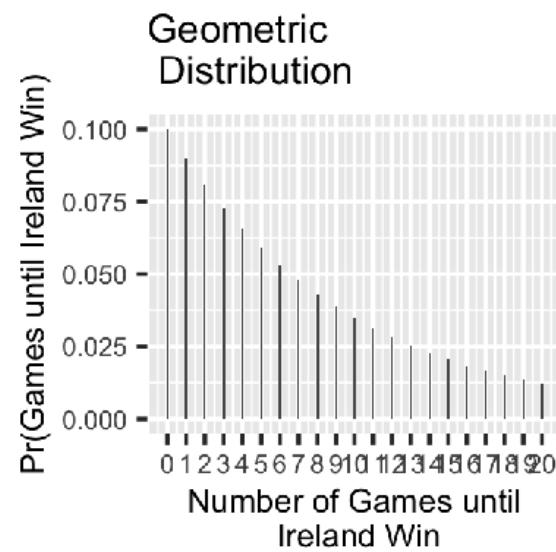
Geometric Distribution

The formula for the Geometric distribution is:

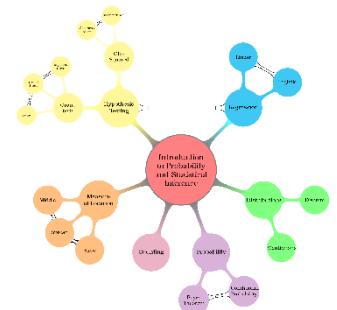
$$\Pr(k) = q^{(k-1)} p, \quad k = 1, 2, \dots$$

$$E[k] = \frac{1}{p}, \quad Var[k] = \frac{q}{p^2},$$

k is the number of events until one "win", p is the probability of a "win", $q = 1 - p$ probability of a "loss".



Week 3-4



Discrete Distribution

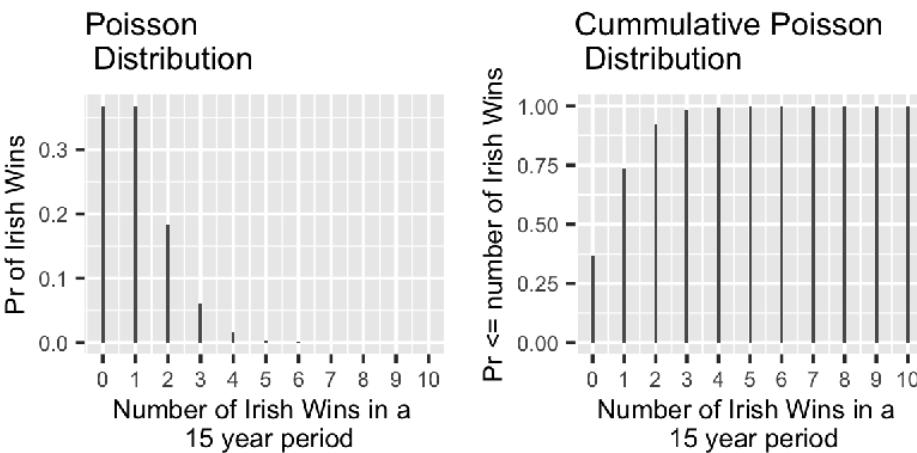
Poisson Distribution

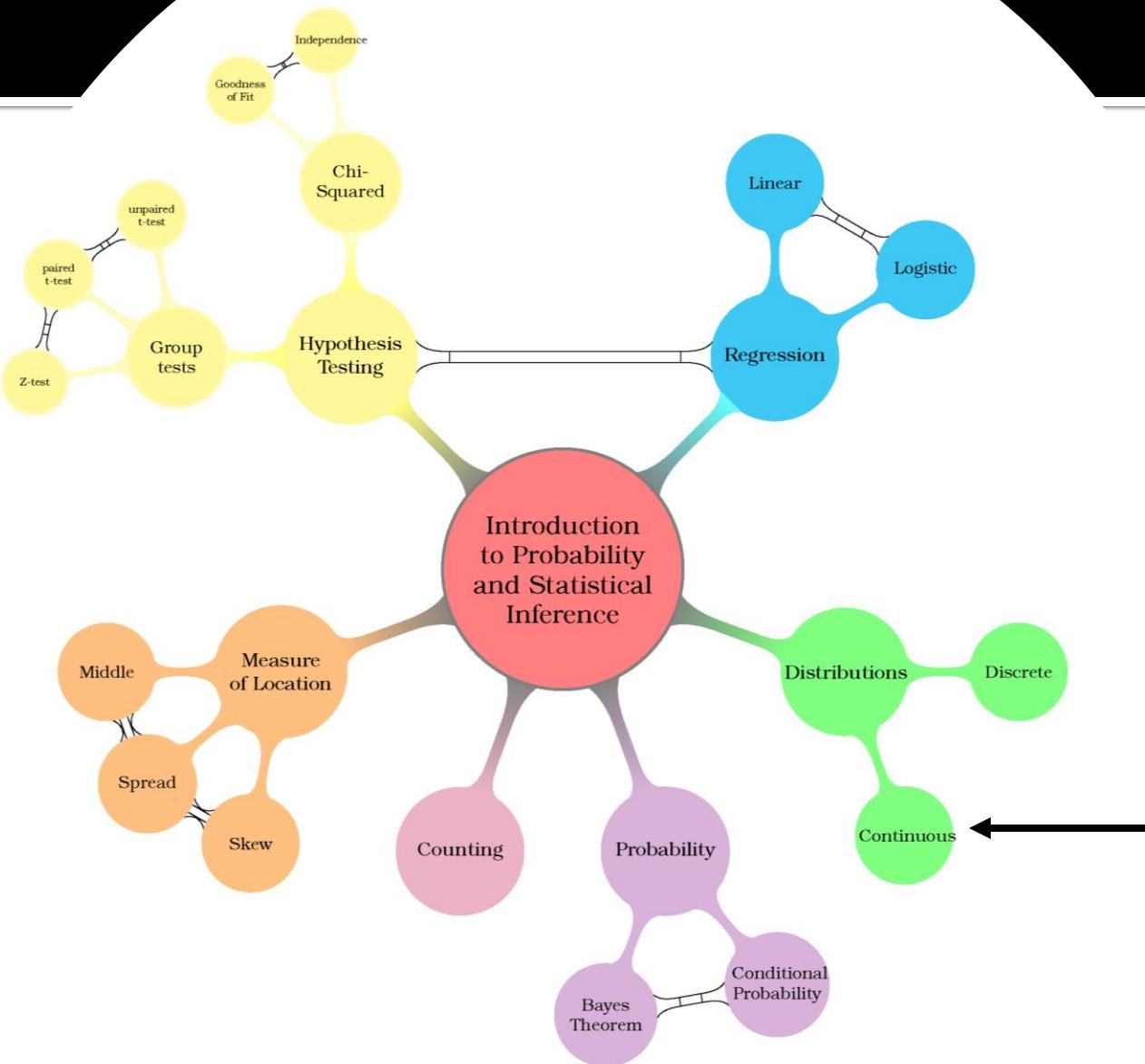
The formula for the Poisson distribution is:

$$\Pr(k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots$$

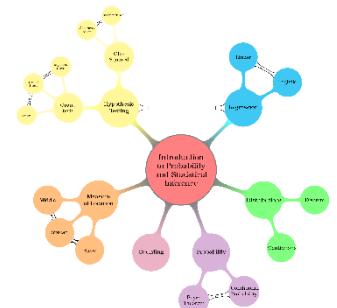
$$E[k] = \lambda, \quad \text{Var}[k] = \lambda.$$

where λ is the mean and standard deviation of the distribution and k is the number of "wins" in a specified time or space.





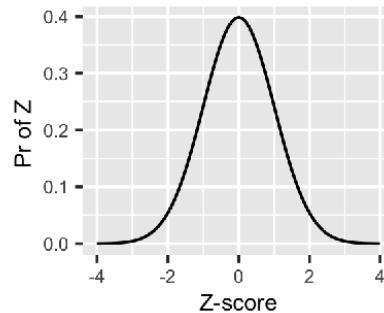
Week 4-5



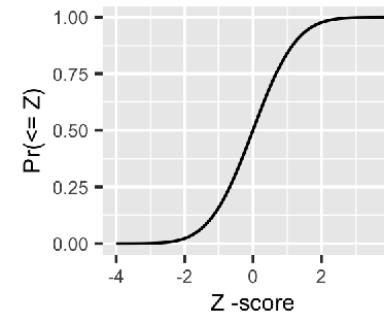
Continuous Distribution

Normal Distribution

Gaussian (Normal) Distribution

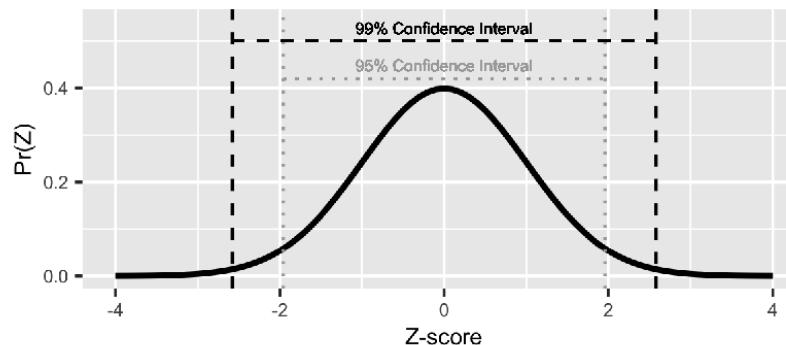


Cummulative Gaussian Distribution



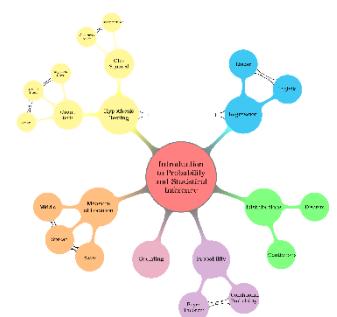
Confidence Intervals

Confidence Intervals





Week 6-8



Hypothesis Testing

Five steps for Hypothesis testings

1. State the Null Hypothesis H_0 ;
2. State an Alternative Hypothesis H_α ;
3. Calculate a Test Statistic (see below);
4. Calculate a p-value and/or set a rejection region;
5. State your conclusions.

z-test

Continuous Data

The test statistic is given by

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1),$$

where \bar{x} is the observed mean, μ is the historical mean, σ is the standard deviation and n is the number of observations. $\mathcal{N}(0, 1)$ is the normal distribution with a mean of 0 and a standard deviation of 1.

Do supplements make you faster?

The effect of a food supplement on the response time in rats is of interest to a biologist. They have established that the normal response time of rats is $\mu = 1.2$ seconds. The $n = 100$ rats were given a new food supplement. The following summary statistics were recorded from the data $\bar{x} = 1.05$ and $\sigma = 0.5$ seconds.

- The rats in the study are the same as normal rats, $H_0 : \mu = 1.2$.
 - The rats are different, $H_\alpha : \mu \neq 1.2$.
 - Calculate a Test Statistic $Z = \frac{1.05 - 1.2}{\frac{0.5}{\sqrt{100}}} = -3$
 - Reject the Null hypothesis H_0 if $Z < -1.96$ and $Z > 1.96$
 - The data suggests that rats are faster with the new food.

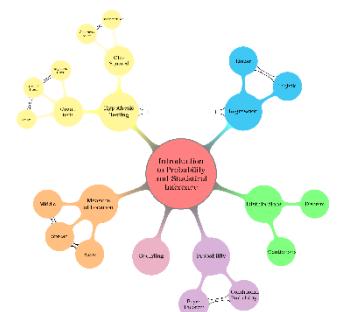
Proportional Data

The test statistic is given by

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \sim \mathcal{N}(0, 1).$$

where \hat{p} is the observed proportion, p_0 is the historical proportion, q_0 is the complement $q_0 = 1 - p_0$, and n is the number of observations.

Week 6-7



Discrete Distribution

Probability Mass Functions

Event Number i	0	1	2	3	4
Event Value x_i	-1	0	1	2	3
Probability of Event $\Pr(x_i)$	0.3	0.1	0.3	0.1	0.2

The expected value of the distribution is:

$$\mu = E[X] = \sum_i x_i \Pr(x_i),$$

$$\sum_i x_i p(x_i) = -1 \times 0.4 + 0 \times 0.1 + 1 \times 0.3 + 0.1 \times 2 + 0.2 \times 3 = 0.7,$$

The variance of the distribution is:

$$Var[X] = \sum_i (x_i - \mu)^2 p(x_i) = \sum_i (x_i - 0.7)^2 p(x_i).$$

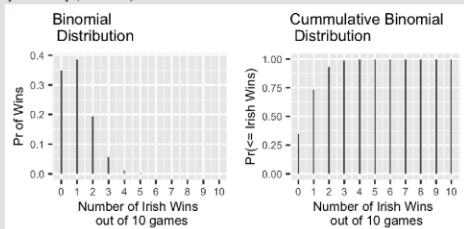
Binomial Distribution

The formula for the Binomial distribution is:

$$\Pr(k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, 2, \dots, n,$$

$$E[k] = np, \quad Var[k] = npq,$$

where n is the total of games, k is the number of 'wins', p is the probability of a 'win', $q = 1 - p$ probability of a 'loss':



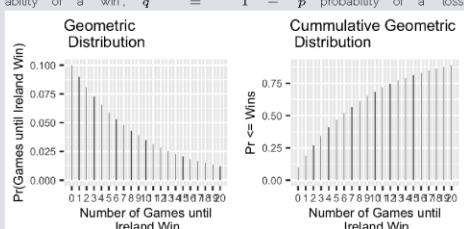
Geometric Distribution

The formula for the Geometric distribution is:

$$\Pr(k) = q^{(k-1)} p, \quad k = 1, 2, \dots$$

$$E[k] = \frac{1}{p}, \quad Var[k] = \frac{q}{p^2},$$

k is the number of events until one 'win', p is the probability of a 'win', $q = 1 - p$ probability of a 'loss'.



Discrete Distribution

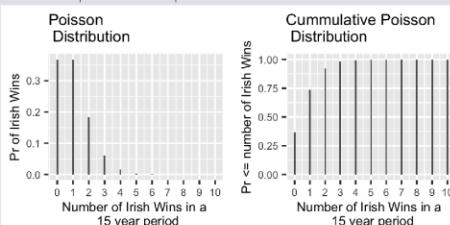
Poisson Distribution

The formula for the Poisson distribution is:

$$\Pr(k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots$$

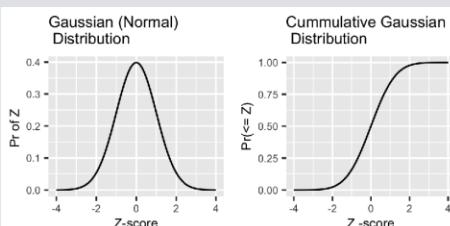
$$E[k] = \lambda, \quad Var[k] = \lambda.$$

where λ is the mean and standard deviation of the distribution and k is the number of 'wins' in a specified time or space.

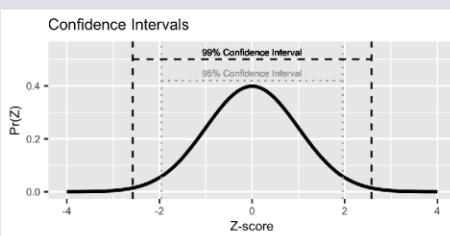


Continuous Distribution

Normal Distribution



Confidence Intervals



Hypothesis Testing

Five steps for Hypothesis testing

- State the Null Hypothesis H_0 :
- State an Alternative Hypothesis H_α :
- Calculate a Test Statistic (see below);
- Calculate a p-value and/or set a rejection region;
- State your conclusions.

z-test

Continuous Data

The test statistic is given by

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1),$$

where \bar{x} is the observed mean, μ is the historical mean, σ is the standard deviation and n is the number of observations. $\mathcal{N}(0, 1)$ is the normal distribution with a mean of 0 and a standard deviation of 1.

Do supplements make you faster?

The effect of a food supplement on the response time in rats is of interest to a biologist. They have established that the normal response time of rats is $\mu = 1.2$ seconds. The $n = 100$ rats were given a new food supplement. The following summary statistics were recorded from the data: $\bar{x} = 1.05$ and $\sigma = 0.5$ seconds

- The rats in the study are the same as normal rats, $H_0 : \mu = 1.2$.
- The rats are different, $H_\alpha : \mu \neq 1.2$.
- Calculate a Test Statistic $Z = \frac{1.05 - 1.2}{\frac{0.5}{\sqrt{100}}} = -3$
- Reject the Null hypothesis H_0 if $Z < -1.96$ and $Z > 1.96$
- The data suggests that rats are faster with the new food.

Proportional Data

The test statistic is given by

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \sim \mathcal{N}(0, 1).$$

where \hat{p} is the observed proportion, p_0 is the historical proportion, q_0 is the complement $q_0 = 1 - p_0$, and n is the number of observations.

t-test

paired t-test

The test statistic is given by

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{\alpha, df}$$

where \bar{x} is the observed mean, μ_0 is the null mean, s is the standard deviation and n is the number of observations. α is the alpha level and df is the degrees of freedom.

unpaired t-test

The test statistic is given by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{\alpha, df}$$

where $s_p = \sqrt{\frac{s_{x_1}^2 + s_{x_2}^2}{2}}$ is the pooled sample standard deviation, \bar{x}_1 and \bar{x}_2 are the sample means, n_1 and n_2 are the sample sizes.

χ^2 Independence test

The test statistic to test if data are independent of group is given by:

$$\chi^2_{Ind} = \sum \frac{(O - E)^2}{E} \sim \chi^2_{(r-1)(c-1)}$$

where O is the observed data, E is the expected data if independent, r is the number of rows and c is the number of columns.

Does ice-cream flavour matter?

An ice-cream company had 500 people sample one of three different ice-cream flavours and asked them to say whether they liked or disliked the ice-cream.

	Vanilla	Chocolate	Strawberry
Liked	130	170	100
Disliked	20	30	50

The χ^2_{Ind} independence test could be used to determine if the enjoyment of the ice-cream depends on the flavour.

χ^2 Goodness of Fit

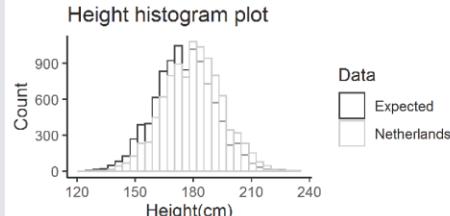
The test statistic to test if data come from a specific distribution is given by:

$$\chi^2_{GoF} = \sum \frac{(O - E)^2}{E} \sim \chi^2_{k-1},$$

where O is the observed data, E is the expected data from a chosen distribution and k is the number of observation bins.

Does it fit?

The χ^2_{GoF} can test if the observed distribution of the height of Dutch people (grey) fits the expected distribution of heights (dark grey).



Linear Regression

A linear regression is used to model a linear relationship of the dependent variable y and the regressors x_1, x_2, \dots

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots,$$

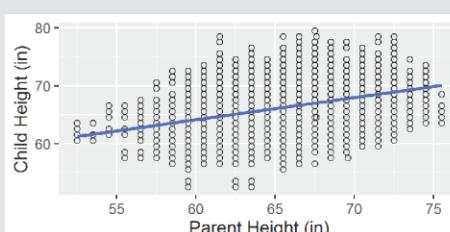
where β_0, β_1 are the slopes of the regressors.

Height Prediction

A simple linear regression (correlation) is used to predict the height of 744 children y using the height of their parent x :

$$y = \beta_0 + \beta_1 x.$$

The plot below shows the fit of the model:



The parents' height x explained 12.7% of the childrens' height y .

Logistic Regression

A logistic regression (or logit model) is used to model the probability of a binary events such as win/lose. The general formula for the Logistic regression is

$$p_i = \frac{e^\eta}{1 + e^\eta},$$

where

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

and β is the slope corresponding to the predictor variable x .

Sexton Conversion Rate

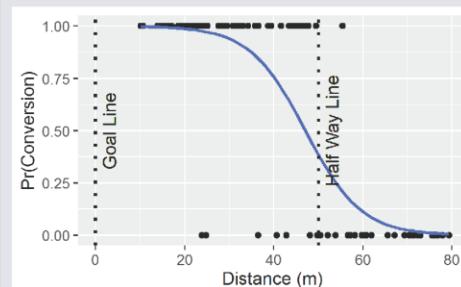
Data from 1000 conversions kicks by Johnny Sexton was acquired; the distance (m) from the goal-line and if the kick was a miss 0 or a conversion 1. The data was fit to a logistic regression. The model was

$$p = \frac{e^\eta}{1 + e^\eta},$$

where

$$\eta = \beta_0 + \beta_1 \text{Distance}$$

and p is the probability of a conversion. The plot below shows the fit of the model:



The model predicts that at the half-way line (50m) Sexton has a 0.375 probability of conversion.

Bibliography

1. Alexander, R. - Telling Stories with Data 2022 [website](#)
2. Devore & Peck - Statistics: The exploration and analysis of data (2011)
3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: Springer [book website](#).
4. Poldrack R. Statistical Thinking in the 21st Century 2020 [website](#).

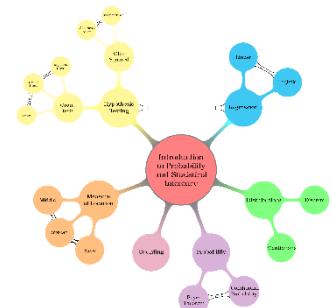
Popular Press

Fry, H. - Hello World: How to be Human in the Age of the Machine. Doubleday, 2018

Resources

Butler, J. S., [Module GitHub Repository](#)

Week 6-7



t-test

paired t-test

The test statistic is given by

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{\alpha, df}$$

where \bar{x} is the observed mean, μ_0 is the null mean, s is the standard deviation and n is the number of observations. α is the alpha level and df is the degrees of freedom.

unpaired t-test

The test statistic is given by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{\alpha, df}$$

where $s_p = \sqrt{\frac{s_{x_1}^2 + s_{x_2}^2}{2}}$ is the pooled sample standard deviation, \bar{x}_1 and \bar{x}_2 are the sample means, n_1 and n_2 are the sample sizes.



χ^2 Independence test

The test statistic to test if data are independent of group is given by:

$$\chi^2_{Ind} = \sum \frac{(O - E)^2}{E} \sim \chi^2_{(r-1)(c-1)}.$$

where O is the observed data, E is the expected data if independent, r is the number of rows and c is the number of columns.

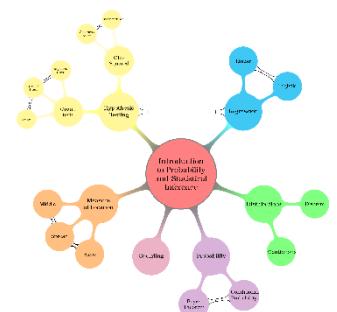
Does ice-cream flavour matter?

An ice-cream company had 500 people sample one of three different ice-cream flavours and asked them to say whether they liked or disliked the ice-cream.

	Vanilla	Chocolate	Strawberry
Liked	130	170	100
Disliked	20	30	50

The χ^2_{Ind} independence test could be used to determine if the enjoyment of the ice-cream depends on the flavour.

Week 7-8



χ^2 Goodness of Fit

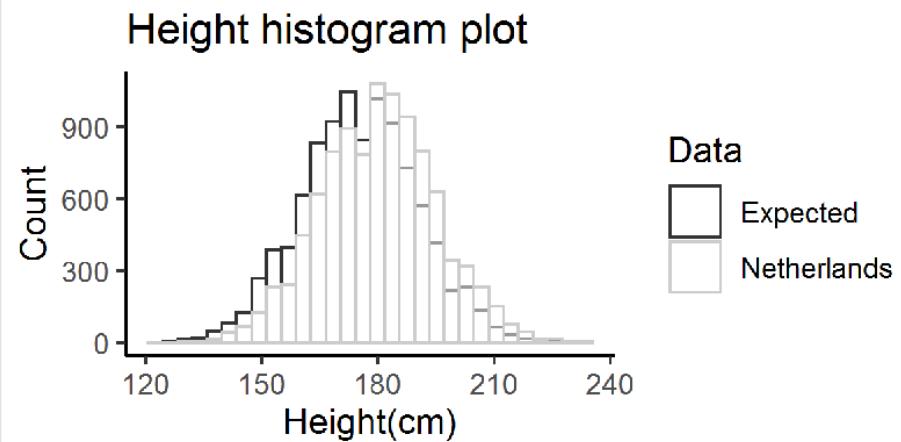
The test statistic to test if data come from a specific distribution is given by:

$$\chi^2_{GoF} = \sum \frac{(O - E)^2}{E} \sim \chi^2_{k-1},$$

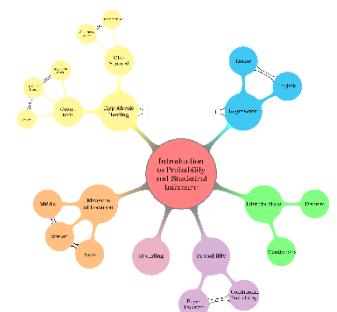
where O is the observed data, E is the expected data from a chosen distribution and k is the number of observation bins.

Does it fit?

The χ^2_{GoF} can test if the observed distribution of the height of Dutch people (grey) fits the expected distribution of heights (dark grey).

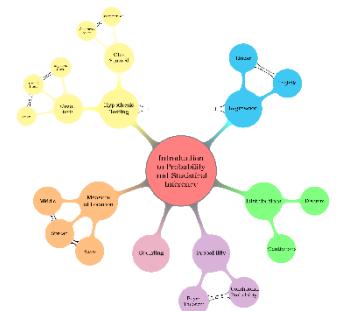


Week 7-8





Week 9-10



Linear Regression

A linear regression is used to model a linear relationship of the dependent variable y and the regressors x_1, x_2, \dots

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots,$$

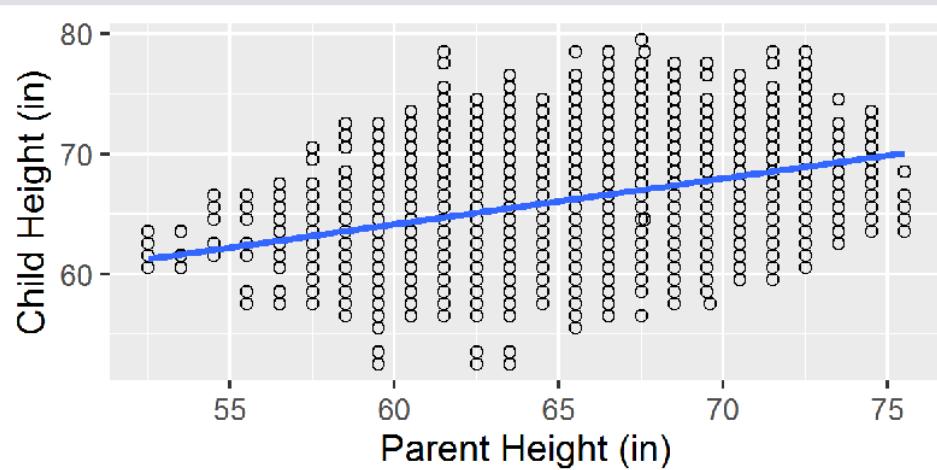
where β_0, β_1 are the slopes of the regressors.

Height Prediction

A simple linear regression (correlation) is used to predict the height of 744 children y using the height of their parent x .

$$y = \beta_0 + \beta_1 x.$$

The plot below shows the fit of the model:



The parents' height x explained 12.7% of the children's height y .

Logistic Regression

A logistic regression (or logit model) is used to model the probability of a binary events such as win/lose. The general formula for the Logistic regression is

$$p_i = \frac{e^\eta}{1 + e^\eta},$$

where

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

and β is the slope corresponding to the predictor variable x .

Sexton Conversion Rate

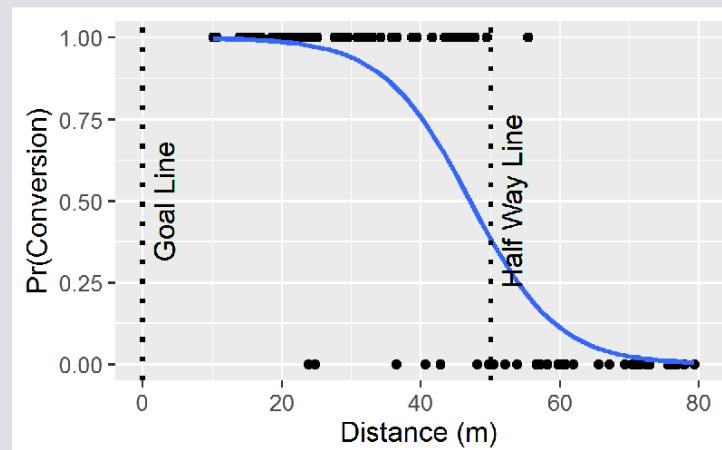
Data from 1000 conversions kicks by Johnny Sexton was acquired; the distance (m) from the goal-line and if the kick was a miss 0 or a conversion 1. The data was fit to a logistic regression. The model was

$$p = \frac{e^\eta}{1 + e^\eta},$$

where

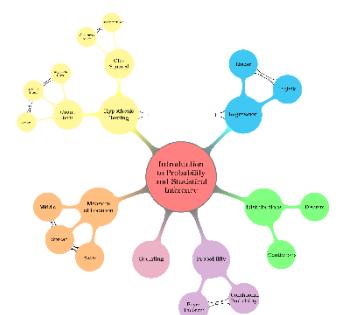
$$\eta = \beta_0 + \beta_1 \text{Distance}$$

and p is the probability of a conversion. The plot below shows the fit of the model



The model predicts that at the half-way line (50m) Sexton has a 0.375 probability of conversion.

Week 10-11



t-test

paired t-test

The test statistic is given by

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{\alpha, df}$$

where \bar{x} is the observed mean, μ_0 is the null mean, s is the standard deviation and n is the number of observations. α is the alpha level and df is the degrees of freedom.

unpaired t-test

The test statistic is given by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{\alpha, df}$$

where $s_p = \sqrt{\frac{s_{x_1}^2 + s_{x_2}^2}{2}}$ is the pooled sample standard deviation, \bar{x}_1 and \bar{x}_2 are the sample means, n_1 and n_2 are the sample sizes.

χ^2 Independence test

The test statistic to test if data are independent of group is given by:

$$\chi^2_{Ind} = \sum \frac{(O - E)^2}{E} \sim \chi^2_{(r-1)(c-1)}$$

where O is the observed data, E is the expected data if independent, r is the number of rows and c is the number of columns.

Does ice-cream flavour matter?

An ice-cream company had 500 people sample one of three different ice-cream flavours and asked them to say whether they liked or disliked the ice-cream.

	Vanilla	Chocolate	Strawberry
Liked	130	170	100
Disliked	20	30	50

The χ^2_{Ind} independence test could be used to determine if the enjoyment of the ice-cream depends on the flavour.

χ^2 Goodness of Fit

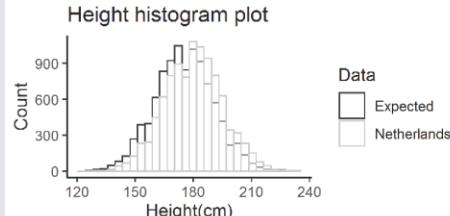
The test statistic to test if data come from a specific distribution is given by:

$$\chi^2_{GoF} = \sum \frac{(O - E)^2}{E} \sim \chi^2_{k-1},$$

where O is the observed data, E is the expected data from a chosen distribution and k is the number of observation bins.

Does it fit?

The χ^2_{GoF} can test if the observed distribution of the height of Dutch people (grey) fits the expected distribution of heights (dark grey).



Linear Regression

A linear regression is used to model a linear relationship of the dependent variable y and the regressors x_1, x_2, \dots

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots,$$

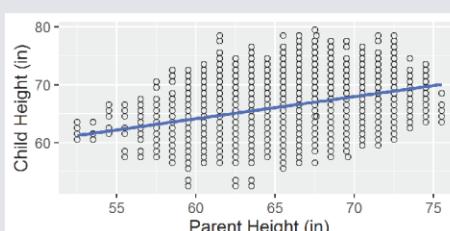
where β_0, β_1 are the slopes of the regressors.

Height Prediction

A simple linear regression (correlation) is used to predict the height of 744 children y using the height of their parent x :

$$y = \beta_0 + \beta_1 x.$$

The plot below shows the fit of the model:



The parents' height x explained 12.7% of the children's height y .

Logistic Regression

A logistic regression (or logit model) is used to model the probability of a binary events such as win/lose. The general formula for the Logistic regression is

$$p_i = \frac{e^\eta}{1 + e^\eta},$$

where

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

and β is the slope corresponding to the predictor variable x .

Sexton Conversion Rate

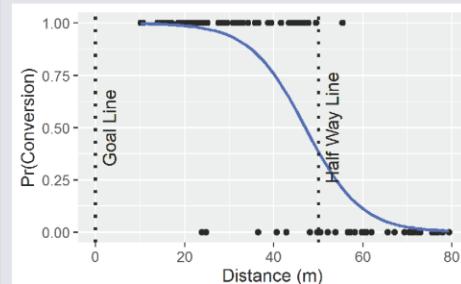
Data from 1000 conversions kicks by Johnny Sexton was acquired; the distance (m) from the goal-line and if the kick was a miss 0 or a conversion 1. The data was fit to a logistic regression. The model was

$$p = \frac{e^\eta}{1 + e^\eta},$$

where

$$\eta = \beta_0 + \beta_1 \text{Distance}$$

and p is the probability of a conversion. The plot below shows the fit of the model:



The model predicts that at the half-way line (50m) Sexton has a 0.375 probability of conversion.

Bibliography

1. Alexander, R. - Telling Stories with Data 2022 [website](#)
2. Devore & Peck - Statistics: The exploration and analysis of data (2011)
3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: Springer [book website](#).
4. Poldrack R. Statistical Thinking in the 21st Century 2020 [website](#).

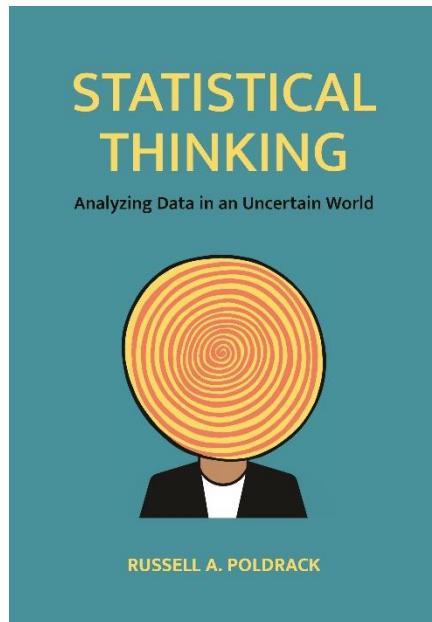
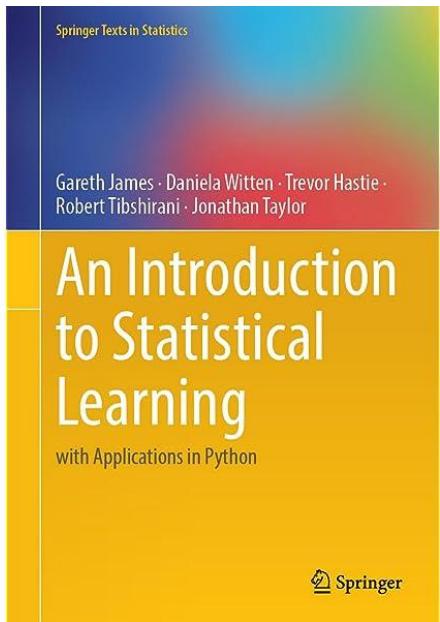
Popular Press

Fry, H. - Hello World: How to be Human in the Age of the Machine. Doubleday, 2018

Resources

Butler, J. S., [Module GitHub Repository](#)

Bibliography



Bibliography

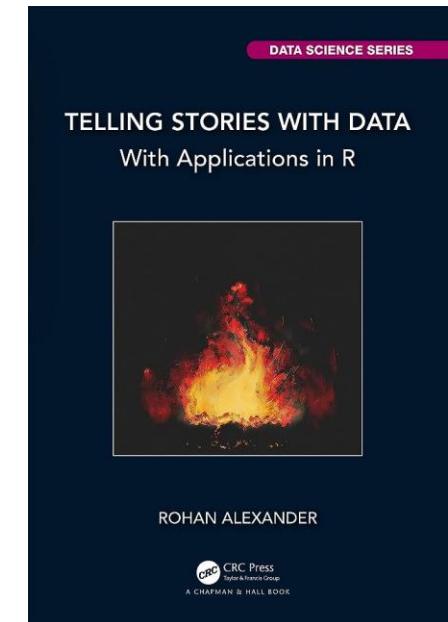
1. Alexander, R. - Telling Stories with Data 2022 website
2. Devore & Peck - Statistics: The exploration and analysis of data (2011)
3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: Springer book website.
4. Poldrack R. Statistical Thinking in the 21st Century 2020 website.

Popular Press

Fry, H. - Hello World: How to be Human in the Age of the Machine. Doubleday, 2018

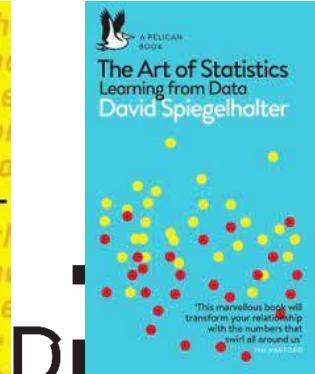
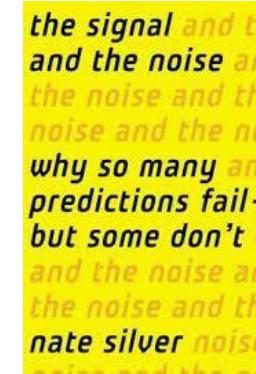
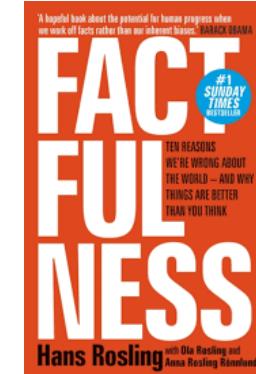
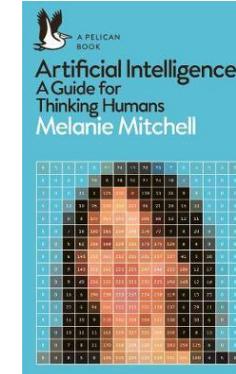
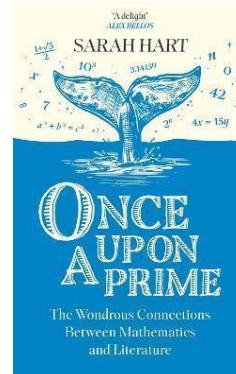
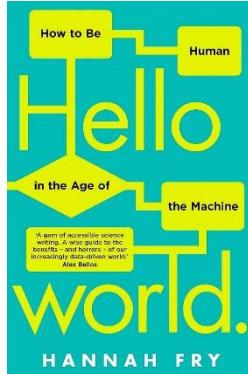
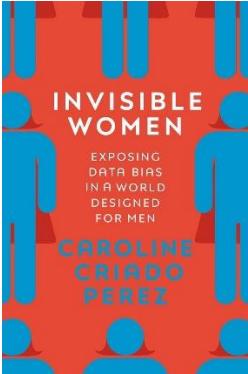
Resources

Butler, J. S. [Module GitHub Repository](#)



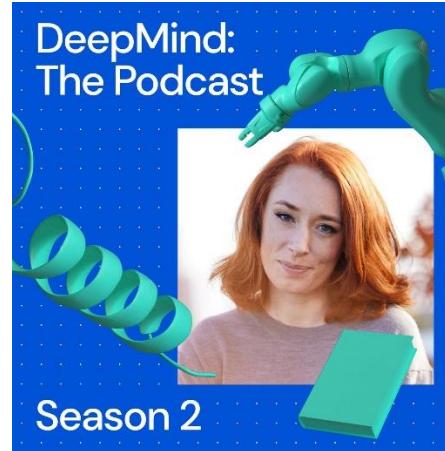
Popular Books

- Criado Perez, C., (2021), Invisible Women: Data Bias in a World Designed for Me
 - Fry, H. (2018), Hello World: How to be Human in the Age of the Machine
 - Hart, S., (2023), Once Upon a Prime
 - Mitchell, M. (2019), Artificial Intelligence: A Guide for Thinking Humans
 - Rosling, H, Rosling, O, Rosling Ronnlund, (2019), Factfulness: Ten Reasons We're Wrong About The World - And Why Things Are Better Than You Think
 - Silver, N, (2013) The Signal and the Noise: The Art and Science of Prediction
 - Spiegelhalter, D. (2019), The art of statistics: Learning from data



Podcasts

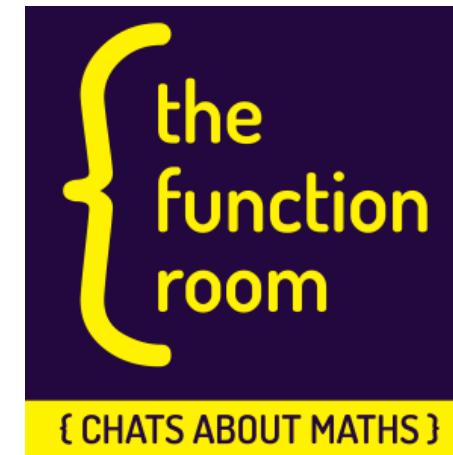
- Deep Mind: Podcast
- Freakonomics Podcast
- More or Less: Behind the Stats
- Visible Women
- Colm O'Regan: The Function Room



FREAKONOMICS RADIO NETWORK



STEPHEN J. DUBNER



DUBLIN
TECHNICAL
UNIVERSITY DUBLIN

Notation

- \bar{x} - mean of a list of numbers x_i
- σ - standard deviation of a list of numbers x_i
- σ^2 - variance of a list of numbers
- $\Pr(A)$ - probability of event A
- $\Pr(\bar{A})$ - probability of not event A
- $\Pr(A|B)$ - probability of event A given event B is known
- $\sum_i^n x_i$ - the sum of a list of number x_i
- $n!$ - n factorial is $n \times (n - 1) \times \dots \times 1$
- $5! = 5 \times (5 - 1) \times (5 - 2) \times (5 - 3) \times (5 - 4) = 5 \times 4 \times 3 \times 2 \times 1 = 120$
- $\binom{n}{k} = {}^n C_k$ - n choose k equals to $\frac{n!}{k!(n-k)!}$
- $\binom{5}{3} = {}^5 C_3$ - 5 choose 3 equals to $\frac{5!}{3!(5-3)!} = \frac{5!}{3!2!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1 \times 2 \times 1} = 10$
- ${}^n P_k$ - n pick k equals to $\frac{n!}{(n-k)!}$
- ${}^5 P_3$ - 5 pick 3 equals to $\frac{5!}{(5-3)!} = \frac{5!}{2!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1} = 60$
- p - p probability of a "win"
- q - q probability of a "loss" $1 - p$
- p^n - p to the power of n is $p \times p \times \dots \times p$
- 0.1^4 - 0.1 to the power of 4 is $0.1 \times 0.1 \times 0.1 \times 0.1 \times 0.1$
- $E[X]$ - the expected value of a probability distribution
- $Var[X]$ - the variance of a probability distribution
- e - is the exponential which is it equal to approximately 2.718 it comes up again and again in mathematics formulas
- H_0 - null hypothesis
- H_α - alternative hypothesis
- μ - real mean (generally never known)
- \bar{x} - observed mean given the data
- \hat{p} - is the observed sample proportion
- \hat{p} - is the observed sample proportion
- $\mathcal{N}(\mu, \sigma)$ - is the Gaussian distribution with mean μ and standard deviation σ
- $\mathcal{N}(0, 1)$ - is a special case of Gaussian distribution known as the Normal Distribution with mean 0 and standard deviation 1
- df-degrees of freedom
- χ^2_{df} - Chi (χ)-squared (χ^2) distribution with degrees of freedom df
- β the coefficient for a regression
- $\hat{\beta}$ the coefficient estimated for a regression from the observed

Preview



Measures of Locations

Measures of Locations

First moment

1. Mean (average)
2. Median
3. Mode

Second Moment (Spread)

1. Standard deviation
2. Variance

Third Moment

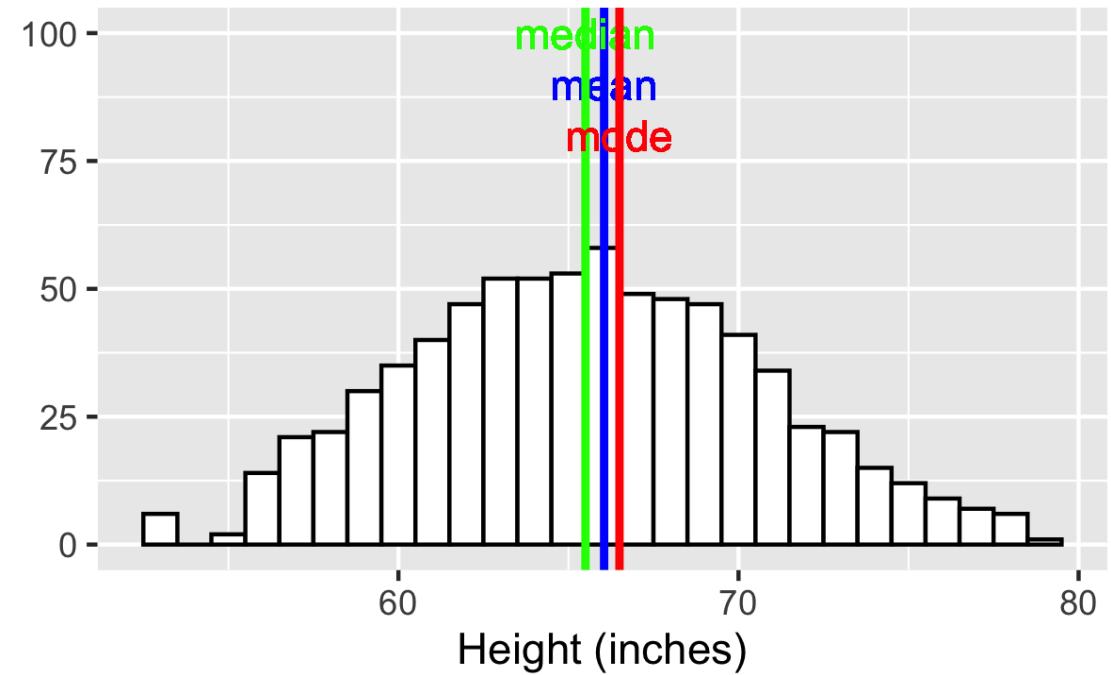
1. Skewness

First Moment of Data (Middle)

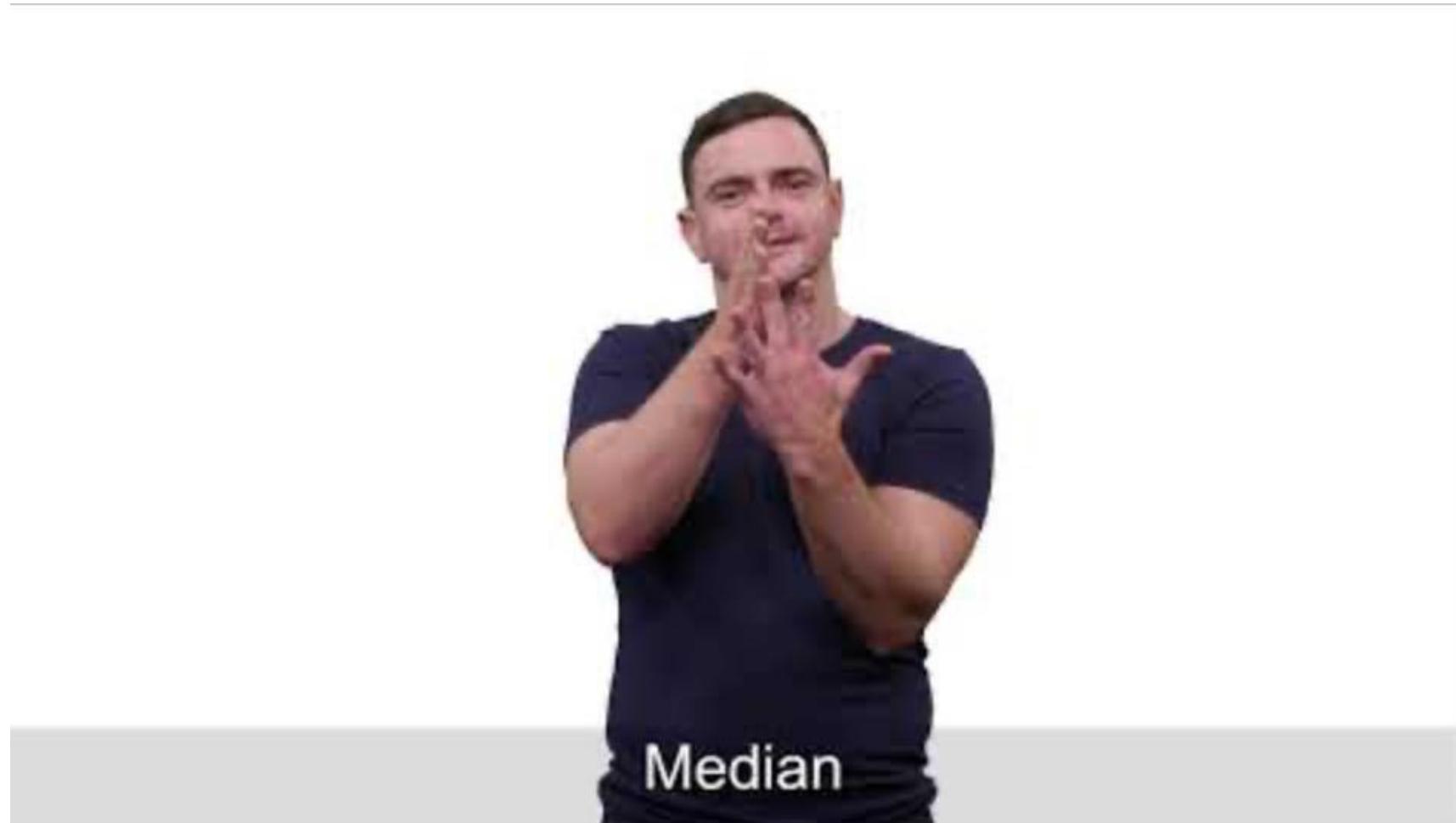
Measures of Locations

First moment

1. Mean (average)
2. Median
3. Mode



Median: Irish Sign Language



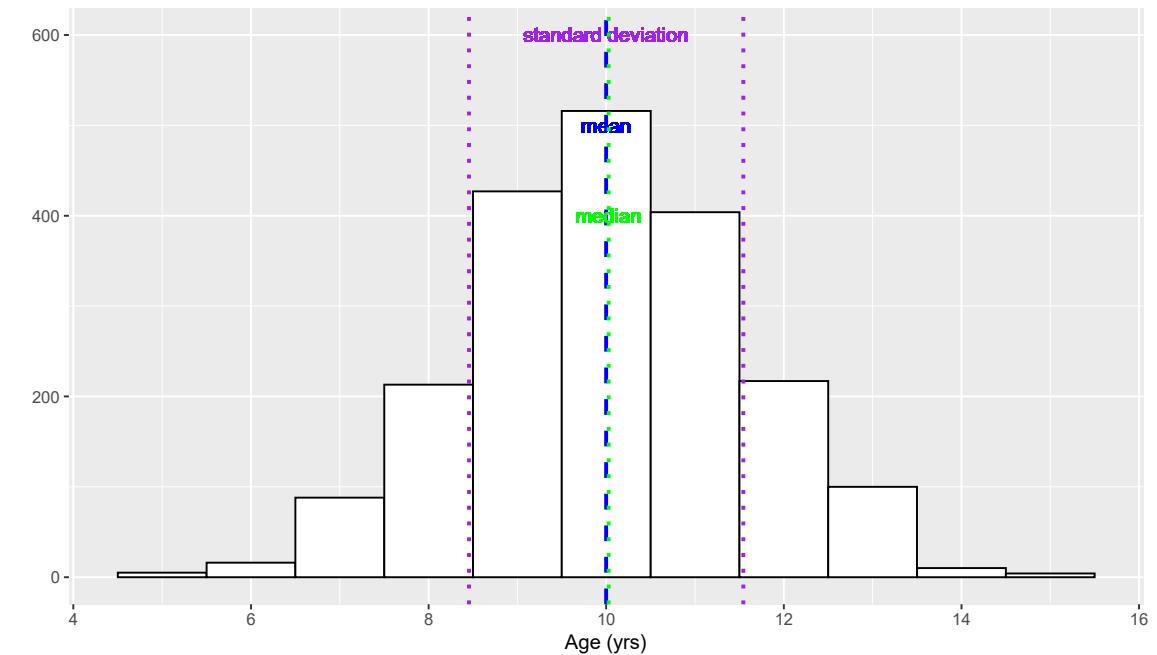
Second Moment

The spread of the data

Measures of Locations

Second Moment (Spread)

1. Standard deviation
2. Variance



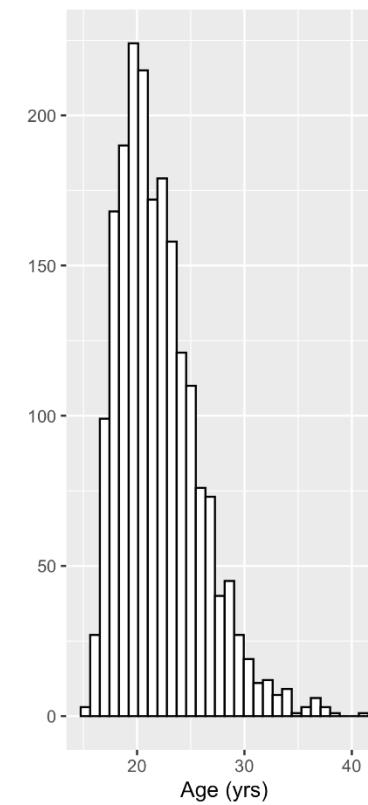
Third Moment of Data (Skewness)

Measures of Locations

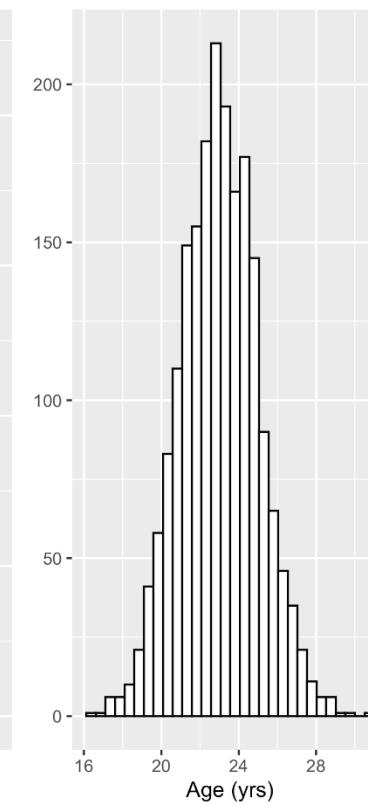
Third Moment

1. Skewness

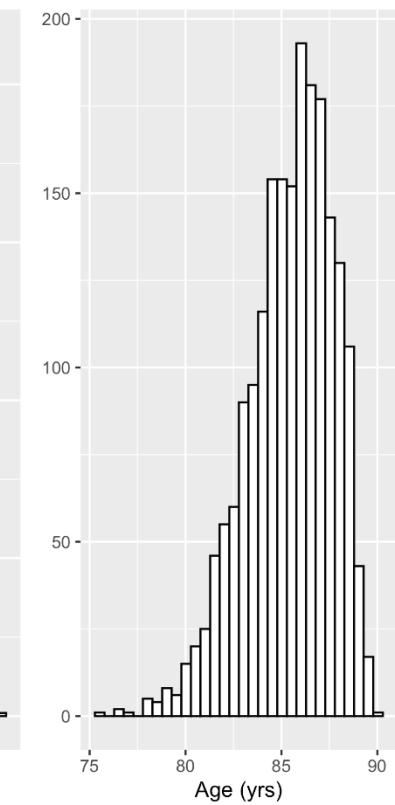
Positive



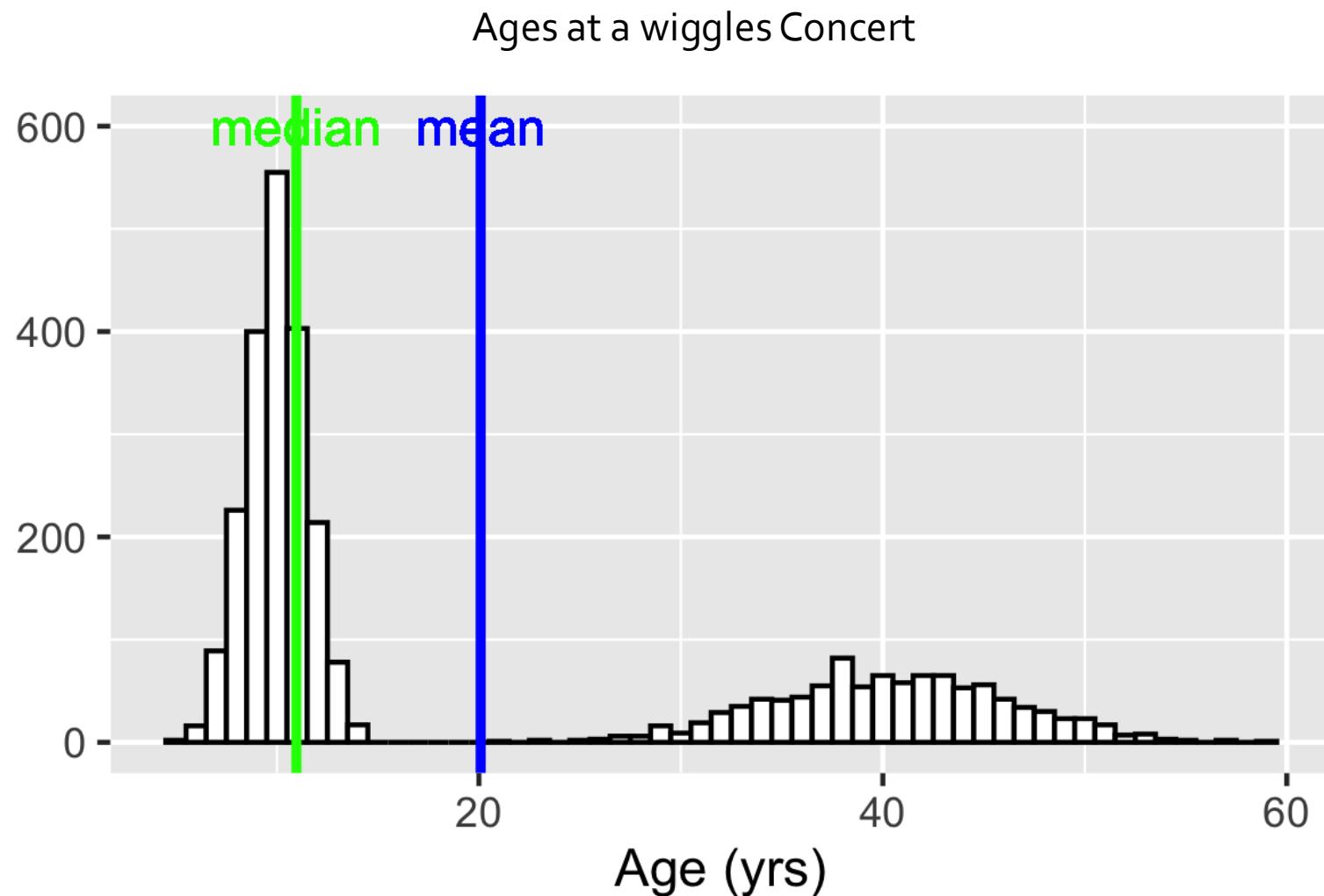
Symmetric



Negative



Problems with Moments of location



Counting Rules

Dr. John S. Butler

Useful counting Rules

- Consider selecting or sampling r objects from a group of n distinct objects, sampling with replacement.
- The total number of possible ordered samples
- $n \times n \times \cdots \times n = n^r$
- The probability is $\frac{1}{n^r}$

Useful counting Rules

- Consider selecting or sampling r objects from a group of n distinct objects, sample without replacement
- $p_r^n = \frac{n!}{(n-r)!}$
- This equals the number of arrangements (permutations) of any r objects selected from a group of n distinct objects

Useful counting Rules

- Consider selecting or sampling r objects from a group of n distinct objects, sampling without replacement. The total number of possible non-ordered samples

$$\frac{p_r^n}{r!} = \frac{n!}{((n - r)! r!)} = C_r^n = \binom{n}{r}$$

- (Binomial coefficient)

Useful counting Rules

- The number of distinct arrangements of n objects of which n_1 are of one kind n_2 of a second kind, ..., n_k are of k th kind is given by multinomial coefficient

$$\frac{n!}{(n_1!n_2!\dots n_k!)} \text{ where } \sum_{i=1}^k n_i = n$$

- Example: What is the probability that in 6 throws of a fair die all faces turn up?