Statistics 1

Summary Sheet

John S Butler (TU Dublin)

Course Twitter Account

Data Type

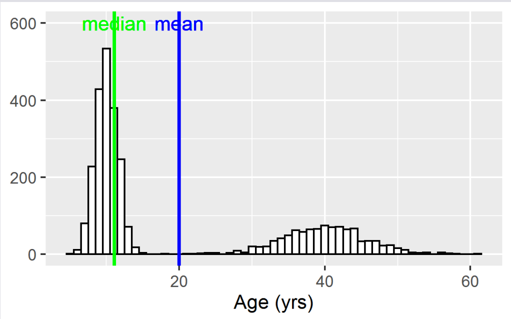
- Categorical
- Interval
- Ordinal
- Ratio

Measures of Location

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

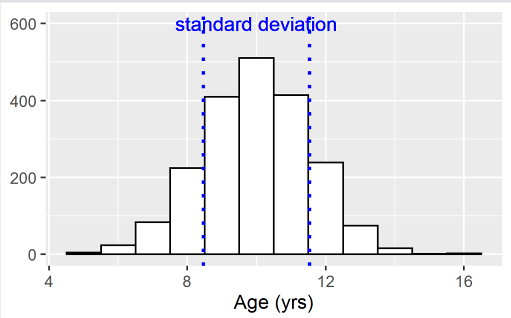
1. The First Moment: Mean, Mode or Median;
2. The Second Moment: Variance, Standard Deviation;
3. The Third Moment: Skewness.

First Moment: Middle



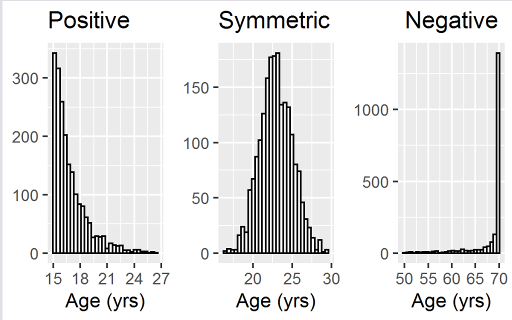
Wiggles Concert Attendance

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

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

Counting Rules

1. Consider selecting r objects from a group of n distinct objects, sampling **with replacement**

$$n \times n \times \cdots \times n = n^r$$

2. Consider selecting r objects from a group of n distinct objects, sampling **without replacement**. The total possible of **ordered** samples is

$$^nP_r = \frac{n!}{(n-r)!}$$

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

$$\binom{n}{r} = ^n C_r = \frac{n!}{(n-r)!r!} \text{Binomial Coefficient}$$

4. The number of distinct arrangement of n objects of which n_1 are of one kind, n_2 are of a second kind, ..., n_k are of a k^th kind is given by the **multinomial coefficient**

$$\frac{n!}{n_1!n_2!\cdots n_k!} \quad \text{where} \quad \Sigma_{i=1}^k n_i = n$$

Conditional Probability

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

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

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

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

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 $\Pr(\text{rain} \bigcap \text{Ireland}) = 0.2$,

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

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

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

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,

$$\Pr(\text{Disease}|+) = \frac{95}{1085} = 0.0875576.$$

Discrete Distribution

Probability Mass Functions

The table of the probability mass function is:

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

The expected value of the distribution is:

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

$$\Sigma_i x_i p(x_i) = -1 \times 0.4 + 0 \times 0.1 + 1 \times 0.3 + 3 \times 0.3 = 0.9,$$

The variance of the distribution is:

$$Var[X] = \Sigma_i (x_i - \mu)^2 p(x_i) = \Sigma_i (x_i - 0.9)^2 p(x_i) =$$

$$= (-1-0.9)^2 0.3 + (0-0.9)^2 0.1 + (1-0.9)^2 0.3 + (3-0.9)^2 0.3$$

$$= 2.49.$$

The table of the cumulative distribution function (cdf) is:

r	< -1	$-1 \le r < 0$	$0 \le r < 1$	$1 \le r < 3$	>=3
$F(r)$	0	0.3	0.4	0.7	1.0

Discrete Distribution(cont.)

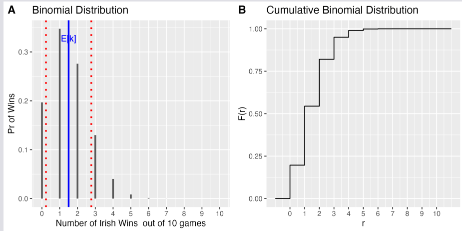
Binomial Distribution

The formula for the Binomial distribution is:

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

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

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".



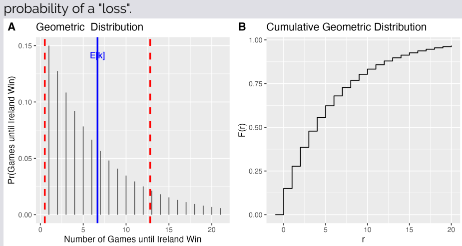
Geometric Distribution

The formula for the Geometric distribution is:

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

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

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



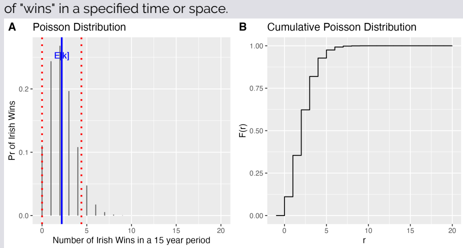
Poisson Distribution

The formula for the Poisson distribution is:

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

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

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



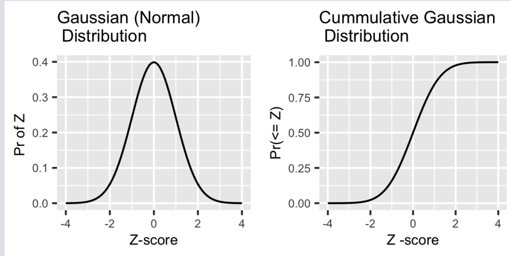
Continuous Distribution

Normal Distribution

The formula for the Normal distribution is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where μ is the mean and σ standard deviation of the distribution, which is denoted as $\mathcal{N}(\mu, \sigma)$ The standard normal distribution, also called the z-distribution, is a special normal distribution where the mean is 0 and the standard deviation is 1, $\mathcal{N}(0, 1)$, as shown below.

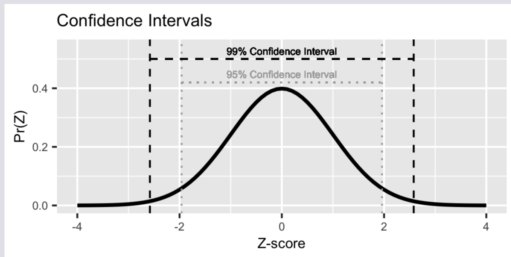


Confidence Intervals

The general formula for confidence intervals is:

$$\text{CI}_{(1-\alpha)\times 100\%} : \bar{x} \pm z_{1-\alpha/2} \times \frac{s}{\sqrt{n}}$$

where α is a value between 0 and 1, $(1-\alpha) \times 100\%$ is the confidence level, $z_{1-\alpha/2}$ is a value from the standard normal distribution, $\bar{x}$ is the observed sample mean and s is the observed sample standard deviation.



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.

The next step is interpretation and discussion of the result.

z-test

Continuous Data

The test statistic is given by

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1),$$

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 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 supplements. The following summary statistics were recorded from the data $\bar{x} = 1.05$ and $\sigma = 0.5$ seconds

1. The rats in the study are the same as normal rats, $H_0 : \mu = 1.2$.
2. The rats are different, $H_\alpha : \mu \neq 1.2$.
3. Calculate a Test Statistic $Z = \frac{1.05 - 1.2}{\frac{0.5}{\sqrt{100}}} = -3$
4. Reject the Null hypothesis H_0 if $Z < -1.96$ and $Z > 1.96$
5. 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

$$t = \frac{\bar{x} - \bar{\mu}_0}{\frac{s}{\sqrt{n}}} \sim t_{\alpha, df}$$

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

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

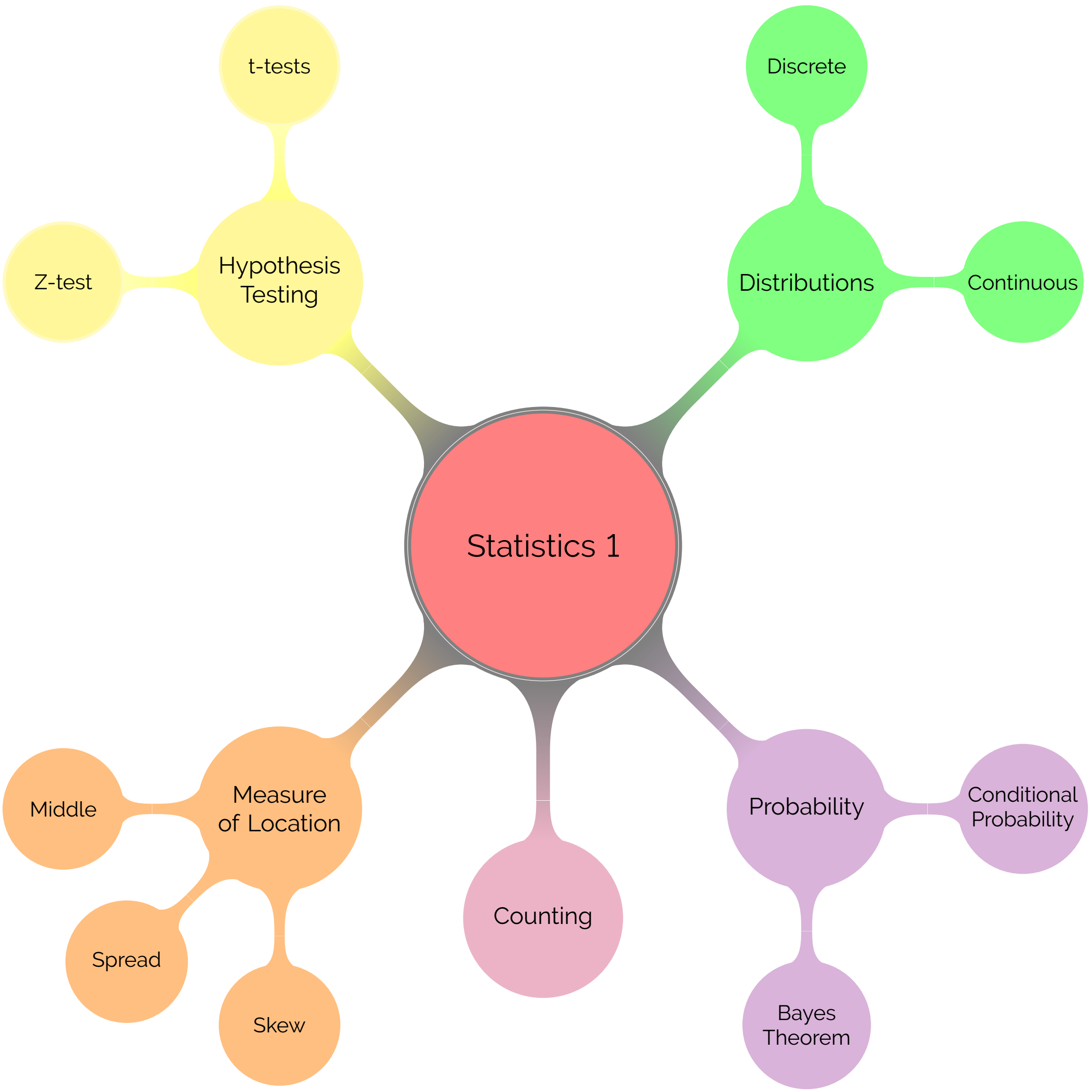
where $s_P = \sqrt{\frac{s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2}{2}}$ is the pooled sample standard deviation, $\bar{x}_1$ and $\bar{x}_2$ are the sample means, n_1 and n_2 are the sample sizes.

Selected Bibliography

1. Montgomery, D. C., & Runger, G. C. (2010). Applied statistics and probability for engineers. John Wiley & sons.
2. Peck, R., & Devore, J. L. (2011). Statistics: The exploration analysis of data. Cengage Learning.
3. Larson, H. J. (1982). Introduction to probability and statistical inference. JOHN WILEY & SONS, INC., 605 THIRD AVE., NEW YORK, NY 10158, 1982, 480.
4. Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (1993). Probability and statistics for engineers and scientists (Vol. 5). New York: Macmillan.
5. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: Springer book website.
6. Poldrack R. Statistical Thinking in the 21st Century 2020 website.
7. Gareth, J., et al. - An introduction to statistical learning. Vol. 112. New York: Springer, 2013.
8. Fry, H. - Hello World: How to be Human in the Age of the Machine, Doubleday, 2018
9. Alexander, R. - Telling Stories with Data 2022 website
10. Butler, J. S., Course GitHub Repository

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
- $\Sigma_i^n x_i$ - the sum of a list of number x_i
- $n!$ - n factorial is $n \times (n-1) \times \cdots \times 1$
- $5!$ - 5 factorial is $5 \times (5-1) \times (5-2) \times (5-3) \times (5-4) = 5 \times 4 \times 3 \times 2 \times 1 = 120$
- $\binom{n}{k} = {}^nC_k$ - n choose k equals to $\frac{n!}{k!(n-k)!}$
- $\binom{5}{3} = {}^5C_3$ - 5 choose 3 equals to $\frac{5!}{3!(5-3)!} = \frac{5!}{3!2!} = \frac{5\times4\times3\times2\times1}{3\times2\times1\times2\times1} = 10$
- ${}^nP_k$ - n pick k equals to $\frac{n!}{(n-k)!}$
- ${}^5P_3$ - 5 pick 3 equals to $\frac{5!}{(5-3)!} = \frac{5!}{2!} = \frac{5\times4\times3\times2\times1}{2\times1} = 60$
- p - p probability of a "win"
- q - q probability of a "loss" $1-p$
- p^n - p to the power of n is $p \times p \times \cdots \times p$
- 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 is comes up again and again in mathematics formulas
- H_0 - null hypothesis
- H_α - alternative hypothesis
- μ - real mean (generally never known)
- μ_0 - historical mean
- p_0 - is the historical proportion
- $\bar{x}$ - observed mean given the data
- $\hat{p}$ - 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 (^{2}) distribution with degrees of freedom df


Statistics 1
Hypothesis Testing
t-tests
Z-test
Distributions
Discrete
Continuous
Measure of Location
Middle
Spread
Skew
Counting
Probability
Conditional Probability
Bayes Theorem