Introduction to Data Collection and Sampling Approaches

Edge Hill University

Edge Hill University

Who Am I?

- Dr Marcello Trovati
- Room THG10
- Email: trovatim@edgehill.ac.uk

- Surgery Hours:
  - Tuesday 1pm–2pm
  - Thursday 4pm–5pm

Outline

Today's lecture will focus on

- Probability
- Descriptive and inferential statistics

Statistics and Probability

There are two main areas in statistics, *descriptive* and *inferential* statistics

- Descriptive statistics refers to the set of techniques and methods to organise, summarise and visualise information
- Inferential statistics aims to reach measurable and testable conclusions regarding a population defined by some hypotheses

Probability

Consider a discrete sample space $X = \{x_1, x_2, \ldots, x_n\}$, consisting of n events. The probability $P(x_i)$ of an event $x_i \in X$ must satisfy the following properties

1. $0 \leq P(x_i) \leq 1$
2. $P(X) = P\left(\bigcup_{j=1}^{n} x_j\right) = 1$
3. If two events are *independent*, that is $x_i \cap x_j = \emptyset$, then $P(x_i \cup x_j) = P(x_i) + P(x_j)$.
4. In general, if $x_i \cap x_j \neq \emptyset$, then $P(x_i \cup x_j) = P(x_i) + P(x_j) - P(x_i \cap x_j)$
5. $P(\emptyset) = 0$
6. $P(\bar{x}_i) = 1 - P(x_i)$, where $P(\bar{x}_i)$ is the *complement* of x_i.

Edge Hill University

Mean

The simplest yet one of the most useful concepts in statistics is the *mean*, or average.

For $X = \{x_1, \ldots, x_n\}$,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_n,$$

Example

Consider the set of observations $X = \{0.5, 0.3, 0.1, 0.1, 0.6, 0.9, 1.3, 1.1, 0.1\}$. Therefore the mean is

$$\bar{X} = \frac{0.5 + 0.3 + 0.1 + 0.1 + 0.6 + 0.9 + 1.3 + 1.1 + 0.1}{9} = 0.55.$$

Mean: Remarks

- The average is affected by
  - Variable with high frequency
  - Those values that lie far from the majority of those of the other variables
- The average does not provide any information on how sparse, or frequent the sample elements are

Mode and Median

- The *mode* refers to the value of a variable which occurs with the greatest frequency within the sample.
  - If each variable occurs only once, that is the frequency is 1 for each of them, then the corresponding sample is said to have no mode.
- The *median* corresponds to the value of the variable that splits the set of observed values into half.
  - If the number of observation is an odd number, the median is the value of the variable which lies in the middle of the (ordered) list.
  - If we have an even number of observations, then the median is defined to be between the two observations in the middle of the (ordered) list.

Example

Let $X = \{0.5, 0.3, 0.1, 0.1, 0.6, 0.9, 1.3, 1.1, 0.1\}$. If we sort X, then we have $\{0.1, 0.1, 0.1, 0.3, 0.5, 0.6, 0.9, 1.1, 1.3\}$

- Note that we have an odd number of values
- The mode is 0.1 and
- The median is 0.5.

Range

- The *range* of a sample X, defined as $R = \max\{X\} - \min\{X\}$, gives an insight on how widely spread a sample is.
- However, this is not sufficient to understand the behaviour of the corresponding sample.
    - How many values are clustered around the midpoint of the range?
    - How many of them are spread towards the endpoints?
    - Are they equally distributed over the range?

Quartiles

Consider a sample $X = \{x_1, x_2, \ldots, x_n\}$, so that $x_i \leq x_{i+1}$, or in other words, they are sorted in increasing order.

- The *first quartile* is $Q_1 = \dfrac{n+1}{4}$
- The *second quartile* is $Q_2 = \dfrac{n+1}{2}$, and finally
- The *third quartile* is $Q_3 = \dfrac{3(n+1)}{4}$.
  - If any of these values is not a whole number, then linear interpolation is commonly utilised.
- The *interquartile range* of X is then defined as

$$IR = Q_3 - Q_1$$

Quartiles

Loosely speaking,

- The *first quartile* splits off the lowest 25% of the sample from the highest 75%
- The *second quartile* splits the sample in half, and
- The *third quartile* splits off the highest 25% of data from the lowest 75%

Standard Deviation

Standard deviation is another measure of how widely spread the sample is, and it is defined as

$$\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2}.$$

In other words, it evaluates how much the observations (elements of the sample) vary with respect to the mean.

Example

As in the previous examples, let $X = \{0.5, 0.3, 0.1, 0.1, 0.6, 0.9, 1.3, 1.1, 0.1\}$. When sorted, we have that $X = \{0.1, 0.1, 0.1, 0.3, 0.5, 0.6, 0.9, 1.1, 1.3\}$

- The range $R = 1.3 - 0.1 = 1.2$,
- The standard deviation is $\sigma_X = 0.45$,
- The first quartile $Q_1 = \dfrac{9+1}{4} \approx 2$, which corresponds to 0.1,
- The second quartile $Q_2 = \dfrac{9+1}{2} = 5$, which corresponds to 0.5,
- The third quartile $Q_3 = \dfrac{3(9+1)}{4} \approx 7$, which corresponds to 0.9.
- Finally, the interquartile range is $IR = 0.9 - 0.1$.

Joint and Conditional Probabilities

- The joint probability $P(A, B)$ of two events A and B refers to the probability of their occurring, or being observed, at the same time
  - Note that $P(A, B) = P(B, A)$
- The conditional probability $P(A|B)$, is the probability of an event A being observed, given B has also being observed. These are linked by the following equation

$$P(A|B) = \frac{P(A, B)}{P(B)} \qquad \text{for} \quad P(B) \neq 0$$

- We can easily see that $P(A|A) = 1$ and that in general, for $A \subset B$, $P(A|B) = 1$
- If $A \cap B = \emptyset$, that is they are independent, then $P(A|B) = A$, as observing B does not influence the probability of observing A

Bayes' Rule

The previous equation can be re-arranged as follows

$$P(A|B)P(B) = P(A,B) = P(B,A) = P(B|A)P(A).$$

This gives the *Bayes' Rule*:

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

The Bayes' Rule is often used to relate $P(A|B)$ with $P(B|A)$, which is especially useful in parameter inference.

Example

Assume that

- The probability of being a Facebook user is $P(A) = 0.6$,
- The fraction of young men in the population is $P(B) = 0.3$, and
- The fraction of Facebook users among young men is $P(A|B) = 0.25$.

What is the fraction of young men among Facebook users, $P(B|A)$?

By the Bayes Rule, we have that

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{0.25 \cdot 0.3}{0.6} = 0.125.$$

Probability Distributions

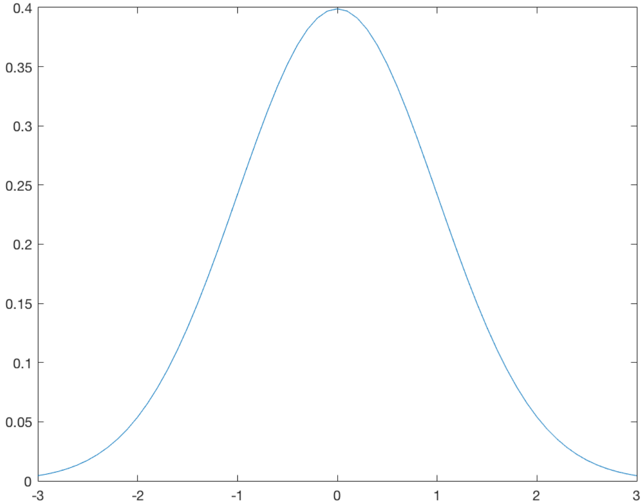
- A continuous random variable X is normally distributed if its density curve is symmetric defined by its mean $\bar{X}$ and standard deviation σ.
- For a number y, the probability associated with the interval $[\bar{X} - y\sigma, \bar{X} + y\sigma]$ is identical for all normal distributions. More specifically, we have the following identities

$$P(\bar{X} - \sigma, \bar{X} + \sigma) = 0.683 \tag{1}$$
$$P(\bar{X} - 2\sigma, \bar{X} + 2\sigma) = 0.954 \tag{2}$$
$$P(\bar{X} - 3\sigma, \bar{X} + 3\sigma) = 0.997 \tag{3}$$

In other words, when $y = 1, 2$ and 3, respectively.

Probability Distributions

Figure: The plot of the normal distribution.



Probability Distributions

- The standard normal distribution tables are usually used to determine the probability $P(X \leq x)$ of a random variable X, by providing an estimation of the area under the density curve to the left of a specific value x
- Similarly, $P(x_1 \leq X \leq x_2)$ gives the probability of X within the interval $[x_1, x_2]$ and it is estimated by finding the difference between the areas of the density curve to the left of x_2 and x_1, respectively
- Note that, we have that $P(X \leq 0) = 1/2$ as we take only half of it

Edge Hill University

Table of Contents

Expectation and Variance

- The expectation is the mean of the associated experiment for many iterations
- The variance measures how widely a set of (random) numbers are spread out from their mean value.
  - In other words, it is the average of the squared differences from the mean
  - Similarly to standard deviation, small values of the variance are associated with values of the observations close to the average value

Covariance

- Covariance $Cov(X, Y)$ measures the joint variability of two random variables X and Y
- The covariance determines the level linearly dependence between X and Y
  - If $Cov(X, Y)) > 0$, then both variables tend to take on relatively high values simultaneously.
  - If $Cov(X, Y) < 0$, then one variable tends to take on a relatively high value at the times that the other takes on a relatively low value and vice versa.
  - Correlation is similar to the covariance, with the different that is normalises the contribution of each variable in order to measure only how much the variables are related

Hypothesis Testing

- A prediction of a property is defined as *hypothesis*, which needs to be tested against a specific sample to ascertain whether this is indeed valid
- Hypothesis testing is based on investigating a random proportion of the sample to assess how much it supports the initial hypothesis.
    - We need to compare the case of a true hypothesis, with the scenario when our prediction does not apply to the sample. The former is denoted as the *alternate hypothesis* H_1, and the latter as the *null hypothesis*, or H_0
- Loosely speaking, there are two possible scenarios: *either* reject H_0 and accept the validity of our hypothesis H_1, *or* accept H_0 due to lack of evidence, which supports the validity of H_1
- However, as in the majority of statistical analysis, both cases do not entail any definite truth, rather a general assessment of the evidence for, or against a hypothesis

Hypothesis Testing

Assume we want to determine whether a new chemical compound is effective in treating a specific disease

- In many medical studies, a placebo is often administered to a portion of patients to understand whether it has the same effect as the tested chemical compound.
  - If so, we can say it has no "real" effect. In this case the placebo is, loosely speaking, comparable to the null hypothesis.
- More specifically, the likelihood of the observed sample occurring if the null hypothesis is true, defined as the $p-$value, is evaluated to assess whether the the outcome is comparable to what H_0 predicts
- The smaller the $p-$value is, the stronger the evidence of contradicting H_0 is

Edge Hill University

Table of Contents

Bayesian Inference

- Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available.
- Bayesian inference derives the *posterior probability* from two antecedents, a *prior probability* and a *likelihood function* based on a statistical model for the observed data

Bayesian Inference

This is defined as

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

- H is the hypothesis whose probability may be affected by data
- E is the the evidence, which corresponds to new data not used to evaluate the prior probability
- $P(H)$ is the prior probability
- $P(H|E)$ is the posterior probability, i.e. the probability of H given E
- $P(E|H)$ is the probability of having E given H
- $P(E)$ is the model evidence, that is the probability of E

Bayesian Inference

Informally, if the evidence does not match the hypothesis, then we should reject the hypothesis.
Consider the following example[1].
Imagine you are at the cinema and somebody drops their ticket. Assume s/he has long hair and that you can not tell their gender.
Do you call out "Excuse me madam!" or "Excuse me sir!"?

[1] Taken from `https://brohrer.github.io/how_bayesian_inference_works.html`.

Bayesian Inference

Assume that

- There are about half men and half women at the cinema, so out of 100 people, 50 are men, 50 are women.
- Out of the women, half have long hair (25) and the other 25 have short hair.
- Out of the men, 48 have short hair and 2 have long hair.

Since there are 25 long haired women and 2 long haired men, assuming that the ticket owner is a woman is a reasonable assumption

Bayesian Inference

Now consider the situation where you spot this person standing in a queue for the men's toilette.

- Out of 100 people in the men's toilette queue, however, there are 98 men and two women keeping their partners company.
- Half the women still have long hair and half have short hair, and there are just one of each.
- The proportions of men with long and short hair are the same too. Since there are 98 of them, there are now 94 with short hair and 4 with long.
- Since there is 1 woman with long hair and four men, now it is reasonable to assume that the ticket owner is a man.

This is an example of Bayesian inference. Knowing that the ticket owner is in the men's toilette queue allows us to make a better prediction about them

Conclusions

- We have discussed some general concepts in probability and statistics
- We don't need to be statisticians!
- We need to be acquainted with the general concepts