Inf2 – Foundations of Data Science 2022
Lecture: Descriptive statistics

FOUNDATIONS
OF
DATA
SCIENCE

25th September 2022

Recommended reading:
Modern Mathematical Statistics with Applications, Sections 1.1, 1.3 and 1.4

1 Introduction to descriptive statistics

Statistics The German word *Statistik* arose in the 18th century and originally referred to "data about the *state*" (country). The first use of "statistical" in the English language was in 1791 in the *Statistical Account of Scotland*. Sir John Sinclair, an elder in the Church of Scotland, sent a questionnaire to ministers in every parish (church district) in Scotland. The questionnaire asked many questions about agriculture, industry, economics, employment, poverty and education, as well as "The state of the manners, the morals, and the religious principles of the people". In fact empires and dynasties have been collecting data about population and trade for much longer than this, going back to the Han dynasty in China and the Roman Empire.

Descriptive statistics It's not humanly possible to make sense of a large raw datasets. For example, suppose we know the salary of every member of staff in the University of Edinburgh - the list would be very long. To make sense of this data we can try to summarise it or visualise it. **Descriptive statistics** refers to methods of summarising data. This topic introduces the notation we'll use for the sample and population mean and variance of univariate data. We will also introduce quantiles and skewed distributions.

2 Sample and population mean

Populations and samples It's important to distinguish between **populations** and **samples**. The population is the set of all the things we are interested in, for example, all 400 Scottish wildcats in the Highlands (Figure 1). The sample is a subset of the population that we observe – for example 10 wildcats that we trap, measure and release again into the wild.

We refer to the size of the population with N and the size of the sample with n. Note that we don't always know N exactly. In the case of the wildcats, $N = 400$ is an estimate – there is no practical way of counting all the wildcats in Scotland. In other cases we do know N exactly, for example if the population were a pile of exam papers.

Figure 1: Scottish wildcats, *Felis silvestris silvestris*, a critically endangered species. Credit: Peter Trimming / CC BY 2.0)

Definition of sample mean For a numeric variable x with n observations or instances sampled from a population, $x_1, x_2 \ldots x_n$, the **sample mean** is defined as:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

Sometimes the sample mean is written informally as $1/n \sum x_i$. When reporting the mean of a set of numbers, one convention is to report to one more decimal place than the accuracy of the x_i's. For example, if the age of 6 cats is $3, 4, 5, 6, 6$ and 7 years, the mean would be reported as 5.2 years, not 5.1666 years. Note that the units of the mean should be quoted; i.e. "5.2 years" *not* just "5.2".

The sample mean is a measure of where the centre of the set of instances is. It is guaranteed to lie between the minimum and maximum x_i. It has the same units as the x_i.

Population mean For a numeric variable x with N instances, $x_1, x_2 \ldots x_N$, the **population mean** is defined as:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2}$$

With bivariate data (two variables) or multivariate data (more than one variable), we can distinguish between the population means of variables x, y, z by using subscripts: e.g. μ_x, μ_y, μ_z.

The population mean is a measure of where the centre of all instances in the population is. It is guaranteed to lie between the minimum and maximum x_i. It has the same units as the the x_i.

The sample mean is an estimate of the population mean. We will consider how good an estimate it is later in the course when we learn about statistical inference. For now it is enough to know that it depends on how the sample is chosen (randomly or by some other method), and the values of n and N.

3 Sample and population median

Definition of median The **sample median** $\tilde{x}$ of a variable x is the "middle" value, when the sampled observations x_i are ordered from smallest to largest. To be more precise, if there are n observations, then the sample median is defined:

$$\tilde{x} = \begin{cases} (\frac{n+1}{2})^{\text{th}} \text{ ordered value, if } n \text{ is odd} \\ \text{mean of } (\frac{n}{2})^{\text{th}} \text{ and } (\frac{n}{2}+1)^{\text{th}} \text{ ordered values, if } n \text{ is even} \end{cases} \tag{3}$$

For example, the median age of 6 cats aged $3, 4, 5, 6, 6$ and 7 is 5.5 years. The median age of 5 cats aged $3, 4, 5, 6$ and 7 is 5 years. Note that we should quote the units, as for the mean.

By analogy with the population mean, the **population median** $\tilde{\mu}$ of a variable is the median of the entire population.

Median and mean The mean and median of a sample or population are generally not the same. For example, the mean age of 6 cats aged $3, 4, 5, 6, 6$ and 7 years is 5.2 years, but the median is 5.5 years.

- If a distribution is **symmetric**, $\overline{x} = \tilde{x}$
- If a distribution is **positively skewed**, $\overline{x} > \tilde{x}$
- If a distribution is **negatively skewed**, $\overline{x} < \tilde{x}$

Suppose the age of the cats had been $3, 4, 5, 6, 6$ and 18. The mean age would now be 7 years, but the median is unchanged at 5.5. An instance that appears to be far away from most of the other numbers is called an **outlier**. The example shows that the median is less affected by outliers than the mean. For this reason, the median can be seen as a better way of measuring a typical value of a variable.

It is often worth checking outliers, to make sure that they are real data. Depending on how the data has been collected, an outlier might be due to a faulty sensor, or a mistake in data entry or in the logic of an automated programme collecting data. However, outliers may well be real data, and should not just be removed as a matter of course.

4 Variance and standard deviation

Definition of sample variance and sample standard deviation For a numeric variable with n observations or instances, $x_1, x_2 \ldots x_n$, the **sample variance** is defined as:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{4}$$

The **sample standard deviation** is defined as:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{5}$$

The sample variance and standard deviation give a measure of how spread out the data is. It's an average of a measure of distance of each point from the sample mean (we'll come to why we divide by $n-1$ rather than n later).

One measure of distance we could use is the magnitude (absolute value) of the **deviation from the mean** of each observation: $x_1 - \overline{x}, x_2 - \overline{x}, \dots, x_n - \overline{x}$. We cannot just use the deviations, since they add up to 0 (think about it!). The magnitude of each deviation $|x_i - \overline{x}|$ is guaranteed to be positive. However the average of magnitudes is not as nicely behaved mathematically as the average of the square of the deviations, as defined in Equation 4.

It's important to quote the units of the standard deviation and variance. The standard deviation has the same units as the quantity in question and the variance has those units squared.

Definition of population variance and population standard deviation By analogy with the population mean, for a numeric variable in a population of N instances, $x_1, x_2 \dots x_N$, the **population variance** is defined as:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{6}$$

The **population standard deviation** is defined as:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{7}$$

Why the divisor $n-1$ in the sample variance? In short, we'd like the sample variance to be an unbiased estimate of the population variance. The squared deviations between the sample mean and the observations $(x_i - \overline{x})^2$ will tend to be smaller than the squared deviations between the population mean and the observations $(x_i - \mu)^2$, so if we divided by n we'd underestimate the sample variance.

A second way of thinking about this is that we know that the sum of the deviations is 0:

$$\sum_{i=1}^{n}(\overline{x} - x_i) = 0 \tag{8}$$

So if we know all but one ($n-1$) of the deviations, we can use the above equation to deduce the deviation we don't know. We say that this means there are $n-1$ **degrees of freedom**, and it turns out that it makes sense to divide by $n-1$.

In practice, when n is large, the difference doesn't matter much. It's worth being aware that various Python packages have different conventions about dividing by $n-1$ or n. `pandas.Series.std` divides by $n-1$ whereas `numpy.std` divides by n. This behaviour can be changed by specifying the `ddof` parameter in either function.

Scaled quantities Sometimes quantities can be scaled, for example if the units change. If a variable $y = cx$, where c is a scaling constant, then the following relationships hold:

$$\begin{aligned} \overline{y} &= c\overline{x} \\ s_y^2 &= c^2 s_x^2 \\ s_y &= c s_x \end{aligned} \tag{9}$$

We'll leave that as an exercise for you to prove.

Another way of writing the variance It's sometimes helpful to rearrange the summation over the squared deviations in the sample or population variance:

$$\begin{aligned} \sum (x_i - \overline{x})^2 &= \sum (x_i^2 - x_i\overline{x} - \overline{x}x_i + \overline{x}^2) \quad \text{expanding} \\ &= \sum x_i^2 - \sum x_i\overline{x} - \sum \overline{x}x_i + \sum \overline{x}^2 \quad \text{splitting up the summation} \\ &= \sum x_i^2 - n\overline{x}\,\overline{x} - n\overline{x}\,\overline{x} + n\overline{x}^2 \\ &= \sum x_i^2 - n\overline{x}^2 \end{aligned} \tag{10}$$

5 Quantiles

Definition of a percentile The yth percentile of a set of numeric observations $x_1, \dots, x_N$ is the value of x_i that is above y% of the values. For example, a baby that is on the 95th percentile for weight will weigh more than 95% of other babies.

The median as the 50th percentile By definition 50% of observations are less than the median, so we could also think of the median as the 50th percentile.

Lower and upper quartiles The 25th percentile is called the **lower quartile** (since it encloses the lower quarter of the distribution) and the 75th percentile is called the **upper quartile**. The difference between the upper and lower quartiles is called the **interquartile range**. The interquartile range is a measure of the spread of the distribution of values.

Quantiles Percentiles and quartiles are all examples of the general concept of **quantiles**. q-Quantiles are the values that divide the population or sample into q separate nearly equally sized groups. Percentiles are 100-quantiles and quartiles are 4-quantiles. Other common uses are deciles (10-quantiles) and quintiles (5-quantiles).

6 The mode

Definition of the mode The mode is the most frequent element in a set of data. In contrast to all the summary statistics described so far, the mode can be computed for categorical data as well as numeric data. For example the most popular name given to baby boys in 1964 was David – David was therefore the mode of baby boys' names in 1964.