Statistics and Machine Learning 1

Lecture 2B: Summary Statistics

Mark Muldoon
Department of Mathematics, Alan Turing Building
University of Manchester

Week 2

Measures of Central Tendency

Often, we are interested in what a *typical* value of the data; here we are going to start with some definitions you may have seen before, but which are a route into a more systematic treatment of what are called *summary statistics*.

- The *mean* of the data is

$$\mathrm{Mean}(x) = \langle x \rangle = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1}$$

- The *median* of the data is the value that sits in the middle when the data are sorted by value. This is a special case of an *order statistic*.
- A *mode* in data is a value of x that is 'more common' than those around it, or a 'local maximum' in the density. For discrete data, this can be uniquely determined as the most common value, but for continuous data modes need to be *estimated*, one aspect of a major strand in data science, *estimating distributions*.

Calculating Means

This goes much as you'd expect!

- In Python:

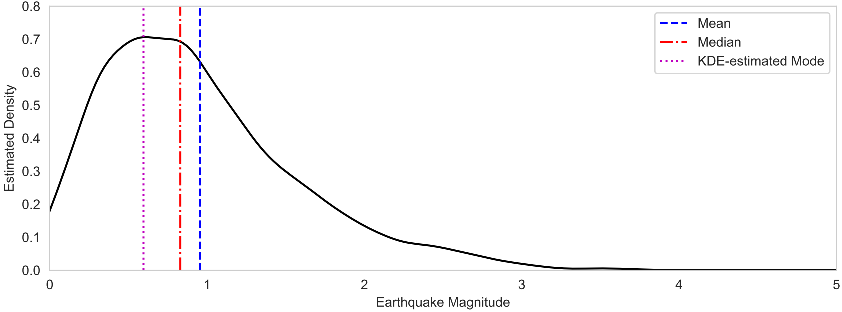
```
>>> np.mean(x)
0.9555100148367953
```

- In R:

```
> mean(x)
[1] 0.95551
```

Visualising Measures of Central Tendency

- For the earthquake data, we estimate from the kernel density that there is one mode, and its location (more on this later) and calculate the mean and median directly.
- The data are *right-skewed*, and as a consequence of this the mode is smallest and the mean is largest – we will consider this further; note that for a normal distribution all would be equal.



Variance

Practice manipulating expectations; the data's *variance* is:

$$
\begin{aligned}
\mathrm{Var}(x) &= \left\langle (x - \langle x \rangle)^2 \right\rangle \\
&= \frac{1}{n} \sum_{i=1}^{n} (x_i - \langle x \rangle)^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} (x_i^2 - 2x_i \langle x \rangle + \langle x \rangle^2) \\
&= \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) - 2 \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)}_{\text{This is } \langle x \rangle} \langle x \rangle + \frac{1}{n} \underbrace{\left(\sum_{i=1}^{n} 1 \right)}_{\text{This is } n} \langle x \rangle^2 \\
&= \frac{1}{n} \left(\sum_{i=1}^{n} x_i^2 \right) - \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2 \\
&= \left\langle x^2 \right\rangle - \langle x \rangle^2 .
\end{aligned} \tag{2}
$$

Unbiased Variance and Computation

- later in the course you will encounter a slightly different formula, the *unbiased estimate of the variance*:

$$\widehat{\mathrm{Var}}(x) = \frac{n}{n-1}\mathrm{Var}(x) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right). \tag{3}$$

- Wikipedia offers a good explanation of the difference.
- Python by default calculates the biased version, and R the unbiased one.
- So, as always, you should *check your software's documentation*, but n is often so large that the distinction is not important.

Calculating Variances

- Biased

```
>>> np.var(x)
0.400257047006528
```

- Unbiased

```
>>> np.var(x,ddof=1)
0.40033129242426246
```

- Unbiased

```
> var(x)
[1] 0.4003313
```

- Biased

```
> library(moments)
> moment(x, order=2, central = TRUE)
[1] 0.400257
```

‘Natural’ units

- Generally our results should not depend on the *units* with which we make measurements, e.g. whether we are working in inches or metres.
- When we have a relevant physical scale, this can be used to define ‘natural units’. For example the mass of the carbon atom is used to define atomic mass.
- For more general data, we do not have such constants but there are two commonly-used quantities that have the same units as the data.
- One is the mean,
$$\mu = \mathrm{Mean}(x), \tag{4}$$
and the other is the *standard deviation*,
$$\sigma = \sqrt{\mathrm{Var}(x)}. \tag{5}$$

Working in 'natural' units

- These two quantities let us define two transformations commonly applied to data.
- The first is *centring*, with the centred data given by

$$y_i = x_i - \mu, \tag{6}$$

  which is defined so that

$$\mathrm{Mean}(y) = 0.$$

- The second is *standardisation*, with the standardised data given by

$$z_i = \frac{y_i}{\sigma}. \tag{7}$$

  This choice means that

$$\mathrm{Var}(z) = 1.$$

Higher moments

- In general, the r-th *moment* of the data is

$$m_r = \langle x^r \rangle . \tag{8}$$

- The r-th *central moment* of the data is

$$\mu_r = \langle (x-\mu)^r \rangle = \langle y^r \rangle , \tag{9}$$

  where the y's are the centred versions of the data.

- The r-th *standardised moment* of the data is

$$\tilde{\mu}_r = \left\langle \left(\frac{x-\mu}{\sigma} \right)^r \right\rangle = \langle z^r \rangle = \frac{\langle (x-\mu)^r \rangle}{\sigma^r} = \frac{\mu_r}{\sigma^r}. \tag{10}$$

- In theory, all higher moments are informative about the data, but in practice those with $r = 3$ and $r = 4$ are most commonly reported.

Skewness

- We define the *skewness* by

$$\mathrm{Skew}(x) = \tilde{\mu}_3. \tag{11}$$

- A larger (more positive) value of this quantity indicates *right-skewness*, meaning that more of the data's variability arises from values of x larger than the mean.
- Conversely, a smaller (more negative) value of this quantity indicates *left-skewness*, meaning that more of the data's variability arises from values of x smaller than the mean.
- A value close to zero means that the variability of the data is similar either side of the mean (but *does not imply* an overall symmetric distribution).

Calculating Skewness

As noted before, the earthquake data is right-skewed:

- In Python:

```
>>> ss = np.sqrt(np.var(x))
>>> sp.stats.moment(x,3)/(ss**3)
1.0710553097009332
```

- In R:

```
> ss = sqrt(moment(x, order=2, central = TRUE))
> moment(x, order=3, central = TRUE)/(ss^3)
[1] 1.071055
```

Kurtosis

- We define
$$\mathrm{Kurtosis}(x) = \tilde{\mu}_4. \tag{12}$$
- A value of this quantity larger than 3 means that more of the variance of the data arises from the tails than would be expected if it were normally distributed.
- A value of this quantity less than 3 means that less of the variance of the data arises from the tails than would be expected if it were normally distributed.
- A value close to 3 is consistent with, though not strong evidence for, a normal distribution.
- The difference between the kurtosis and 3 is called the *excess kurtosis*.

Calculating Kurtosis

The earthquake data has a positive value of excess kurtosis and so is *leptokurtic*:

- In Python:

```
>>> (sp.stats.moment(x,4)/(ss**4)) - 3
1.4368380067616835
```

- In R:

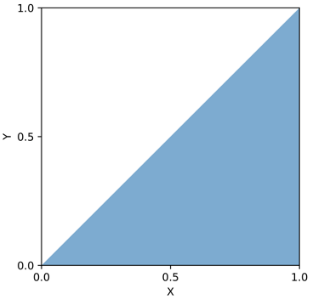
```
> (moment(x, order=4, central = TRUE)/(ss^4))-3
[1] 1.436838
```

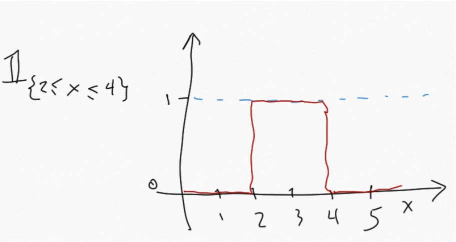
Mathematical aside: indicator functions

- The *indicator function* of a logical proposition A

$$\mathbb{1}_{\{A\}} = \begin{cases} 1 \text{ if } A \text{ is true} \\ 0 \text{ if } A \text{ is false} \end{cases} \tag{13}$$

- Two examples of indicator functions: at left, an example from last week's quiz, where the blue shaded area shows that part of the unit square where $\mathbb{1}_{\{y \leq x\}} = 1$ and at right, the function $\mathbb{1}_{\{2 \leq x \leq 4\}}$, which is nonzero on the interval $2 \leq x \leq 4$.



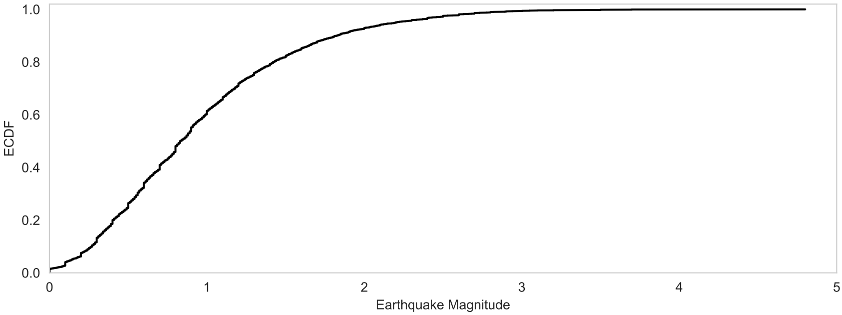


The ECDF

- The *empirical cumulative distribution function* (ECDF) is

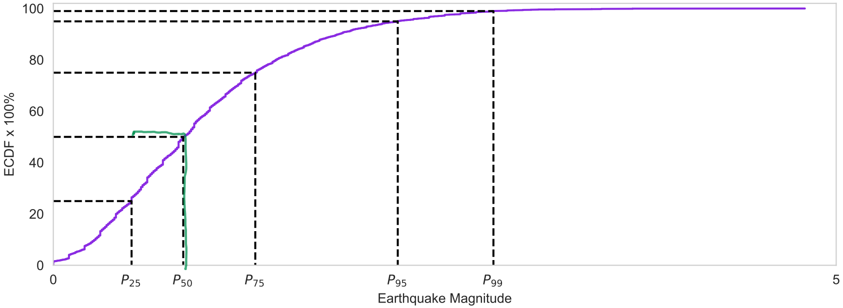
$$E(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i \,\leq\, t\}} = \left\langle \mathbb{1}_{\{x \,\leq\, t\}} \right\rangle . \tag{14}$$

- The plot of this function is a *lossless* visualisation of the data.



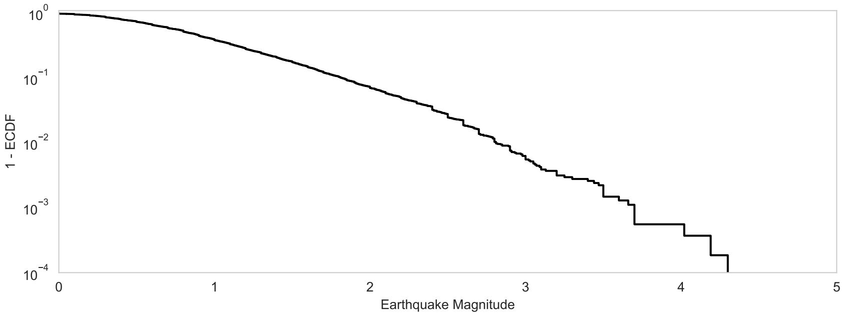
Quantiles and Order Statistics

- The z-th *percentile*, P_z is the value of x for which $z\%$ of the data is $\leq x$.
- So the *median* is $\mathrm{median}(x) = P_{50}$.
- This is related to the ECDF as illustrated below.
- A measure of *dispersal* of the data is the *inter-quartile range*, $\mathrm{IQR}(x) = P_{75} - P_{25}$.



The Tail

- The upper tail of the data—*i.e.* the largest values of x—are often of most interest, but hardest to visualise.
- For example, low-magnitude earthquakes are of much less interest than high-severity ones!
- A commonly used Measure is $1 - E(x)$, which is sometimes called the *survival function*, and which can be plotted using a logarithmic y-axis to make the behaviour of the tail clearer, e.g.:



Multimodality

- The first definition of the *mode* that most people see is the most frequent value in a dataset. According to this definition, the mode of both $(2, 3, 1, 2, 2)$ and $(2, 3, 1, 2, 12)$ is 2.
- For continuous data, there aren't typically identical observations (and if there are, they aren't typical) so we will need to *estimate* the modes, which we define as local maxima (peaks) of the probability density function.
- The location and number of modes is typically the most relevant measure of central tendency and variability for *multimodal* data.
- We will see that different estimation procedures give different modes, and even for simulated data like the below that is 'obviously' bimodal, they will give different answers about mode locations.