Week 5

Data Science Laboratory 1

Dr John Evans
j.evans8@herts.ac.uk

University of Hertfordshire UH

Plan for today

More Advanced Statistics

Uncertainty and significance

Types of Data

More advanced statistics (Skew and kurtosis)

- Skewness and kurtosis are further useful descriptors of the shape of a distribution.
- Skewness is a measure of how asymmetric a distribution is about its midpoint.
- Kurtosis measures the ‘tailedness’ of a distribution, relative to a normal distribution.
- A high kurtosis means that there is more weight in the tails compared to a normal distribution (more outliers), and a low kurtosis means the tails are ‘lighter’ than that expected for a normal distribution (few outliers).

More advanced statistics (Skew and kurtosis)

- Skewness and kurtosis are further useful descriptors of the shape of a distribution.
- Skewness is a measure of how asymmetric a distribution is about its midpoint.
- Kurtosis measures the 'tailedness' of a distribution, relative to a normal distribution.
- A high kurtosis means that there is more weight in the tails compared to a normal distribution (more outliers), and a low kurtosis means the tails are 'lighter' than that expected for a normal distribution (few outliers).
- (**Fisher-Pearson coefficient** of skewness) For a sample $x_1, x_2, x_3, ..., x_N$ skew is defined as follows:

$$g = \frac{\Sigma_{i=1}^{N}(x_i - \bar{x})^3/N}{\sigma^3},$$

where $\bar{x}$ and σ are the mean and standard deviation of the distribution, and N is the total number of samples.

Kurtosis definition

- Kurtosis is then defined as follows:

$$k = \frac{\Sigma_{i=1}^{N}(x_i - \bar{x})^4/N}{\sigma^4}.$$

- For a normal distribution, $k = 3$; this is called **Pearson's** definition of kurtosis.
- Often, kurtosis is calculated as $k - 3$ such that a Normal distribution will yield a result of zero.
- This is called **Fisher's** definition.

Pearson's --> for a normal distribution k = 3.
Fisher's --> for a normal distribution k = 0.

Using Python

- In practice, we can use existing Python functions to calculate kew and kurtosis. For example, using the Scipy module we have the following:

```
from scipy import stats
import numpy as np

sample = np.random.normal(0,1,size=100)
print(stats.skew(sample),stats.kurtosis(sample))
#-0.03476607216695398 -0.4636003530458841
```

- The reader is recommended to always check the documentation to see how these values are being calculated. For skew, the calculation is sometimes modified for low-number statistics.
- There are also alternative definitions of skewness that are sometimes used. In the above example, it is clear that Scipy is using the Fisher definition of kurtosis for example, as the result is close to 0.

General rule of thumb in interpretation

- Positive skewness shows that the mean is larger than the median and there is a tail of high values.
- Negative skewness shows that the mean is smaller than the median and there is a tail of small values.
- Positive kurtosis corresponds to a ‘thin and pointed’ distribution.
- Negative kurtosis corresponds to a ‘broad and flat’ distribution.

More advanced statistics (Spearman's rank correlation coefficient)

- Imagine if we 'rank' our observations x and y in ascending order. Spearman's rank correlation assesses the correlation between these ranked variables.
- The key reason for doing this compared to Pearson's coefficient is that r_{xy} assesses linear relationships. By ranking the data and calculating Spearman's coefficient we can tell if the variables are correlated by any general monotonic function[1].

[1] A monotonic function is one that does not change (or reverses) the rank order of a variable.

More advanced statistics (Spearman's rank correlation coefficient)

- Imagine if we 'rank' our observations x and y in ascending order. Spearman's rank correlation assesses the correlation between these ranked variables.
- The key reason for doing this compared to Pearson's coefficient is that r_{xy} assesses linear relationships. By ranking the data and calculating Spearman's coefficient we can tell if the variables are correlated by any general monotonic function[1].
- If we call the ranked data $\tilde{x}$ and $\tilde{y}$, then Spearman's rank correlation coefficient is given by

$$\rho_{xy} = \frac{\mathrm{cov}(\tilde{x}, \tilde{y})}{\sigma(\tilde{x})\sigma(\tilde{y})}.$$

- See the following for an example:
  `https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php`

[1] A monotonic function is one that does not change (or reverses) the rank order of a variable.

In Python

- For us, we note that in Python we can calculate this value easily, as follows:

```
from scipy import stats

# Assuming x and y are numpy arrays containing the samples
r,p = stats.spearmanr(x,y)
```

- With the return values representing the equivalent to those described above for `stats.pearsonr(x,y)`.

Uncertainty and significance

- When making a measurement with some instrument there is a certain margin of ‘error’ involved. This refers to the range of confidence we have, i.e. the probability distribution that tells us the likelihood that a given value is the true answer.
- If we weigh some flour, we will probably get somewhere close to the true weight of flour. However, perhaps some of the markings are worn away, or the calibration of the scales is off... these are all sources of error.

Uncertainty and significance

- When making a measurement with some instrument there is a certain margin of ‘error’ involved. This refers to the range of confidence we have, i.e. the probability distribution that tells us the likelihood that a given value is the true answer.
- If we weigh some flour, we will probably get somewhere close to the true weight of flour. However, perhaps some of the markings are worn away, or the calibration of the scales is off... these are all sources of error.
- So, maybe we are fairly confident of our reading (let us say it is 100 g), but it could be off by 10 g either way. That ‘plus or minus’ 10 g is our margin of error, or confidence range. When writing it down we would say the weight is 100 ± 10 g.
- If we were to plot the weight on a graph, we would show as a point the best guess value of 100 g, but also show an ‘error bar’ indicating the ± 10 g range.

Uncertainty and significance

- When making a measurement with some instrument there is a certain margin of 'error' involved. This refers to the range of confidence we have, i.e. the probability distribution that tells us the likelihood that a given value is the true answer.
- If we weigh some flour, we will probably get somewhere close to the true weight of flour. However, perhaps some of the markings are worn away, or the calibration of the scales is off... these are all sources of error.
- So, maybe we are fairly confident of our reading (let us say it is 100 g), but it could be off by 10 g either way. That 'plus or minus' 10 g is our margin of error, or confidence range. When writing it down we would say the weight is 100 ± 10 g.
- If we were to plot the weight on a graph, we would show as a point the best guess value of 100 g, but also show an 'error bar' indicating the ± 10 g range.
- An error resulting from imperfections in an instrument, or a human's inability to read the exact value is called a random error. An error resulting from, for example, a fixed offset in the calibration of an instrument is called a systematic error.

Probability distribution functions

- Ideally, for any measurement we would have the full probability distribution showing for all possible values what the probability density is for a given value being the true one.
- The integral of the probability distribution function (PDF) will then be the unity, meaning that the true value is definitely contained somewhere within the full range of values.

Probability distribution functions

- Ideally, for any measurement we would have the full probability distribution showing for all possible values what the probability density is for a given value being the true one.
- The integral of the probability distribution function (PDF) will then be the unity, meaning that the true value is definitely contained somewhere within the full range of values.
- Generally we do not have the full PDF, which is arguably the most complete description of uncertainty for a given observation or measurement.
- Typically, we assume that the PDF can be described by a Gaussian distribution, characterised by some mean μ and standard deviation σ, where μ and σ represent the measurement and error bar.

The Gaussian probability distribution

- The Gaussian probability distribution is defined to be

$$G(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

- We can think of integrating $G(x)$ over ranges about the mean value μ in multiples of σ.

The Gaussian probability distribution

- The Gaussian probability distribution is defined to be

$$G(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

- We can think of integrating $G(x)$ over ranges about the mean value μ in multiples of σ.
- For example:

$$\int_{\mu-\sigma}^{\mu+\sigma} G(x)dx = 0.682.$$

This means that 68% of the probability is contained within $\pm 1\sigma$ of the mean. Therefore we often quote '1σ' errors because this range contains the majority of the probability density.

More examples

- If we go to 2σ then we get

$$\int_{\mu-2\sigma}^{\mu+2\sigma} G(x)dx = 0.954.$$

So, there is now a 95% chance that the true value lies in the range $\mu \pm 2\sigma$.

More examples

- If we go to 2σ then we get

$$\int_{\mu-2\sigma}^{\mu+2\sigma} G(x)dx = 0.954.$$

So, there is now a 95% chance that the true value lies in the range $\mu \pm 2\sigma$.

- Even more extreme, we have the following:

$$\int_{\mu-3\sigma}^{\mu+3\sigma} G(x)dx = 0.997.$$

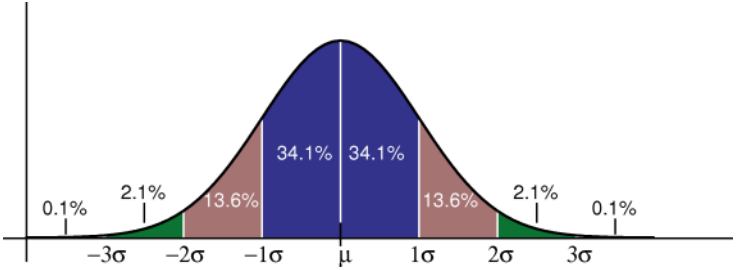
99.7% chance that the true value lies in the range mu +- 3sigma

Significance

- Characterising the uncertainty in this way also gives us a means of quantifying the significance of an observation.
- For a sample drawn from a distribution $G(x)$, the probability of finding a value with $x = \mu \pm 5\sigma$ is vanishingly small, since

$$\int_{\mu-5\sigma}^{\mu+5\sigma} G(x)dx = 0.999999426697.$$

- Therefore there was a 1 in 1,744,278 chance of it occurring.
- You will no doubt often hear about a '5σ' result in science – this is what that is referring to. A 5σ deviation is *significantly* different to the mean.


0.1%
2.1%
13.6%
34.1%
34.1%
13.6%
2.1%
0.1%
−3σ
−2σ
−1σ
μ
1σ
2σ
3σ

Significance example

- Suppose we have two measurements $x = 13$ and $y = 17$ and we want to assess whether they are significantly different. On the face of it, you would say $x - y = -4$ and so yes the observations are different. But to determine significance we need to compare the confidence ranges, or the error bars.
- Assuming Gaussian uncertainties, if we have $x = 13 \pm 2$ and $y = 17 \pm 3$ we would calculate the difference and combine the uncertainties in quadrature: $x - y = (13 - 17) \pm \sqrt{2^2 + 3^2} = -4 \pm 3.6$.
- The margin of error is similar in size to the magnitude of the value, or the 'signal-to-noise' ratio, and is $4/3.6 = 1.1$.
- In other words, the uncertainties on the measurements are too large to determine if the results are different, because the 1σ confidence range (nearly) includes the result $x - y = 0$.

Example continued

- If we made our measurements again, using better instruments or observations, and got $x = 13 \pm 0.1$ and $y = 17 \pm 0.5$, we would have a different $x - y = -4 \pm 0.5$.
- This is *eight* standard deviations (8σ) away from the result $x - y = 0$, so the two values are indeed significantly different.

Evaluating Gaussian significances in Python

```
from scipy import stats

# Define the mean and s.d. of a distribution G(x)
mu = 0
sigma = 1

x = 2

# Probability density for a given x
print(stats.norm.pdf(x,mu,sigma))
#0.053991

# Cumulative probability density from -infty to x
print(stats.norm.cdf(x,mu,sigma))
#0.977250

N = 3
# Integrated probability between +/-N*sigma
print(stats.norm.cdf(N*sigma,mu,sigma) - stats.norm.cdf(-N*sigma,mu,sigma))
#0.9973002
```

Types of data

- Now that we have an idea about *how* to apply statistical tools, we next need to ensure we have an understanding about *when* to apply statistical tools.
- The key idea here is to understand that data can come in a variety of types, and certain types are less suitable to statistical measures than others.

Types of data

- Now that we have an idea about *how* to apply statistical tools, we next need to ensure we have an understanding about *when* to apply statistical tools.
- The key idea here is to understand that data can come in a variety of types, and certain types are less suitable to statistical measures than others.
- To discuss this, we need some definitions.

Some definitions

- A *data set* is a set of measurements taken from some environment or process. This can be viewed as a collection of *data objects*[2].
- In turn, data objects are typically described by their *attributes*. These capture some characteristic of the object, such as the mass of a physical object, or the time at which an event occurred.

[2] Data objects are also called records, points, vectors, events, samples, instances, observations...

Some definitions

- A *data set* is a set of measurements taken from some environment or process. This can be viewed as a collection of *data objects*[2].
- In turn, data objects are typically described by their *attributes*. These capture some characteristic of the object, such as the mass of a physical object, or the time at which an event occurred.
- If the data objects are stored in a database, they are *data tuples*; that is, the rows of a database correspond to the data objects, and the columns correspond to the attributes.
- For example, a database could consist of the entire student body of the University of Hertfordshire. In this case, each row corresponds to a student, and each column an attribute, such as Student ID, Year of Study, Course, Modules Taken, Grades, and so on.

[2] Data objects are also called records, points, vectors, events, samples, instances, observations...

Examples

The following are examples of attributes:

- $\{$Blue, Brown, Green, Hazel$\}$ is a set of possible eye colours.
- Temperature is an attribute of an object that varies over time.
- $\{$Student ID, Name, Home Address$\}$ are attributes of a student at UH.
- Height, weight and sex are all attributes of a patient at a hospital.

N.B. It should be clear that the properties of an attribute need not be the same as the properties of the values used to measure it. In other words, the values used to represent an attribute can have properties that are not properties of the attribute itself, and vice versa. We can illustrate this with two examples.

Example 1

- Consider the attributes of Employee Age and Employee ID for any given employee.
- Both of these attributes can be represented by integers (if age is taken in years) and for any set of integers, we can compute the average.
- While it makes sense to talk of the average age of employees, it does not make sense to talk of the average Employee ID.
- Instead, for this latter attribute, we are only really interested in whether two given IDs are distinct or not.

Example 2

- Now consider the attributes of Length of line segments.
- There are two ways we could do this, we could either let the attributes belong to some discrete set $\{$short, medium, long$\}$, or we could assign numbers.
- However, with the latter, there are many numbers we could assign.

Example 2

- Now consider the attributes of Length of line segments.
- There are two ways we could do this, we could either let the attributes belong to some discrete set $\{\text{short, medium, long}\}$, or we could assign numbers.
- However, with the latter, there are many numbers we could assign.
- We could assign an actual value of length (in cm, say), or we could assign numbers that only capture the order of these line segments (so 1 for shortest, 2 for next shortest, and so on).
- Again, for the latter, things like average or sum make little sense.
- If the line segments are actually something like a plank of wood, then we can also talk about an upper bound on the length, whereas there is no upper bound on the set of real numbers, say.

The point of all this

- The point of the above is that we need to know the type of an attribute because it tells us which properties of the measured values are consistent with the underlying properties of the attribute, and therefore allows us to avoid foolish actions such as computing the average Employee ID.
- The type of an attribute is determined by the set of possible values the attributes can have.
- Usually, we classify them into one of four categories: nominal, ordinal, interval and ratio (NOIR).

Properties of attributes

- To discuss these classifications, we first need to identify the properties of numbers that correspond to the underlying properties of the attribute.
- The following properties (operations) of numbers are typically used to describe attributes:
  1. **Distinctness**, $=$ and $\neq$
  2. **Order**, $<$, $\leq$, $>$ and $\geq$
  3. **Addition**, $+$ and $-$
  4. **Multiplication**, $\times$ and $/$

Properties of attributes

- To discuss these classifications, we first need to identify the properties of numbers that correspond to the underlying properties of the attribute.
- The following properties (operations) of numbers are typically used to describe attributes:
  1. **Distinctness**, $=$ and $\neq$
  2. **Order**, $<$, $\leq$, $>$ and $\geq$
  3. **Addition**, $+$ and $-$
  4. **Multiplication**, $\times$ and $/$
- The four types mentioned above (NOIR) are defined in the following table, along with statistical operations that are valid for each type. Each attribute type possesses all of the properties and operations of the attribute types above it. Consequently, any property or operation that is valid for nominal, ordinal and interval attributes is also valid for ratio attributes, i.e. the attribute types are cumulative.

Qualitative attributes

(discrete)

Attribute Type	Description	Examples	Operations
Nominal can distinguish differences / naming things	The values of a nominal attribute are just different names, i.e. nominal values provide only enough information to distinguish one object from another ($=, \neq$)	postcodes, employee ID, eye colour, sex	mode, entropy, contingency, correlation, χ^2 test
Ordinal can be ordered	The values of an ordinal attribute provide enough information to order objects ($<, >$)	hardness of minerals, {good, bad}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests

Quantitative attributes

(continuous)

Attribute Type	Description	Examples	Operations
Interval can add or subtract	For interval attributes, the differences between values are meaningful, i.e. a unit of measurement exists ($+$, $-$)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, t and F tests
Ratio can multiple and divide	For ratio variables, both differences and ratios are meaningful ($\times$, $/$)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Nominal attributes

- **Nominal** (meaning “relating to names") attributes are symbols or *names of things*.
- Each value represents some kind of category, code or state, and so nominal attributes are also referred to as categorical.
- They do not have any meaningful order and things like mean (average) or median (middle) have no meaning here.
- One thing that may be of interest, however, is the attribute’s most commonly occurring value, i.e. the mode.

Ordinal attribute

- **Ordinal** attributes are differentiated from nominal values by the fact that ordering is now meaningful.
- In this way, we can rank ordinal values.
- For example, *drink_size* could correspond to the size of drinks in a fast-food restaurant. It could have three possible values: *small, medium* and *large*. These values have a meaningful sequence corresponding to increasing drink size, and so are ordinal.
- Other examples could be grades, military rank or certain job titles.
- Ordinal attributes are useful for registering subjective assessments of qualities that cannot be measured objectively. They are therefore often used in surveys for ratings (e.g. a survey on customer satisfaction).
- The central tendency of an ordinal attribute can be represented by its mode and its median, but the mean cannot be defined.

Interval attributes

- **Interval**-scaled attributes are measured on a scale of equal-size units.
- The values of interval-scaled attributes have order and can be positive, zero or negative.
- Thus, not only can we provide a ranking of values, but we can also compare and quantify the difference between values. This makes them quantitative, rather than qualitative.
- They are measurable quantities, presented as integer, rational or real values.
- For interval-scaled attributes, we can compute their mean value, median and mode.

Example

- A typical example is temperature. If we measure the temperature outside our window each day for a month, then we can order these values, obtain a ranking of which days were hottest/coldest, and quantify the swing in temperature(s).
- However, there is a subtlety. When measured on the Kelvin scale, a temperature of 2° is, in a physically meaningful way, twice that of a temperature of 1°. This is not the case when temperature is measured in Fahrenheit or Celsius where, physically, a temperature of 1° Fahrenheit (Celsius) is not much different than a temperature of 2° Fahrenheit (Celsius).

Example

- A typical example is temperature. If we measure the temperature outside our window each day for a month, then we can order these values, obtain a ranking of which days were hottest/coldest, and quantify the swing in temperature(s).
- However, there is a subtlety. When measured on the Kelvin scale, a temperature of 2° is, in a physically meaningful way, twice that of a temperature of 1°. This is not the case when temperature is measured in Fahrenheit or Celsius where, physically, a temperature of 1° Fahrenheit (Celsius) is not much different than a temperature of 2° Fahrenheit (Celsius).
- The reason for this is the somewhat arbitrary zero point for Fahrenheit and Celsius, i.e. 0° (Fahrenheit or Celsius) does not indicate 'no temperature'.
- As such, for Fahrenheit and Celsius we can compute the *difference* between temperature values, but we cannot talk of one temperature value as being a multiple of another. For this reason, it is interval-scaled, and not ratio-scaled.

Ratio attribute

- Finally, a **ratio**-scaled numeric attribute is differentiated from an interval-scaled numeric attribute by the presence of an inherent zero-point.
- This allows us to make statements such as ‘Attribute A is twice that of Attribute B’.
- So these values are ordered and we can compute the difference between values, multiples of values, as well as the mean, median and mode.

Ratio attribute

- Finally, a **ratio**-scaled numeric attribute is differentiated from an interval-scaled numeric attribute by the presence of an inherent zero-point.
- This allows us to make statements such as ‘Attribute A is twice that of Attribute B’.
- So these values are ordered and we can compute the difference between values, multiples of values, as well as the mean, median and mode.
- As we saw above, temperature measured in Kelvin is ratio-scaled as 0° ($= -273^\circ C$) is a true zero-point. It is the point at which the particles that comprise matter have zero kinetic energy.
- Another example is that of *years_of_experience* or *number_of_words*.
- We could also include things like height, weight etc.

Summary

- We have seen some more advanced statistics, including skew and kurtosis, as well as Spearman's rank correlation coefficient
- We have seen the ideas of uncertaity and significance.
- We have seen ways to categoraise different data attributes: NOIR.