**Statistics Training – Session 1**

These are your notes to supplement the PowerPoint presentations and practical sessions in Excel and R. Hopefully these notes will be a useful resource to reference back to if you want to refresh on your statistics after the training. There will be more images in the PowerPoint than in this file – this is mostly for notes - so remember to look back through the PowerPoint as well.

**Why do we need to understand different data types? (slide 2)**

Before you can start to analyse your data, you need to understand the data. Part of understanding the data includes figuring out what type of data you have. The type of data helps to determine how you should summarise and visualise the data.

You can then go onto to select the most appropriate statistical tests and analysis methods (covered in later sessions) based on your data type. Using inappropriate methods for your data type can lead to incorrect assumptions and conclusions being drawn.

Throughout this session, we will go through the different types of data, how to identify these, and how you can carry out basic statistical analysis for each data type. There are practical sessions throughout the in-person training and a couple of summary tables at the end of the PowerPoint, so you will hopefully come away with good insight and resources that you can take away into your own role.

**Types of data (slide 3)**

We are going to go over discrete and continuous data. Discrete or continuous is a way of categorising quantitative data – this means data that we can count or measure. There are other types of data, like qualitative (this is more descriptive and requires interpretation e.g. a staff survey with free text), but we won’t go over these in this session.

**Discrete Data**

Discrete data can, broadly, be categorical or numerical.

In categorical data, the values fall into distinct groups or categories, and the value can only be one of these categories, there’s no in-between values. A really clear example is IMD Decile – these only range from 1 to 10. You can’t have in-between values like 1.5, 5.42 etc.

There are some slightly more confusing examples of categorical data, usually ones that involve numbers, but the numbers are really categories. For example, a patient pain rating scale that goes from:

**5**

**4**

**3**

**2**

**1**

**No Pain**

**Some Pain**

**Moderate Pain**

**Significant Pain**

**Extreme Pain**

This is discrete because it has 5 distinct groups. You might think it falls under discrete numerical due to the numbering, but it’s categorical because the numbers are just acting as labels. The numbers themselves don’t mean anything, it’s the pain rating that we’ve assigned to them that matters. If the patient rated their pain as 1, and you went to a doctor in a completely different team and told them the patient has pain of 1, they would have no idea what this means. So, the number itself is not significant here, it’s the label of no pain.

There are also binary categorical data that you might see quite often at work. The outcome can be one of only 2 things, usually a 1 or 0. This might be patients that do (1) or do not have (0) a disease. This might make the distinction between categorical discrete and numerical discrete data a bit easier to understand. Even though the values are assigned 1 and 0, the numbers themselves don’t actually mean anything as numbers, they’re just labels, so this is categorical rather than numerical.

If we compare this to nurses in a workforce for example, we are actually dealing with the numbers themselves, we’re counting the number of people, so this would be discrete numerical.

**Continuous Data**

Then we have continuous data which is also numeric, **but it is not** discrete. Earlier we said discrete data values fall into distinct groups or categories, and the value can only be one of these categories. This is not the case in continuous data. As the name suggests, continuous data exists over a continuous scale – we can divide the value into smaller and smaller proportions. Think of ambulance handover time in minutes. We can divide this down into seconds, hundredths of a second, thousands of a second etc. so it is impossible to divide this data into distinct categories.

Birthweight is another common example of continuous data in healthcare. The measurement of weight (pounds, kg etc.) is a continuous measurement.

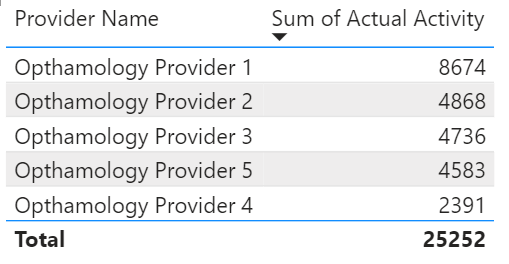
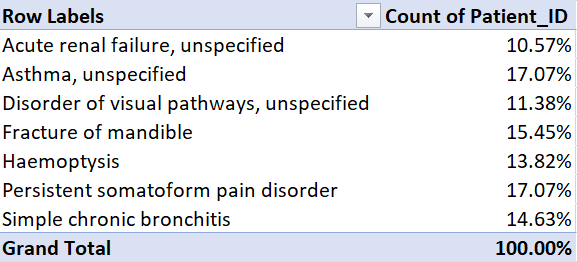
Length of stay (LOS) is also a measurement of time, making it continuous. If we know exactly how long a patient has been in hospital, we can drill down to even the number of seconds they spent in hospital. However, you might more commonly come across LOS data that has been rounded and grouped into whole days. In this case, it has actually become discrete numerical data as we are just looking at whole numbers rather than a continuous scale. We know how many days they have stayed in hospital, but we are not able to divide this into smaller fractions.

**Discrete data visualisations (slides 4-6)**

We will go over frequency tables (counts and proportions), column and bar charts, and pie charts.

**Frequency Tables**

If you have categorical data, because the data is sorted into categories or groups, you can simply count the number of observations in each group as a first point of call. These counts are called frequencies. Once you have the counts, or frequencies, you can further analyse them by calculating them as a percentage or proportion of the whole data set – this can help give a more meaningful value than the raw counts in some cases.

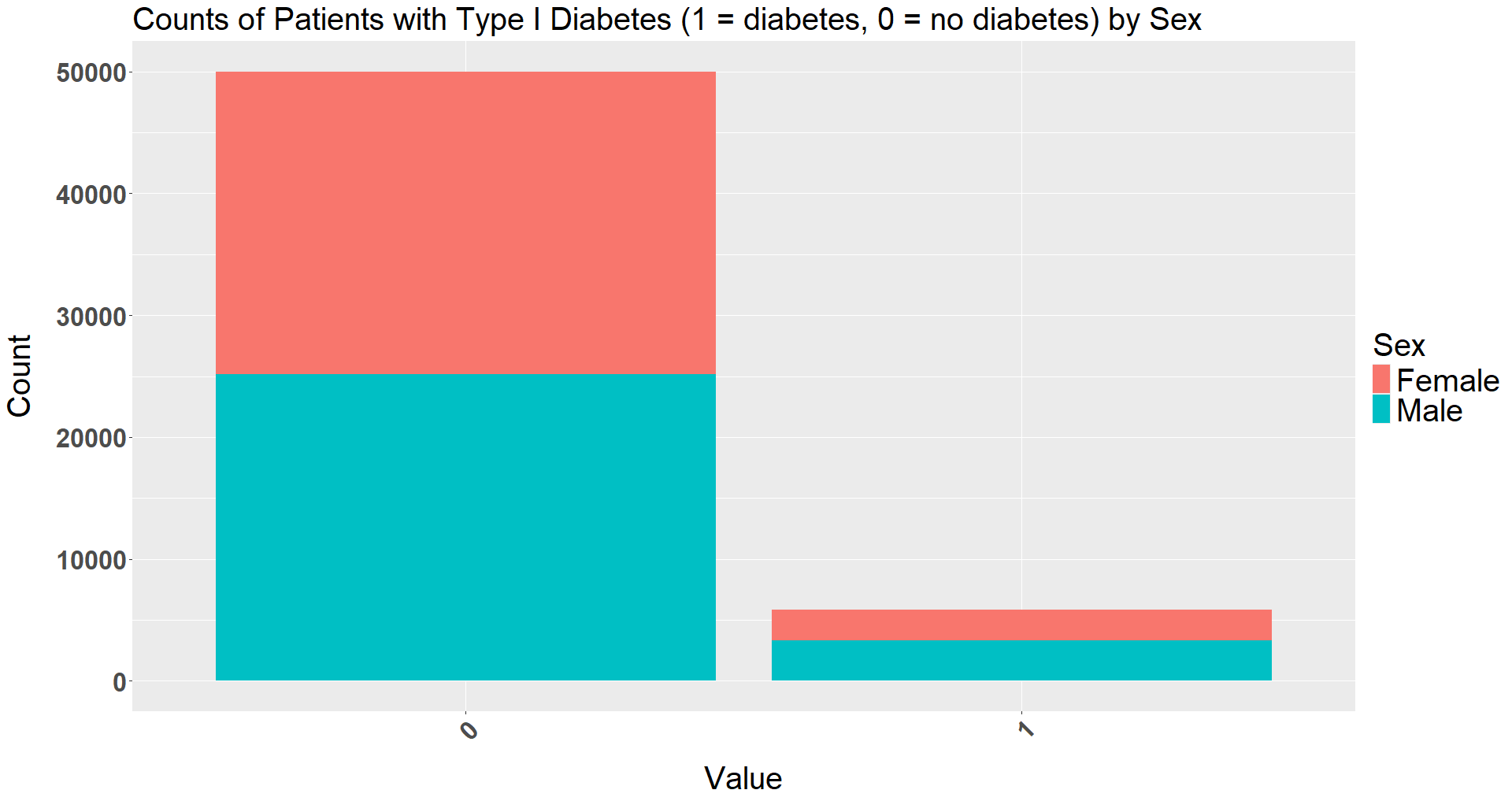
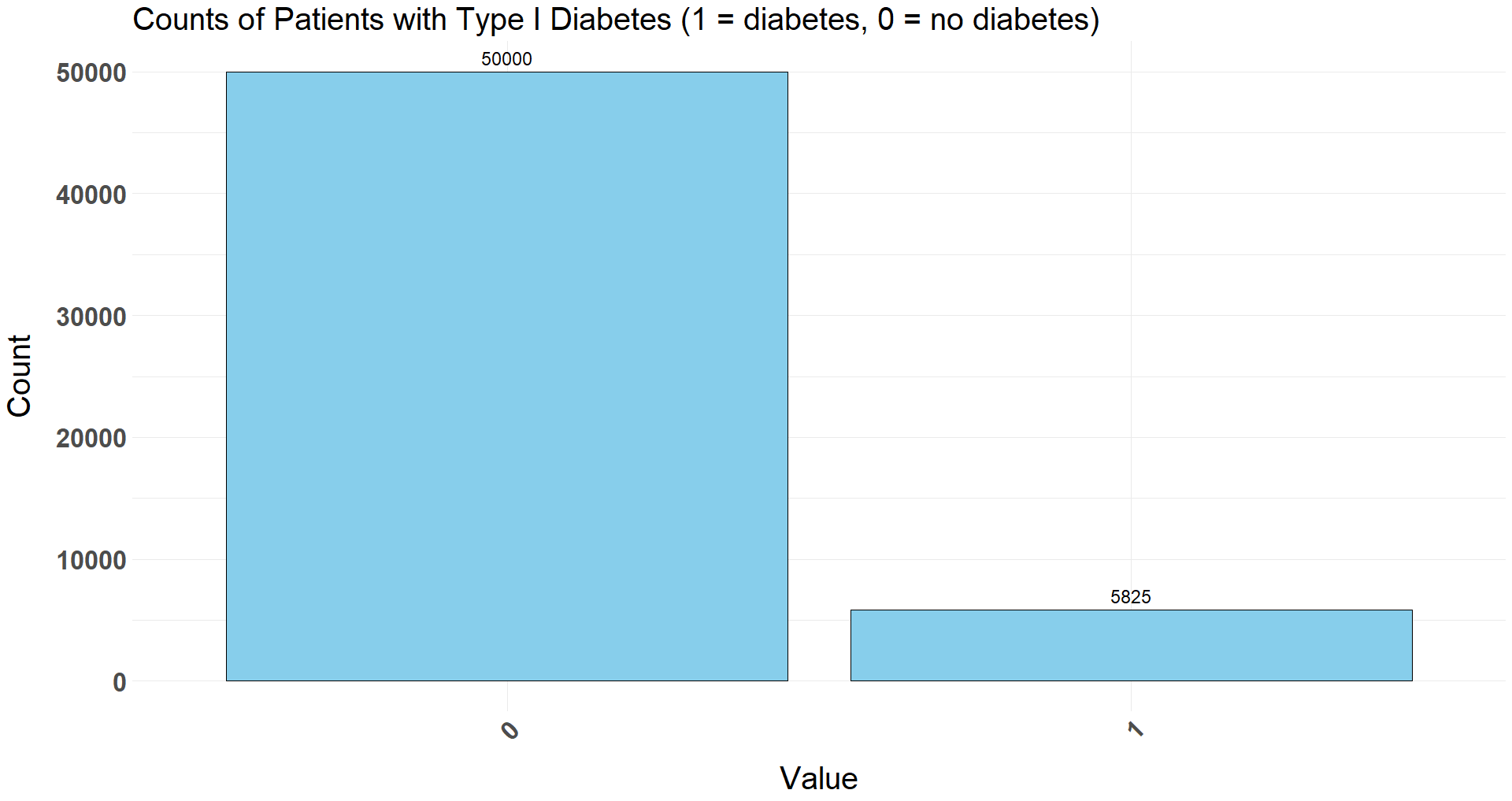


Using this data, we can go on to make bar/column charts and pie charts. These might make the data easier to interpret at a first glance.

**Bar & Column Charts**

Bar and column charts are useful because you can visualise multiple discrete variables together, so in the image on the below left you can see the count of patients with diabetes. You can also see the number of people with and without diabetes split by sex on the below right – this is a stacked column chart. You can split the data by lots of things e.g. ethnicity, those with and without comorbidities etc. So, it can be a very powerful visualisation method.

**Question: Do you know the difference between a column and bar chart?** These images show column charts because the bars are vertical. A bar chart has horizontal lines.

**BETWEEN A BAR CHART AND COLUMN CHART?**

**Continuous data visualisations (slides 13-15)**

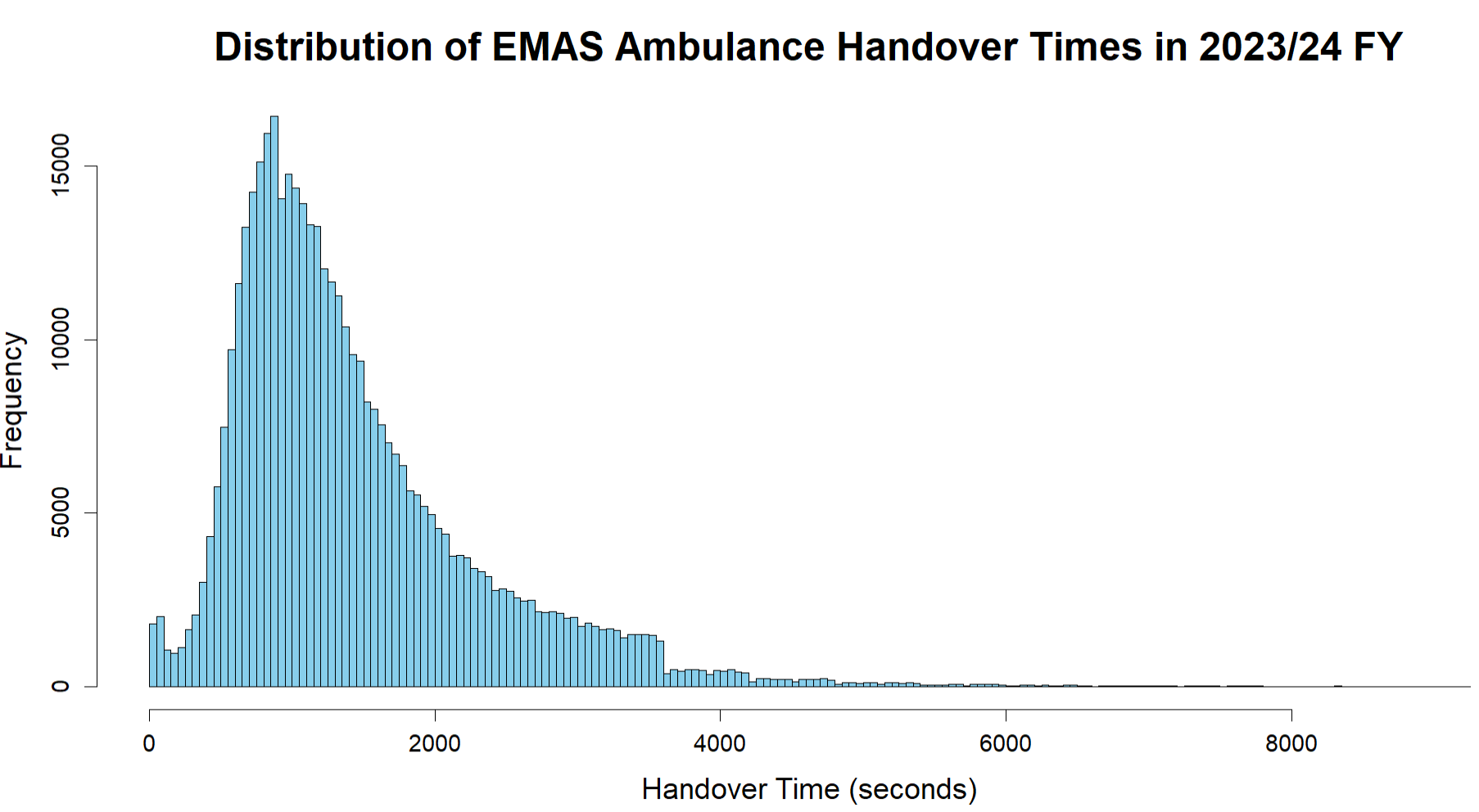
We’ll go over some popular ways to visualise continuous data, mainly scatter graphs and histograms, and we’ll also understand what a histogram shows us and the importance of normally distributed vs skewed distributions.

**Scatter Graphs**

Scatter graphs, or scatter plots, are a good way of visualising 2 continuous variables. In the PowerPoint (slide 13) we have used Ambulance Handover Time against the hour of the day on a specific day. This can help visualise a trend over a period of time.

**Histograms**

You might be thinking, histograms look really similar to column charts, and you would be correct. However, they are actually very different when you get into the details, and as you know from earlier, column charts are suitable for discrete data, whereas histograms are suitable for continuous data. We will go over the key differences between histograms and column charts in the next section, but first let’s understand what a histogram is.

****

This histogram shows the count of ambulances (frequency) that handover a patient within each of the buckets of time represented by the column widths on the x axis.

**Question: What do we mean by data buckets/bins?** Buckets are basically the data sorted into groups of different value ranges, in this case, based on the handover time. This is so that we can visualise a smaller number of data points, making it easier to se the data on the histogram.

But wait. Groups of data??? This might sound suspiciously like discrete data, but it’s not, and here’s why. Remember that with discrete data, we can’t have in-between values. If there are 10 categories of IMD Decile, or we’re counting people in a workforce, there are no values existing in between that data – we can’t divide the data into smaller fractions. But we know that with continuous data, such as ambulance handover time, we can. So, even though we are now grouping the continuous data to make it easier to visualise the data on our histogram, these groups are not distinct categories and the data is still continuous. This concept can be particularly difficult to understand theoretically, but it will become easier to understand as you work with different data types.

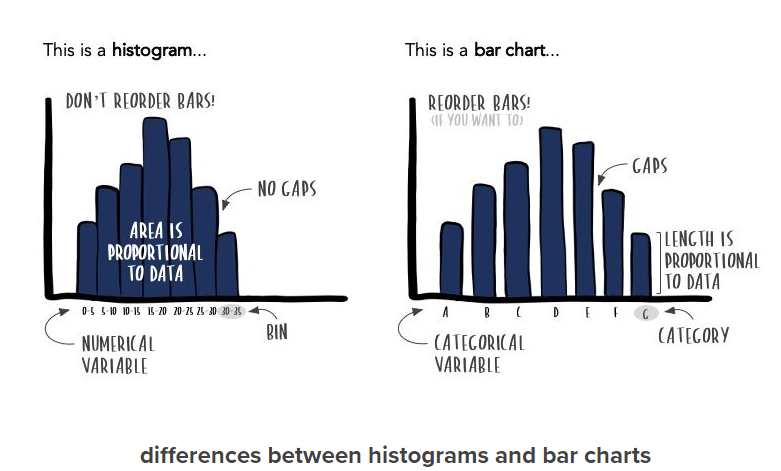
Because the x axis is still continuous, the bars of a histogram are always touching, there’s no gap in between them. We also would never re-order the bars (this is really important for when we look at distribution too).

**Histogram Bin Widths**

Now we understand why bins are necessary, but it’s also important to select the right number of bins. Too many or too few and patterns in the data may be lost. You can set bin widths manually in R and Excel, but you can also use established methods such as Scotts’ Rule, Sturges’ Rule, or Freedman-Diaconis Rule.

**The Difference Between Histograms and Column Charts?**

It can be difficult to understand the difference between the 2 at first glance, as they can look quite similar. But, they do have very different uses and properties as we have discussed. Let’s summarise these.



StorytellingWithData, accessed 03.07.24 https://www.storytellingwithdata.com/blog/2021/1/28/histograms-and-bar-charts

There are 3 main differences between column charts and histograms, as shown in the table below.

|  |  |  |
| --- | --- | --- |
| Difference | Bar Chart | Histogram |
| Data | Discrete (distinct categories/non-divisible data) | Continuous (continuous scale but sorted into bins) |
| Bar Position | Not touching to represent discrete scale | Touching to represent continuous scale |
| Bar Ordering | Reorder if you like. Usually in descending height order | Cannot reorder as x axis is a continuous scale. Not reordering allows us to observe distribution |

**So, We Only Use Histograms for Continuous Data, Right?**

Yes and no. Generally yes, we only use histograms when we have continuous data, and that’s a good rule of thumb to follow when you start out in statistics. However, as you become more experienced, there are situations where data falls into a bit of a grey area between continuous and discrete data, and you may have to use your own discretion. You might also sometimes see that the data leans more towards discrete, but you may still find some use in creating a histogram. For example, the LOS example we talked about earlier.

LOS should be continuous data as it is a measurement of time. However, in my experience, the database that you get LOS from has normally already been rounded into whole days, and you don’t have access to any more specific data. Therefore, this ‘should-be-continuous-data’ has now become discrete. But really, in practice, we want to understand LOS in terms of time, so we might still use a histogram to look at the distribution. Another example might be data where you have age measured in years. These are the cases where data type is not always black and white, and you should use your own judgement to analyse it.

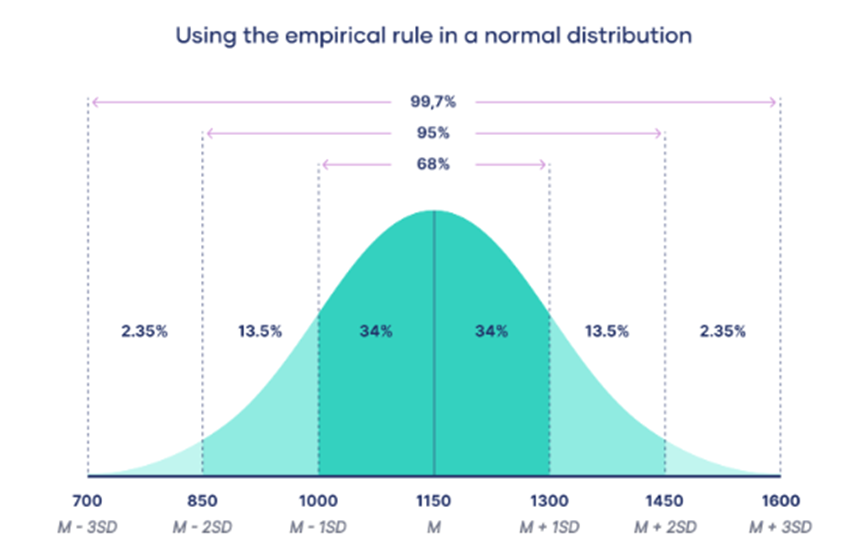
**Histogram distribution (slides 17-20)**

When you have a histogram, the shape of the histogram curve can tell you quite a bit about the data. The distribution can be normal, skewed left, or skewed right. This is the main reason we would want to create a histogram, to see the distribution of the data. Things tend to be a lot easier when data is normally distributed, so we often try and manipulate skewed data to make it behave more like normally distributed data. We’ll go over the types of distribution and why normal distribution is preferred in the following sections.

**Normal Distribution**

The bell-shaped curve below is a normal distribution. So, if you create a histogram with your data and see this shape, congratulations!! This will make use of statistical analyses and tests easier, as you will find in later sessions. If your data doesn’t look normally distributed, don’t worry. There are ways to try and fix this which we will go through in the below sections.

Quite a lot of healthcare data tends to be normally distributed e.g. BMI, blood pressure, weight etc. but you might not always have access to this kind of data, especially if you work in a higher level of the healthcare system (like an ICB).



Scribbr, accessed 01.07.24 https://www.scribbr.co.uk/stats/the-normal-distribution/

The normally distributed bell-shaped curve is symmetrical, with its highest point being in the middle. This middle point will be the mean, median, and mode of the data, which will all be equal. The sides of the curve lower on either side from the middle point. The symmetry on either side of the middle point implies that the probability of an observation falling on either side of the mean/median/mode is equal.

Some other characteristics of a normally distributed curve are described in the table below. For some helpful visualisations and animations, check slide 17 of the PowerPoint.

|  |  |
| --- | --- |
| Characteristic | Explanation |
| Extends indefinitely in both directions along the x axis | You can use the normal distribution to calculate the probability of observing any number from negative to positive infinity. The normal distribution will get very close to the x ais as you extend closer to positive or negative infinity, but it will never actually touch it |
| The empirical rule (see image above), also known as the 68-95-99.7 rule, tells you where values lie in a normal distribution | Around 68% of values are within 1 standard deviation from the mean, round 95% of values are within 2 standard deviations from the mean, and around 99.7% of values are within 3 standard deviations from the mean |
| Symmetry of the curve | The highest point of the curve is in the middle, and this is the mean , median, and mode which are all equal. The sides of the curve slope down from the middle, and the probability of an observation falling on either side are also equal |

**Why Is Normal Distribution Important?**

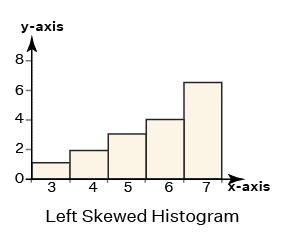
Despite what you might personally experience in the data sets you use at work, a lot of data in life is normally distributed. And these don’t just include healthcare data, they also include exam scores, job satisfaction, reading ability etc.

Because we commonly find normally distributed data, a lot of statistical tests like ANOVA and t-test are modelled around this and assume your data is normally distributed. Therefore, it can make drawing meaningful conclusions from these tests much easier if you have normally distributed data.

Analysis is also often carried out on a sample of a population, rather than the entire population, usually because you either don’t have access to all of the data or it’s too expensive to collect it all. If the data is normally distributed, you can use inferential statistics to make educated guesses about the larger population. This doesn’t just require a normal distribution however, you also have to make sure the sample you have is genuinely representative of the entire population.

**Left-Skewed Distribution**

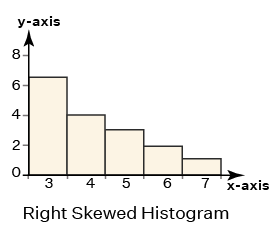
When data is not normally distributed, we can say it is skewed. Left-skewed data will look something like the image below, where the tail is closer to the left side of the x axis. Left-skewed data is also known as negatively skewed as the tail extends closer to the negative side of the axis.



cuemath, accessed 01.07.24 https://www.cuemath.com/data/right-skewed-histogram/

**Right-Skewed Distribution**

Right-skewed data will look something like the image below, where the tail is closer to the right side of the x axis. Right-skewed data is also known as positively skewed as the tail extends closer to the positive side of the axis.

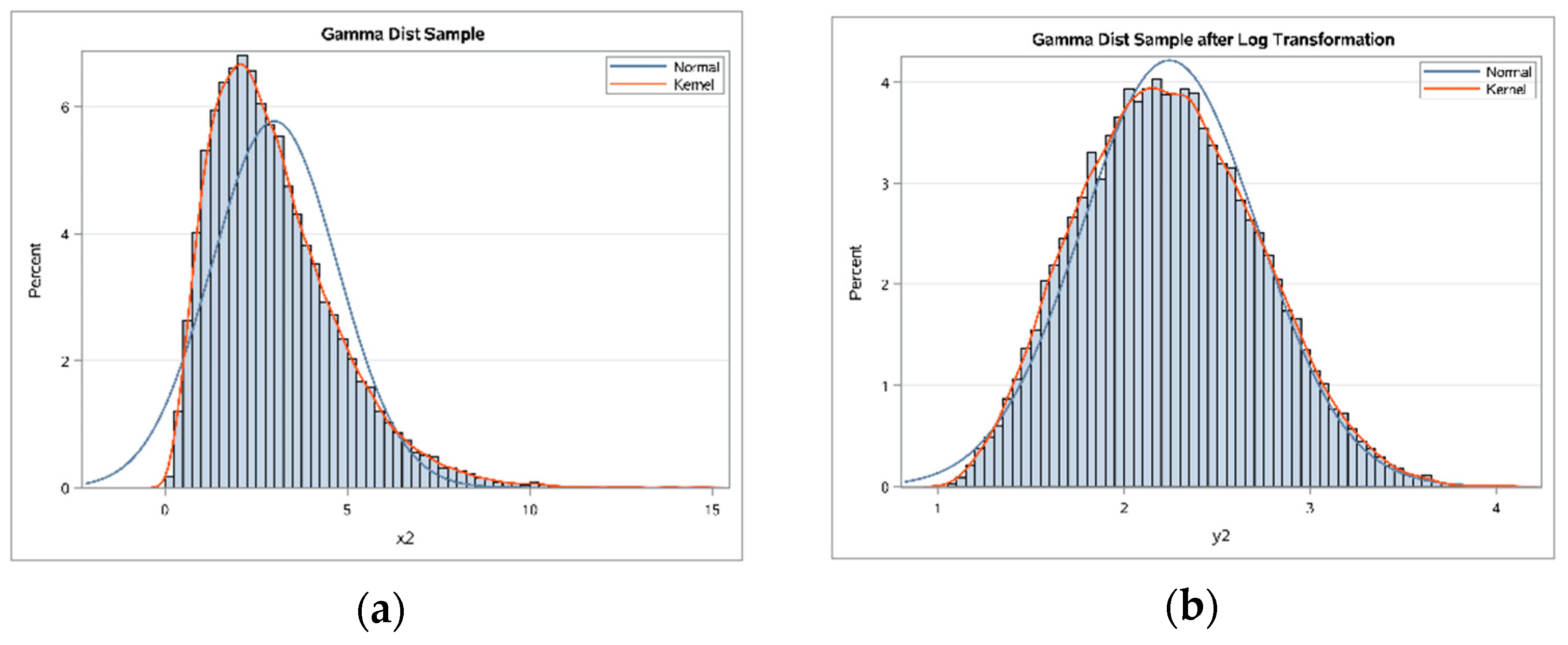


cuemath, accessed 01.07.24 https://www.cuemath.com/data/right-skewed-histogram/

**Log Transformation**

So what can we do if our data is not normally distributed, but we want to use tests like ANOVA? We can try to transform the data and see if this makes it normally distributed. One transformation method is log transformation. This involves taking the log of the value rather than the raw values themselves.

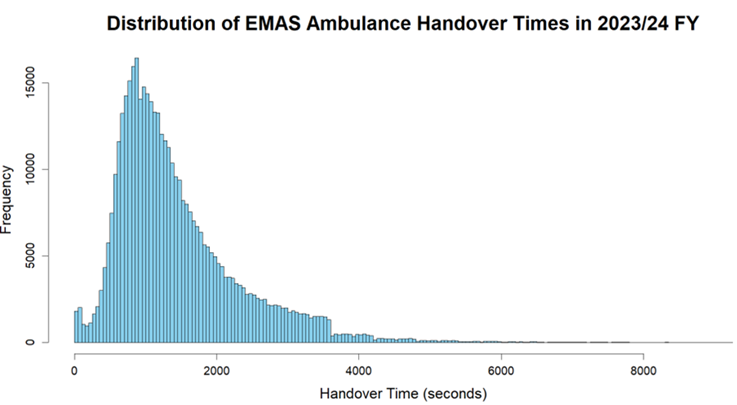
This is quite easy to do in both R and Excel (please refer to the practical session on slide 27 of the PowerPoint for more details), and you’ll want to ensure you use a log base 10. You can sometimes change this depending on your requirements e.g. to log base 2, but base 10 is the generic natural logarithm. This can help transform your data as shown below.

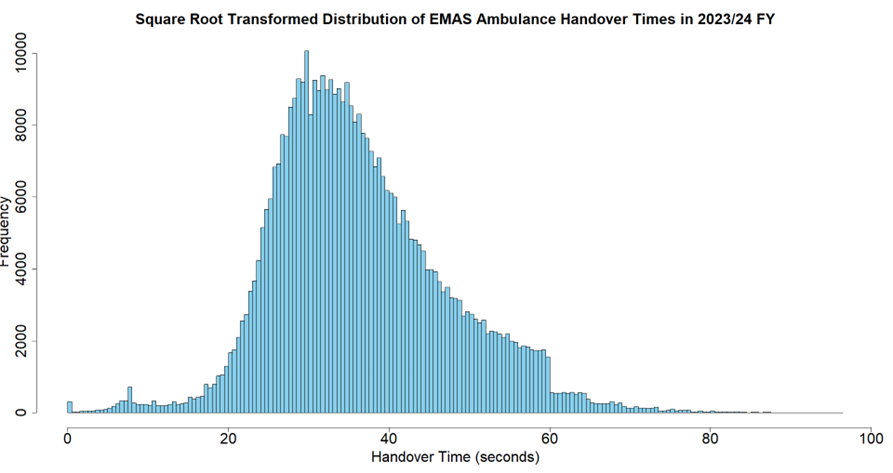


Handling Skewed Data: A Comparison of Two Popular Methods accessed 01.07.24 https://www.mdpi.com/2076-3417/10/18/6247

**Square Root Transformation**

You can also take the square root of your values rather than using the raw values. Have a look at the practical on slide 29 of the PowerPoint for instructions on how to do this in R and Excel. Here’s an example of when taking the square root can help transform skewed data to be normally distributed:





These methods don’t always transform your data as you would like, however (have a look at slide 28 of the PowerPoint). It’s a good idea to try both transformations as a starting point and see if either help make your data more normally distributed. Don’t be disheartened if they don’t, you can have a look at some other transformation methods, or you might just need to use statistical tests and analysis methods that don’t rely on normal distribution. Also, don’t forget to transform your data back if you do go ahead with a transformation method (e.g. when giving your audience figures).

**Summary measures (slides 21-22)**

After you have understood whether your data is discrete or continuous, you can use summary measures to further understand your data. Summary measures are especially useful when we have a large dataset, as we can summarise the data in one value that is typically representative of the data set – if you use the right measures for your data. We will go through which measures to use for discrete or continuous data below. Summary measures can generally be split into measures of central tendency and measures of data spread.

**Measures of Central Tendency**

Measures of central tendency help us understand the location of the data. They include things like mean, median, and mode.

The mean can help indicate where the centre of a data set is and is a commonly used measure of the average. It is calculated by summing all the values and then dividing this by the total count of values. This can be useful for looking at average ambulance handover times in a particular hour, for example. However, the mean can be greatly skewed by outliers – imagine there were just 1 or 2 particularly bad handovers that took 4 hours. This would make the average look a lot higher than we might want considering the majority of the data.

The median shows you the centre of the data set when the values are arranged in ascending order. This can be particularly useful when you don’t have normally distributed data, such as with LOS. LOS can fluctuate greatly due to the variability in health conditions and reasons for requiring an inpatient stay. Finding the median, rather than the mean, can show you the centre of the data, without being skewed by outliers as much, therefore representing a more typical value for that data.

The mode is the most frequent value. This is probably the least used summary measure. It might be more useful in fields such as epidemiology to detect the most common virus during an outbreak, for example. You probably won’t need to use this measure in your work.

These measures can be used for various continuous and discrete data – see the table below for a complete breakdown.

|  |  |
| --- | --- |
| Measure of Central Tendency | Suitable Data Types |
| Mean | * Normally distributed continuous * Numerical discrete that does not have very high or very low outliers * E.g. Birthweight, square-root transformed ambulance handover times |
| Median | * Skewed continuous * Discrete numerical with outliers in data * E.g. LOS |
| Mode | * Not used very often * Useful in epidemiology e.g. to find the most common virus during an outbreak |

**Measures of Data Spread**

Measures of data spread describe how similar or varied the data is. The most common measures of data spread include the range, interquartile range, and standard deviation.

The range is the highest value minus the smallest value. This helps us understand the variability in the data. However, the range can also be heavily skewed if the data has even 1 very high or very low outlier.

Interquartile range is the Q3 of the data minus the Q1 of the data (look at slide 22 for a helpful animation whilst reading the following text). Quartiles are calculated by arranging the values in ascending numerical order and dividing the data into 4 equal parts. Quartiles will just be a single value, not a range – you can think of the quartile value as the cut off point for that quarter of the data.

After ordering the values, you find the median and this is the Q2 value. Then find the median of each half of the data on either side of the median you just calculated. The lower median is Q1 and the upper median is Q3, and you can just subtract these.

Box plots (again, refer to slide 22 for a helpful animation) can be really useful to visualise the quartiles and interquartile range. There is an additional section below to explain the different features of box plots.

Interquartile range describes that middle 50% of values when they are ordered highest to lowest. It can work better for data with outliers as it is not affected by these.

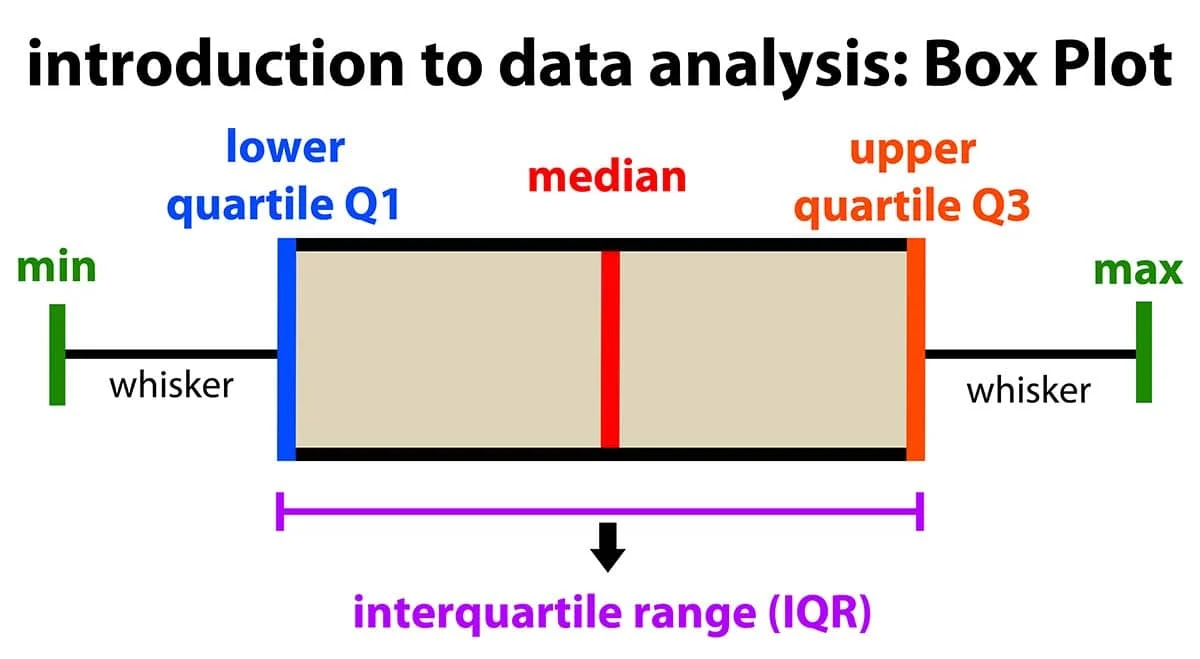
Standard deviation (SD) measures how spread out the data is, relative to the mean. It’s calculated by taking the square root of the variance. It can show you how volatile the data is i.e is the data fairly consistent with the mean or does the data vary from the mean a lot. A higher SD indicates more volatility. You can use it to compare the consistency between 2 different data sets. 2 data sets could have the same mean, median, and range, but actually be quite different. SD would help indicate this as one would likely have a higher SD.

These measures can also be used for various continuous and discrete data – another table has been provided below to demonstrate this.

|  |  |
| --- | --- |
| Measure of Central Tendency | Suitable Data Types |
| Range | * Normally distributed continuous * Numerical discrete without outliers |
| Interquartile Range | * Skewed continuous * Numerical discrete with outliers |
| Standard Deviation | * Normally distributed continuous * Compare between data sets – 2 sets of data could have same mean, median, range, but have very different data. SD would help highlight this |

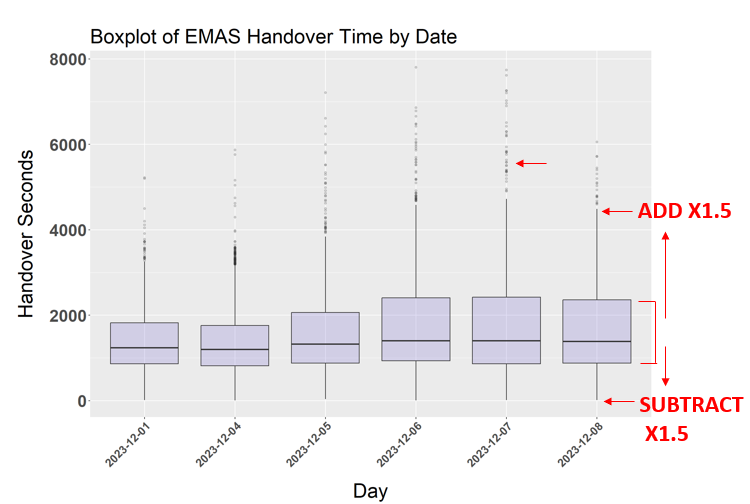
**Box Plots**

Box plots can be a bit complicated to look at initially, but they’re really useful and quite easy to break down once you understand the different components. Let’s take a look.



SimplyPsychology, accessed 01.07.24 https://www.simplypsychology.org/boxplots.html

The whiskers show the range of the data from the minimum point to Q1 and from Q3 to the maximum point. It’s worth noting that the whiskers generated during the practical session (again, take a look at slide 24 for a helpful animation to supplement this text) might look a bit different than you were expecting due to the method used to create the box plots – the Tukey method. The Tukey method multiplies the IQR width by 1.5 and adds this to the top of the box plot to get the upper limit of the whisker, and it subtracts it from the bottom of the IQR to get the lower whisker, and then the dots are considered outliers (see image below).

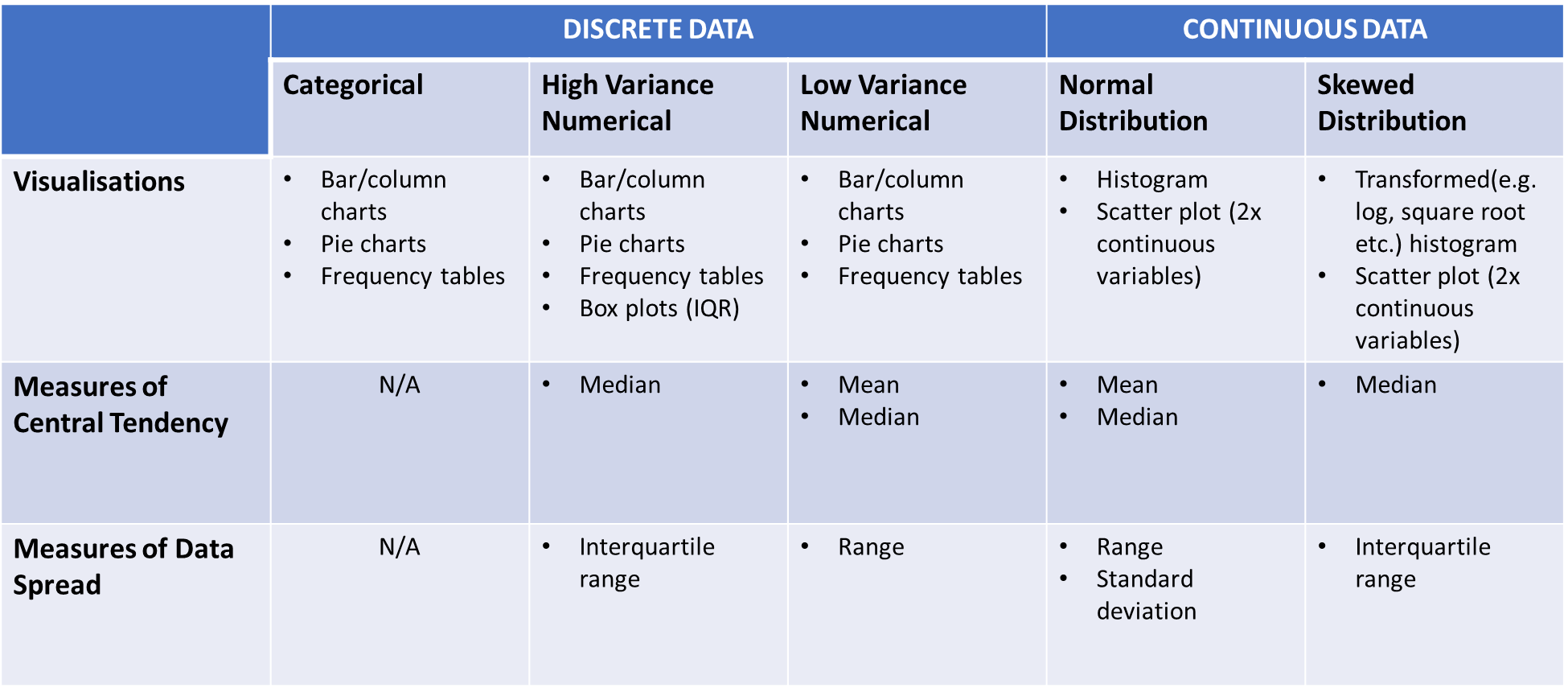


This is why all of our minimum values look like they’re 0, even though we don’t actually have any 0 values in our data. By subtracting the width multiplied by 1.5, we would actually go into the negative axis. This is likely to be again because the handover data is right skewed because the target is 15 minutes, so there are a lot more lower values than higher values. This is a fairly standard way to calculate the whiskers, so it’s not something to worry about, but you should have an understanding of your data and be aware of these discrepancies, especially if you want to give a user accurate figures.

Going back to the first image, we also have the Q1, Q2 (median), and Q3 values. Refer to the section above (Measures of Data Spread) for further information on how these quartiles are calculated. We can also visualise the middle 50% of the data, which is the interquartile range, and then the whiskers also illustrate data that falls outside of this middle 50%. This means that the whisker on the left/bottom shows you the lowest 25% of scores and the right/upper whisker shows you the highest 25% of scores.

So, you can see that box plots can tell you a lot about your data in quite a compact visual, so they can be a really useful tool.

**Summary (slide 26)**

This is a lot of information to take in, so don’t worry, you are not expected to remember all of this! This training was designed to provide a detailed and (hopefully) easy to understand reference guide to basic statistics. You can always refer back to these materials at a later date, and you will find the concepts and ideas presented here come more quickly to you as you practice statistics more and more in your roles. It will also help provide a solid background understanding before you undertake any of the more advanced statistics training sessions. We have provided a table below that tries to neatly summarise the learning you have undertaken so far.

**Contact & Resources (slide 31)**

We really hope you enjoyed this training and found it useful. I would welcome any feedback (especially constructive) regarding how this session could be improved in the future. Please also feel free to reach out with any questions. You can contact me with feedback and/or questions on [**amy.makawana1@nhs.net**](mailto:amy.makawana1@nhs.net)

Here’s a few resources you also might find useful for your statistics journey:

**Discrete Data:**

* https://thirdspacelearning.com/us/math-resources/topic-guides/statistics-and-probability/discrete-data/
* <https://www.indeed.com/career-advice/career-development/what-is-discrete-data#:~:text=Discrete%20data%20is%20a%20type,of%20customers%20in%20a%20store>.
* <https://www.open.edu/openlearn/mod/oucontent/view.php?id=85587&section=1>
* <https://www.bbc.co.uk/bitesize/guides/zyqc9qt/revision/4>
* https://thirdspacelearning.com/us/math-resources/topic-guides/statistics-and-probability/quantitative-data/

**Continuous Data:**

* <https://www.indeed.com/career-advice/career-development/what-is-discrete-data#:~:text=Discrete%20data%20is%20a%20type,of%20customers%20in%20a%20store>.
* <https://www.open.edu/openlearn/mod/oucontent/view.php?id=85587&section=1>
* <https://www.bbc.co.uk/bitesize/guides/zyqc9qt/revision/4>
* https://thirdspacelearning.com/us/math-resources/topic-guides/statistics-and-probability/quantitative-data/

**Frequency Tables:**

* <https://www.geeksforgeeks.org/frequency-table-in-r/> **(R)**
* <https://www.indeed.com/career-advice/career-development/excel-frequency-distribution-table> **(Excel)**

**Column/Bar Charts:**

* https://r-graph-gallery.com/218-basic-barplots-with-ggplot2.html **(R)**
* <https://support.microsoft.com/en-gb/office/create-a-bar-chart-14832c6e-0a66-458d-82e2-7fd3bce4d05a#:~:text=In%20the%20ribbon%2C%20select%20Create,the%20chart%20for%20better%20readability>. **(Excel)**

**Scatter Graphs:**

* <https://www.datacamp.com/tutorial/scatterplot-in-r> **(R)**
* https://r-graph-gallery.com/scatterplot.html **(R)**
* <https://support.microsoft.com/en-gb/office/present-your-data-in-a-scatter-chart-or-a-line-chart-4570a80f-599a-4d6b-a155-104a9018b86e> **(Excel)**

**Histograms:**

* <https://www.datacamp.com/tutorial/make-histogram-basic-r> **(R)**
* https://r-graph-gallery.com/histogram.html **(R)**
* [Histogram in Excel (In Easy Steps) (excel-easy.com)](https://www.excel-easy.com/examples/histogram.html) **(Excel)**
* <https://support.microsoft.com/en-gb/office/create-a-histogram-85680173-064b-4024-b39d-80f17ff2f4e8> **(Excel)**
* https://careerfoundry.com/en/blog/data-analytics/how-to-create-a-histogram-in-excel/ **(Excel)**

**Distribution:**

* <https://www.atlassian.com/data/charts/histogram-complete-guide#:~:text=A%20histogram%20is%20a%20chart,value%20within%20the%20corresponding%20bin>.
* https://asq.org/quality-resources/histogram

**Log Transformation:**

* <https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/log> **(R)**
* <https://minhngocda.medium.com/logarithmic-scale-in-excel-237db7563322> **(Excel)**

**Measures of Central Tendency:**

* https://www.scribbr.co.uk/stats/measures-of-central-tendency/#:~:text=a%20data%20set.-,The%203%20most%20common%20measures%20of%20central%20tendency%20are%20the,the%20total%20number%20of%20values.

**Measures of Data Spread:**

* https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/measures-spread#:~:text=Measures%20of%20spread%20describe%20how,range%2C%20variance%20and%20standard%20deviation.

**Box Plots:**

* <https://www.atlassian.com/data/charts/box-plot-complete-guide>
* <https://r-graph-gallery.com/boxplot.html> **(R)**
* <https://www.indeed.com/career-advice/career-development/how-to-make-box-plot-in-excel> **(Excel)**