This chapter will discuss basic univariate analysis and summary statistics from the survey results, alongside what could be inferred from these. We will look at each section individually and perform multiple initial comparisons whereby we subset for various factors, such as the language used to make the plots, and the order in which plots have been presented.

First we will look at the summary statistics for each of the survey questions for the whole population, alongside box plots and density plots to gain an overview of the shape of the data and spread of values. Each table of summary statistics presents columns for the three plot types; the control plot, the truncated plot, and the logarithmic plot, respectively in that order. Note that sample size, n, and summary statistics are given after removing NA values. In the chapter 3 we will delve deeper into the data, sub-setting for various groups and performing comparisons between these groups to ascertain whether different demographic or population factors have an effect on responses.

Ninja Warrior - Part 1

The first part of the survey consisted of showing the respondents three bar plots representing data regarding how many times four obstacles were used throughout 10 seasons of American Ninja Warrior. The three presented visualisations all showed the same raw data, but used three different y-axis scalings in order to assess whether changing this scale in these ways affects viewer interpretation. The questions asked were designed to test the effect of scale on both reading off exact values and gauging differences in values. Each respondent was asked four questions; two free form answer and two multiple choice. The use of free form answers did result in occasional non-valid answers, such as statements along the lines of "Don't know" when a number was required. Many people also opted to write a number between 0 and 1 when a percentage was required, but these will not be considered invalid, as there were a large number of responses of this type, but rather we will assume that any number between 0 and 1 is considered as the corresponding percentage. Ie. an answer of 0.5 will be considered as 50%.

Approximately many times would you say the 'Salmon Ladder' was used?

This question, the first of the survey, asked participants to type the how many times Salmon Ladder was used, based on the bar plot. The 'correct' answer, or rather the true height of the corresponding bar, was 42. There were three invalid answers in these responses; one for the R versions of the survey and two for the Python versions. The invalid response in the R survey was '41/42', which we will take to be 41.5, and the invalid Python responses were given as 'Don't know' and 'Next to none.'. These two will be considered as 'NA' responses and thus discounted from the analysis of this question. These responses will. however, still be useful in our investigation; both were entered for the logarithmically scaled plot made in Python. The default log scaling in Python uses standard form notation, which perhaps these two participants were less familiar with. Similarly, there were two answers of '10^15' and '10^9', again potentially pointing towards the respondents being less familiar with this notation.

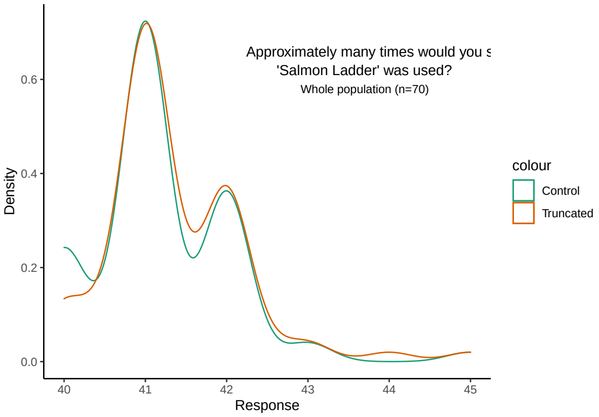
Below we see the summary statistics for the overall population, where we have taken the '41/42' response as 41.5 and omitted the two invalid text answers as well as an additional NA response and thus obtain a sample size of 67.

The table below presents the total population summary statistics for the for the first question.

We can see that for the control and truncated plots we have means 41.21 and 41.35 respectively and both have median 41, which at first glance do not appear significantly different from the true value of 42. To investigate this further we will run some statistical tests. These two sets also both have a range of 5, with minimums of 40 and maximums of 45, giving a fairly compact spread of data around the medium, and meaning that all respondents were in roughly the correct area. We also note that the means and medians of this data lie just below the true value, from which could be inferred that respondents tended to under-estimate. Although, the interquartile range does contain 42, so this may be a small underestimation. These two sets of responses also have very similar variances, at 0.742 and 0.753 respectively.

The mean of the responses for the log plot, on the other hand, is much higher at 1.493e+13 with the median much smaller in magnitude than this. In fact, the median is lower than for the control and truncated plots despite the mean being higher. We also have a very big range of $[9, 1 \times 10^{15}]$, meaning some respondents inferred very different values from the plot than others. We see, however, that the interquartile range is much smaller, given as $[30, 40]$. This is still larger than the range for the control and truncated plot, but shows that the middle 50% of the data lies in this range. We also note that the 42 does not lie in this range but rather above it, perhaps once signifying a more significant respondent underestimation, despite the values at the upper end of the range being many orders of magnitude higher than the true value. In statistical terms, these will be considered outliers, but will still provide important insight with regard to the reasoning behind these outliers occurring. The large values at the extreme ends of the distribution also means the variance is very large, at 1.49×10^{28}.

To decide if t-tests are applicable here we will first look at the distribution of these variables before running Shapiro-Wilk tests of normality. Below we can see a density plot depicting the distributions of the values for the control and truncated plots.



The two distributions are very similarly shaped, and neither appears similar to a Gaussian curve, hence it is likely that they will violate the normality condition of a t-test. To confirm this hypothesis, Shapiro-Wilk tests for normality are performed.

For both the control plot and truncated plot responses, the Shapiro-Wilk tests give $p << 0.05$, and thus we reject the hypothesis that these data are normal, and so they do, in fact, violate the normality condition required for a one-sample t-test.

One alternative to using a t-test is to use a Wilcoxon-Mann-Whitney (WMW) test. Note however that this test requires a symmetric distribution with even spread of values about the median. We see below, and from the density plot, that this isn't the case.

We now move on to consider the one-sample sign test. This has a significantly lower power than the t-test and WMW test, but is required as the data violates the conditions for these two tests.

We will first look at the two sided tests to decipher if the sample medians differ significantly from the true value of 42.

The p values both being << 0.05 signifies that the medians do in fact differ from the true value of 42. In fact we see that, for the control plot, the 95 confidence interval has both a lower and upper bound of 41, meaning we have a 95 chance that the true median of this population is 41. Similarly, the 95 confidence interval for the truncated plot is given as $[41, 41.25]$, meaning once again we have a 95 chance of the true median lying in this range. Thus we can infer that participants tended to slightly underestimate the height of this bar, no matter whether the axis was truncated or not. This makes sense as the top of the bar will appear to be at the same point on both scales. The underestimation may be as a result of the default scales going up to only 40; the respondents may have seen the bar being slightly above the 40 mark, and thus taken this as 41 rather than the true value of 42. Here this doesn't have a significant impact, but say the scale was in £ billions for example, then an underestimation of one is significant. It may then be important to specify the exact numerical values of the bars alongside the bars themselves in order for the interpretation to be accurate. This again could be used to either deliberately or accidentally mislead consumers, such as into believing a rival company may be doing worse than they actually are.

To confirm that it is in fact underestimation we are dealing with, we perform one-sided sign tests.

From this series of tests we see p-values of << 0.05 when considering the median as less than 42, and p-values of 1 when considering it as greater than 42. Thus, we can deduce that respondents tended to underestimate by a small but statistically significant margin.

We could also consider taking a set of samples from a normal distribution of mean and variance the same as our set of responses, and performing a t-test on the set of means of these samples. Below is the result of taking the means of 100 samples, each of size 100, and performing two sided t-tests on the set of means.

This follows our previous testing in the conclusion that we differ from the true value of 42, and once again we can observe in the 95 CI as well as the very negative t statistics that there was a tendency to underestimate the height of the bar.

Now, we haven't yet considered the plot with logarithmic scaling. Looking back at the summary statistics we see wildly different values to the responses in relation to other two plots. As previously mentioned there are answers of '10ˆ9' and '10ˆ15'. Even discounting these two values, as seen below by setting these values to NA, there is a wide range in the responses, from 9 all the way up to 1000. Again, in some situations the responses that are orders of magnitude greater than the others could be considered outliers, but here they provide important insight. See below the summary statistics for the data after removing the two high magnitude values.

It seems that the mean is still much higher than the true value, at 67.07, and the IQR has stayed the same. Note that there is still a value of 1000 included in the data, which will be causing a skew to the right, and this again means there is a fairly large variance, this time of 28809.56.

The below density plot shows the two curve from before, but with the curve of the logarithmic curve added.

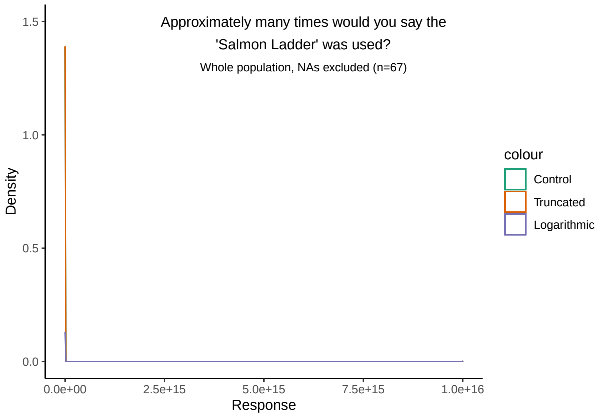
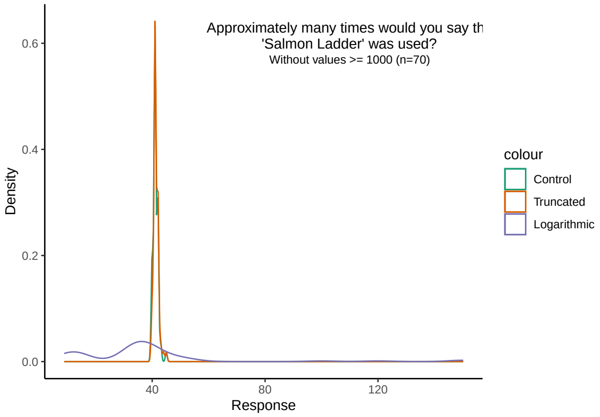


Figure 1: Density plot showing distributions of responses regarding all three plots

This plot does not give us much information about the distribution of the responses for the logarithmically scaled plot due to the values with high magnitude causing the x-axis to extend beyond a point for which we can discern any meaningful distributions. Consider removing values ≥ 1000 to look at the distribution of the lower values. Removing the four values ≥ 1000 we obtain the following density plot;



This gives a better idea of the distribution. we see that almost all of the density of the control and truncated plot is around the 40 mark, whereas the logarithmic has a much greater spread with much lower densities. This, again, could reflect confusion or less familiarity regarding this scaling. This should be considered when designing visualisations; the creator of the visualisations may find the logarithmic scale more effective in showing the data, but they should consider the target audience. Are the audience going to be familiar with this? If, for example, visualisations are being published in a paper targeted at academics in a subject likely to use such scalings often and understand them, this may be a good way to depict the data. However, using this in something such as an advertising campaign could mislead the public, causing them to either over or under estimate values. As previously discussed, however, this is often done deliberately in order to push the message the creator wishes to sell.

While it at first glance appears that the logarithmic scale alone causes these drastic range of values, we must also consider the notation used. As mentioned prior, the default notation in Python for a logarithmic scale is standard form; in our case we have three tick values on the y-axis of 0, 10^0 and 10^1. To explore this, we can split the surveys by language to obtain two sets of data; one for R and one for Python. Now on closer inspection of the separate languages (below), the large range is almost fully attributed fully to the python versions of the plot.

Consider first the summary statistics for the R version.

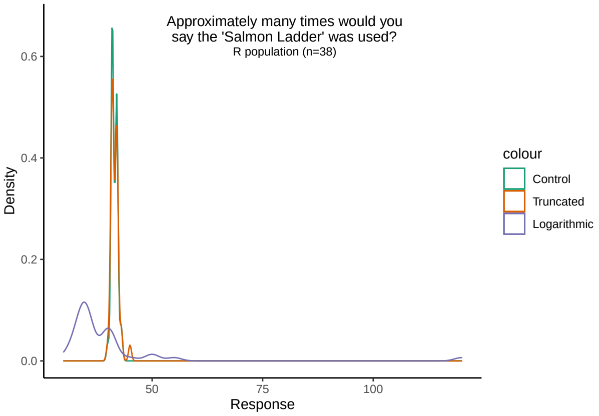
Summary statistics of the R versions

We see that, based on the median and mean, the responses for log plot were on average lower than that of the control or truncated plots, conversely to the whole population in which the mean is magnitudes greater, although the median here is the same as for the whole population, with a slightly smaller interquartile range of size 5. We again see that the upper limit of the interquartile range is 40; below the true value. The whole range here is from 30 to 120, so does still show that there may be some large discrepancies between responses, however much less than the whole population.

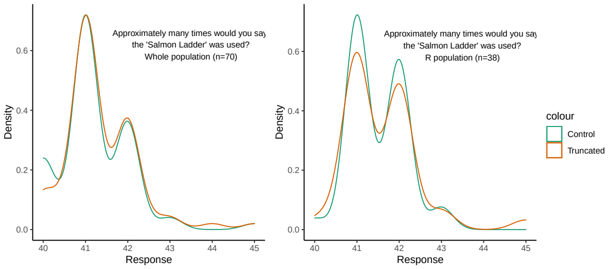
Looking at the control plot now, the values seem very similar to those for the whole population. With a similar mean and identical median. The whole range is slightly smaller, but the interquartile range is the same as before, signifying that the spread in the centre of the data is the same, with slightly less spread towards the upper tail. This could potentially show a slightly better gauging of the value for the R plot than the Python, but we can also calculate that there is only one response higher than 43, which is the 45 at the upper end of the range for the Python population. We note that there is also only one value of 43; all other values are in the range $[40, 42]$. The values of 43 and 45 could possibly be considered as outliers, and then we see minimal differences in the distributions of the data for each of the languages.

Similarly to the control plot, we see that the distribution for the Python version of the truncated plot has an almost identical distribution to the total population or R population.

Below see the density plot for the R data.

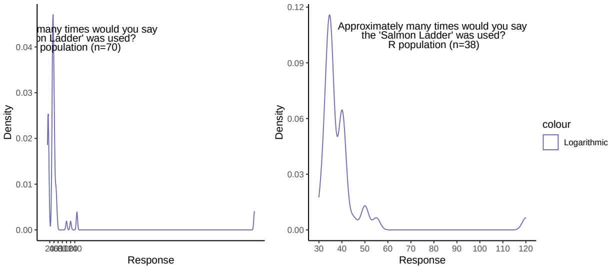


As expected, the distributions of the control and truncated plots so far look similar to the plots for the whole population. To get a clearer picture of the distributions we also plot these side-by-side without the log data.



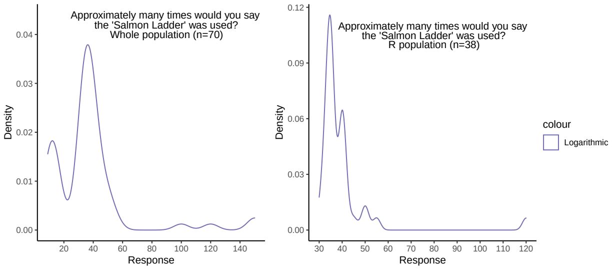
We can in fact see that the distributions of each population are fairly similarly shaped. However, the two curves on the R density plot appear to differ from each other slightly more than on the density plot for the whole population, and the second peak is higher, with greater density at 42 than the whole population, signifying a higher proportion of the values lying around this point than in the total population. we also see that the peak for the control plot are slightly higher than the truncated plot, signifying that there are perhaps more responses around the 41/42 mark than for the truncated plot, which makes sense as we also see the right tail of the truncated plot curve lifts slightly as a result of the 43 value. Overall it doesn't look like the responses for the control and truncated plots for the R population differ much from the overall population.

Now consider the two density curves for the responses relating to the logarithmically scaled plot.



Once again, we see the plot for the whole population, which we discussed before as showing little information about the lower end of the distribution. The density plot for the R population shows the spread up to 120. It appears that the majority of the density lies over the range $[30, 50]$, with a peak around the 35 mark. This signifies the underestimation of values using the logarithmic plot as compared to the other two. For the R version, any differences can be fully attributed to the scale itself, unlike for the Python plot whereby familiarity with standard form notation has an impact. It was hypothesised prior to performing the survey that there may be underestimations as looking at the value of 30 on the axis and comparing it with the top of the bar, it appears the bar falls just above 30.

Now look at the same density plots when removing values ≥ 1000.



We see for the whole population a peak at 40 with drops to a tail at around 60, with a very small density after this point. This follows the R population, where the majority of density is below 60.

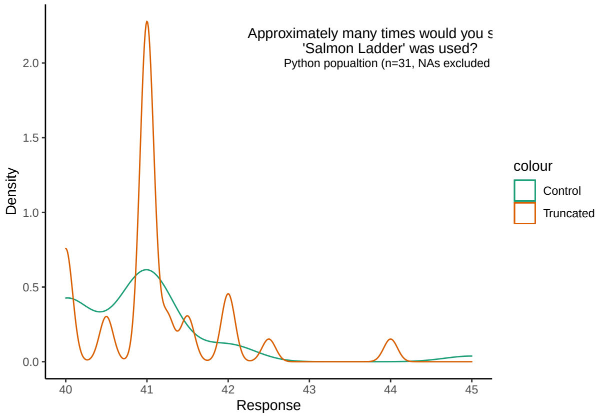
Now we consider the same for the responses corresponding to the Python visualisations.

Summary statistics of the Python versions

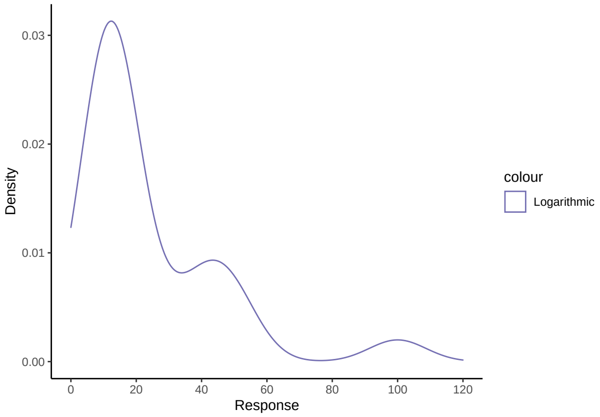
The statistics for the control and truncated plots again appear to centre around 41, based on the median and mean values as well as the IQR. The range for the truncated plot responses is slightly smaller than that of the overall population, at size 4 rather than 5. These results are consistent with the statistics for the prior analysed populations, whereby there appears to be a slight underestimation in gauging the height of the bar. This perhaps depicts that the plotting package defaults don't significantly affect interpretation when considering these two scales.

As expected, the statistics for responses corresponding to the logarithmic plot are much different than those of the R population. Looking at the data set as a whole, we can again see that the values contributing heavily to the large mean were the two given as 10ˆ15 and 10ˆ9, alongside two responses of 1000. This lends to the idea that using matplotlib's default of standard form notation for log scales may have misled some participants who perhaps are less familiar with standard form. Adding to this conclusion are the invalid text responses, “Don’t know” and “Next to None.”, which show that there was confusion regarding the height of the bar.

The density plot below for the control and truncated responses looks very different to those for the whole population and R population. We see significant differences between the control and truncated density curves, as well as between this plot and the previously discussed density plots. the control plot curve appears smoother than that of the whole and R populations, whereas the curve for the truncated plot responses is much less smooth with much more variability, despite the summary statistics seeming relatively similar for both. Both curves once again reach a maximum density at 41 with a large amount of the density falling below 43.



Once again, we now consider the density plot of the responses to the logarithmic plot, where we have once again excluded the values ≥ 1000.

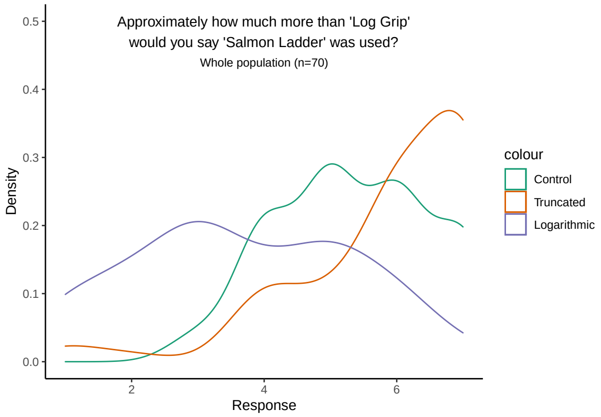


There appears to be a high density peak around 20, which then gradually decreases as we tend towards 120. Although, this peak is large only relative the rest of this plot as opposed to the density curves of the control and truncated curves, where we have maximum densities of around 2.5 and 0.6 respectively. Overall, it appears that the use of the truncated scale had little impact on judging the height of an individual bar as compared to a control with no scale alterations, and these observations are consistent through both R and Python-created visualisations. It has also been observed that the scale by default ending at 40 while the bar height is at 42 leads to some underestimation in the height for both the control and truncated bar plots.

The log scale, when using R’s default of non-standard form notation, also leads to underestimation of values and to a greater level than the other two scales, suggesting this scale itself has an impact on interpreting the height of an individual bar, potentially as a result of seeing the value of 30 on the axis and subconsciously extrapolating this in a linear manner, and thus misjudging the actual position of the bar when seeing it appears just higher than 30. However, we do still have some high values in this set, which again could be down to extrapolation errors. The python default of standard form notation appears to have confused certain respondents, who are perhaps not as used to seeing this notation, and there was a very large range in the responses along with one person not even entering a number, but rather stating that they "Don't know", and another stating they believed the value was "Next to none". The "Nest to none" entry is very subjective, but could potentially be be assumed as a value close to 0, once again maybe as a result of standard form being less well known.

Approximately how much more than 'Log Grip' would you say 'Salmon Ladder' was was used?

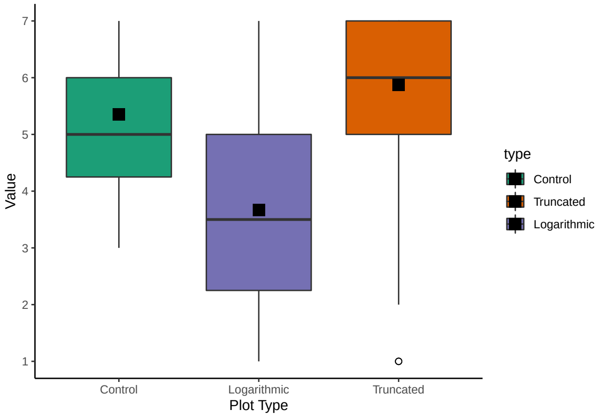
Now we consider the results from the second question, in which the participants were asked to respond on a scale from 1-7. The density plot below depicts the distribution of results for each of the three plot types. We see that the distributions appear to once again be non-normal.



The distribution for the logarithmic plot values has a fairly wide, flat curve, showing that the subjective view appeared to vary a fair amount from respondent to respondent. The distribution for the truncated plot seems very skewed to the right, depicting that the subjective view on the difference between the bar heights was that the difference was on the larger side.

An initial look at the table of summary statistics reveal means of 5.375, 3.671 and 5.871 respectively for the control, log and truncated plots, meaning that for the 'baseline' control plot participants, on average, judged the difference to be moderately significant, with the perceived difference being smaller for the log plot and marginally larger for the truncated plot. This appears to be consistent with results from [[[CITE chrome-extension://cbnaodkpfinfiipjblikofhlhlcickei/src/pdfviewer/web/viewer.html?file=file:///C:/Users/Katie/Downloads/YangVargasRestrepoStanleyMarsh%20(2020).pdf]]], in which the researchers, similar to this survey, showed participants a series of control bar plots alongside those with a truncated axis, and concluded that the difference in values for the truncated axis were perceived to be larger than those of the control plots. However, the average perceived difference here is fairly small, much smaller than initially hypothesised, so tests will be needed to decipher whether this is significant. The logarithmic plot causing the average perceived difference to be smaller follows the hypothesis from prior to running the survey.

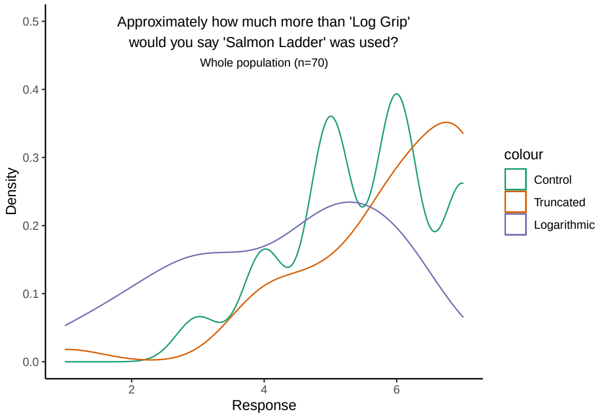
The box plot shows these results for each plot type.



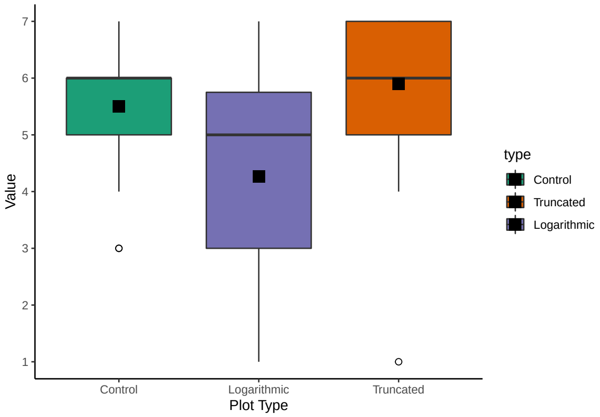
We see that the interquartile range for the control plot is smallest of the three at 1.75, followed by the truncated plot at 2, and then the log plot at 2.75. This depicts that overall, there was more of a consensus in the subjective perception of the difference for the control plot than the other two, and less agreement between participants for the logarithmic scale.

The black squares represent the means here, and we can see that for the control and truncated boxes, the mean is higher than the median, perhaps signifying a positive skew, with a slightly negative skew for the logarithmic plot.

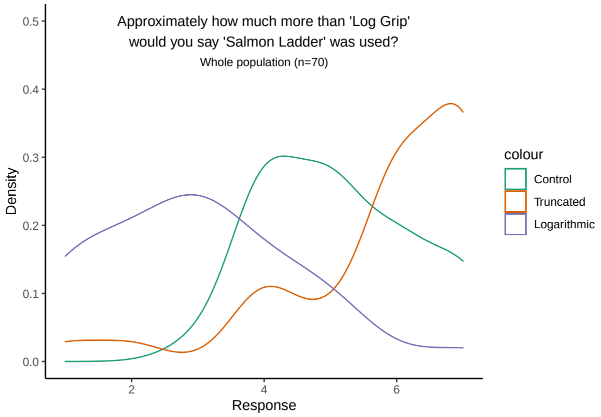
The summary statistics for the R population are given below.



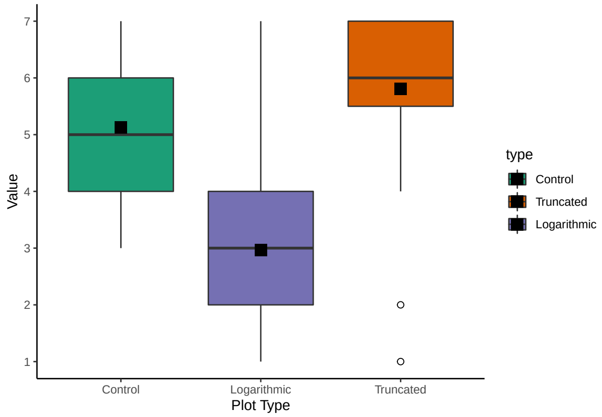
The means of the R population are very similar to the whole population for the control and truncated plots, with the mean for the logarithmic plot is slightly higher. The control and truncated values lie on the upper half of the scale, depicting that the difference in heights between the two bars was perceived to be considerable. The mean for the logarithmic responses is 4.243; almost directly in the centre of the scale. Thus, this difference was considered moderate but not too considerable. The medians for both the control and truncated plots are 6, with the median of the logarithmic slightly lower at 5, showing that these data centre around 6 and 5 respectively. All of these do, however, have large ranges, with the truncated and logarithmic responses having a range over the whole scale of $[1, 7]$ and the control with a range over $[3, 7]$. We therefore had some respondents who felt the differences between bar heights were very small, although the IQRs of the three, in order, are $[5, 6]$, $[5, 7]$ and $[3, 5.75]$ respectively, showing that 50% of the responses lie fairly close to the means, and the IQRs for the control and truncated plot responses encompass values in the upper half of the scale, furthering the conclusion that these height differences were perceived to be larger.



The control plot has the smallest IQR, with the value of 3 picked up as an outlier. The boxes for the log and truncated responses, as expected, sit slightly below and above the control box, respectively. Although, we also see that the upper end of the IQR for the log responses almost matches that of the control responses, and the lower end of the IQR for the truncated responses matches the lower end of the control plot. This shows that all three had a number of responses in the range $[5, 6]$, but the logarithmic responses also contained a fair amount of responses below this, and the truncated responses contained a fair amount above the $[5, 6]$ range. It can also be seen that the value of 1 is picked up as an outlier for the truncated responses, although not for the logarithmic responses. Thus it appears that 1 is in fact a response that can be considered as a valid response. It can also be noted that the boxplot of the logarithmic responses covers the whole range $[1, 7]$ with no outliers, implying that there was a large amount of variability in responses between participants. This is also highlighted in the summary statistics, where we have a variance of 2.523, which is higher than for the control and truncated.



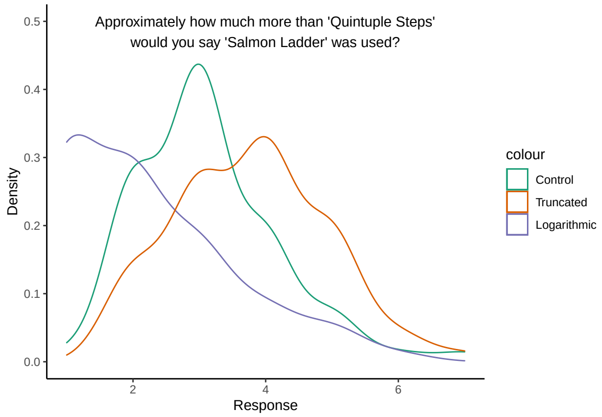
The statistics for the control and truncated plots are once again similar to those of the whole population and the R population, with the mean of the logarithmic responses slightly lower again at 2.968, signifying that the difference between the bar heights was perceived as very minimal. The IQR for the control responses place the bulk of these values in the range $[4, 6]$, which tells us that, on average, respondents felt the height difference was moderate. The truncated plot responses IQR places these values in the range $[5.5, 7]$, showing a larger perceived difference, and conversely the perceived differences in the log scaled plot appear on the smaller side, given the IQR of $[2, 4]$.



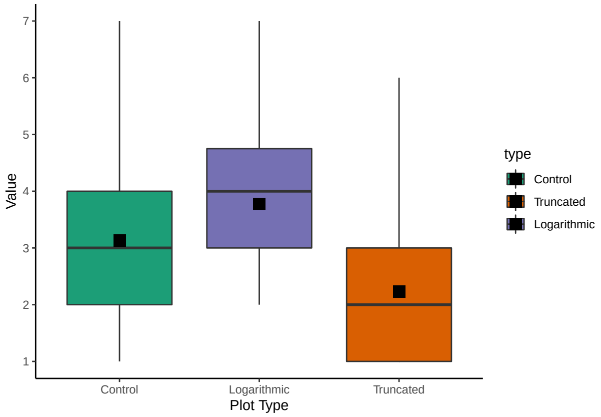
The box plots for the python population are much more distinct than for the R population, with little crossover of boxes. The log responses box is very obviously sat below the control, and the truncated lies above. The lower values of 1 and 2 are counted as outliers here for the truncated plot, although the box plot for the log responses once again spreads over the whole range, but the box itself is shorter than that of the R population.

Approximately how much more than ‘Quintuple Steps’ would you say ‘Salmon Ladder’ was used?

This is a similar question to the one prior, but the purpose was to see if there was a difference in perceived difference for bars next to each other vs bars on opposite sides of the plot.



```
   control          logarithmic        truncated
 Min.   :1.000   Min.   :1.000   Min.   :2.000
 1st Qu.:2.000   1st Qu.:1.000   1st Qu.:3.000  
 Median :3.000   Median :2.000   Median :4.000  
 Mean   :3.129   Mean   :2.229   Mean   :3.771  
 3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:4.750  
 Max.   :7.000   Max.   :6.000   Max.   :7.000  
```



Ninja Warrior - Part 2

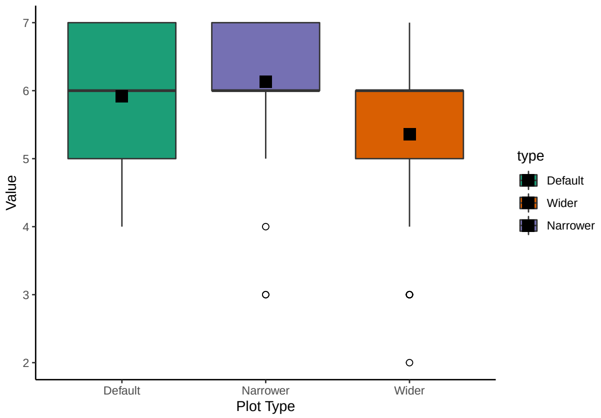
The second section of the survey tests whether altering aspect ratio of plots affects interpretation. The purpose of this is to mirror what my occur when visualisations are published, and may be resized to fit the section of the page they sit on. As in [[[CITE: http://perceptualedge.com/articles/visual_business_intelligence/bar_widths.pdf]]], it was hypothesised prior to the survey that an aspect ratio that narrows the bars may cause overestimation in values, and vice versa, using a ratio that widens bars could lead to underestimation. In the paper, the author discusses how increasing the widths of bars could distract from the bar height as well as take up excessive space on a page. It is also mentioned that wider bars may be "aesthetically displeasing". This section tests both how bar width alters perceived difference between bars as well as opinions on the aesthetics. The method in the paper also involves altering spaces between bars, including bar plots with spaces at 50% of the bar widths and then reducing the width of the space by a third. Conversely to this, we will not be considering different width of spaces between bars, but only the widths of the bars themselves. The author concludes that a length-to-width ratio of 10:1 appears to suffer from perceptual imbalance, but increasing this such that the bars become narrower and longer does not appear to have as much of an impact; the ratio can be increased relatively far with out causing much perceptual imbalance. In our version of this investigation, we have three bar plots of 7 obstacles, each with a different aspect ratio. The control, or default, plot is given as the plot for which the aspect ratio has not been altered, the plot with narrow bars has a doubled aspect ratio, and the plot wide bars a halved aspect ratio.

How large would you say the difference between 'Jumping spider' and 'Salmon Ladder' is?

This question once again uses the 7-point scale to gain a subjective view on the degree to which respondents felt the heights between the two bars corresponding to 'Jumping Spider' and 'Salmon Ladder' differed for three bar plots of 7 obstacles, where 'Salmon Ladder' is furthest to the left, and 'Jumping Spider' furthest to the right.

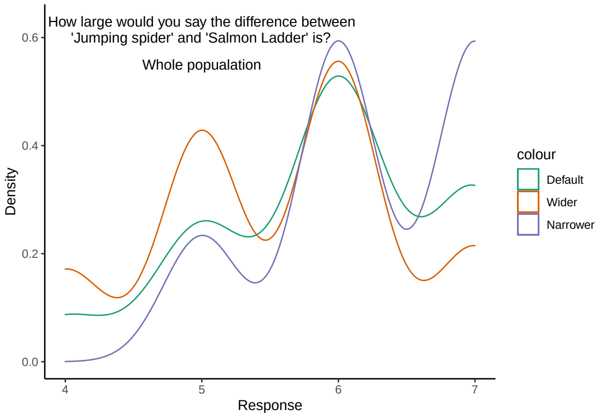
Looking at the means and medians here, it doesn't seem like there is that much of a difference in perception of the differences between the three aspect ratios. If anything, based on the means, we could say that the wider plot gave a slightly smaller perception of the difference, and the narrower plot slightly larger. However, these differences appear very marginal, although they do agree with the hypothesis formed from [[[CITE AS ABOVE]]].

To discuss the ranges, see the box plot below.



From these box plots, it appears that the IQRs for the two plots with altered aspect ratios have very little, if any, overlap, despite the means being similar and medians being identical. The narrower plot shows a tendency for the responses to lie more towards the upper end of the scale than the wider plot, which also ranges over the upper half of the scale but between roughly 5 and 6 rather than 6 and 7. The default plot covers the entire IQR of both of the other plots, and the box plots then show that, even though the means and medians are very similar, the center bulk of the values for the narrower plot tended to be more towards the upper end of the default's IQR, whereas the central points of the wider plot sat on the lower half of this IQR. Additionally, we see that there are two outliers each for the narrower and wider plots, with values 2, 3 and 4. Perhaps excluding these from the data and re-checking the summary statistics we will see a more marked difference.

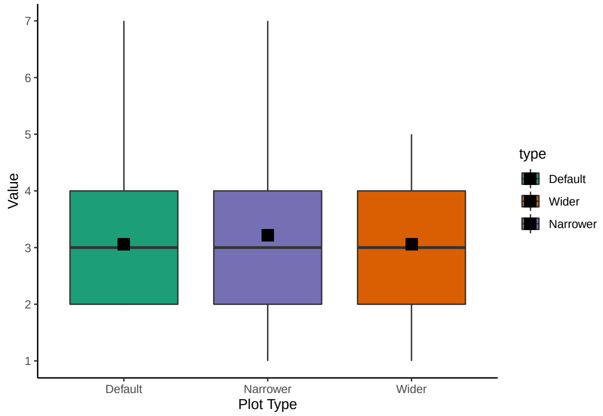
It seems that removing these outliers didn't actually have a huge effect on the means and medians, although as expected, the ranges and variances are lower, meaning the spread over values over the range is smaller, however this only furthers the point that altering the axis ratio appears to have minimal effect.This is again confirmed by looking at the distributions of the three plot types, which are very similar to one another. We can however see that the plot with the density for the wider plot is highest of the three for the values of 4 and 5, but is the lowest of the three for upper end of the distributions, and vice versa for the densities narrower plot. The default mostly stays in between the other two curves. For a third time, this gives way to the observation that the wider bars have a small lessening effect on gauging differences in height, and using narrower bars has a mild increasing effect on difference perception.



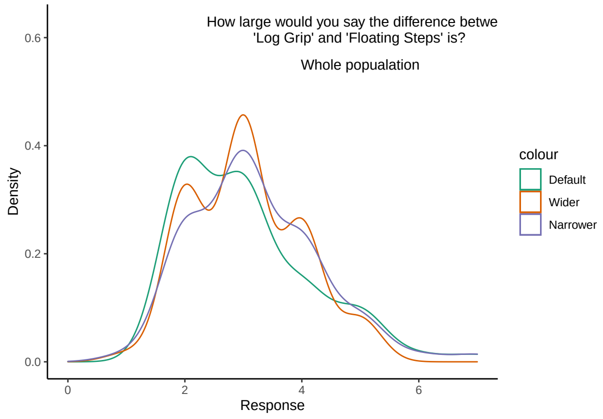
How large would you say the difference between 'Log Grip' and 'Floating Steps' is?

Similar to part 1, we have two questions for gauging differences between bars, for which one asks about bars far away from each other, and one about bars next to each other. In the case of this section, the first question contained bars on opposite ends of the x-axis, and this question asks about two bars that sit adjacent to one another.

From these statistics we see that altering the axis ratio appears to have even less of an effect than in the first question, with the means of the responses for the default and wider plots being identical, with the mean of the narrower plot responses only 0.157 greater. The medians are identical for all three, along with the IQRs. The variances, however, appear to differ more than the other statistics.



As anticipated from the table of summary statistics, all three IQR boxes are identical with the exception of the slightly higher mean of the responses from the narrower bars as well as differences in the overall range.

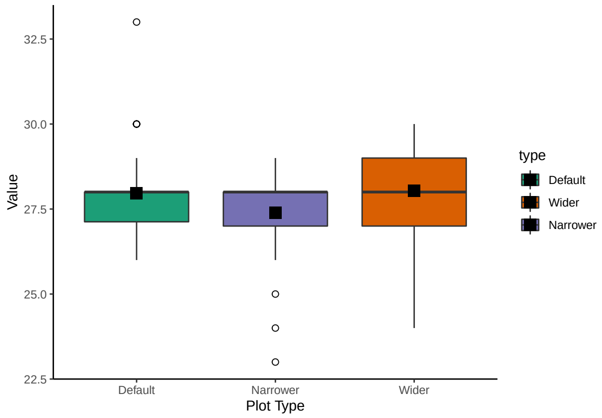


We see all three distributions are very similar, and almost appear to form bell curve shaped distributions, albeit with some irregularities.

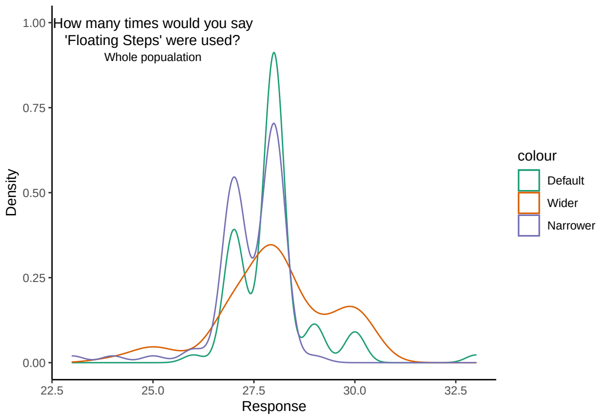
How many times would you say 'Floating Steps' were used?

This is again similar to question 1 of part 1, where participants were asked to state what they believed to be the height of the bar for 'Salmon Ladder', however this time we choose the third bar from the axis. This is to ascertain whether the distance of the bar from the axis may have an effect alongside any potential perceived distortion of values. Note that the true value was 28.

We see that the means of each of the three sets of responses are very close to the true value, and the medians are exactly equal to the true value. Based on the means and medians it appears that, once again, altering the axis ratio had minimal, if any, effect on interpretation of the data value. The value for the default plot also appear to be closer to the true value than the control plot in part 1, question 1.



Looking at the box plots, we see very small ranges in the values, signifying that there was a large consensus between respondents in terms of what they perceived the height to be. It can also be seen that there are three outliers below the box plot for the narrower plot responses, and two above for the default plot responses. There is very little overlap between the boxes, and it appears again that there altering the aspect ratio of the bar plot has little to no impact on reading the height of the bar. Additionally, there was less agreement between respondents for the wider plot than for the other two, although this doesn't seem to be too significant.



The distributions for the default and narrower plot responses are very similar, both seeming to be fairly centred on the mean with a steep decrease in density on either side of the mean to very shallow tails within the range $[25, 30]$. The responses for the wider plot appear to be more spread with lower density function values, with a slight negative skew.

Consider now the summary statistics after removing outliers.

After removing the outliers the medians have stayed the same, and the mean has obviously decreased for the default and increased for the narrower, however, these means are all still fairly similar to each other and at a first glance prior to testing it again seems that changing the aspect ratio, at least to the degree tested here, is inconsequential to interpretation of the actual value. As expected as well, the variances for the outlier-removed sets have decreased.

Comparisons

The last set of questions in part 2 show respondents all three of the bar plots presented in this section and ask them to select which they find most aesthetically pleasing, and which they find easiest and hardest to interpret. Below a table is laid out giving the number of respondents that selected each plot for each of the three questions.

For the first question, relating to how aesthetically pleasing respondents found each plot, just over half of the respondents chose the default aspect ratio as the most aesthetically pleasing, with 37 out of the 69 who responded selecting this.

Similarly, 37 out of the 70 that responded to the second question found the plot with the default aspect ratio easiest to read and interpret. Perhaps the people that preferred this aspect ratio aesthetically did so because they found it easiest to interpret. Investigating this, we find that 27 who chose the default for question 1 also chose this for question 2.

The plot judged hardest to read and interpret by the most respondents was the one with the wider bars, with 30 selecting this and 20 selecting each of the other two. While a significant number chose the default and narrower bars, the slightly higher amount selecting the plot with wider bars matches the previously stated hypothesis formulated from following the Stephen Few paper, which discusses that an ratio of greater width to length could suffer from perceptual imbalance. While we don't see this imbalance in the numbers from the previous questions, the result here does give some indication that the aspect ratio producing wider bars may impact on ease of interpretation.

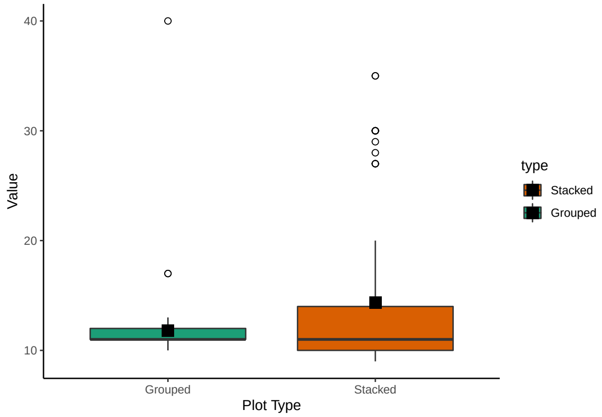
Ninja Warrior - Part 3

The third and final part of the questions about the American Ninja Warrior data discusses stacked bars and colour schemes. The questions asked in this part are used to decipher how data with multiple categories may be best represented in a bar plot. The plots presented use the same bars as in part 1, but this time we highlight the number of times each obstacle was used in each stage of the competition for each bar. Each participant was shown both a stacked and a grouped bar plot in one of three colour schemes; the default for the language, viridis, and greyscale. For three versions of the survey, the stacked bars were shown first, and for the other three versions the first shown was the grouped bars. The final question of this part also asked respondents to compare two colour schemes, and through the 6 surveys we have comparisons of every colour scheme against every other colour scheme.

How many times would you say 'Floating Steps' were used in the Finals (Regional/City) round?

Again we start with the less subjective question regarding the reading of a numerical value off the axis. In this question we ask about 'Floating Steps', which is the bar third along from the y-axis. The question asks respondents to view the bar plot, where the bars will either be grouped of stacked, and decipher how many times this obstacle was used in the specified round of the competition. The true value for this was 11. The hypothesis for this question is that the respondents will more accurately gauge the value for the grouped bar than the stacked, which as we see below appears to be the case.

The mean for the values estimated by respondents using the stacked bars is 14.32, a fair bit larger than the true value of 11, and the mean estimated value for the grouped bars was closer to the true value, at 11.8. The IQR for the grouped bars is also smaller than for the stacked, and comprises of the range $[11, 12]$, insinuating that the estimated values tended to be fairly accurate but with some respondents perhaps slightly overestimating. The IQR for the stacked bars on the other hand covers the interval $[10, 14]$, which does contain the true value, but shows a tendency for both over and underestimation of respondents. Additionally to this, there is a very large variance in the responses to this question, at 54.8 compared to the variance of 13.1 for the responses regarding the grouped bar plots. This adds to the picture that there was much less agreement between respondents, with many straying away from the mean of 14.3. We do see however that the median for both the stacked and grouped bars is 11, showing that the higher mean of the stacked bars may be a result of an influential value at the upper end of the distribution, and that many observations do actually sit around 11. The fact that many values actually sit around 11 could be contributing to the higher variance, as variance is simply the sum of the squared distances from the mean, and so will be elevated if there are many values that sit some distance away from the mean. The higher mean could be reflected in the maximum of the stacked responses being 35, although the maximum of the grouped responses is 40, so there may be more than one influential point in the stacked responses. We can check for outliers by looking at the box plots for this data.



We do in fact see that the box for the grouped responses is very short and centered around 11. The box for the stacked responses shows many high valued outliers that could be causing the mean to be higher, although the IQR is still a fair bit larger than that of the responses for the grouped bars. The mean for this also sits above the IQR, and thus the outliers may be having a significant influence. Now we will remove the outliers assuming, from the box plot, that outliers are any values above or equal to 25 for the stacked responses and above or equal to 20 for the grouped.

We see that removing the outliers as specified by the box plot, the mean of the stacked responses is now just above 11, and actually closer to the true value than the mean of the other set of responses, and the median has decreased to 10. From this one could infer that there is no difference between each type of bar plot in terms of gauging the size of the bars. However, we see that there are 12 outliers in the stacked responses, which leads to the idea that these are not in fact all outliers and may be valid responses that just sit on the upper end of the distribution. However, it seems the cause of the high values could be respondents taking the whole height of the bar, which has an actual height of 28, rather than the section of interest. Many of the potentially influential values fall around the range $[25, 30]$, with all but 2 of the 12 potential outliers sitting in this interval, with the remaining two both being 35. Looking below at the summary statistics for only the values picked up as outliers, we see a mean of 29.83, which is higher than the true value of 28, and interestingly goes against the analysis from part 1, question 2 whereby respondents were asked to judge the height of this bar and on average underestimated. The fact that so many participants misinterpreted this plot and signify that stacked bar plots may not be the best way to present data to general public, as there may be the potential to misread the height of the whole bar as the size of the top category.

As a result of this, we will discount this set of 12 values from the analysis, and thus come to the conclusion that, for the respondents that appear to have judged the height of the correct section, there was little to no impact when using stacked vs grouped bar charts, and most of the difference comes from misinterpretation of the plot itself, as opposed to a poorer judgment of size.

To see if either of these values are significantly far from the true value, we once again run tests. Firstly, run Shapiro tests to test for normality.

The small p-values signify that this data is not normal, and a look at the summary statistics tells us neither is symmetric about the mean. Thus, as in question 1, we run sign tests, alongside the t-tests on samples from a normal distribution with mean and variance equal to out data. We will test against a median of 11 for the sign tests, and a mean of 11 for the t-tests.

This test gives a high p value of 0.5258, showing that for the stacked bar plot responses (after removing the values as priorly specified), the participant estimated values do not differ significantly from the true value.

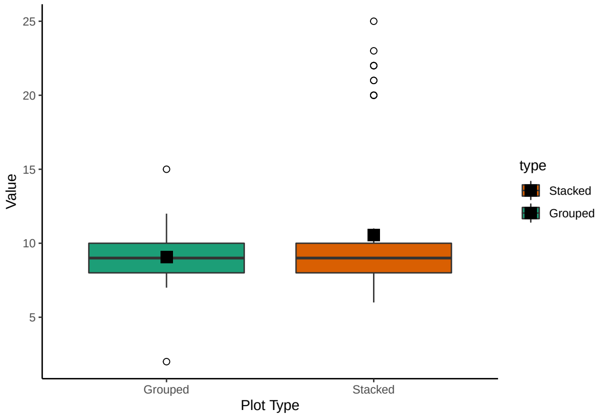
For the grouped bar plot we have a p value of $0.009 < 0.05$, and thus these responses are statistically significantly different from the true value.

Running t-tests on the means, however, we see both sets of responses differ statistically significantly from the true value.

How many times would you say ‘Log Grip’ was used in the Finals (Regional/City) round?

This question is similar the above, but for the next bar to the right. The purpose of this question was to test the same hypothesis as the previous question, and also to lead into the following question, where respondents were asked to compare the ‘Floating Steps’ and ‘Log Grip’. Additionally, the bar in the previous question had only two categories, of which the respondents were asked to judge the size of the category on the top of the bar in the stacked plot, whereas the bar for 'Log Grip' has 5 categories, of which the category of interest sits above 4. The true value of this was 9.

Similarly to the previous question, we see that the mean response for the stacked bar plots are higher than that of the grouped, and the mean of the stacked also slightly overestimates the value. Once again however, we appear to see a selection of respondents judging the full height of the bar rather than the category as asked. Looking at the data, the interval for these responses seems to be $[20, 25]$, as the next response below 20 is a value of 10, seeming to separate the data into two separate subsets. This can be confirmed by a box plot.



We indeed see that the distribution of values for each of the two response sets appears to be almost identical with the exception of outliers at and above 20 for the box plot of responses for the stacked bar plot. Thus we view the sets of summary statistics for the two but with these values removed.

Here we see that there tended to be a slight underestimation in the value for the stacked bar plot, however this is approximately 0.46 away from the true value, and unlikely to be significant. This can again be tested as above, where it is less clear whether the data are symmetric, so we will also run symmetry test.

Once again we see that the response sets are non-normally distributed and asymmetric, and so sign tests are once again applicable.

Here we see a p value of around 0.04, which shows a statistically significant difference in the responses from the true value of 9 at the 0.05 level of significance. However, this would very easily become insignificant by slightly lowering the significance level to, say, 0.035.

This p value is » 0.05, as expected given that the median of the data sits at the true value.

The t-tests show that the differences in the means from the true value are statistically significant, although not considering the tests we can see by eye that the means are relatively close to 9.

Please select the statement you feel applies to the bar chart above. This question asked respondents to judge whether log grip was used more, less, or an equal amount in the Finals (Regional/City) and Qualifying(Regional/City) rounds. This was to see how well differences between sizes of categories are judged when relating to the same variable, and are in the same bar. The results for this are given in the table below.

The table shows overwhelmingly that significantly more people accurately judged that the two values were the same for the grouped bars than for the stacked bars. This was the hypothesised result, and has presented to an even greater extent than previously anticipated. All but 7 of the respondents who responded to this question correctly judged from the grouped bars that the obstacle was used an equal number of times in each of the two rounds, whereas the responses for the grouped bar seemed fairly well split between the three options. It may be interesting in the multivariate analysis section to compare responses depending on whether respondents were shown the stacked or grouped bars first. Perhaps a reason for the incorrect judging with the stacked

Which obstacle do you think was used MORE in Finals (Regional/City) rounds, 'Log Grip' or 'Floating Steps'? Similar to the previous question, this asks for a comparison between the size of two categories, but this time about how many times two different obstacles were used in the round Finals (Regional/City), where these two obstacles are those discussed at the start of this part of the survey.

This was a potentially poorly formulated question, as the respondents had already been asked to specify how many times each of these obstacle was used in this round and respondents mostly judged this accurately with regard to both plots, but this could have been impacted by the previous questions. However, this does follow from the results from the past questions showing that respondents mostly accurately judged the values correctly, aside from those who instead judged the height of the whole bar.

Which bar chart do you feel is easiest to read and interpret? Here was simply assess the perceived ease of interpretation of both bar plots. This is to gain an understanding in how data may best be presented in an easily understandable, easily readable manner. This is an important factor in visualisation, as a main aim in creating visuals is to provide an aid for the viewer to simply and quickly see the message. The opposite may be beneficial in certain applications however; based on the misreadings in the question regarding judging the number of times 'Log Grip' was used in the specific round, viewers of the visualisations could be easily mislead by incorrectly interpreting the plot. The people being shown the plot in, for example, an advert, may only take a fleeting look and not go beyond to analyse the plot to see accurate differences between values, and thus it is important to produce a plot that gives the easiest interpretation.

The large majority of participants found the grouped bar chart easier to read and interpret, as predicted.

Which colour scheme do you find most aesthetically pleasing? This question and the one following it are asked with the purpose of assessing the colour scheme that gives the greatest aesthetic pleasure, or effectively which colour palette the respondents feel is subjectively the 'prettiest' or 'nicest'. It is important to note here that aesthetics and readability do not always go hand-in-hand; a plot that is made to look very aesthetically pleasing may sacrifice readability, and vice versa. For each of the two languages we created six pairings of three different colour palettes, whereby the first colour was the one displayed for the main questions, and the second used only for the comparison questions. As previously discussed, the three colour schemes considered are viridis, greyscale, and each language's default plotting colour palette. The colour palette pairings are outlined below.

This table shows that when it came to the default/viridis pairings, displayed in the first two rows, the respondents tended to have no preference overall, although this may differ between languages, which will be explored later on. Comparing this to the bottom two rows, in which we put viridis against greyscale, only 1 respondent out of the 23, a proportion of 0.04, found the grey more aesthetically pleasing, as hypothesised. When considering greyscale/default, there was still a majority preferring the non-greyscale palette, but a higher proportion preferred this as compared to the viridis/greyscale, with 4 out of the 22, or a proportion of 0.18, preferring the grey.

Do you feel that one of the colour schemes makes it easier to read and interpret? If so, please select which one. Complementing the aesthetic preferences, this question assesses the colour preference with regard to readability and ease of interpretation. As mentioned before, this will be used to test both the colour palette preference itself alongside whether this preference matches up with aesthetic preference.

Interestingly here, we see that the top two rows appear to give opposing results; the respondents who were presented with viridis for the main questions and the default as a secondary palette stated that they found either viridis easier to interpret or had no preference, whereas those presented with the default first and viridis second tended to find the default easier. Once again looking at the comparisons with the greyscale, there were some respondents that found this easier to read, but the majority chose the alternative, whether this is viridis or the default.

Sales - Part 1

Now we move on to the sales part of the survey. In this section data was taken from a the BJsales data set in R, which is a time series data set containing 150 observations. This data set constitutes a single vector of values with no specified timings, and the visualisation data was formed by taking subsets of size 12 this and setting a month between each point to give a year of fictional sales data.

How much would you say sales of each company increased between January and December? [Company A] This question was included for the purpose of testing whether, again, axix scaling impacts the precieved differences between values.

How much would you say sales of each company increased between January and December? [Company B]

How large would you say the drop in sales between April and July of Company A is?

Sales - Part 2

Based on the above graph, how large would you say the difference is between the number of sales Company C makes and the number of sales Company D makes?