Are Certain Visualizations Better for Certain Tasks?

Purpose

We want to determine whether certain visualizations were better for certain tasks.

Method

23 MDS students were given 2 surveys in a random order: survey A and survey B. Survey A had 2 histogram plots, one representing the number of intentional walks, and one representing the number of home runs. Survey B had one scatter plot representing the home runs as the explanatory variable and intentional walks as the response variable.

Data was sourced from the Batting table of the Lahman R package for the Toronto Blue Jays team for all years in the dataset (1977 - 2015).

Each survey had 12 identical questions designed to determine if certain visualizations were better for certain tasks. The first 3 questions asked the participants to respond with a rank from 1-5, questions 5 and 6 required that participants select an answer from a drop-down list of options, and the rest of the questions allowed the participants to input any response.

The results were then compiled and read into R, where plots were generated and tests were conducted in order to determine if certain visualizations were better for certain tasks. In order to analyze the data, all entries that did not conform to the data type that we were looking for were replaced with NA values. NA values were added by default when the classes of the columns were set.

The plots were created and the tests were determined based on the class of the data in the data frame. We assigned 3 different classes:

Column Class	Plot	Test	Questions
Factor	2 types of bar plots	chi-squared	5, 6
Ordered Factor	2 types of bar plots	wilcox	1-3
Numeric	A histogram and a boxplot	t	4, 7-12

Permutation tests were also conducted for further rapore.

We are assuming that the data from each survey has an equal variance and that a sample of N > 20 is large enough to perform both the chi-squared test and the t-test.

Results

Hypothesis testing:

For each question:

- H_0: There is no difference in the average results of surveys A and B
- H_A: There is a difference in the average results of surveys A and B

With $\alpha = 0.05$, we determined the following p-values and hypothesis conclusions:

number	question	test	p.value	result
1	This visualization makes it easy to see the distribution of the data	chi-squared	0.0001663	We reject the null hypothesis
2	This visualization makes it easy to identify outliers	chi-squared	0.0082396	We reject the null hypothesis
3	This visualization makes allows you to easily identify trends in the data	chi-squared	0.4534141	We fail to reject the null hypothesis
4	From this visualization, estimate the average number of intentional walks for a player:	t	0.6368289	We fail to reject the null hypothesis
5	From this visualization, estimate the distribution of the intentional walks:	wilcox	0.1729272	We fail to reject the null hypothesis
6	From this visualization, estimate the distribution of the home runs:	wilcox	0.4177438	We fail to reject the null hypothesis
7	From this visualization, estimate the number of intentional walks that are outliers:	t	0.3189416	We fail to reject the null hypothesis
8	From this visualization, estimate the number of home runs that are outliers:	t	0.6496092	We fail to reject the null hypothesis
9	From this visualization, estimate the most frequent number of intentional walks observed:	t	0.3902744	We fail to reject the null hypothesis
10	From this visualization, estimate the most frequent number of home runs observed:	t	0.0021712	We reject the null hypothesis
11	From this visualization, estimate the smallest number of intentional walks observed for any player:	t	0.0433640	We reject the null hypothesis
12	From this visualization, estimate the smallest number of home runs observed for any player:	t	0.0049474	We reject the null hypothesis

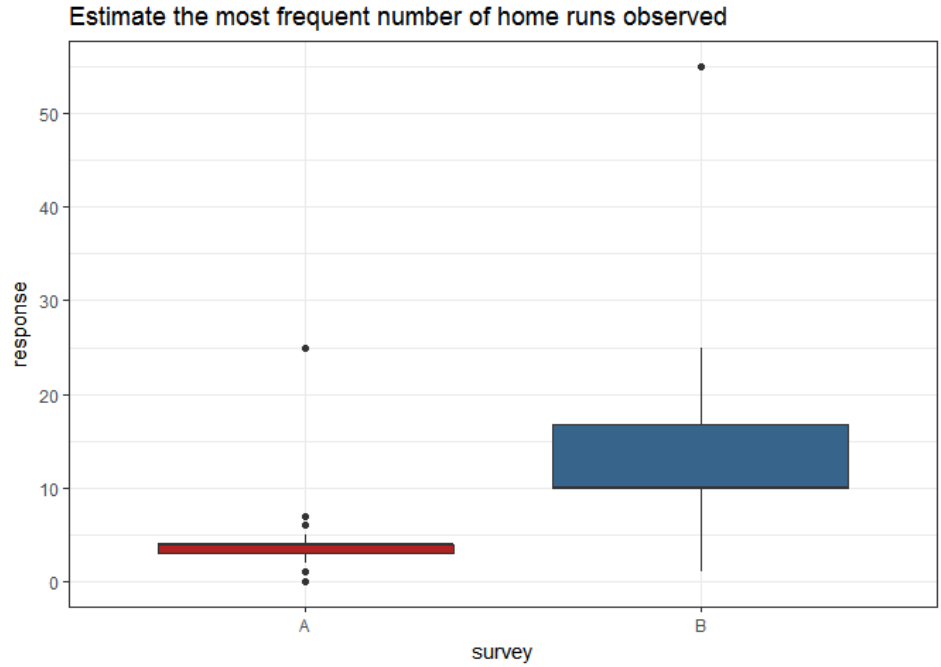
Based on our hypothesis testing, we rejected the null hypothesis in 5 of the 12 questions, meaning that those questions suggest a difference in the average results of surveys A and B (1-2, 10-12). Based on our testing, the rest do not suggest a difference in the average results therefore we cannot conclude that one visualization is better for a task than the other.

We can look at the plots in order to determine which visual, A or B, was better for the task for the questions in which the null hypothesis was rejected.

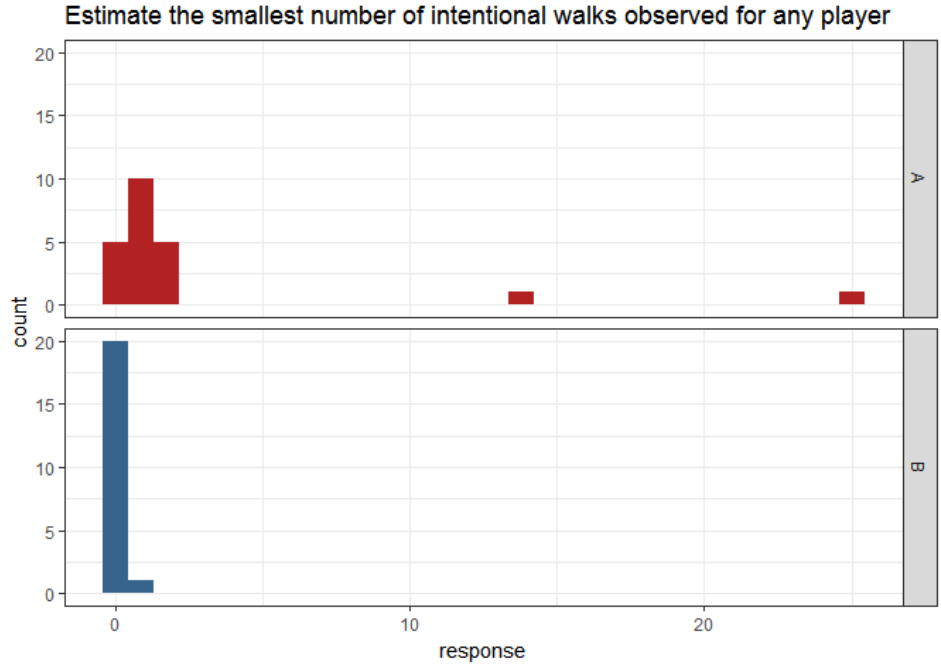
Plots

From the bar plot above, we can clearly see overall higher rankings in survey A with 7 participants ranking ease at '4' (vs 3 in survey B) and 12 participants ranking ease at '5' (vs 4 in survey B). We can also see that nobody found it difficult to see the distribution in survey A, while some participants did in survey B (rankings of '1' and '2') Clearly, based on the survey results, visualisation A (the histogram) is better at showing the distribution of the data than visualisation B (the scatterplot). Based on our survey results, we can conclude that histograms make it easy to see the distribution of the data. This is expected as scatterplots plot two variables against one another making it extremely difficult to visualise distributions.

From the bar plot above, we can clearly see overall higher rankings in survey B with 10 participants ranking ease at '4' (vs 4 in survey A) and 8 participants ranking ease at '5' (vs 5 in survey A). We can also see that nobody found it difficult to identify outliers in survey B, while some participants did in survey A (ranking of '1') Clearly, based on the survey results, visualisation B (the scatterplot) is better at identifying outliers than visualisation A (the histogram). This is expected as scatterplots plot all the points while histograms plot counts and unless the binwidth is set to 1 it would be more difficult to spot the outliers.



Based on the snippet of the Batting dataset that was used in the plots, the most frequent number of homeruns was 2 with 72 occurences. We can use this information to determine which visual was better at estimating the most frequent number of home runs observed. We can see from the boxplots that survey A responses, in general were a lot closer to 2 than survey B responses. As there are some outliers in this case, using the quartiles as metrics is better than relying on the mean as they are less prone to outliers. Based on this boxplot, we can conclude that visualisation A is better for estimating the most frequent number of home runs than visualisation B. We can also infer that histograms are better at estimating frequencies than scatterplots. This is expected as histograms are essentially plots of frequencies.



Questions 11 and 12 are analyze the same thing, but for different factors (home runs vs intentional walks); therefore, we will only discuss one of them (11). Based on the snippet of the Batting dataset that was used in the plots, the smallest number of intentional walks observed for any player is 0. We can clearly see from this histogram that most participants selected 0 in survey B, while there was a larger variation in results in survey A responses. While some selected 0 in survey A, many selected other results such as 1 or two (some even more). In this case, based on our results, we can conclude that visualisation B (the scatterplot) is better at estimating the smallest number of intentional walks for any player and infer that scatterplots are better than histograms at estimating minimums (this is a wrong conclusion and we will discuss possible reasons why below).

Conflicts

In question 11, and presumably question 12 too, our results suggest that scatterplots are better at estimating counts than histograms. Both of these plots should have been able to easily answer those questions. Upon further inspection, I noticed that the histogram in survey A cuts out the results where intentional walks/home runs is equal to 0. Upon inspecting the code, this is because the `xlim` function cuts off the data at the bounds.

Based on concepts discussed in class, we should have been able to conclude that scatterplots are better at recognizing trends than histograms. This is because we are plotting two factors against each other in scatterplots. I believe that this may be due to ambiguity in the question.

In questions 5 and 6, we should have been able to conclude that histograms are better at recognizing distributions than scatterplots. This is because histograms plot the frequencies or how often possible values occur, which is by definition what a distribution is. Scatterplots plot 2 factors against one another making it very difficult to infer distributions. This could be due to the forced participation nature of the question. There was no NA option in the drop-down list.

Finally, in questions 7 and 8, we should have been able to conclude that visual B (scatterplot) is better at detecting outliers than visual A (histogram) for the reasons discussed above. Since our tests rely on the mean, and the mean is not very robust against outliers, we may have come to our results due to a few extreme outliers. This could be combatted by removing some outliers from the data.

Recommendations

In conclusion, our survey results do not all agree with concepts we have been taught. These could be minimized by making some modifications. In general, the survey could be improved by modifying some of the questions and using a larger sample. We should use `coord_cartesian` instead of `xlim` and `ylim` to avoid cutting out important information. Also, adding a model (such as a linear model for example) to a scatterplot makes it easier to identify the trends. Changing the widths and sizes of bins in histograms makes a significant difference as to whether or not the distribution is easy to infer. Sometimes simply changing the scale (zooming in and out of a plot) makes it significantly easier to get the message or task they are trying to portray across (for example, makes it easier to see outliers). In scatterplots, changing to color of outliers is also a good way to improve the visualisation.