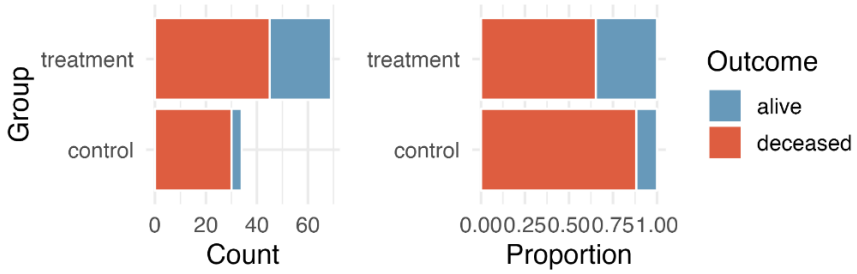
SDS 220 - Lecture 6 Handout

IMS Chapter 4, Chapter 5, Section 7.1.4

1. **[IMS 4.5] Heart transplant data display**. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was officially designated a heart transplant candidate, meaning that he was gravely ill and might benefit from a new heart. Patients were randomly assigned into treatment and control groups. Patients in the treatment group received a transplant, and those in the control group did not. The visualization below displays two different versions of the data. (Turnbull, Brown, and Hu 1974)



   (a) Provide one aspect of the two-group comparison that is easier to see from the stacked bar plot (left)?

   (b) Provide one aspect of the two-group comparison that is easier to see from the filled (also called standardized) bar plot (right)?

   (c) For the Heart Transplant Study which of those aspects would be more important to display? That is, which bar plot would be better as a data visualization?

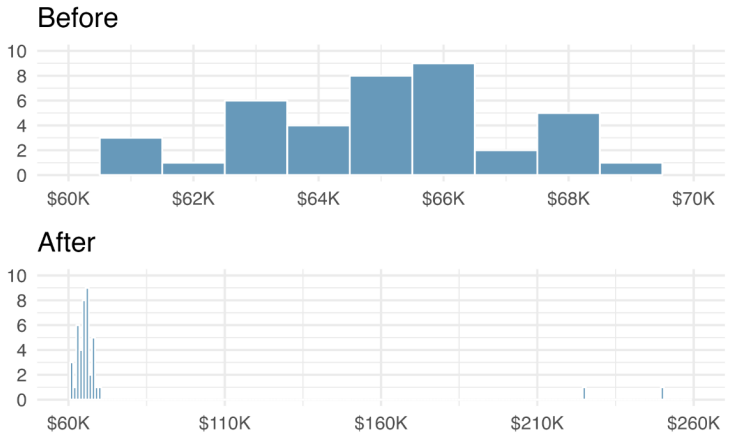
   (a) In the stacked bar plot, it is easier to see the number of participants in each of the two treatment groups. (b) In the standardized bar plot, it is easier to see which treatment group had a higher proportion of survival. (c) The focus is most likely to be on the proportion of people who survive, so the standardized bar plot should be displayed as a way to visualize the survival improvement in the treatment versus the control group.

2. **[IMS 5.22] Income at the coffee shop.** The first histogram below shows the distribution of the yearly incomes of 40 patrons at a college coffee shop. Suppose two new people walk into the coffee shop: one making $225,000 and the other $250,000. The second histogram shows the new income distribution. Summary statistics are also provided, rounded to the nearest whole number.

   (a) Would the mean or the median best represent what we might think of as a typical income for the 42 patrons at this coffee shop? What does this say about the robustness of the two measures?

   (b) Would the standard deviation or the IQR best represent the amount of variability in the incomes of the 42 patrons at this coffee shop? What does this say about the robustness of the two measures?

	n	Min	Q1	Median	Mean	Max	SD
Before	40	$60,679	$60,818	$65,238	$65,089	$69,885	$2,122
After	42	$60,679	$60,838	$65,352	$73,299	$250,000	$37,321



(a) The median is a much better measure of the typical amount earned by these 42 people. The mean is much higher than the income of 40 of the 42 people. This is because the mean is an arithmetic average and gets affected by the two extreme observations. The median does not get effected as much since it is robust to outliers. (b) The IQR is a much better measure of variability in the amounts earned by nearly all of the 42 people. The standard deviation gets affected greatly by the two high salaries, but the IQR is robust to these extreme observations.

3. Most years have 365 days. The first day of the calendar year is January 1st, February 1st is the 32nd day of the calendar year, and so on. Determine whether a relationship exists between the number of the day of the year and the temperature high for that day for the first of each month.

   (a) Collect the data for 2022 (you will have to look up historical data online).

Date	Day of Year	Temp High
Jan 1	1	
Feb 1	32	
Mar	60	
Apr 1	91	
May 1	121	
Jun 1	152	
Jul 1	182	
Aug 1	213	
Sept 1	244	
Oct 1	274	
Nov 1	305	
Dec 1	335	

   (b) Create a scatterplot for *Day of Year* and *Temp High*.

(c) Does there look like there is a relationship? What would you suspect the correlation to be between the two variables? Explain.

(d) If you could add a third (or fourth) variable to the plot, what would you want to add? How would you incorporate this variable to the plot?

(e) Can you think of another variable that would have a stronger correlation with *Day of Year*?