Lecture 4:
Summarizing and Displaying Categorical Data
STAT 310, Spring 2023

Frequency Tables and Bar Plots

- Recall that a **categorical variable** takes on values that fall into distinct categories. For example, gender, education level, and marital status are categorical variables.

- A **frequency table** summarizes data for a single categorical variable. It shows the *counts* for each category.

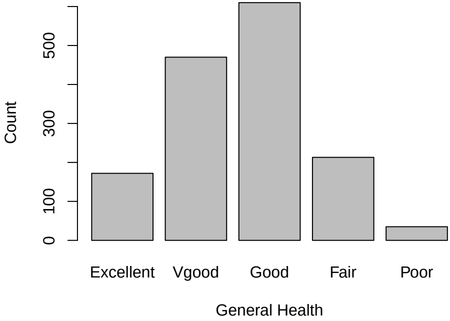
- A **bar plot** is a common way of visualizing the distribution of a single categorical variable.

Frequency Tables and Bar Plots

A frequency table and bar plot for the categorical variable `HealthGen` (self reported health rating), from the `nhanes` data set (discussed in lab 2).

```
> table(nhanes$HealthGen)
Excellent     Vgood      Good      Fair      Poor
      172       470       610       213        35
```

```
> barplot(table(nhanes$HealthGen),
   xlab = "General Health", ylab = "Count")
```



Relative Frequency Table

A **relative frequency table** shows the *proportions*, instead of counts, for each category.

```
> dim(nhanes)
[1] 1500   11

> table(nhanes$HealthGen) / 1500
Excellent     Vgood      Good      Fair      Poor
    0.115     0.313     0.407     0.142     0.023
```

Contingency Tables

A **contingency table** summarizes data for two categorical variables. Each value in the table represents the number of times a particular combination of variable outcomes occurred. For example, below is a contingency table between the variables `PhysActive` and `HealthGen`.[1]

```
> table(nhanes$PhysActive, nhanes$HealthGen)

      Excellent Vgood Good Fair Poor
  No         48   169  279  150   31
  Yes       124   301  331   63    4
```

[1] The variable `PhysActive` indicates whether the respondent does moderate or vigorous-intensity sports, fitness or recreational activities (No / Yes).

Contingency Tables

```
# include row and column totals
> addmargins(table(nhanes$PhysActive, nhanes$HealthGen))

      Excellent Vgood Good Fair Poor  Sum
  No         48   169  279  150   31  677
  Yes       124   301  331   63    4  823
  Sum       172   470  610  213   35 1500
```

- What proportion of respondents reported being in excellent health?
  $172/1500 = 0.115$
- What proportion of respondents reported being physically active?
  $823/1500 = 0.549$
- What proportion of respondents are both physically active **and** reported being in excellent health?
  $124/1500 = 0.083$

Column Proportions

In a contingency table of column proportions the counts are divided by the corresponding column totals. So the columns sum to 1.

```
> prop.table(table(nhanes$PhysActive, nhanes$HealthGen), margin = 2)

      Excellent Vgood  Good  Fair  Poor
  No      0.279 0.360 0.457 0.704 0.886
  Yes     0.721 0.640 0.543 0.296 0.114
```

- What does 0.721 represent in the table above?
  $124/172 = 0.721$ represents the proportion of respondents in excellent health who are physically active.
- What does 0.114 represent in the table above?
  $4/35 = 0.114$ represents the proportion of respondents in poor health who are physically active.

Row Proportions

Similarly, in a contingency table of row proportions the counts are divided by the corresponding row totals. So the rows sum to 1.

```
> prop.table(table(nhanes$PhysActive, nhanes$HealthGen), margin = 1)
```

```
      Excellent Vgood  Good  Fair  Poor
  No      0.071 0.250 0.412 0.222 0.046
  Yes     0.151 0.366 0.402 0.077 0.005
```

- What does 0.222 represent in the table above?
  $150/677 = 0.222$ represents the proportion of **not** physically active people who are in fair health.

- What does 0.077 represent in the table above?
  $63/823 = 0.077$ represents the proportion of physically active people who are in fair health.

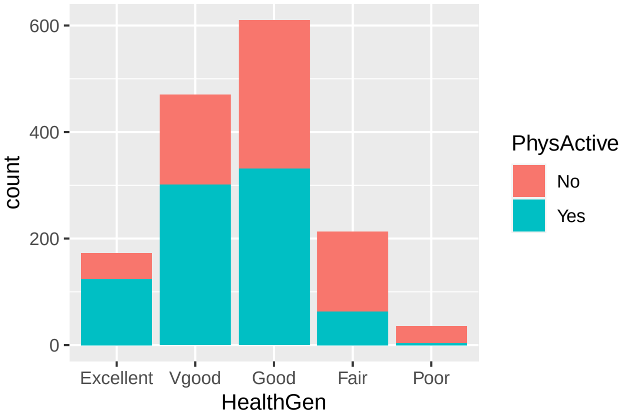
Bar Plots with Two Variables

Some ways to visualize contingency table information:

- Stacked bar plot
- Side-by-side bar plot
- Standardized bar plot
- Mosaic plot

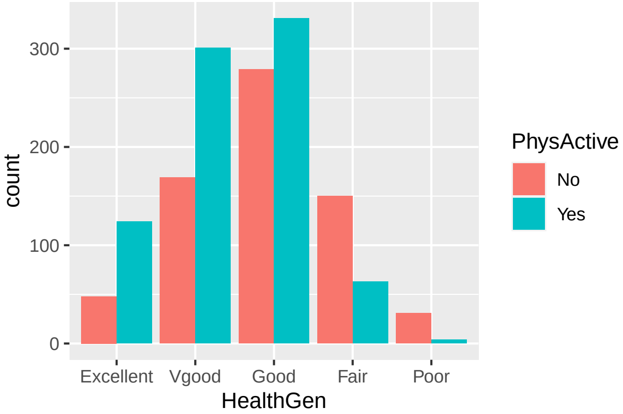
Stacked Bar Plot

A **stacked bar plot** is a graphical display of contingency table information, for counts.



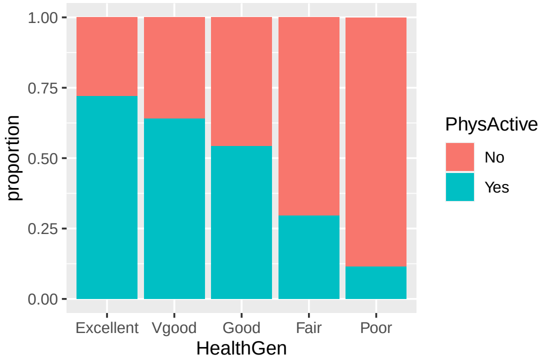
Side-by-Side Bar Plot

A **side-by-side bar plot** places bars next to, instead of on top of, each other.



Standardized Bar Plot

A **standardized bar plot** is a graphical display of a contingency table of column proportions.

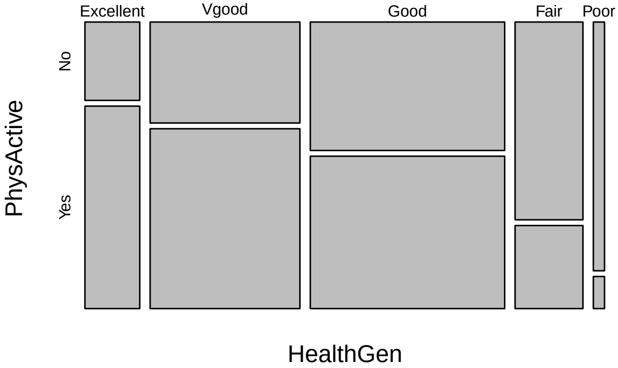


Based on the plot above, does there appear to be a relationship between `HealthGen` and `PhysActive`?

Yes, as general health goes from poor to excellent, the proportion of respondents who are physically active increases. That is, respondents who are in better health are more likely to be physically active.

Mosaic Plot

A **mosaic plot** is similar to the standardized bar plot, except the column widths correspond to the proportion of respondents who are in each of the general health categories.


Excellent
Vgood
Good
Fair
Poor
No
Yes
PhysActive
HealthGen