Exam 1 Prep

Define variation in your own words and explain why it is central to statistics.

In the context of a data frame, what do rows represent? What do columns represent?

Which of the following is a *quantitative* variable?

A) Height
B) Favorite color
C) Gender
D) Eye color

Why is sampling necessary in statistics? Provide one reason.

What does it mean for variables to be categorical versus numerical? Give an example of each from a dataset.

In modeling notation, what do the symbols `~` and `data=` indicate? Offer a brief explanation.

The vector `x <- c(2,1,3,3,2,3,1,2,1)` is given.

A) After sorting `x`, what pattern becomes visible?
B) What does the frequency table of `x` show?

Using the `Fingers` dataset (from class):

A) What do boxplots of `Index ~ Gender` visually display about variability?
B) Describe center, spread, overlap, and any unusual features.

Consider the following R code:

```
gf_histogram(~Score, data = Data, binwidth = 2)
```

A) Interpret what the following R code does and what you would expect the plot to show:

B) What distribution characteristics would the histogram reveal?

Explain what the five-number summary tells you about a numerical variable and relate it to variation.

Write one sentence comparing the distributions displayed by:

```
gf_histogram(~Index, data = Fingers, binwidth = 0.25)
gf_histogram(~Index, data = Fingers, binwidth = 0.5)
```

Focus on how the choice of **binwidth** affects the appearance.

Explain in plain language what an outlier is and how the 1.5×IQR rule identifies outliers.

What does it tell you if two boxplots (for `Index` by two groups) show a large difference in medians but overlapping boxes? What does that say about variation within and between groups?

consider the following R output:

```
favstats(Fingers$Pinkie)
```

```
 min Q1 median Q3 max     mean       sd   n missing
  33 55     58 63  98 59.41252 9.080594 157       0
```

A) What can you say about the *center* and *spread* of the `Score` distribution?

B) Use 1.5IQR rule to find if there are any outliers.

Explain why it is important to identify the response variable before choosing a plot to visualize a relationship.

Given these R commands:

```
mean(Pinkie ~ Gender, data = Fingers) # Returns means for Pinkie in different genders
gf_boxplot(Pinkie ~ Gender, data = Fingers)
```

Explain what each line does and how the two results complement each other.

Describe in words what the residual represents in the context of a simple model predicting `Pinkie` using `Gender`, and why smaller residuals imply a better model.

The following R code calculates proportions:

```
tally(Gender~Job, data=Fingers)
```

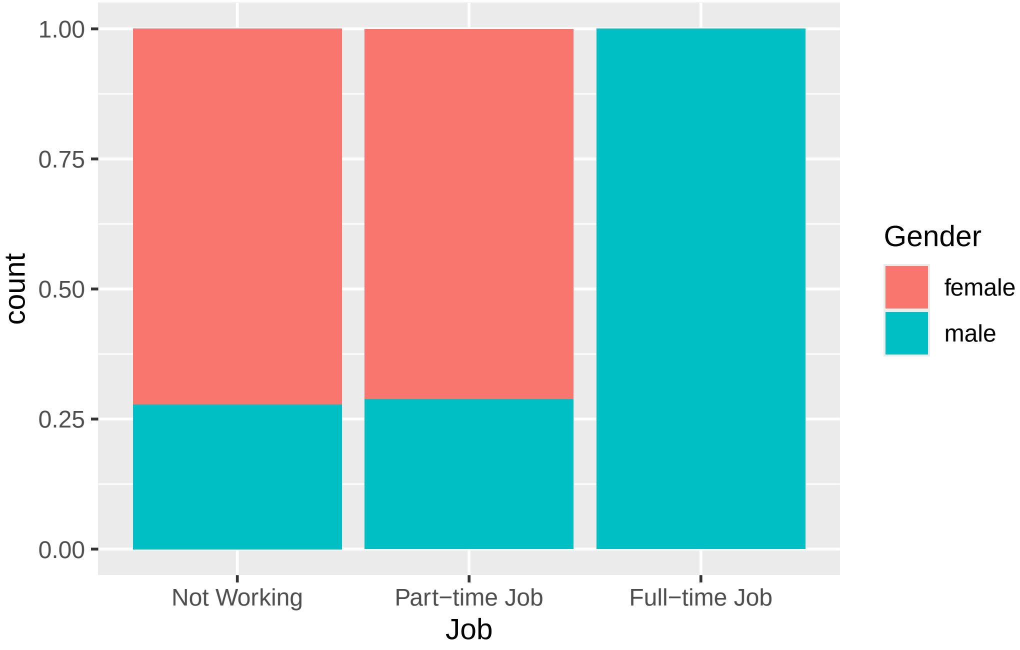
```
        Job
Gender    Not Working Part-time Job Full-time Job
  female           65            47             0
  male             25            19             1
```

Give an example of a conditional probability and compute it.

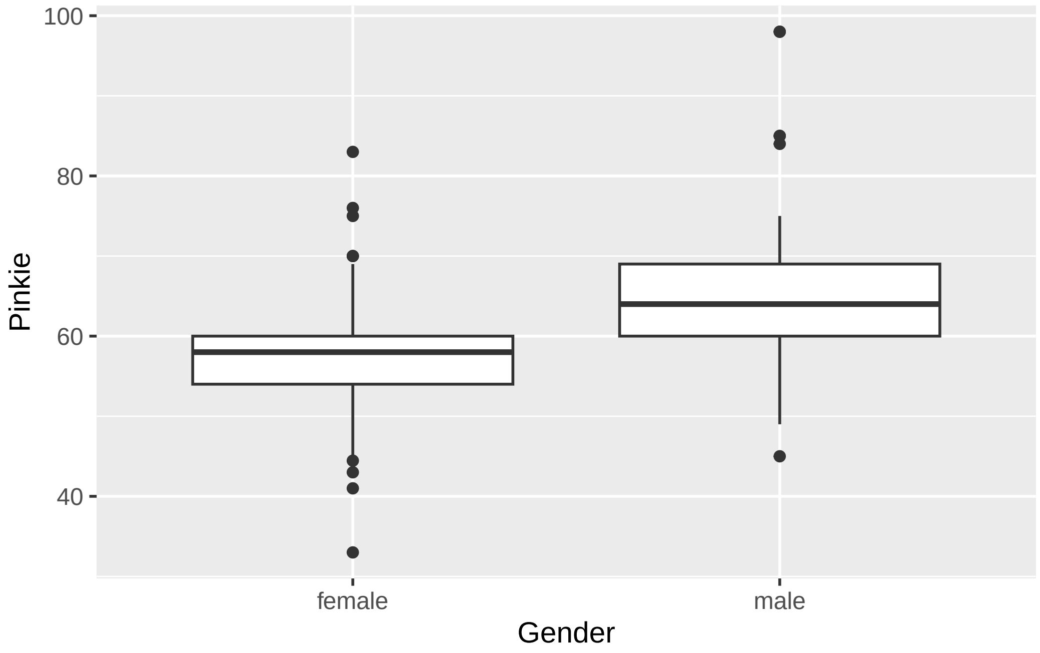
Below is R code that generates a conditional proportion bar chart. Explain how this visualization helps you *explain variability in the response*.

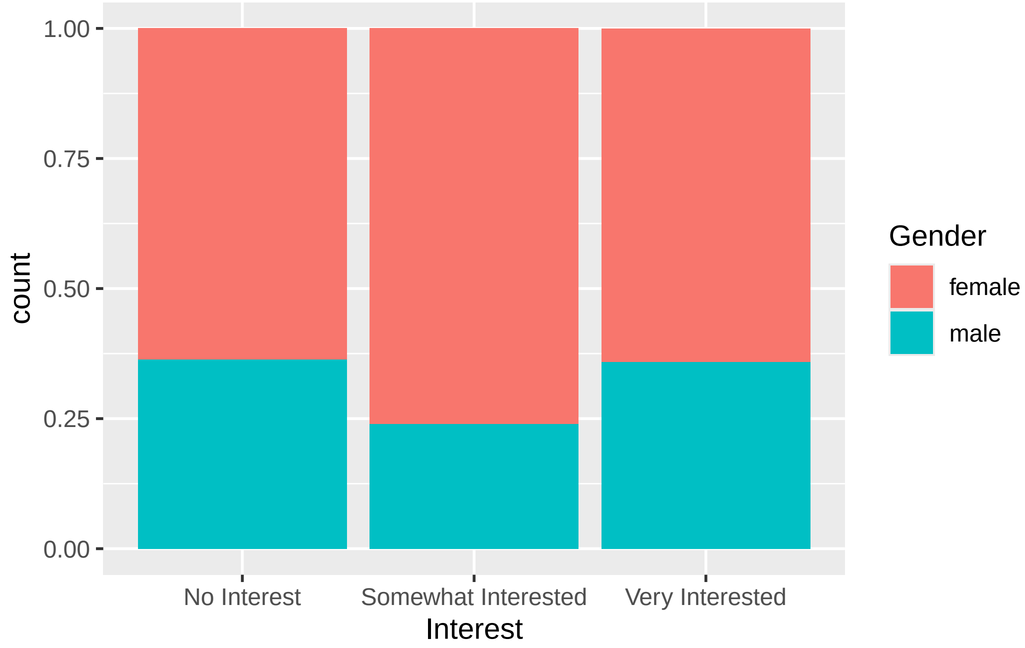
```
gf_bar(~Job, data = Fingers, fill = ~Gender, position = "fill")
```

```
gf_bar(~Job, data = Fingers, fill = ~Gender, position = "fill")
```



Consider the two plots below:





Compare and contrast how these two visuals explain variability in the *response* variable. In your answer, mention:

- What the plot shows
- How variation is partitioned or summarized
- What conclusions might be drawn about the role of the explanatory variable