Describing Distributions

With Graphs or Tables

April 16, 2012

- Distribution of a Variable
- Scales of Measurement
- Distribution of a Qualitative Variable
- Histograms
- What to look for in a Histogram
- Variation on Frequency Histograms
- Stem and Leaf Plot

1.0 Distribution of a Variable

Definition

The distribution of a variable tells us what values it takes and how often it takes these values.

1.0 Distribution of a Variable

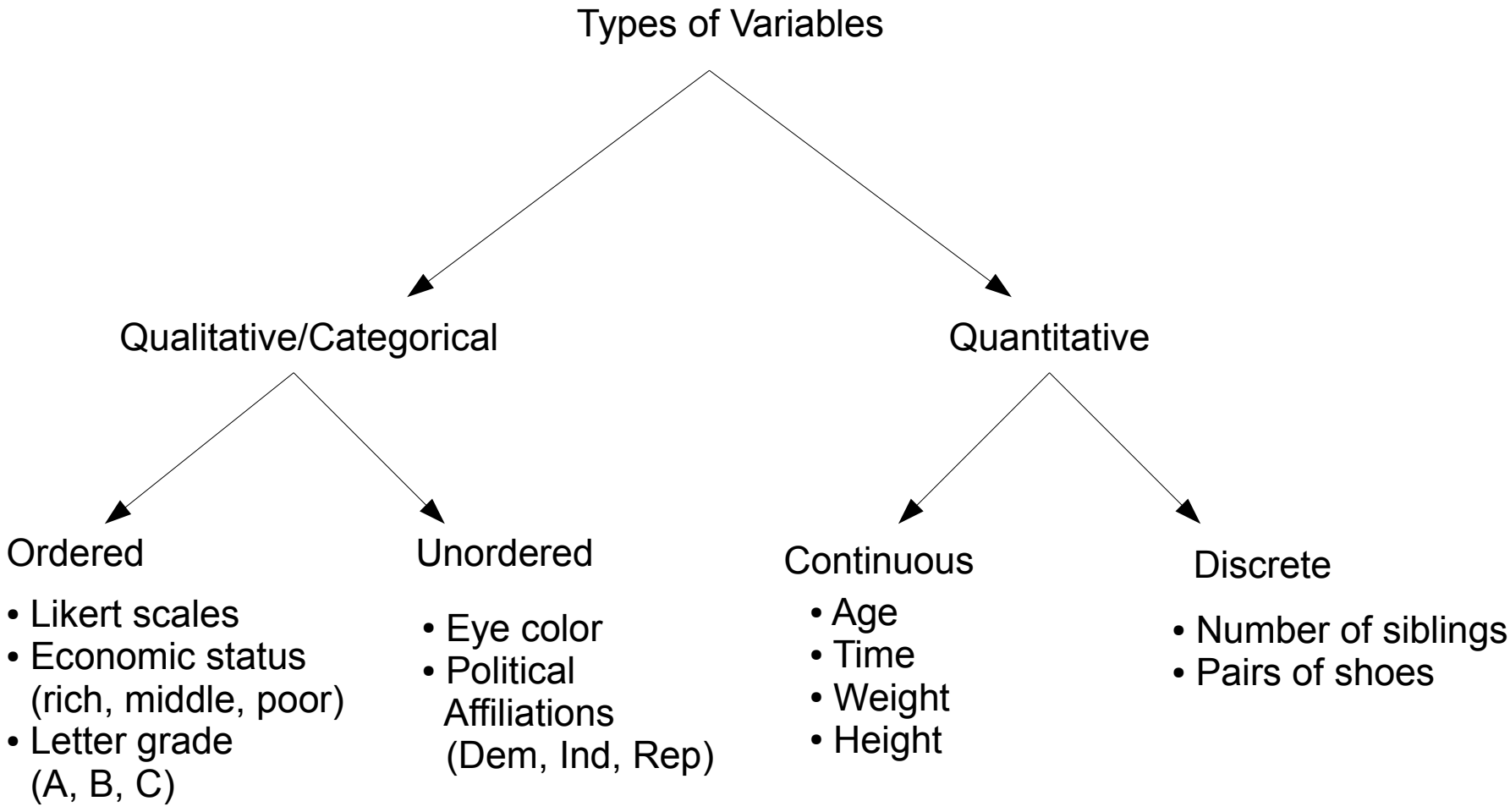
Table: On a scale of 1-10, how nervous are you about taking STAT 220? (1=very nervous, 10=not nervous)

Level of Nervousness	Number of Students	Percent of Total
1	3	0.04
2	3	0.04
3	5	0.07
4	7	0.09
5	15	0.20
6	4	0.05
7	9	0.12
8	11	0.15
9	7	0.09
10	10	0.13
Total	74	0.98

2.0 Scales of Measurement

- A categorical or qualitative variable places an individual into one of several groups or categories.
- The categories may have ordering in some cases. ordered categorical
- A quantitative variable takes numerical values for which arithmetic operations such as adding and subtracting makes sense.
- Quantitative variables can be continuous or discrete. For a continuous variable, the values can differ by any amount. For a discrete variable, the values can only differ by fixed amounts.

2.0 Scales of Measurement



2.1 Categorizing a Quantitative Variable

A study of the age distribution of the audience for social networking sites categorized age as under 25 years, 25 to 34 years, 35 to 49 years and over 49 years. Is age a quantitative or a qualitative variable in this context?

Age group	Facebook visitors	MySpace visitors
under 25 years	26.8%	44.4%
25 to 34 years	23.0%	22.7%
35 to 49 years	31.6%	23.5%
over 49 years	18.7%	9.4%

3.0 Distribution of a Qualitative Variable

The distribution of a categorical or qualitative variable lists the categories and gives either the count or the percent of individuals who fall in each category.

- Common ways to display the distribution of a categorical variable are:
  - Tables
  - Pie charts
  - Bar graphs (or plots)

3.1 Practicing Making a Distribution Table

- A survey of college freshmen in 2001 asked what field they planned to study. The results: 12.6%, arts and humanities; 16.6%, business; 10.1%, education; 18.6%, engineering and science; 12.0%, professional; 10.3%, social science; and 19.8%, other.

- What are the observational units? What is the variable on which data has been collected? What is its scale of measurement?

- Make a table showing the distribution of the variable.

3.1 What Makes a Clear Table?

- A caption that tells the content of the table.
- Labels within the table identify the variable clearly.
- The distribution of the variable is shown in numbers and also percents (or rates), if possible.
- The source of the data at the foot of the table adds credibility.

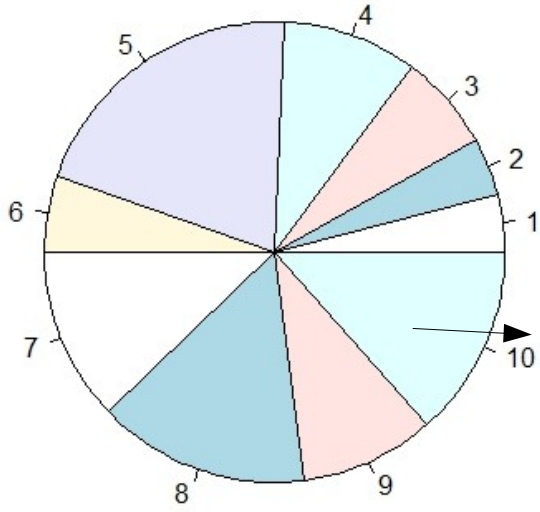
3.1 Distribution Table for Two Qualitative Variables

- In the survey of college freshmen, suppose in addition to the field of study, gender of the freshmen was also recorded.
- A CROSS-TAB shows the distribution of choice of field by gender.

	Gender		
Field of Study	Male	Female	Total Percent
Arts and Humanities	4.3%	8.3%	12.6%
Business	7.3%	2.8%	10.1%
Education	5.5%	13.1%	18.6%
Engineering and Science	8.4%	3.6%	12.0%
Professional	5.3%	5.0%	10.3%
Other	10.1%	9.7%	19.8%
Total %	40.9%	59.1%	100%

3.2 Pie Charts for Qualitative Variables

On a scale of 1-10, how nervous are you about taking STAT 220? (1=very nervous, 10=not nervous)



This area is 13% of the pie because 13% of students gave the answer ``10".

- Pie charts show the distribution of a categorical variable as slices of a "pie".
- Use a pie chart only when you want to emphasize each category's relation to the whole.

3.2 Pie Charts for Qualitative Variables

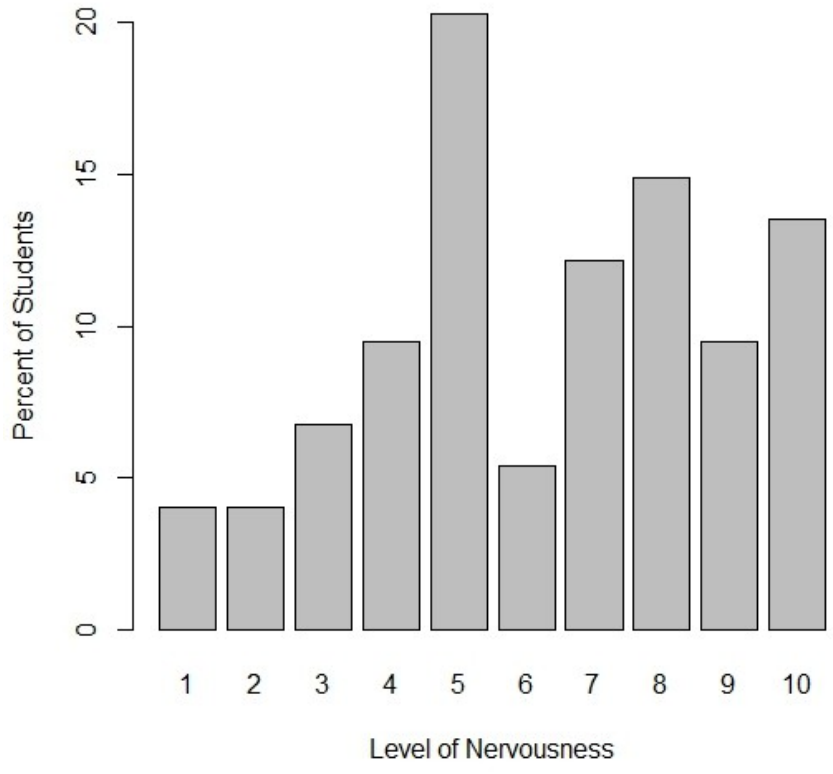
The Higher Education Research Institute's Freshman Survey includes over 200,000 full-time freshmen who entered college in 2009. The survey reports the following data on the sources students use to pay for college expenses.

Source for college expenses	Students
Family resources	78.2%
Student resources	62.8%
Aid – not to be repaid	70.0%
Aid – to be repaid	53.4%
Other	6.5%

Explain why it is not correct to use a pie chart to display this data.

3.3 Bar Graphs for Qualitative Data

On a scale of 1-10, how nervous are you about taking STAT 220? (1=very nervous, 10=not nervous)



- Bar graphs represent each category as a bar.
- The bar heights show the category counts or percents.
- Bar graphs can compare quantities that are not part of a whole.
- A Pareto bar graph has the bars ordered from tallest to shortest.

4.0 Histograms

- Tables work for categorical variables because these variables take relatively few values.
- Quantitative variables can take so many values that any meaningful display must group nearby values together into class intervals.
- The most common graph of the distribution of a quantitative variable is a HISTOGRAM.

4.1 Bar Graphs versus Histograms

- A histogram displays the distribution of a quantitative variable. The horizontal axis is marked in marked in the units of measurement of the variable.
- Draw histograms with no space, to indicate that all values of the variable are covered.

- A bar graph compares the sizes of different quantities. The horizontal axis need not have any measurement scale.
- Draw bar graphs with space in between the bars.

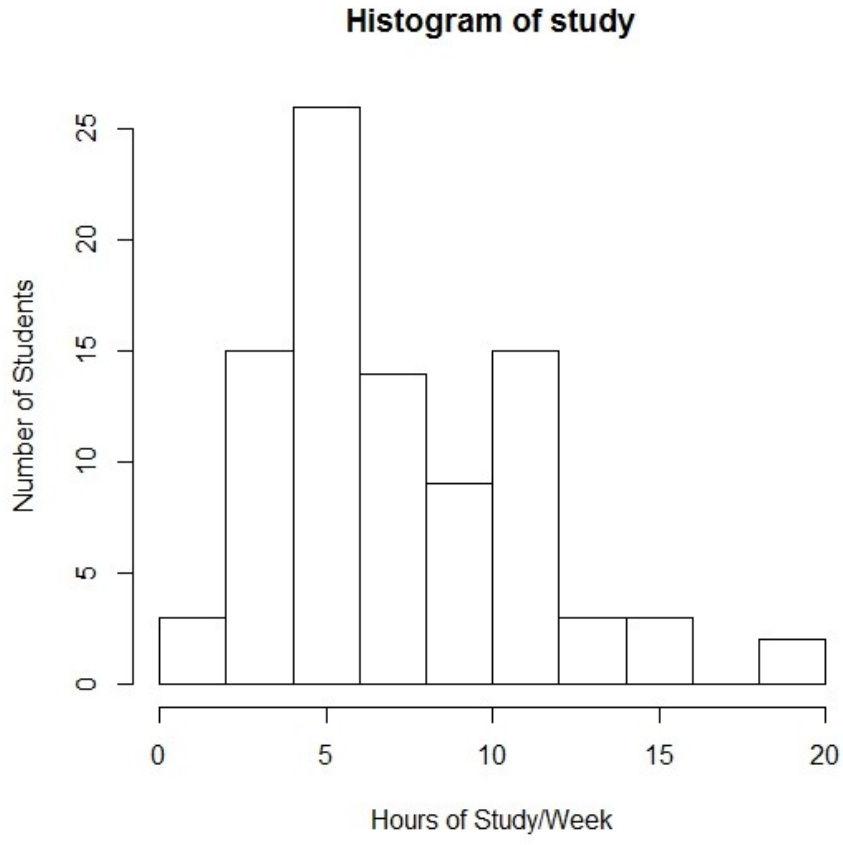
4.2 How to Make A Histogram

How many hours per week (h) will you study for STAT 220?
quantitative

Hours	Count	Proportion
$0 \leq h < 2$	3	0.03
$2 \leq h < 4$	15	0.17
$4 \leq h < 6$	26	0.29
$6 \leq h < 8$	14	0.15
$8 \leq h < 10$	9	0.10
$10 \leq h < 12$	15	0.17
$12 \leq h < 14$	3	0.03
$14 \leq h < 16$	3	0.03
$16 \leq h < 18$	0	0.00
$18 \leq h \leq 20$	2	0.02
Total	90	1.00

4.2 How to Make a Histogram

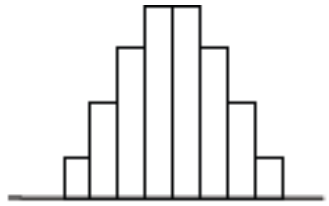
- The HEIGHT of each bar in a histogram corresponds to the count in each bin. This is called a FREQUENCY histogram.



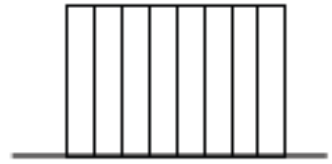
5.0 What to Look For in a Histogram

- Detect OUTLIERS, if any.
- Look at the SHAPE.
  - Is it symmetric? skewed?
- Where is it CENTERED? Where is the mid-point?
- How SPREAD out is it?

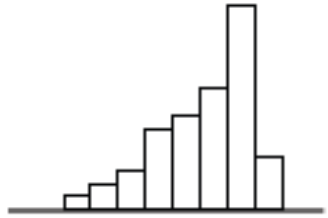
5.1 Basic Shapes of a Histogram



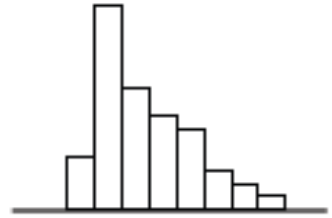
Normal, Triangular, Symmetric



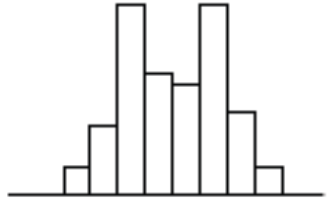
Uniform, Rectangular



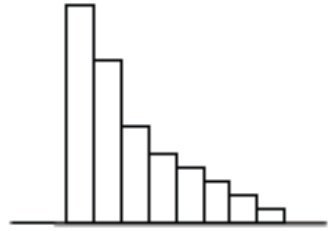
Skewed to left



Skewed to right



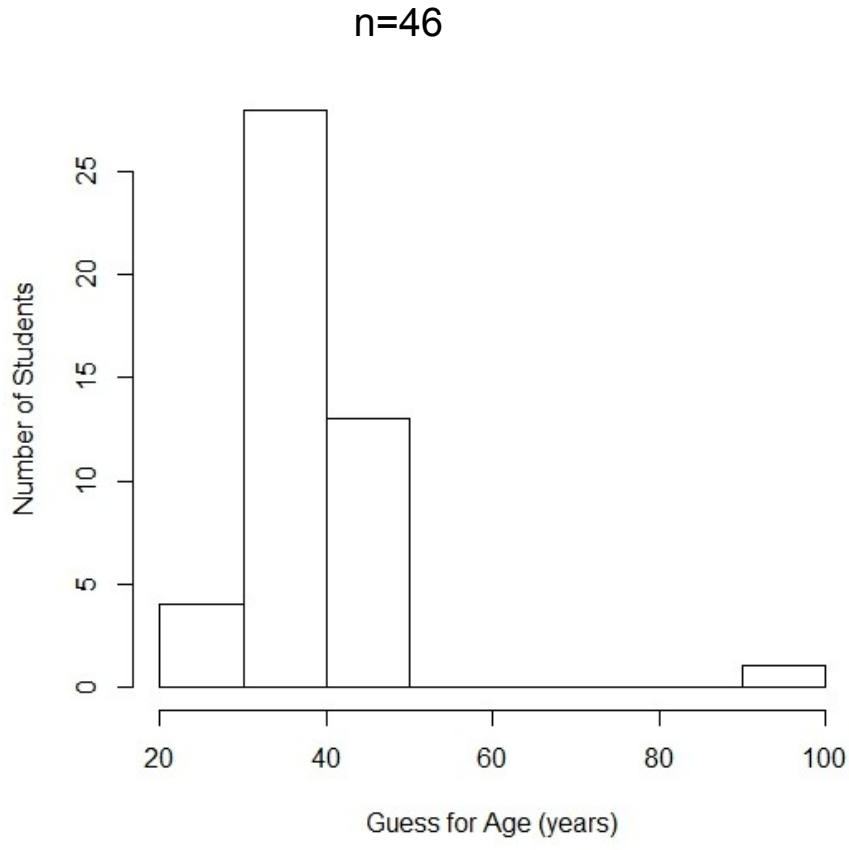
Bimodal



J-shaped

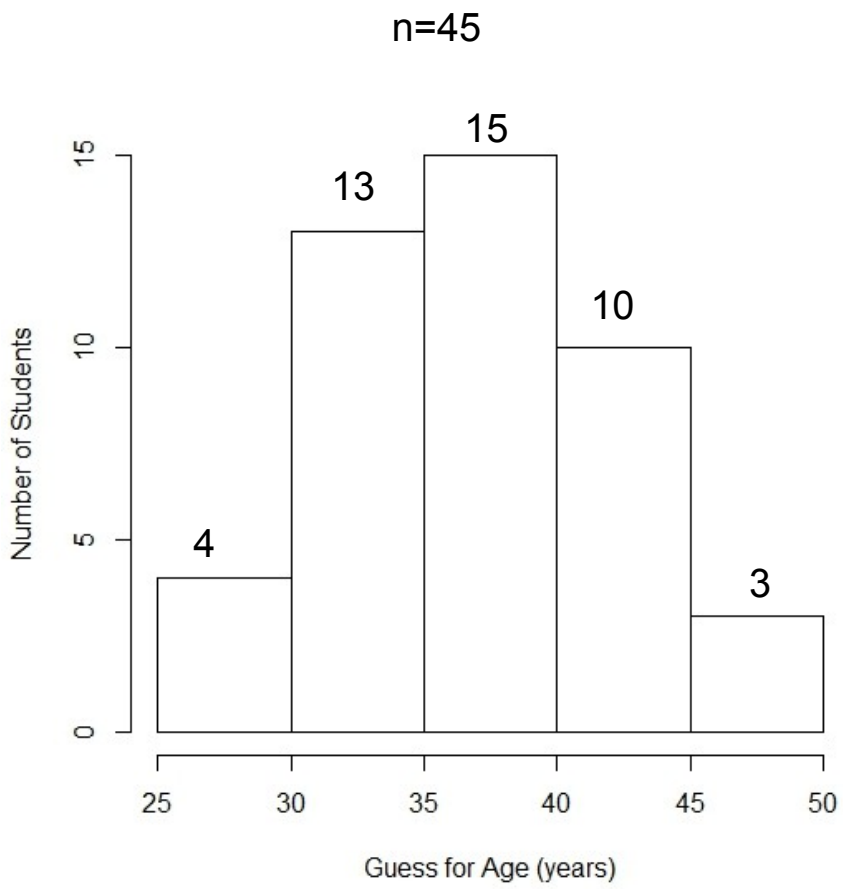
5.2 Describing A Distribution Ex. 1

Mystery Question: How old do you think Prof. Grove is?



5.2 Describing a Distribution Ex. 1

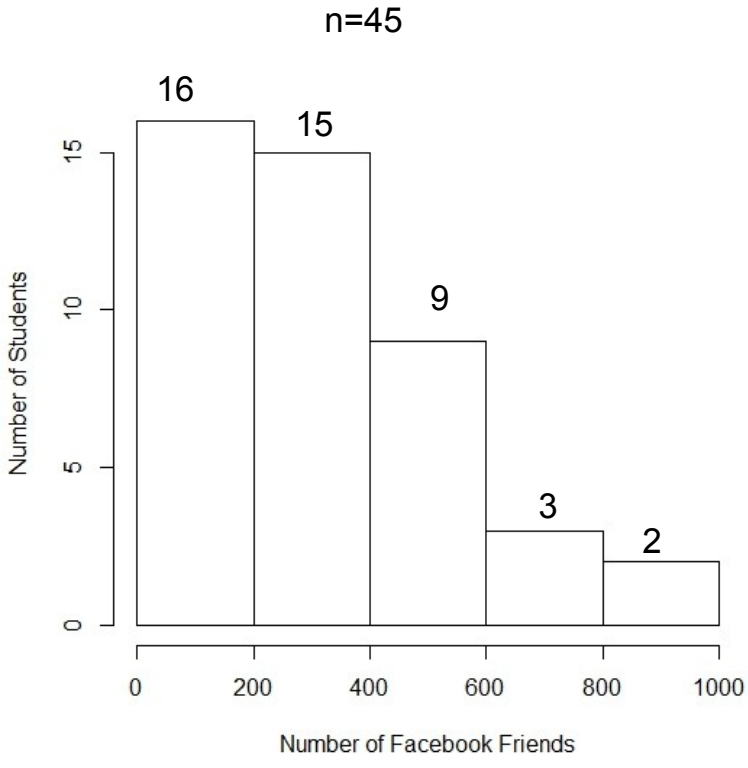
Mystery Question: How old do you think Prof. Grove is? (minus the outlier)

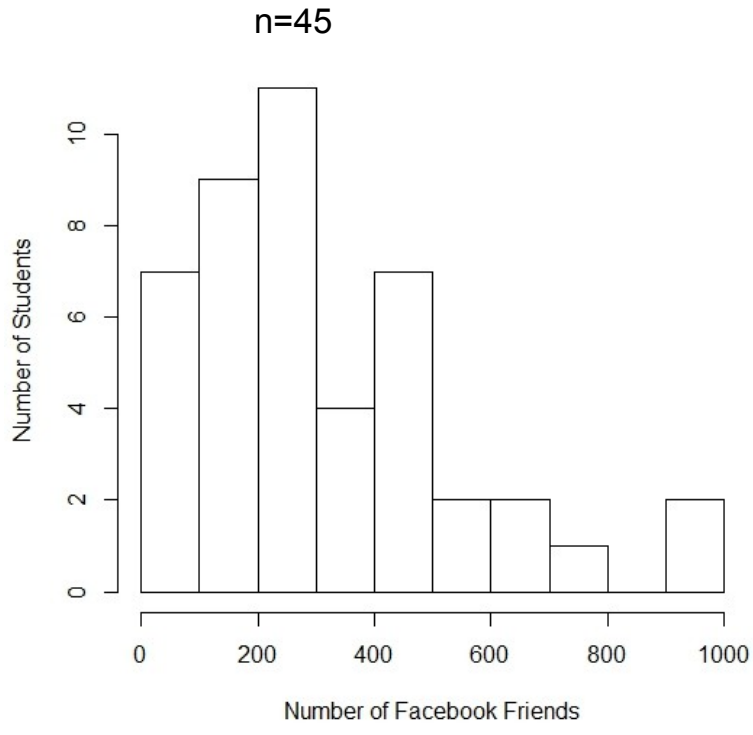


- **Shape:** The distribution has a single peak in the middle representing guesses of 35-40 years. It seems roughly symmetric.
- **Center:** Arranging the observations in order of size shows the mid-point is roughly 37 years.
- **Spread:** The range is from 27 years to 50 years.

5.2 Describing a Distribution Ex. 2

Mystery Question: How many facebook friends do you have?





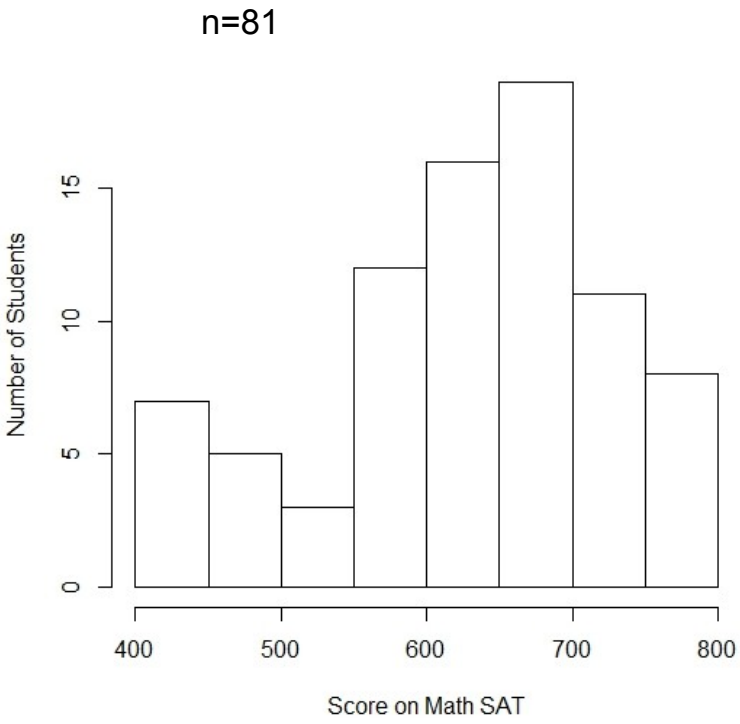
5.2 Describing a Distribution Ex. 2

Mystery Question: How many facebook friends do you have?

- When the class width is halved (figure to right), the individual with 1000 friends stands a bit apart.
- Are they an outlier or just the largest observation?
- We will only flag "strong" outliers.
- Describe the distribution of the number of facebook friends.

5.2 Describing a Distribution Ex. 3

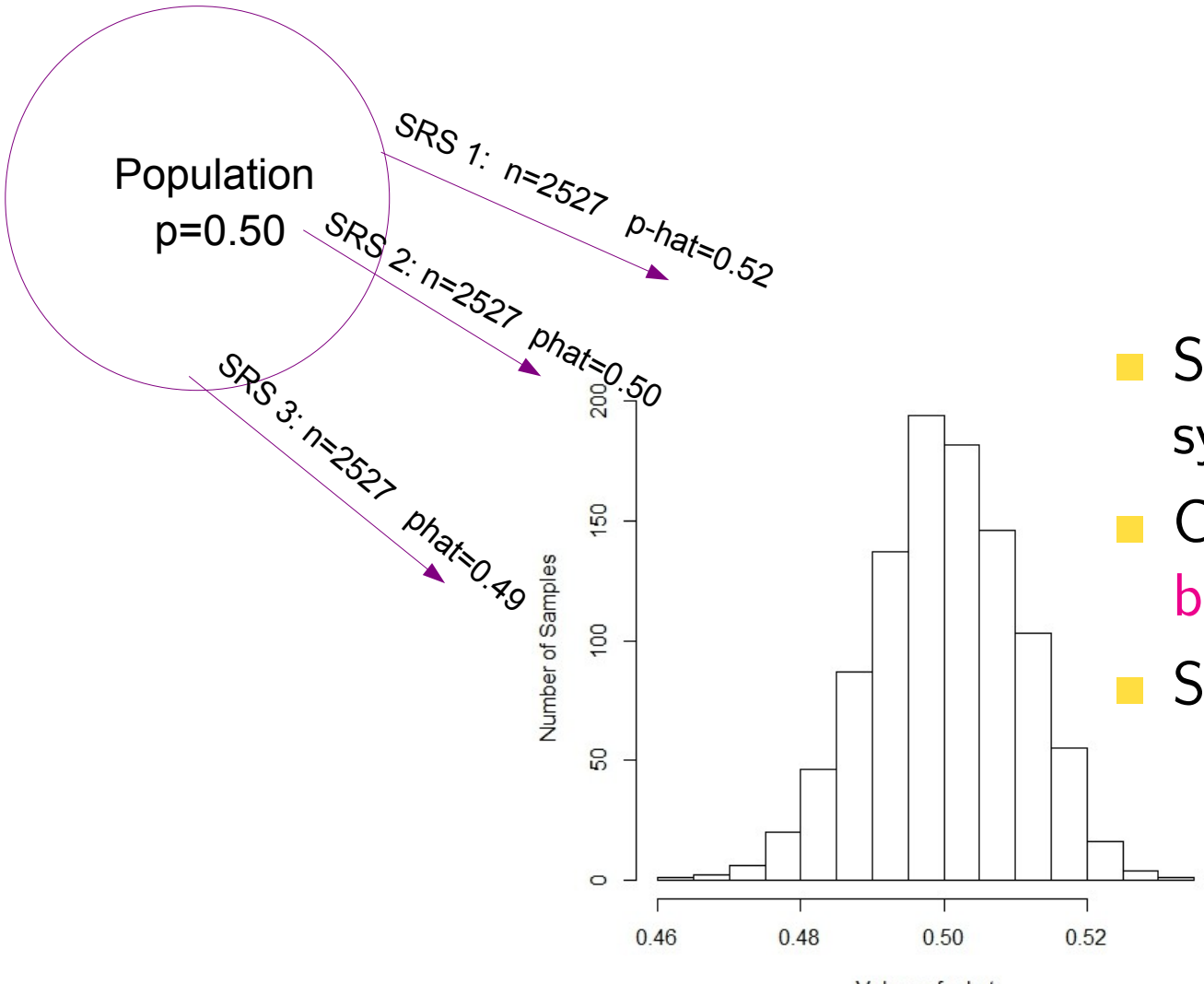
Mystery Question: What was your Math S.A.T. score?



- Single peak, somewhat skewed to the left.
- Center at a score of 650.
- Spread is 400-800.

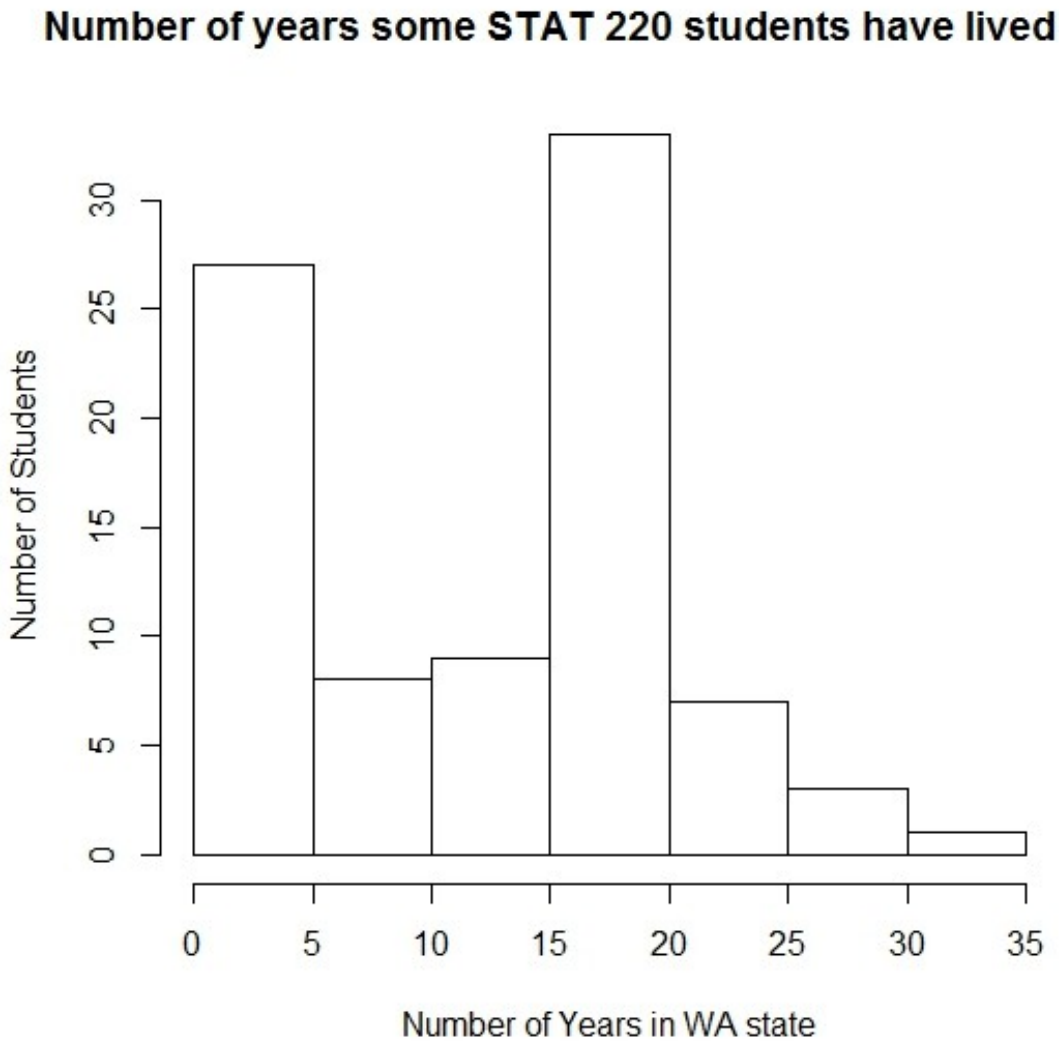
5.2 Describing a Distribution Ex. 4

- The values that a statistic takes in many random samples from the same population form a distribution with a pattern. sampling distribution



- Single peak and symmetric.
- Center at 0.5. lack of bias
- Spread is 0.46 - 0.54.

5.2 Describing a Distribution Ex. 5



- Two distinct peaks.
- Centers at around 3 and 18 years.
- Spread is 0-35 years.

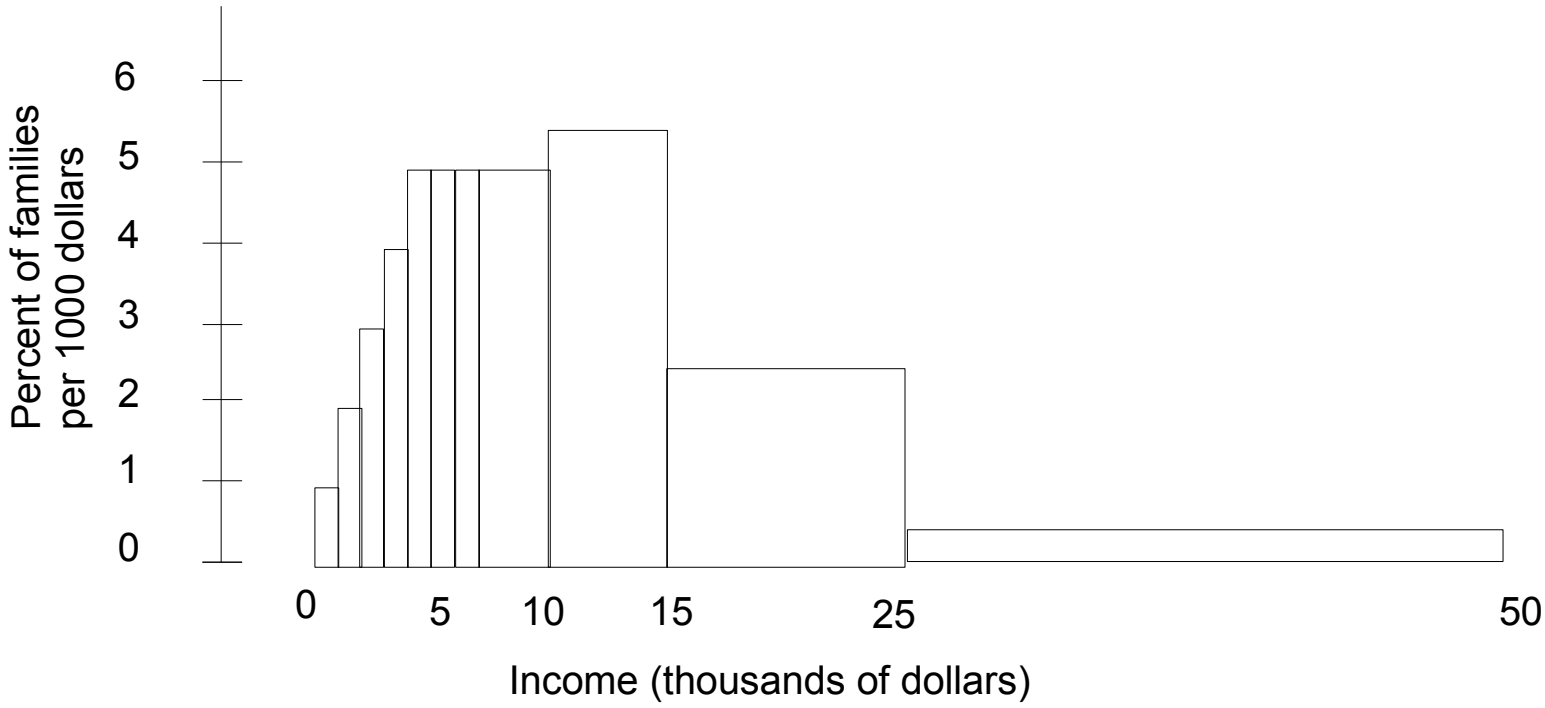
6.0 Variation on Frequency Histogram

Table: Distribution of family income in the U.S., 1973

Income level	Percent
$0 - $1,000	1
$1,000 - $2,000	2
$2,000 - $3,000	3
$3,000 - $4,000	4
$4,000 - $5,000	5
$5,000 - $6,000	5
$6,000 - $7,000	5
$7,000 - $10,000	15
$10,000 - $15,000	26
$15,000 - $25,000	26
$25,000 - $50,000	8
$50,000 and over	1

6.0 Variation on Frequency Histograms

- Since the class intervals are not of equal width, it is misleading to plot the counts versus class intervals.



- Instead, we plot the percent of families per thousand dollars on the vertical axis. density histogram

7.0 Stem and Leaf Plots

Mystery Question: How many hours do you plan to study per week for STAT 220?

The decimal point is at the |

```
 0 | 000
 2 | 000000000000000
 4 | 000000000000000000000000000
 6 | 00000000000000
 8 | 000000000
10 | 000000000000000
12 | 000
14 | 000
16 |
18 |
20 | 00
```

Describe the overall pattern. Are there any outliers?

7.1 Constructing a Stem and Leaf Plot

Separate each observation into a STEM, consisting of all but the final (rightmost) digit, and a LEAF, the final digit. Stems may have as many digits as needed, but each leaf only has a single digit.

Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column. Be sure to include all the stems needed to span the data, even when some stems will have no leaves.

Write each leaf in the row to the right of its stem, in increasing order out from the stem.