MATH 10

INTRODUCTORY STATISTICS

Ramesh Yapalparvi

Week 1

- **Chapter 1 – Introduction**

What is Statistics? Why do you need to know Statistics?

Technical lingo and concepts: Sampling, variables, percentiles, scales, distributions, summation, linear transformations, logarithms.

- **Chapter 2 – Graphing Distributions** ← **this lecture**

Visualizing data containing qualitative and quantitative variables. Histograms etc.

- **Chapter 3 – Summarizing Distributions** ← **this lecture**

Central tendency: mean, median, mode. Variability: variance.

Addendum 1 – Populations and Samples

Simple random sampling, Stratified sampling example.

E.g. Population = everyone in New Hampshire.

E.g. Simple random sample = pick 1000 at random.

E.g. Stratified = pick 500 male, 500 females at random (assuming 50/50 male/female in NH)

We hope the sample is representative of the population in the sense that if we want to study the distribution of heights or income in the population, we hope the sample would give a good estimate of that.

Our sample of 1000 is like a "mini" New Hampshire. We can use it study more than heights or income if we have collected the data.

Addendum 1 – Populations and Samples

Hypothetical Populations – Treatment and Control Group

One way of looking at it: population = everyone with disease taking the drug. Treatment group is the sample from this population. (we use the control group to deduce the effect of the drug)

Another way of looking at it: two populations = everyone with disease and taking drug, everyone with disease and **not** taking drug. Treatment and control group are samples from these two populations.

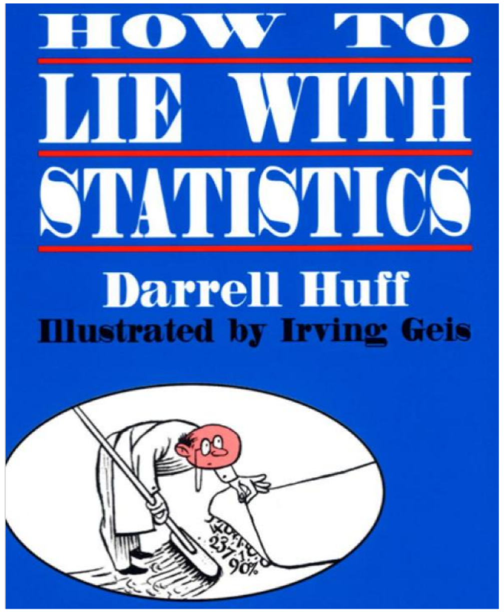
"Representative" in a restricted and precise sense: samples are used only for treatment efficacy. Unlike the "mini" New Hampshire in the previous slide, it is not meaningful to study other aspects of the hypothetical population here.

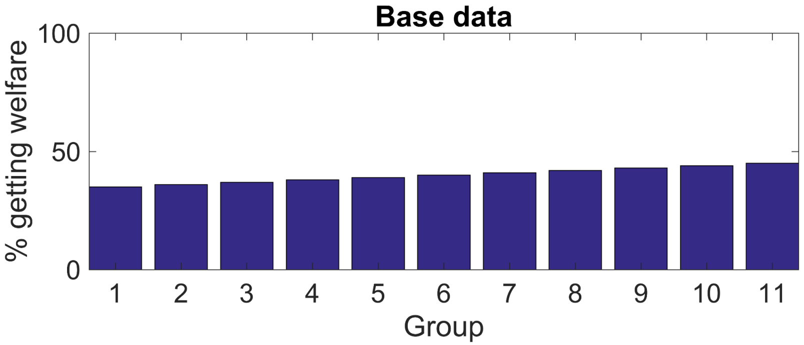
Chapter 2 – Graphing Distributions

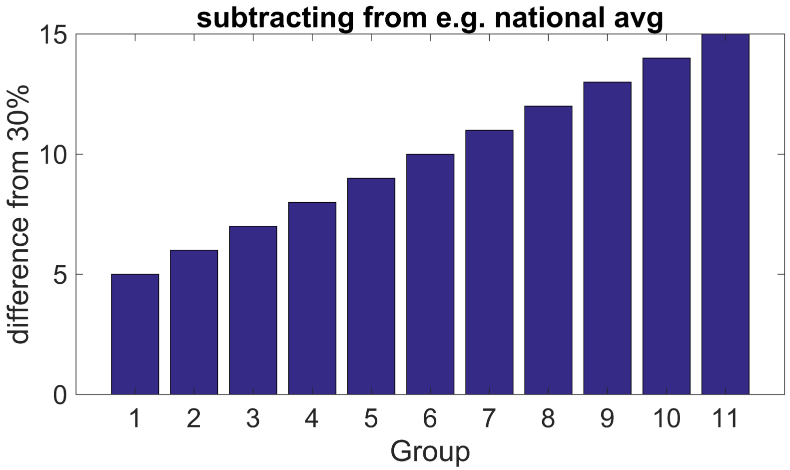
- Visualizing data helps us see patterns, support our conjectures, and can also help *sell* our ideas.
- https://www.youtube.com/watch?v=jbkSRLYSojo
- Hans Rosling, Statistician, on the BBC YouTube channel.

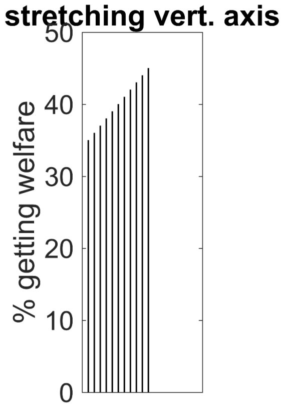
Sometimes, visualization sell our ideas a little *too well*.

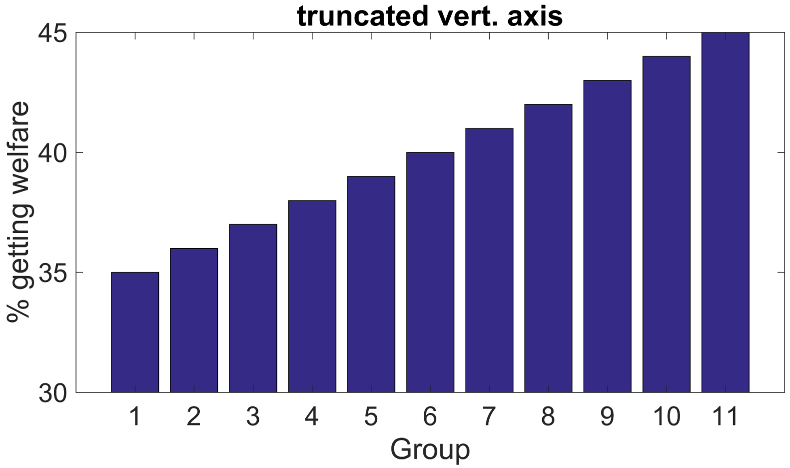
(chapter 2, section 2)









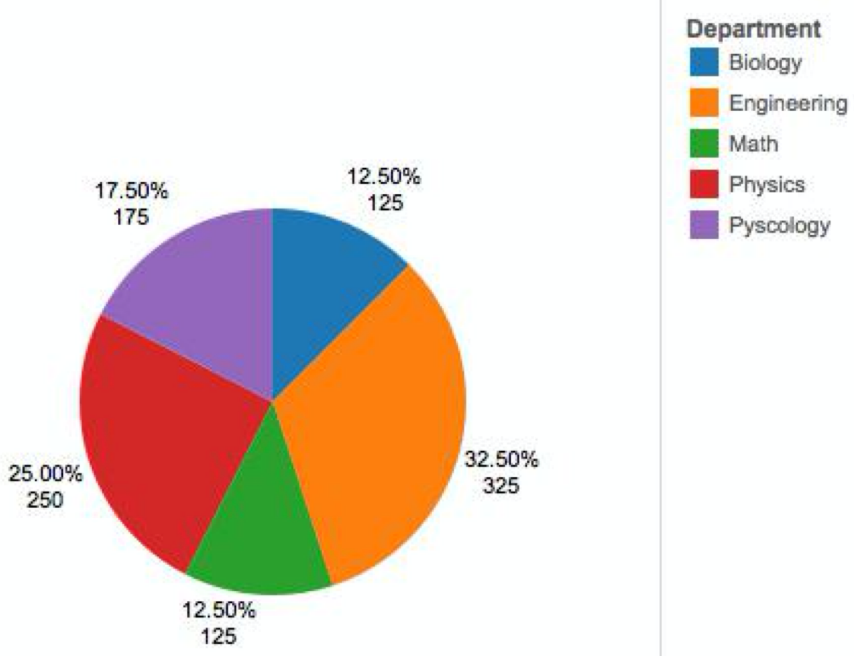


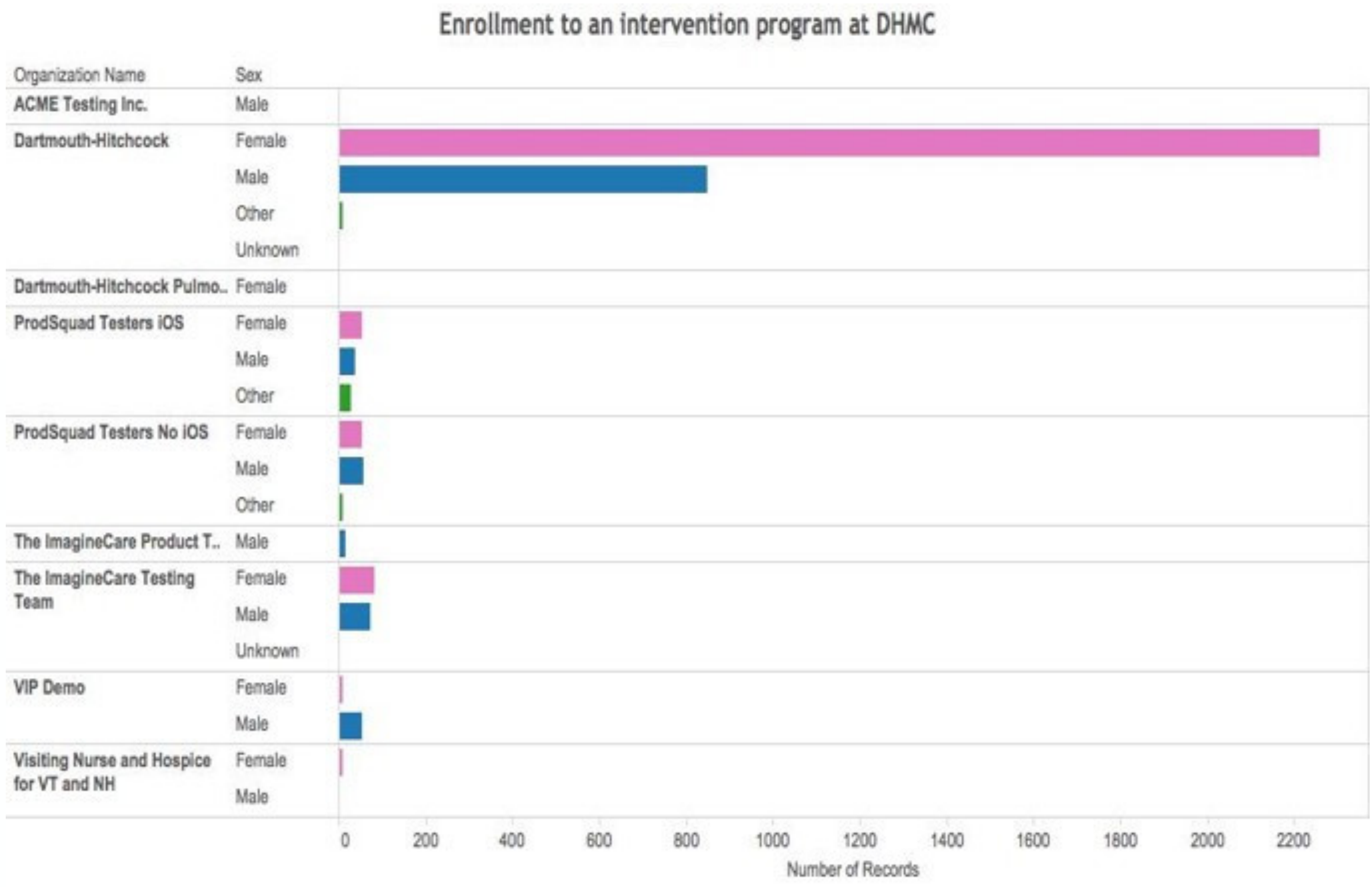
Chapter 2

- **Qualitative variables**
- Not numerical. Usually categories. E.g. hair color, favorite movie etc.
- Use frequency tables, pie charts, bar charts to visualize.

Department	Enrollment
Physics	250
Math	125
Engineering	325
Biology	125
Phycology	175

- **Qualitative variables**





Chapter 2

- **Qualitative variables**
- Not numerical. Usually categories. E.g. hair color, favorite movie etc.
- Use frequency tables, pie charts, bar charts to visualize.

- **Quantitative variables**
- Numbers! → we will see a lot of these in this course.
- Use stem and leaf, histograms, frequency polygons, box plots, bar charts, line graphs, dot plots.
- We will talk about histograms first then go on to Chapter 3 before giving a quick tour of the rest.

Quantitative variables – Histogram Part 1

- Very useful for visualizing the shape of a distribution when number of observations is large.
- **Textbook's example** : 642 students, scores ranged from 46 to 167. A simple frequency table will contain over 100 rows.
- Sort the N obervations into bins or classes intervals. For this course, we'll stick to the same width for each class.
- How many classes or bins? Trial and error → a.k.a. "the eyeball method".
- **Sturges' Rule** : as close to $(1 + log_2(N))$ classes as possible.
- **Rice Rule** : $2\sqrt[3]{N}$ classes. → can differ greatly from Sturge's Rule.
- These good to know, but don't worry, we won't ask you to memorize and state these in the exam.

Table 1. Grouped Frequency Distribution of Psychology Test Scores

Interval's Lower Limit	Interval's Upper Limit	Class Frequency
39.5	49.5	3
49.5	59.5	10
59.5	69.5	53
69.5	79.5	107
79.5	89.5	147
89.5	99.5	130
99.5	109.5	78
109.5	119.5	59
119.5	129.5	36
129.5	139.5	11
139.5	149.5	6
149.5	159.5	1
159.5	169.5	1

Quantitative variables – Histogram Part 2

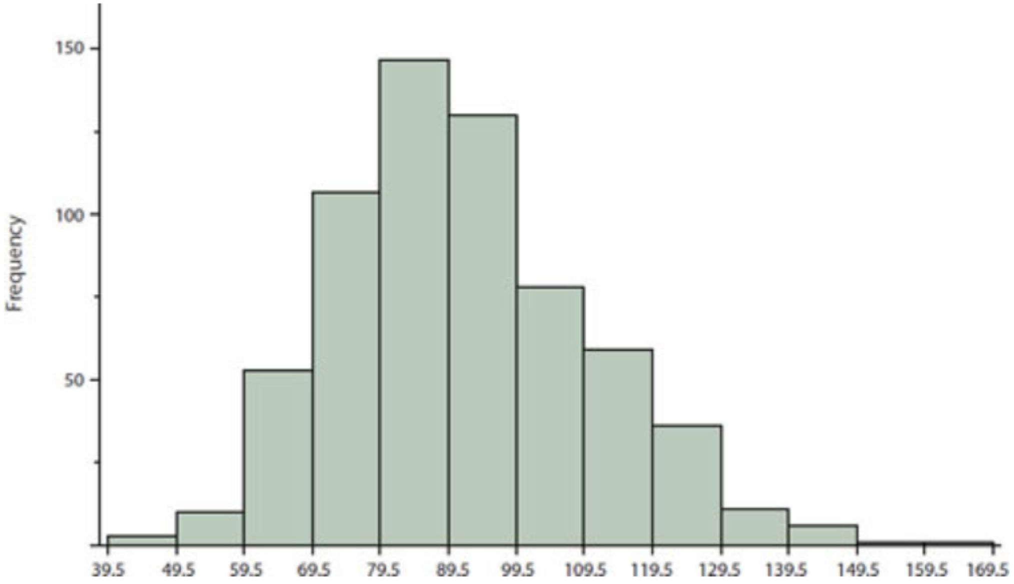


Figure 1. Histogram of scores on a psychology test.

- Vertical axis is the frequency or count for each class/bin.
- We can divide frequency by total number of observations, to get relative frequencies or proportions instead.
- E.g. Relative frequency or proportion of scoring between 69.5 and 79.5 is $107/642 = 0.1667$.

Table 1. Grouped Frequency Distribution of Psychology Test Scores

Interval's Lower Limit	Interval's Upper Limit	Class Frequency
39.5	49.5	3
49.5	59.5	10
59.5	69.5	53
69.5	79.5	107
79.5	89.5	147
89.5	99.5	130
99.5	109.5	78
109.5	119.5	59
119.5	129.5	36
129.5	139.5	11
139.5	149.5	6
149.5	159.5	1
159.5	169.5	1

Chapter 3 – Summarizing Distributions

- Summary statistics.

- How would you summarize your data?
- Perhaps visualize the data with a distribution. Say using a histogram.
- How can we summarize the distribution?

- **Central tendency** : is there "central" value that the data is centered around?
- **Variability** : how "spread" out is the data? How much variation?

Chapter 3, Section 2 to 9 – Central Tendency

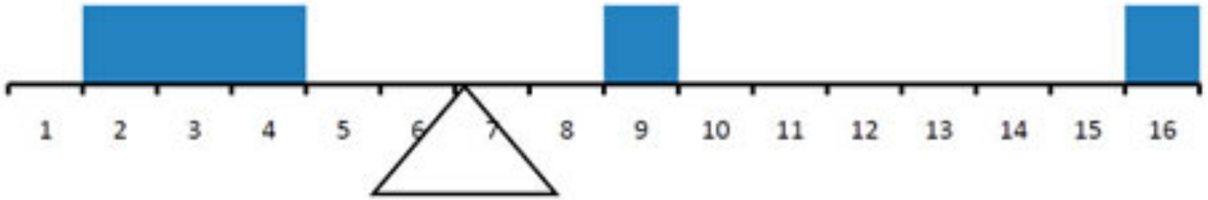
- We want some kind of estimate for the “center” of a distribution.
- We start with three heuristic notions to think about the “center”.
- Then, we will tie them to precise mathematical formulas.

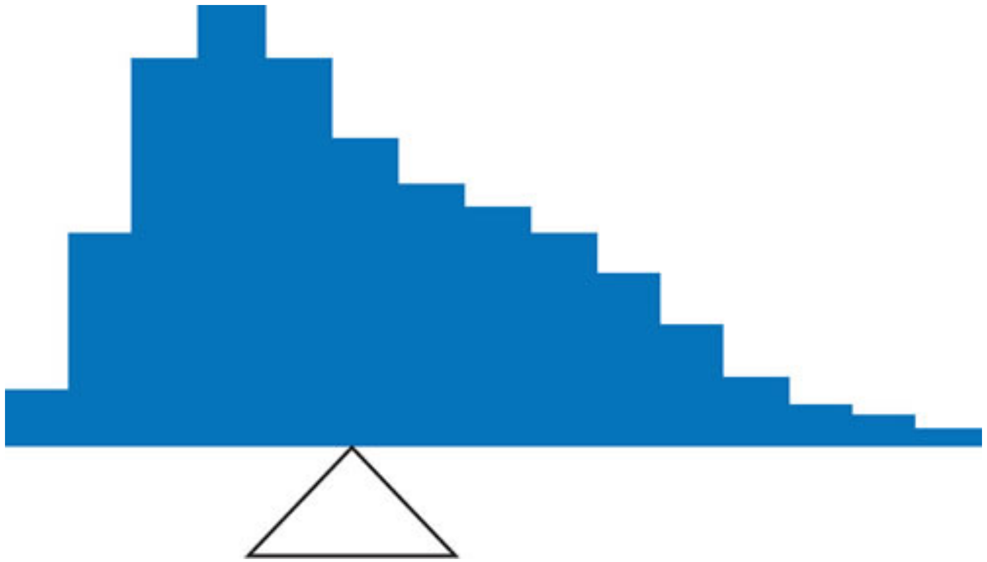
- **Balance Scale**

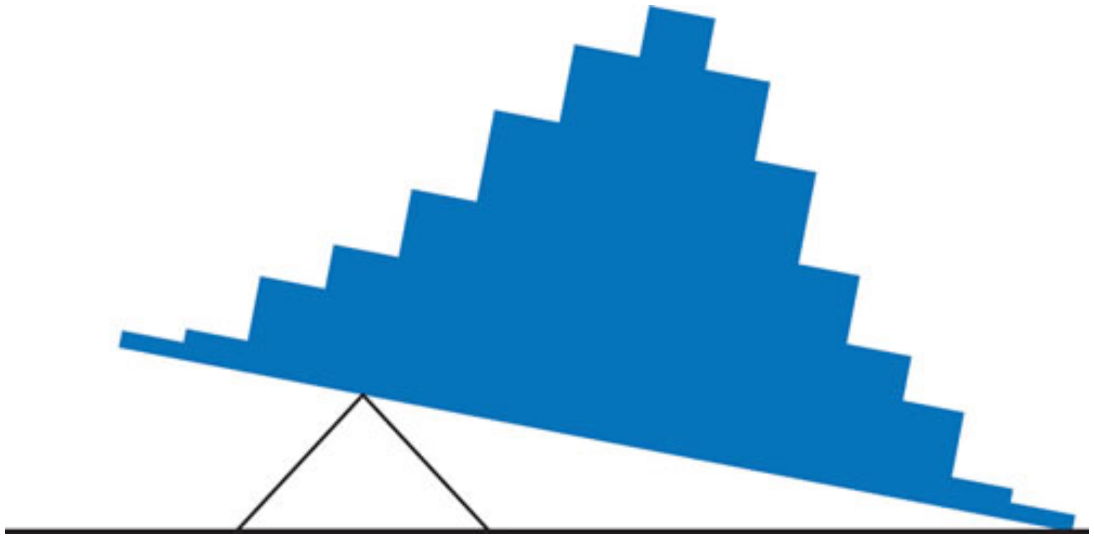
- **Sum of Absolute Deviations**

- **Smallest Squared Deviations**

Balance Scale







Sum of Absolute Deviations

- The “balance scale” heuristic is not very practical as we do not have an easy way to calculate the point at which the distribution balances.
- Another idea/heuristic : what if we look at the sum of absolute deviations from a particular number?
- Then find the number that minimizes this?

Table 2. An example of the sum of absolute deviations

Values	Absolute Deviations from 10
2	8
3	7
4	6
9	1
16	6
Sum	28

Sum of Squared Deviations

- A final idea/heuristic : what if we look at the sum of **squared** deviations from a particular number?
- Then find the number that minimizes this?
- Intuitively, how does this differ from the previous sum of absolute deviations?

Table 3. An example of the sum of squared deviations.

Values	**Squared Deviations from 10**
2	64
3	49
4	36
9	1
16	36
Sum	186

Chapter 3, Section 2 to 9 – Central Tendency

- We started with three heuristic notions to think about the “center”.
- Now, we can tie them to precise mathematical formulas.

- **Mean :** sum all the numbers, divide by number of observations.

- **Median :** order the numbers, find the observation in the middle. For even number of observations, we take the average of the middle two observations.

- **Mode :** the most frequent value. Can have no mode or multiple modes.

Chapter 3, Section 2 to 9 – Central Tendency

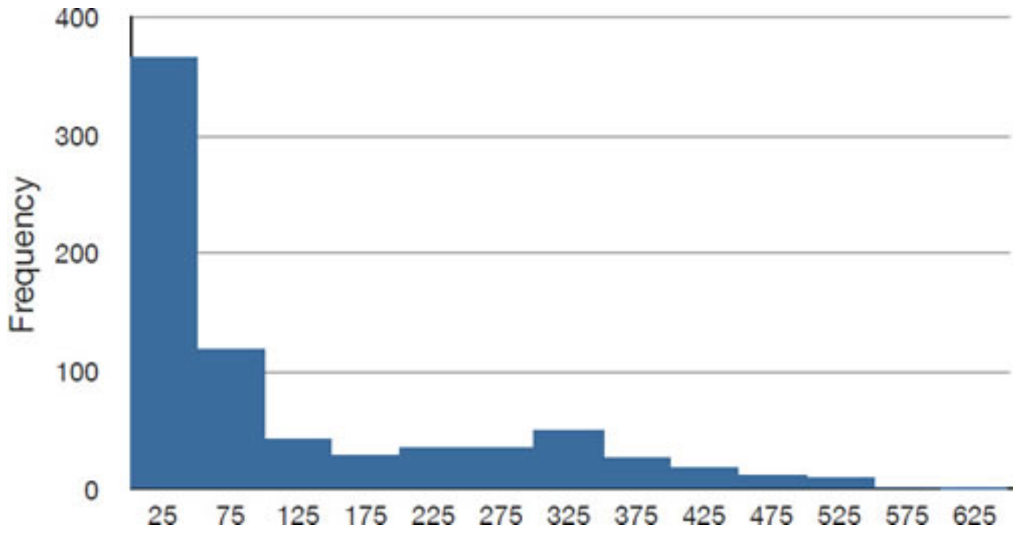
- How does the mean and median connect to our heuristics about central tendency?
- Chapter 3, section 8.

- **Mean :** balances the scale, minimize the sum of squared deviations.

- **Median :** minimizes the sum of absolute deviations only.

Chapter 3, Section 11 – Comparing mean, median and mode.

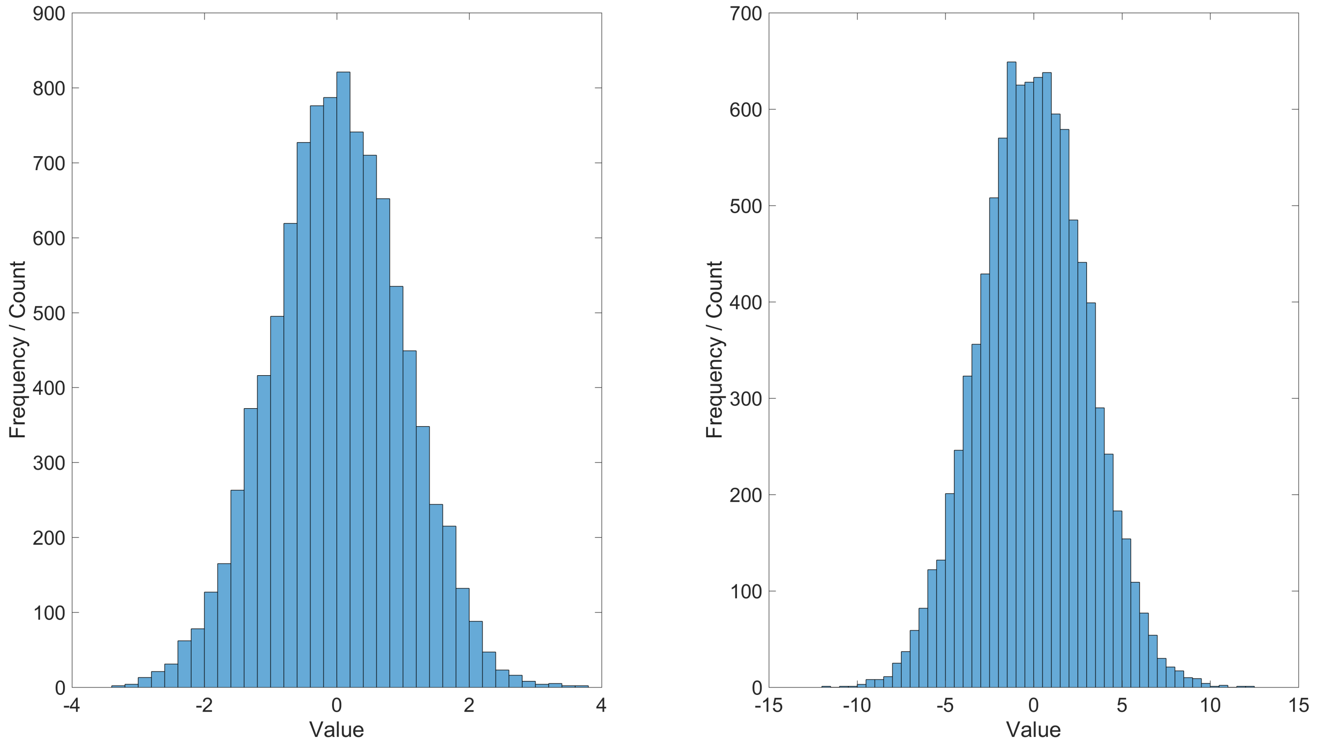
- How do they relate to the shape of the distributions?
- **Symmetric distributions** : mean, median, mode are equal or close to equal.
- **Skewed distributions** : depends. E.g. large positive skew shown below.



- In this case, the mean is greater than the median.
- Intuition : the large numbers in long positive tail pushes the mean up.

Variability – Chapter 3, Section 12 & 13

- We have a notion of central tendency or the "center" of a distribution.
- The two distributions below are both centered at zero, but can you see the difference in the values on the horizontal axis? How can we mathematically quantify this difference?



Variability – Chapter 3, Section 12 & 13

- **Range** : highest minus lowest value
- **Interquartile Range** : 75th percentile minus 25th percentile
- **Population Variance** :

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$$

- **Estimate of the Variance :**

$$s^2 = \frac{\sum(X-M)^2}{N-1}$$

- There are alternative formulas for these in the textbook section.
- **Standard deviation (SD) :** square root of the variance (for population or estimate of the SD)

Two Misc. Things…

1) Estimator of Variance: Why divide by $(N-1)$?

Dividing by just N underestimates the variance. Intuition : the sample mean M was chosen to minimize the sum of square deviations in the first place.

2) What is the mean deviation from the mean? (for population and sample)

$$\frac{\sum(X-\mu)}{N}, \qquad \frac{\sum(X-M)}{N}$$

What do linear transformations do to mean and variance?

- Chapter 3, section 18.

- Y is a linear transformation of X, if it is of the form $Y = bX + a$, where a, b are fixed numbers.
- Y is now a new variable.

- If variable X has mean μ, then the new variable Y will have mean b μ + a.
- If variable X has variance σ^2, then the new variable Y will have variance $b^2\sigma^2$.

Summing Variances - Chapter 3, section 19

- If X and Y are **uncorrelated** then we can apply the sum formula :

$$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y$$

- Examples from the textbook :

1) Thousands of samples of randomly paired unrelated measurements.

2) Thousands of samples of verbal and quantitative SAT scores, each pair belonging to the same person. → correlated!

Will be revisiting this formula when we do “bivariate” data in Chapter 4.

Quantitative variables – Stem and Leaf Display

Table 1. Number of touchdown passes.

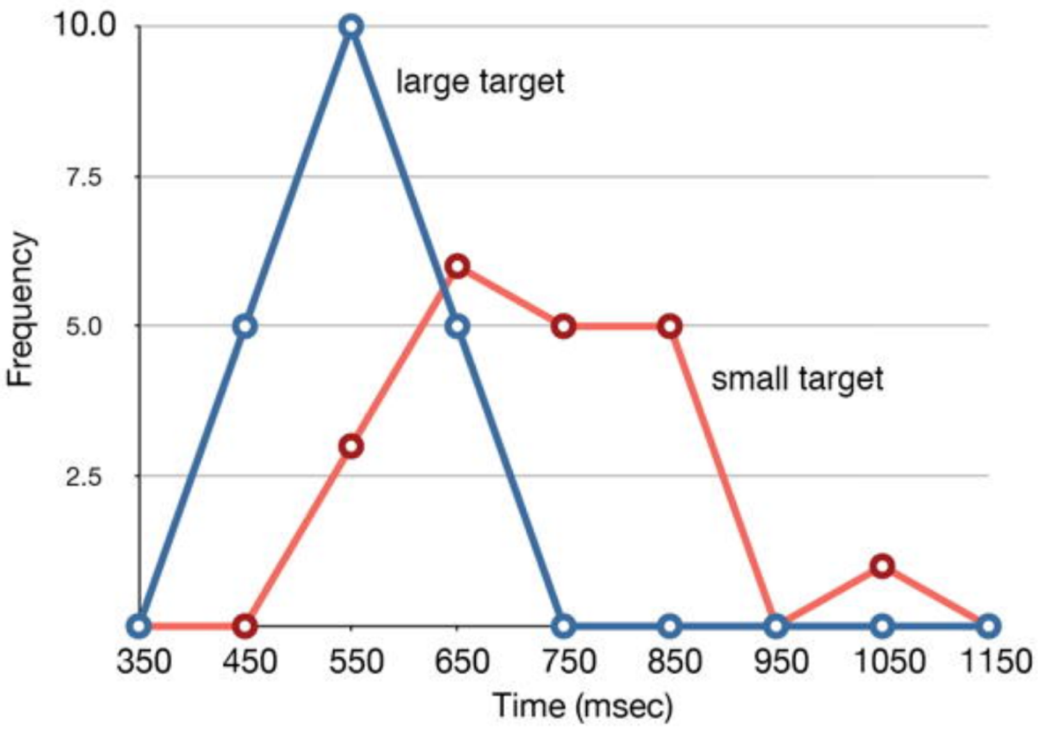
```
37, 33, 33, 32, 29, 28, 28, 23, 22,
22, 22, 21, 21, 21, 20, 20, 19, 19,
18, 18, 18, 18, 16, 15, 14, 14, 14,
12, 12, 9, 6
```

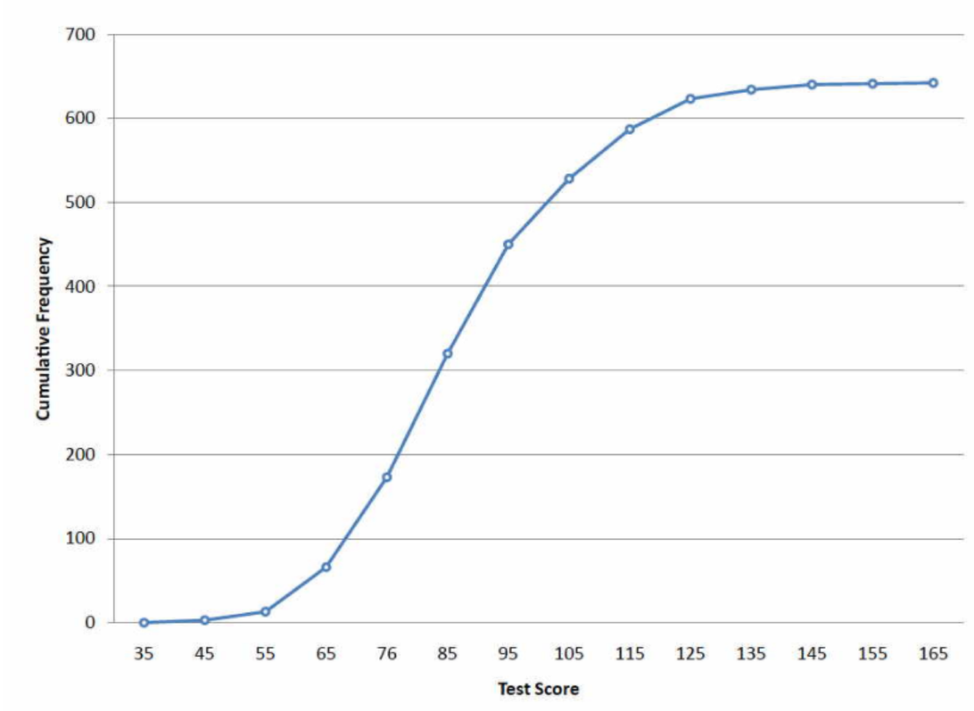
```
3|2337
2|001112223889
1|2244456888899
0|69
```

Figure 1. Stem and leaf display of the number of touchdown passes.

Quantitative variables – Frequency Polygon

- Frequency polygon → start with a histogram, then connect the midpoint of each bar.
- Good for overlapping two set of data for comparison.





← Cumulative frequency polygon

Quantitative variables – Box Plot

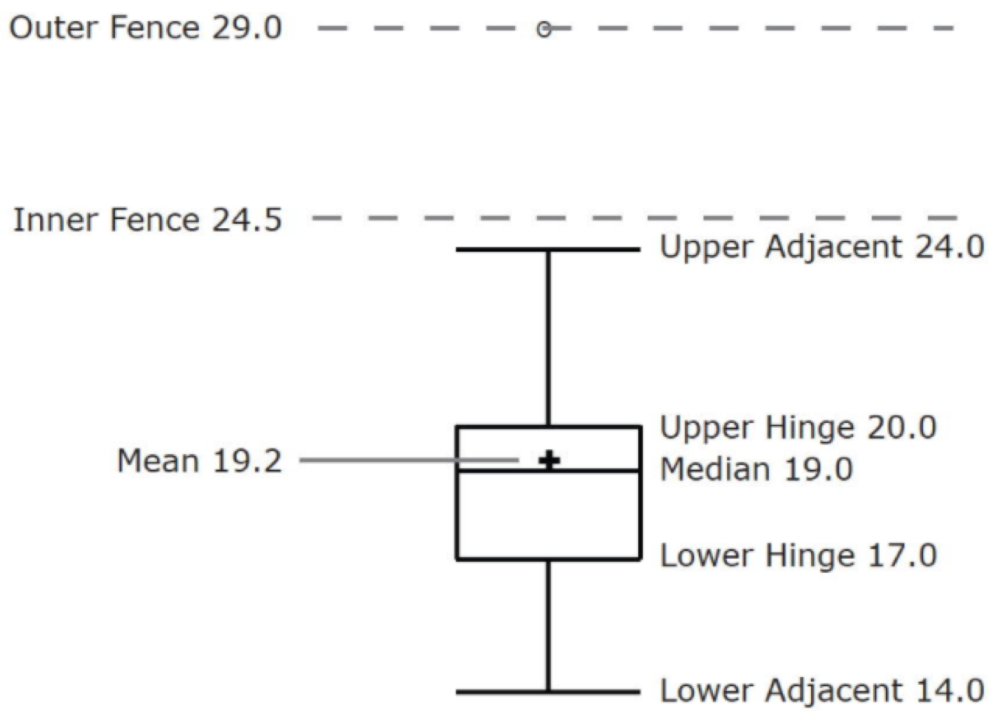


Figure 5. The box plot for the women's data with detailed labels.

- For the exam : Median, upper/lower Hinge, upper/lower adjacent, mean.
- This subset of feature form a “box and whiskers plot”.

Name	Formula
Upper Hinge	75th Percentile
Lower Hinge	25th Percentile
H-Spread	Upper Hinge - Lower Hinge
Step	1.5 x H-Spread
Upper Inner Fence	Upper Hinge + 1 Step
Lower Inner Fence	Lower Hinge - 1 Step
Upper Outer Fence	Upper Hinge + 2 Steps
Lower Outer Fence	Lower Hinge - 2 Steps
Upper Adjacent	Largest value below Upper Inner Fence
Lower Adjacent	Smallest value above Lower Inner Fence
Outside Value	A value beyond an Inner Fence but not beyond an Outer Fence
Far Out Value	A value beyond an Outer Fence

Quantitative variables – Bar Charts

- We have seen this numerous times. Refer to your textbook Chapter 2, Section 9 for more details.
- Examples from textbook to show how bar charts can be used to display data that are not just frequencies/counts.

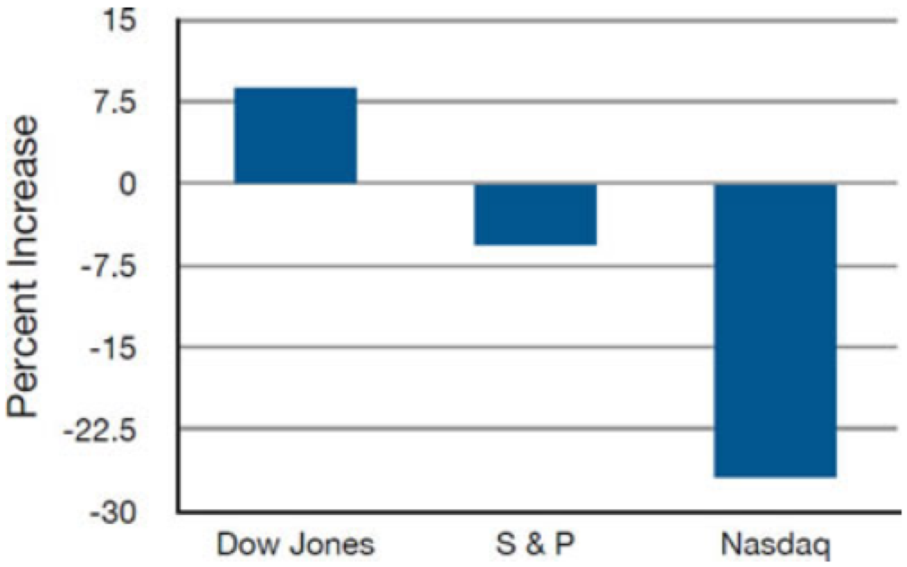


Figure 2. Percent increase in three stock indexes from May 24th 2000 to May 24th 2001.

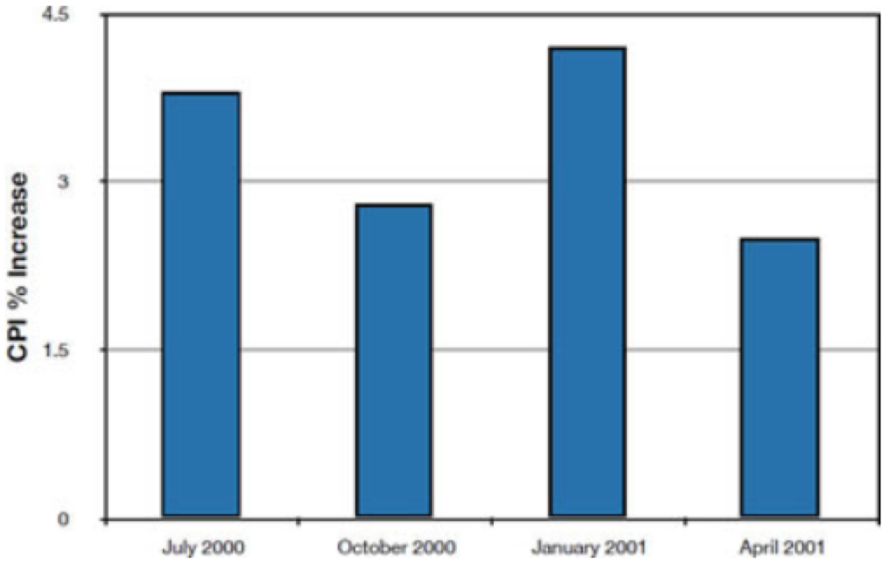
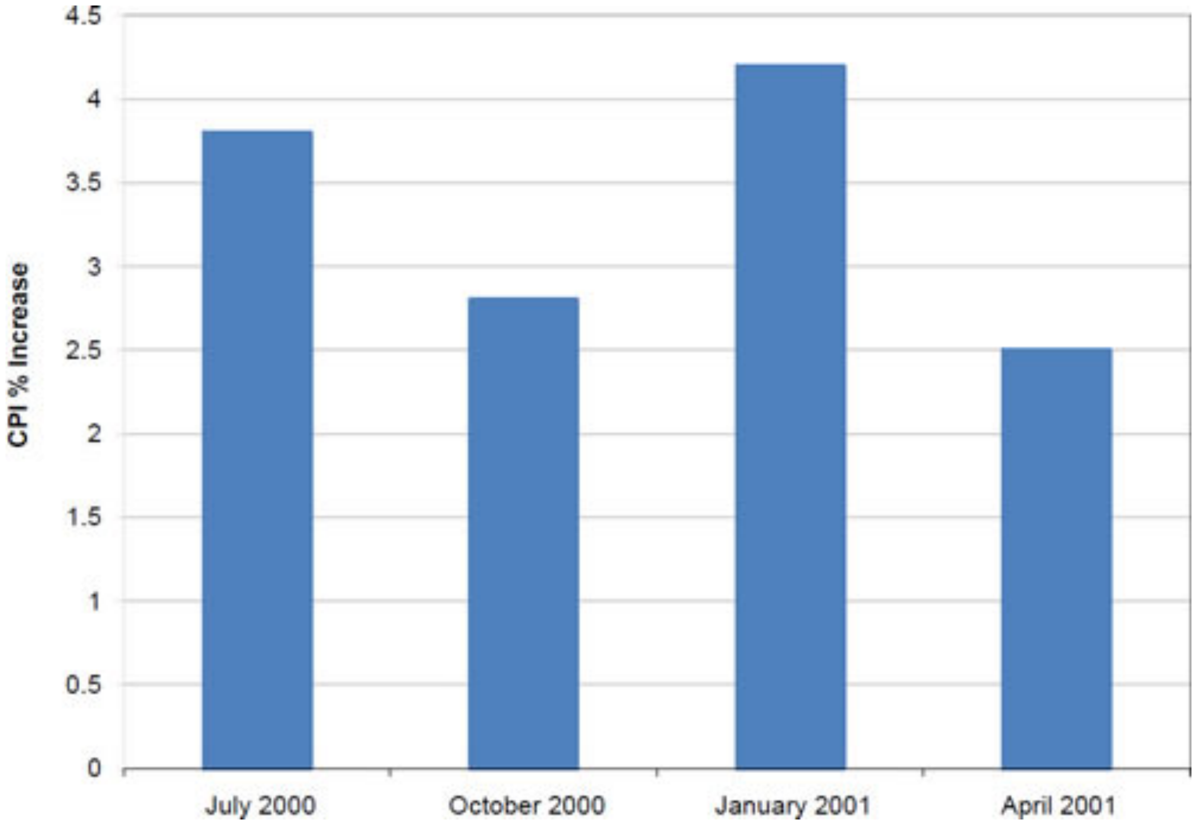
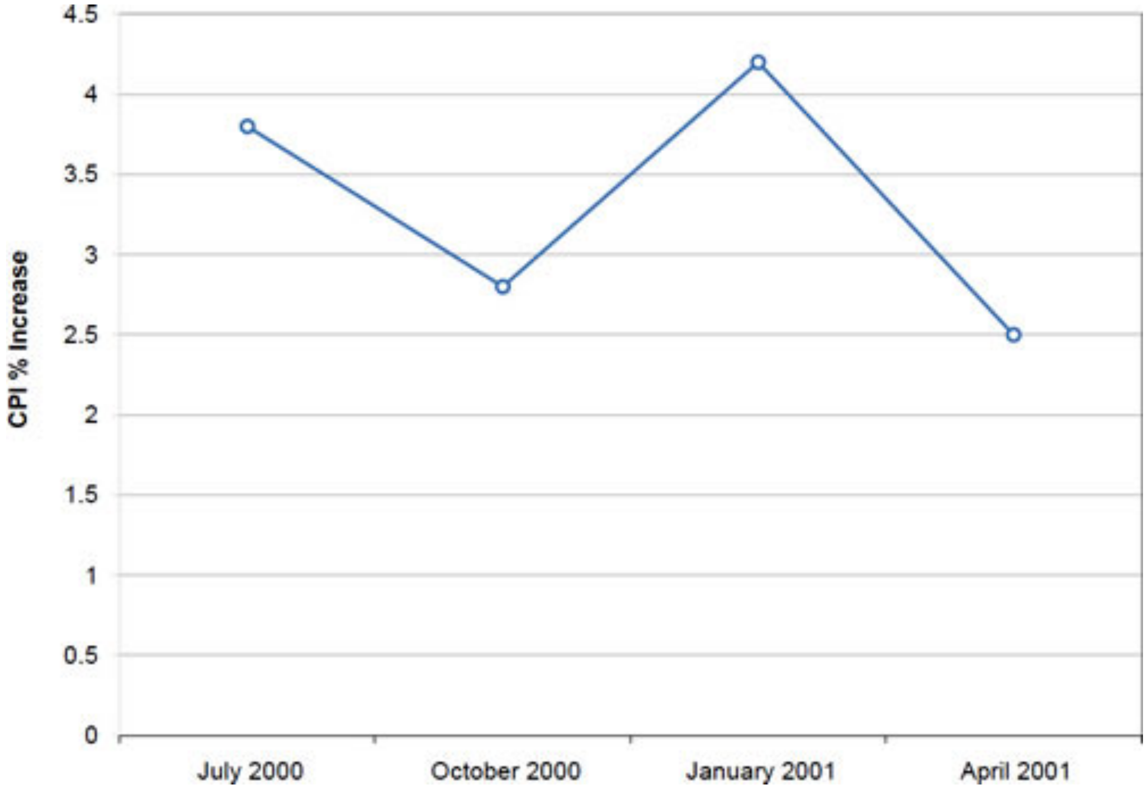


Figure 3. Percent change in the CPI over time. Each bar represents percent increase for the three months ending at the date indicated.

Quantitative variables – Line Graph

- Simply bar graphs with the top of the bars represented by points (jointed by lines) instead of bars.





Quantitative variables – Dot/Scatter Plot

- Very useful for Chapter 4 – Bivariate Data

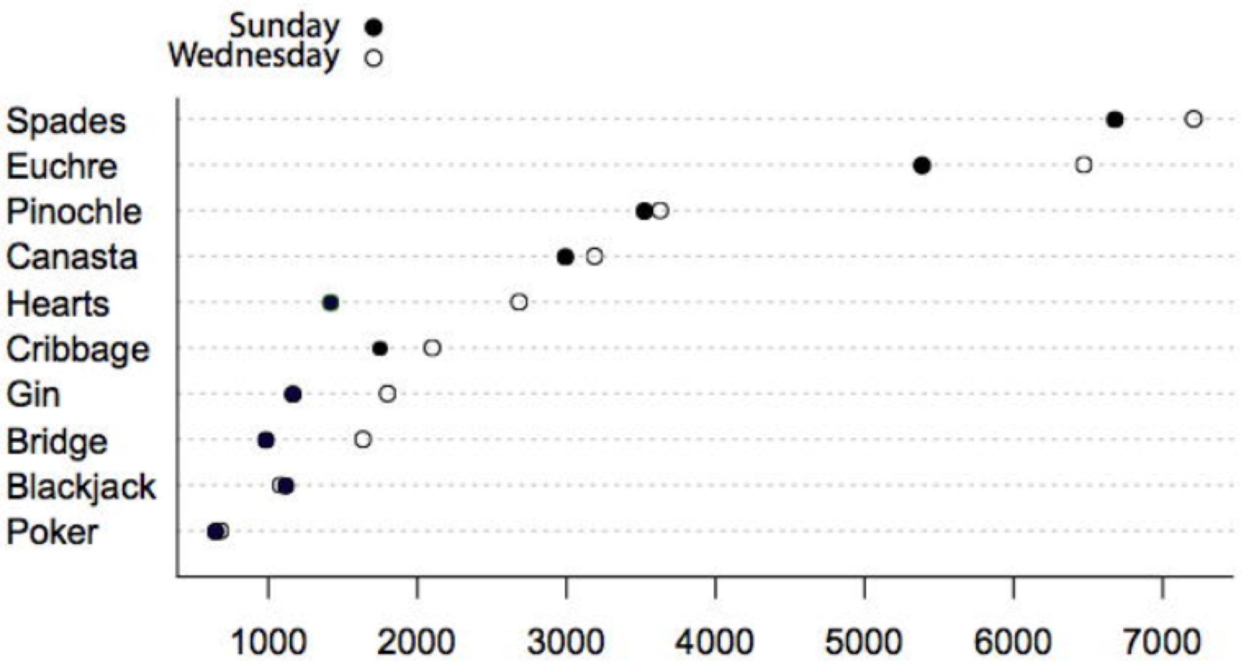


Figure 4. An alternate way of showing the number of people playing various card games on a Sunday and on a Wednesday.

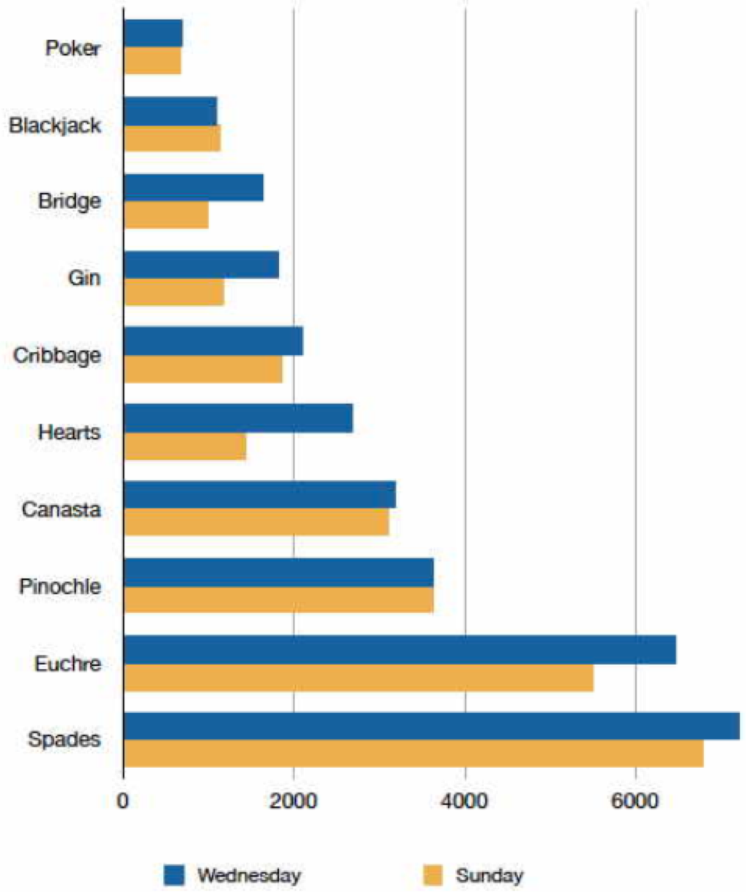


Figure 3. A bar chart of the number of people playing different card games on Sunday and Wednesday.