

## II. Describing Data

PBAF 527u  
Winter 2005

---

---

---

---

---

---

---

---

## Describing Data

1. Keeping In Touch/Policy Report
2. Who's in this class?
3. Tools for Discrete Variable Values
  1. Frequencies (counts)
  2. Relative Frequencies
  3. Pie Charts
  4. Bar Charts
4. Tools for Continuous Variable Values
  1. Frequency Distribution
  2. Histogram
  3. Measures of Central Tendency
  4. Measures of Variability
  5. Numerical Measures of Relative Standing
5. Presenting and Understanding Data

---

---

---

---

---

---

---

---

## Keeping in Touch

Class website  
<http://courses.washington.edu/pba527a/>  
<http://course.washington.edu/pbafrgk/527>

Class Listserv  
[pba527u\\_wi05@u.washington.edu](mailto:pba527u_wi05@u.washington.edu)

Rachel Kleit's email  
[kleit@u.washington.edu](mailto:kleit@u.washington.edu)  
Siri Erickson-Brown's email  
[Sirieb@u.washington.edu](mailto:Sirieb@u.washington.edu)

Policy Report and  
Homework  
Assignment 1

---

---

---

---

---

---

---

---

## Describing Data

1. Keeping In Touch/Policy Report
2. Who's in this class?
3. Tools for Discrete Variable Values
  1. Frequencies (counts)
  2. Relative Frequencies
  3. Pie Charts
  4. Bar Charts
4. Tools for Continuous Variable Values
  1. Frequency Distribution
  2. Histogram
  3. Measures of Central Tendency
  4. Measures of Variability
  5. Numerical Measures of Relative Standing
5. Presenting and Understanding Data

---

---

---

---

---

---

---

---

---

---

---

## Who is in this class?

OBSERVATION #	Count	If Yes	n (if respondents)																						
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Q1	22	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Q2	19	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Q3	4	2	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	
Q4	6	2	1	2	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	
Q5	8	2	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	
Q6	11	2	2	1	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Q7	8	2	1	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Q8	14	2	2	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Q9	21	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Q10	6	2	2	2	1	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Q11																									
			23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	

---

---

---

---

---

---

---

---

---

---

---

## Who is in this class?

TOPIC	Proportion	n
Q1   I have used computer spreadsheets (such as Excel)	96%	23
Q2   I find mathematics fun	52%	23
Q3   I have taken an introduction to SPSS	17%	23
Q4   I have used statistics in my work previously	26%	23
Q5   I have an undergraduate degree in the humanities	39%	23
Q6   I have an undergraduate degree in the social sciences	48%	23
Q7   I have an undergraduate degree in a technical subject or the social sciences	35%	23
Q8   I took a statistics class at some time prior to coming to the Evans School	61%	23
Q9   I like chocolate chip cookies	91%	23
Q10   I would rather eat my dirty socks than do math	26%	23

---

---

---

---

---

---

---

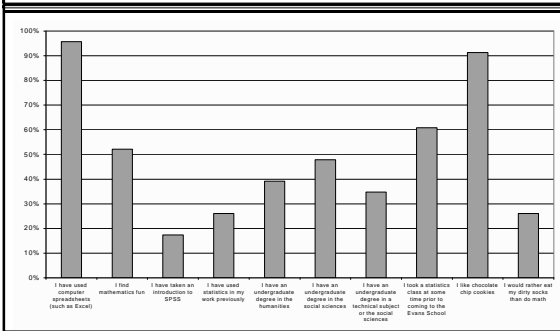
---

---

---

---

## Who is in this class?




---

---

---

---

---

---

---

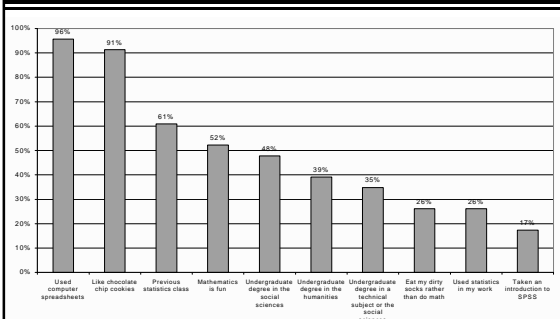
---

---

---

## Who is in this class?

**N=23**




---

---

---

---

---

---

---

---

---

---

## Your Comments

I hope:

...[this class] will work [themes like] overstatements of accuracy-ie census, intransitivity of preference into the curriculum.

...this course is interesting and does not become as tortuous as microeconomics was first quarter

...that the course is taught with the knowledge that the vast majority of Evans School graduates do not go on to be statisticians.

Most of the organizations we will work for will hire a statistician to do heavy analysis work. I hope we are taught to interpret and understand statistics as well as perform fairly basic analyses ourselves.

...the course is fun despite the seemingly dry subject. I hope the course covers how to collect data (primary research), where to find data (secondary research), in addition to the number crunching.

---

---

---

---

---

---

---

---

---

---

### More Comments

Not only would I rather eat dirty socks, I would rather eat dirt than take this course... I'm nervous about doing well - statistics is such a struggle for me. I will work diligently but even that does not guarantee that I will pass the course.

Statistics intimidate me....

Unfortunately, I am not well versed in statistics and analytical methods. But, I will do my best and try very hard.

While I have taken an introduction to SPSS, it was a while back so can't say I'm proficient with it by any means.

I've been on leave for a year. I need to find a class on SPSS so I can do the class work better.

---

---

---

---

---

---

---

---

### Describing Data

1. Keeping in touch/Policy Report
2. Who's in this class?
3. Tools for Discrete Variable Values
  1. Frequencies (counts)
  2. Relative Frequencies
  3. Pie Charts
  4. Bar Charts
4. Tools for Continuous Variable Values
  1. Frequency Distribution
  2. Histogram
  3. Measures of Central Tendency
  4. Measures of Variability
  5. Numerical Measures of Relative Standing
5. Presenting and Understanding Data

---

---

---

---

---

---

---

---

### 2000 Presidential Election Florida Election Results

In the 2000 presidential election, the popular vote in the State of Florida was very close. So close, that it was unclear whether there was really a winner of the popular vote.  
(kind of like Washington State now!)

---

---

---

---

---

---

---

---

## Summary Table

Tally:  
 |||| ||||  
 |||| ||||

Row Is Category

Candidate	Frequency		Relative Frequency	
	Vote Count	Proportion	Percent	
Bush	2,910,078	0.4885	48.85%	
Gore	2,909,117	0.4883	48.83%	
Nader	97,416	0.0164	1.64%	
Buchanan	17,465	0.0029	0.29%	
Browne	16,396	0.0028	0.28%	
Hagelin	2,273	0.0004	0.04%	
Moorehead	1,803	0.0003	0.03%	
Phillips	1,368	0.0002	0.02%	
McReynolds	618	0.0001	0.01%	
Harris	558	0.0001	0.01%	
Totals	5,957,092	1.0000	100.00%	

---

---

---

---

---

---

---

---

---

---

## Relative Frequencies

Proportion of Cases with a Specific Variable Value

$$f(x) = \frac{\text{number of cases with the value of } x}{\text{total number of cases}}$$

**Bush**

$$\frac{2,910,078}{5,957,092} = 0.4885$$

**Gore**

$$\frac{2,909,117}{5,957,092} = 0.4883$$

---

---

---

---

---

---

---

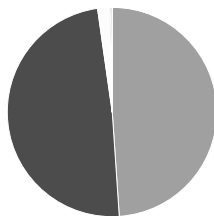
---

---

---

## Pie Chart

- Shows Relative Frequencies
- Shows Breakdown of Total Quantity into Categories
- Useful for Showing Relative Differences




---

---

---

---

---

---

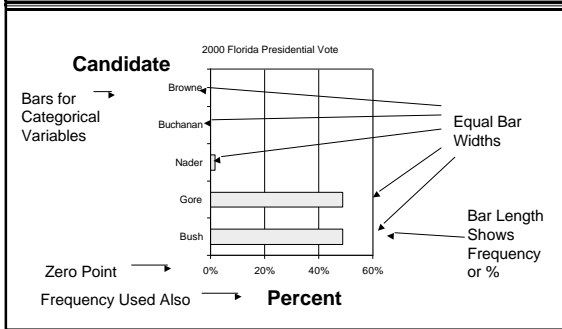
---

---

---

---

## Bar Chart




---

---

---

---

---

---

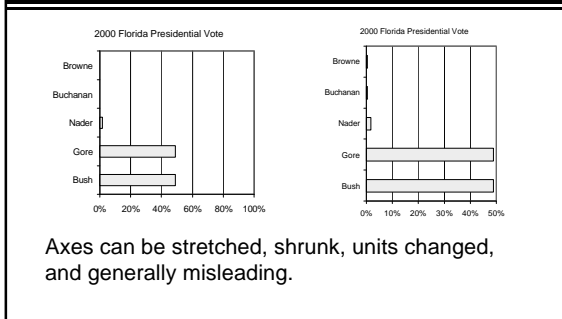
---

---

---

---

## Careful!




---

---

---

---

---

---

---

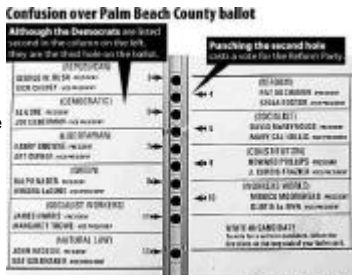
---

---

---

## More on Florida and Election 2000

West Palm Beach County, Florida, was the locale for another drama, this time about the validity of the votes cast using "butterfly ballots".



Source: Ask Top, Jan 2001. <http://www.asktop.com/columns/0420/butterflyBallot.html>, accessed Jan 6, 2004.

---

---

---

---

---

---

---

---

---

---



### Frequency Distribution Steps

1. Determine Range
2. Select Number of Classes
  - Usually Between 5 & 15 Inclusive
3. Compute Class Intervals (Width)
4. Determine Class Boundaries (Limits)
5. Count Observations & Assign to Classes

---

---

---

---

---

---

---

---

### Frequency Distribution Table Example

Class	Frequency
60 but <63	2
63 but <66	6
66 but <69	7
69 but <72	7
72 but <75	1

Width {

Boundaries ↙ ↘

---

---

---

---

---

---

---

---

### Relative Frequency & % Distribution Tables

Class	Relative Frequency Distribution	Percentage Distribution
	Proportion	Percent
60 but <63	0.09	9%
63 but <66	0.26	26%
66 but <69	0.30	30%
69 but <72	0.30	30%
72 but <75	0.04	4%
	1.00	100%

---

---

---

---

---

---

---

---

### Cumulative Percentage Distribution Table

Class	Cumulative Percentage
60 but <63	9%
63 but <66	35%
66 but <69	65%
69 but <72	96%
72 but <75	100%

Percentage Less than **Upper Class Boundary**

Upper Class Boundary

9% + 26%

35% + 30%

---

---

---

---

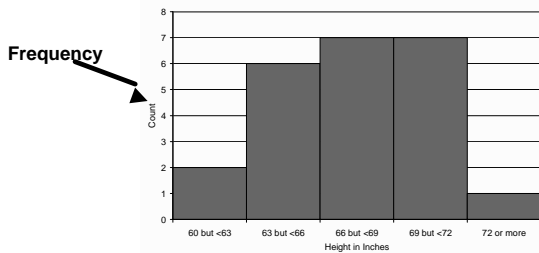
---

---

---

---

### Class Height Histogram




---

---

---

---

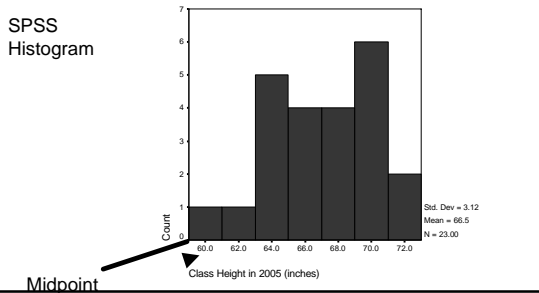
---

---

---

---

### Class Height Histogram




---

---

---

---

---

---

---

---

### Why Histograms?

The distribution that histograms provide tell us about the data

- Shape—peaked or uniformly flat
- Symmetric—one side of the distribution mirrors the other
- Center (rough location)
- Outliers
- Skewness

---

---

---

---

---

---

---

---

### BEWARE

Scales of plots may be compressed or stretched, may not begin at zero, may not cover all observed values of a variable. Pictures are useful but they also can be deceiving.

---

---

---

---

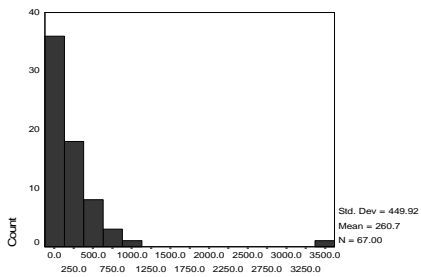
---

---

---

---

### Example: Votes for Buchanan in Florida Counties in 2000



---

---

---

---

---

---

---

---

## Describing Data

1. Keeping in touch/Policy Report
2. Who's in this class?
3. Tools for Discrete Variable Values
  1. Frequencies (counts)
  2. Relative Frequencies
  3. Pie Charts
  4. Bar Charts
4. Tools for Continuous Variable Values
  1. Frequency Distribution
  2. Histogram
  3. Measures of Central Tendency
  4. Measures of Variability
  5. Numerical Measures of Relative Standing
5. Presenting and Understanding Data

---

---

---

---

---

---

---

---

---

---

## Measures of Central Tendency

Propensity of the data cluster or center on certain numerical values

- Mean
- Mode
- Median

---

---

---

---

---

---

---

---

---

---

## Mean

Most Common Measure  
Acts as 'Balance Point'  
Affected by Extreme Values ('Outliers')

Buchanan Vote Count Average  
With Palm Beach County: 260  
Without Palm Beach County: 213

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = (x_1 + x_2 + x_i \dots + x_n) / n$$

$\bar{x}$  = Sample Mean (also called the average and the expected value)  
 $\mu$  (say "mu") = Population Mean  
 $n$  = sample size (this is the denominator for a sample mean)  
 $N$  = population size (this is the denominator for a population mean)

---

---

---

---

---

---

---

---

---

---

## Median

- Middle Value In Ordered Sequence

- If Odd n, Middle Value of Sequence
- If Even n, Average of 2 Middle Values

- Not Affected by Extreme Values

$$\text{Positioning Point} = \frac{n+1}{2}$$

50% of values are above, 50% are below  
(The 50% percentile)

Buchanan Vote Count Median  
With Palm Beach County: 120  
Without Palm Beach County: 117

---

---

---

---

---

---

---

---

## Median

### (odd number of observations)

$$\text{Positioning Point} = \frac{n+1}{2} = \frac{67+1}{2} = 34$$

Counts of Buchanan Votes in FL Counties

9	10	22	23	27	29	29	29	29	30
33	36	37	38	39	43	46	47	65	67
71	73	76	120	88	89	90	90	102	105
108	112	114	120	122	124	127	145	148	182
186	194	229	242	248	263	267	270	271	282
289	305	305	311	446	496	502	532	560	563
570	570	652	788	847	1013	3407			

---

---

---

---

---

---

---

---

## Median

### (even number of observations)

$$\text{Positioning Point} = \frac{n+1}{2} = \frac{66+1}{2} = 33.5$$

Counts of Buchanan Votes in FL Counties

9	10	22	23	27	29	29	29	29	30
33	36	37	38	39	43	46	47	65	67
71	73	76	117	83	88	89	90	90	102
108	112	114	117	120	122	124	127	145	148
186	194	229	242	248	263	267	270	271	282
289	305	305	311	446	496	502	532	560	563
570	570	652	788	847	1013	3407			

---

---

---

---

---

---

---

---

### Mode

- Value That Occurs Most Often
- Not Affected by Extreme Values
- May Be No Mode or Several Modes
- May Be Used for Numerical & Categorical Data

---

---

---

---

---

---

---

---

### Mode

Counts of Buchanan Votes in FL Counties

9	10	22	23	27	29	29	29	29	30
33	36	37	38	39	43	46	47	65	67
71	73	76	83	88	89	90	90	102	105
108	112	114	120	122	124	127	145	148	182
186	194	229	242	248	263	267	270	271	282
289	305	305	311	446	496	502	532	560	563
570	570	652	788	847	1013	3407			

---

---

---

---

---

---

---

---

### Summary of Central Tendency Measures

Measure	Equation	Description
Mean	$\sum X_i / n$	Balance Point
Median	$(n+1) / 2$ Position	Middle Value When Ordered
Mode	none	Most Frequent

---

---

---

---

---

---

---

---

### SPSS Descriptive Statistics Output Examples

SPSS Output: (Analyze>Descriptive Statistics>Descriptives)

	N	Minimum	Maximum	Mean	Std. Deviation
BUCHANAN	67	9.00	3407.00	260.6716	449.9242
Valid N (listwise)	67				

#### Frequencies

SPSS Output: (Analyze>Descriptive Statistics>Frequencies)

then Click on Statistics, choose Mean, Median, Mode)

Statistics		
BUCHANAN		
N	Valid	67
	Missing	0
Mean		260.6716
Median		120.0000
Mode		29.00

---

---

---

---

---

---

---

---

---

---

---

---

### Measures of Variability (spread)

- Range
- Variance
- Standard Deviation

---

---

---

---

---

---

---

---

---

---

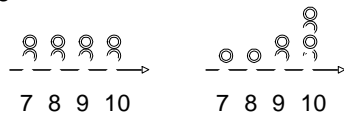
---

---

### Range

1. Measure of Dispersion
2. Difference Between Largest & Smallest Observations  

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$
3. Ignores How Data Are Distributed




---

---

---

---

---

---

---

---

---

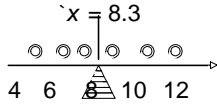
---

---

---

## Variance & Standard Deviation

1. Measures of Dispersion
2. Most Common Measures
3. Consider How Data Are Distributed
4. Show Variation About Mean ( $\bar{x}$  or  $\mu$ )




---

---

---

---

---

---

---

---

## Sample Variance Formula

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

$n - 1$  in denominator!  
(Use  $N$  if Population Variance)

$$= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

The average of the squared deviations from the mean.



Symbols:  
 $s^2$  for sample  
 $\sigma^2$  for population

---

---

---

---

---

---

---

---

## Sample Variance Another Form

Convenient form for calculator

$$s^2 = \frac{\sum (x_i^2) - \frac{(\sum x_i)^2}{n}}{n-1}$$

---

---

---

---

---

---

---

---

## Sample Standard Deviation Formula

$$s = \sqrt{s^2}$$

$$= \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}}$$

The standard deviation has a meaningful physical interpretation and it's in the same units as the mean.

Symbols  
s for a sample  
σ for a population

---

---

---

---

---

---

---

---

$x_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
60	-6.5	42.5
62	-4.5	20.4
63	-3.5	12.4
63	-3.5	12.4
63	-3.5	12.4
64	-2.5	6.3
64	-2.5	6.3
65	-1.5	2.3
66	-0.5	0.3
66	-0.5	0.3
66	-0.5	0.3
67	0.5	0.2
67	0.5	0.2
68	1.5	2.2
68	1.5	2.2
69	2.4	5.7
69	2.5	6.2
69	2.5	6.2
69	2.5	6.2
69	2.5	6.2
70	3.5	12.1
71	4.5	20.1
72	5.5	30.1
		<b>213</b>

### Calculating the Variance and Standard Deviation

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \rightarrow 9.7$$

$$s = \sqrt{s^2} \rightarrow 3.1$$

---

---

---

---

---

---

---

---

## Calculating the Variance and Standard Deviation

SPSS:

ANALYZE>  
DESCRIPTIVE  
STATISTICS>  
FREQUENCIES.

Statistics		
Height 527 2005		
N	Valid	23
	Missing	0
Mean		66.52
Median		67.00
Mode		69
Std. Deviation		3.117
Variance		9.715
Range		12
Minimum		60
Maximum		72
Percentiles	25	64.00
	50	67.00
	75	69.00

---

---

---

---

---

---

---

---

## Interpreting the Standard Deviation

Think about intervals with the mean at their center

$[\bar{x} - s, \bar{x} + s]$	$[m - s, m + s]$
$[\bar{x} - 2s, \bar{x} + 2s]$	$[m - 2s, m + 2s]$
$[\bar{x} - 3s, \bar{x} + 3s]$	$[m - 3s, m + 3s]$

Chebyshev's Rule (abridged)

FOR ANY DATASET, regardless of the shape of the frequency distribution:

No useful information is produced on the fraction of measurements that fall within 1 SD of the mean.

At least  $\frac{3}{4}$  of the measurements will fall within 2 SD of the mean.

At least  $\frac{8}{9}$  of the measurements will fall within 3 SD of the mean.

---

---

---

---

---

---

---

---

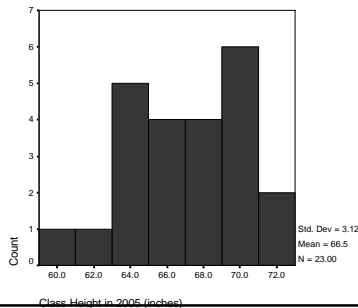
---

---

---

---

## Class Height PBAF 527




---

---

---

---

---

---

---

---

---

---

---

---

## Interpreting the Standard Deviation (2)

Empirical Rule for Mound-Shaped Distributions

- Approx. 68% of the measurements will fall within 1 SD of the mean.
- Approx 95% of the measurements will fall within 2 SD of the mean.
- Approx 99.7% of the measurements will fall within 3 SD of the mean.

---

---

---

---

---

---

---

---

---

---

---

---

### Numerical Measures of Relative Standing

- Percentiles
- Quartiles
- Z-scores

---

---

---

---

---

---

---

---

### Percentiles and Quartiles

The *p*th **percentile** of a group of numbers is the value at which *p* percent of the values (when ordered from smallest to largest) are below this number.

eg: The median is the 50<sup>th</sup> percentile.

A **quartile** is the position where 25%, 50%, 75% or 100% fall below.

In other words, 25% of students have heights of \_\_\_\_ or less.

---

---

---

---

---

---

---

---

#### Statistics

Height 527 2005

N	Valid	23
	Missing	0
Mean		66.52
Median		67.00
Mode		69
Std. Deviation		3.117
Variance		9.715
Range		12
Minimum		60
Maximum		72
Percentiles	25	64.00
	50	67.00
	75	69.00

---

---

---

---

---

---

---

---

**Z-scores**

This measure of relative standing uses the mean and the standard deviation to specify the relative location of a measurement.

For a measurement x:

Sample z-score  $z = \frac{x - \bar{x}}{s}$

Population z-score  $z = \frac{x - \mu}{\sigma}$

---

---

---

---

---

---

---

---

**Z-scores**

- If z is positive and large, then the measurement is larger than nearly every other value.
- If z is negative and large, then the measurement is less than nearly every other value.
- If z is 0 or near 0, then it is at the mean or near the mean.

---

---

---

---

---

---

---

---

**Z-score Example**

So, if we take the county that voted for Buchanan with more frequency than the others (Palm Beach):  
Vote Count=3407

Mean Vote for Buchanan=261  
s=450

$$z = \frac{x - \mu}{\sigma} = \frac{3407 - 261}{450} = 6.99$$

Palm Beach is positive and large, larger than every other measurement.

---

---

---

---

---

---

---

---

### For mound-shaped data

Approx 68% of the measurements will have a z-score between -1 and 1.  
 Approx. 95% of the measurements will have a z-score between -2 and 2.  
 Approx. 99.7% of the measurements will have a z-score between -3 and 3.

---

---

---

---

---

---

---

---

### Summary of Variation Measures

Measure	Equation	Description
Range	$X_{\text{highest}} - X_{\text{lowest}}$	Total Spread
Standard Deviation (Sample)	$\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$	Dispersion about Sample Mean
Standard Deviation (Population)	$\sqrt{\frac{\sum (x_i - \mu_x)^2}{N}}$	Dispersion about Population Mean
Variance (Sample)	$\frac{\sum (x_i - \bar{x})^2}{n-1}$	Squared Dispersion about Sample Mean

---

---

---

---

---

---

---

---

### Some Calculations

In your small group, ask each person their height. Calculate and interpret the:

- Mean
- Median
- Variance
- Standard Deviation
- Range

---

---

---

---

---

---

---

---

## Describing Data

---

---

1. Keeping in Touch/Policy Report
2. Who's in this class?
3. Tools for Discrete Variable Values
  1. Frequencies (counts)
  2. Relative Frequencies
  3. Pie Charts
  4. Bar Charts
4. Tools for Continuous Variable Values
  1. Frequency Distribution
  2. Histogram
  3. Measures of Central Tendency
  4. Measures of Variability
  5. Numerical Measures of Relative Standing
5. Presenting and Understanding Data

---

---

---

---

---

---

---

---

## Eye Cue Test

---

---

Read through the hand out and answer these questions:

- e. What are the drawbacks to using graphical displays of sample data to infer the nature of the population?
- g. Which are more informative, the graphs of the qualitative measures or the graphs of the quantitative measures?
- h. For each group of subjects, interpret the numerical descriptive measures for the quantitative measure of task performance. What do you conclude from this analysis?

---

---

---

---

---

---

---

---

## Conclusion

---

---

1. Keeping in Touch/Policy Report
2. Who's in this class?
3. Tools for Discrete Variable Values
  1. Frequencies (counts)
  2. Relative Frequencies
  3. Pie Charts
  4. Bar Charts
4. Tools for Continuous Variable Values
  1. Frequency Distribution
  2. Histogram
  3. Measures of Central Tendency
  4. Measures of Variability
  5. Numerical Measures of Relative Standing
5. Presenting and Understanding Data

---

---

---

---

---

---

---

---

**End of Chapter**

Any blank slides that follow are  
blank intentionally.

---

---

---

---

---

---

---

---