MPA 6010: Research Methods for LPA

Anirudh V. S. Ruhil

September 13, 2016

The Voinovich School of Leadership and Public Affairs

Table of Contents

Statistics 101

Statistics versus a Statistic

Definition

Statistics involves methods for describing and analyzing data and for drawing inferences about phenomena represented by the data

- Inflation forecasts
- Unemployment data
- Nielsen Weekly U.S. Television ratings
- Per capita income

Definition

A statistic, on the other hand, is the result of applying a computational algorithm to a set of data

- Average height of adult males in the U.S.
- Median household income
- Percent living below the poverty line
- Modal number of stories devoted to covering an electoral candidate
- Mean hours spent on facebook

Populations versus Samples

Definition

A population is the universe (or set) of all elements of interest in a particular study
A sample is the subset of cases drawn for analysis from the population

Example (1)

- Population: All first-time enrollee Freshmen at Ohio University in 2009-2010
- Sample: Freshmen selected for study from the OU Registrar's list of all first-time enrollee Freshmen at Ohio University in 2009-2010

Example (2)

- Population: All national public radio (NPR) members
- Sample: NPR members selected for telephone survey from NPR's list

Example (3)

- Population: All *Productivity* Apps in Apple's App Store
- Sample: 100 *Productivity* Apps drawn at random from App Store

Data and its Components

- Data are the facts and figures collected, analyzed and summarized
- A data set comprises all data collected in the course of a particular study
- A Variable is a characteristic of interest for the elements of a data set
- An Observation is the set of measurements obtained for an element

	row.names	gender	ethnicity	birth
1	1122	female	afam	1979 Q3
2	1137	female	cauc	1980 Q1
3	1143	female	afam	1979 Q4
4	1160	male	cauc	1979 Q4
5	1183	male	afam	1980 Q1
6	1195	male	cauc	1979 Q3
7	1215	male	afam	1979 Q1
8	1224	female	cauc	1979 Q4
9	1246	male	cauc	1979 Q2

Cross-Sectional versus Time-Series Data

Definition

Cross-Sectional data are observations recorded for a single point in time
Time-Series data are observations recorded over time
Panel data combine the two preceding forms (cross-sectional + time-series)

- Cross-Sectional: Ohio's public school districts' ratings for the 2013-2014 school-year
- Time-Series: Athens (OH) public school district's ratings for the 2008-2009 through 2013-2014 school-years
- Panel: Ohio's public school districts' ratings for the 2008-2009 through 2013-2014 school-years
- Primary data are collected by the researcher(s)
- Secondary data have been collected previously by some individual(s) and/or organization(s). Some common secondary sources ...
  1. The U.S. Bureau of the Census
  2. UNICEF
  3. WHO
  4. COMS data sources
  5. The Interuniversity Consortium for Political & Social Research
  6. The Gallup Poll; CBS/NYT Polls, ABC/Washington Post Poll

Measurement

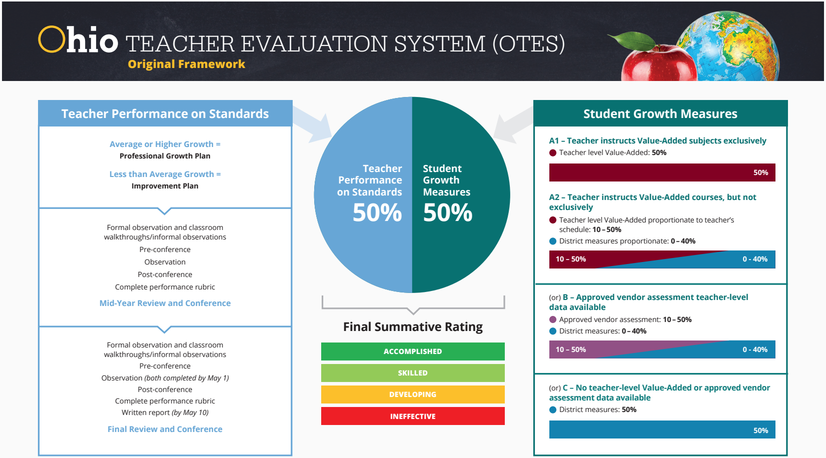
Measurement

Definition

Measurement is the act of assigning numerical values to phenomenon we want to analyze.

- Few things can be measured directly so we start with an operational definition that, when applied, leads to an indicator. For example:
    - Teacher Effectiveness could be defined as control over the classroom, motivating students to work harder, having organized lesson plans, or improving students' performance on standardized tests
    - Control over the classroom could be measured by observing class sessions; student performance could be measured by her/his students' scores on the standardized test, and so on
- A good analyst starts by making explicit how something will be measured
- A good analyst also recognizes that every indicator has built-in error
- Often the phenomenon may need multiple indicators (see next slide)

An Example of Multiple Indicators



Measurement: Validity

The best indicators reflect both validity and reliability

- Validity ...
  - Valid indicators have convergent validity, i.e., they measure the target concept without error
  - Valid indicators also exhibit discriminant validity, i.e., indicator A used to measure concept A is unrelated to indicator B used to measure concept B
  - Analysts use four yardsticks when considering validity ...
    (a) Face Validity – The analyst thinks indicator A has face value as a measure of concept A
    (b) Consensual Validity – Several analysts would agree indicator A is an acceptable measure of concept A
    (c) Correlational Validity – Multiple indicators that measure concept A are highly correlated so we could pick any one indicator
    (d) Predictive Validity – Indicator A predicts outcome Y (e.g., SAT scores and academic success)

Measurement: Reliability

- Reliability means indicator A will measure concept A the same way time after time if concept A is unchanged. Alternatively, indicator A returns the same numerical value regardless of who is conducting the measurement
- Reliability is influenced by two things ...
  (a) Subjectivity – the analyst's biases, prejudices, lack of knowledge, etc. shape what is recorded (for e.g., perceived neighborhood safety). You can minimize subjectivity if you
     - Train the analyst before they set out to measure
     - Use multiple analysts to carry out the measurement
  (b) Imprecision – small samples, hard to measure phenomenon, etc. can lead to imprecise measurement. You can minimize this if you
     - Use larger samples
     - Use more carefully developed fine (rather than crude) indicators
     - Use multiple indicators

Levels of Measurement

Four levels of measurement ...

(a) Nominal – Categorizes the observations without drawing conclusions about A is better or worse than B. For e.g.: Sex, Race/Ethnicity, Democrat/Independent/Republican

(b) Ordinal – Categorizes the observations on the basis of some hierarchy. For e.g.: Freshman/Sophomore/Junior/Senior, Level of Satisfaction, Agree/Neutral/Disagree

(c) Interval – Categorizes the observations with numerical values such that two or more observations can be numerically compared. For e.g., SAT/ACT Scores, Celsius/Fahrenheit temperature scales

(d) Ratio – An interval level measure for which the value of zero implies complete absence of the phenomenon. For e.g., Age; velocity; Distance; Income; Years of formal schooling

Note:

- Numerical values of Nominal and Ordinal variables have no intrinsic value
- Nominal & Ordinal variables are also called categorical or qualitative variables

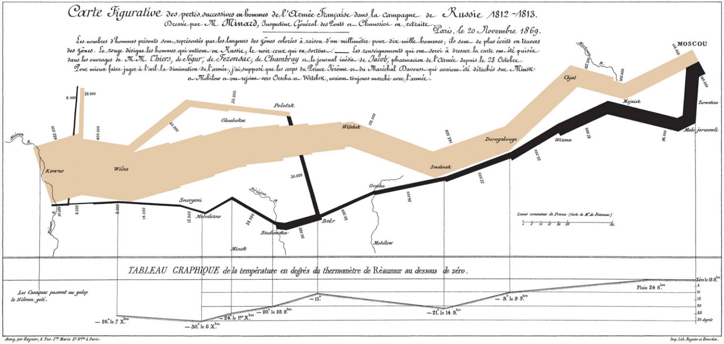
Descriptive Statistics

Descriptive Statistics

Definition

Descriptive Statistics are statistical methods of summarizing and organizing data in a meaningful way

- Tabular representations
- Graphical representations



A Frequency Table

Definition

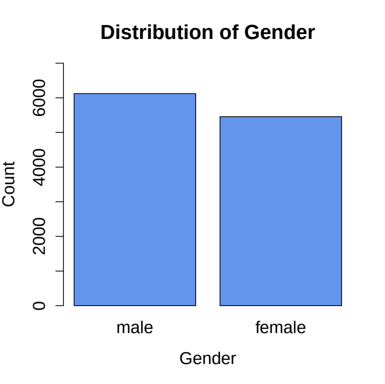
Frequency: A count of cases with a certain value of the variable X
Frequency Distribution: Summary depiction (tabular or graphical) of the values of the variable X

Table 1: Distribution of Students (by Gender)

Sex	Frequency (f)	Rel. Freq	% Freq
Male	6,122	0.5287614	52.87614
Female	5,456	0.4712386	47.12386
Total	11,578	1	100

Frequencies as Graphics

Figure 1: Barchart of Distribution of Students (by Gender)



Frequency Distribution

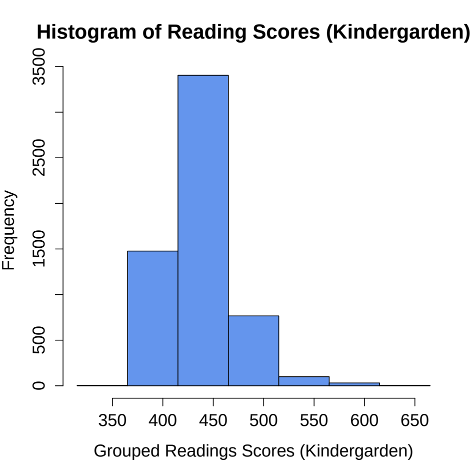
Summarizing/describing quantitative data via tabular & graphical methods is more cumbersome because these variables typically assume a large number of values

Example

Audit Times (in days)			
12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

- If we had to communicate the frequency distribution of a variable such as that in the table, it could hardly be called a summary.
- So we construct “classes" (i.e., “group the data") such as 12-17, 17-22, 22-27, 27-32, 32-37,
- Technically, groups are called bins (for e.g., I have 5 groups so 5 bins, etc.)

Grouped Frequency Distributions via Histograms



Symmetric versus Asymmetric Distributions

Definition

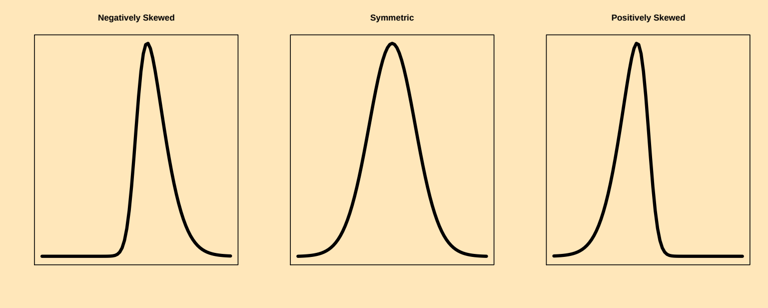
A distribution is said to be symmetric if the left tail mirrors the right tail. That is, if you fold the distribution in half, each half will be a perfect replica of the other half.

Definition

A distribution is said to be asymmetric or skewed if the distribution bunches up towards one end and tapers (i.e., thins) off at the other end.

- Distributions tapering left are “negatively skewed" or skewed left;
- Distributions tapering right are “positively skewed” or skewed right.

Stylized Examples of Symmetric/Asymmetric Distributions



Note:

- Symmetric distributions are easier to work with (statistically)
- Seeing whether something is symmetrically/asymmetrically distributed provides valuable information about the phenomenon under study

Crosstabulation

Definition

A crosstabulation is a tabular summary of two variables

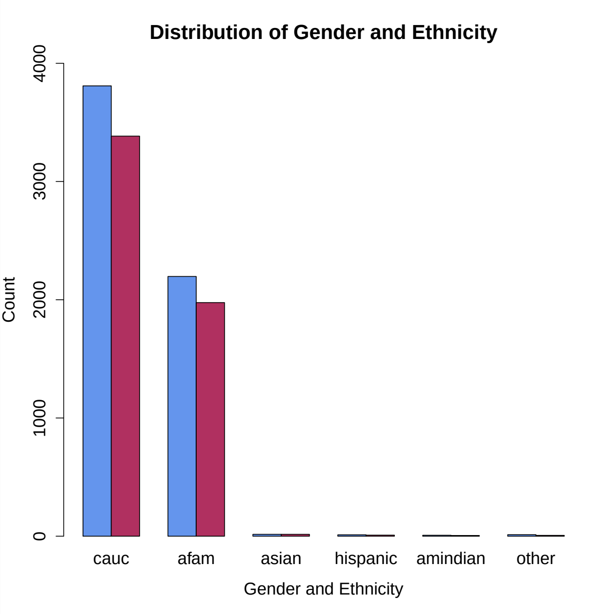
	cauc	afam	asian	hispanic	amindian	other
male	3809	2197	16	12	8	13
female	3384	1976	16	9	6	7

	cauc	afam	asian	hispanic	amindian	other
male	0.33	0.19	0.00	0.00	0.00	0.00
female	0.30	0.17	0.00	0.00	0.00	0.00

	cauc	afam	asian	hispanic	amindian	other
male	0.63	0.36	0.00	0.00	0.00	0.00
female	0.63	0.37	0.00	0.00	0.00	0.00

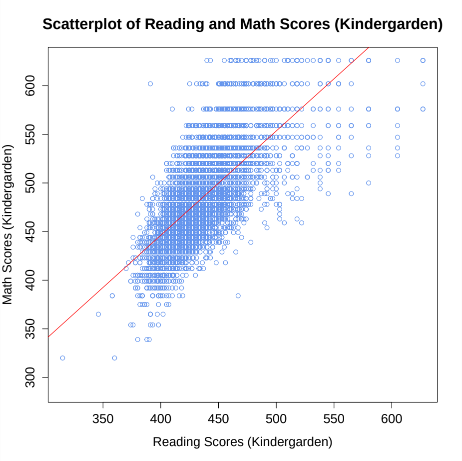
	cauc	afam	asian	hispanic	amindian	other
male	0.53	0.53	0.50	0.57	0.57	0.65
female	0.47	0.47	0.50	0.43	0.43	0.35

Stacked Barcharts



Scatter Diagram and Trendline

- A scatter plot or scatter diagram illustrates the obvious relationship between two quantitative variables
- A trendline is an approximation of this relationship



Definition

Simpson's Paradox refers to instances where conclusions drawn on the basis of aggregate data are reversed or otherwise unjustified by the disaggregated data.

Example

Let there be two justices, Justice 1 and Justice 2, hearing cases in both the Common Pleas Court and the Municipal Court. Their decisions may be appealed and thus ultimately reversed. Aggregate Cross-tabulation is shown below.

	Judge		
Verdict	1	2	Total
Upheld	129 (86%)	110 (88%)	239
Reversed	21 (14%)	15 (12%)	36
Total	150 (100%)	125 (100%)	275

Simpson's Paradox (Continued ...)

Judge 1			
Verdict	Common	Municipal	Total
Upheld	29 (91%)	100 (85%)	129
Reversed	3 (9%)	18 (15%)	21
Total	32 (100%)	118 (100%)	150

Judge 2			
Verdict	Common	Municipal	Total
Upheld	90 (90%)	20 (80%)	110
Reversed	10 (10%)	5 (20%)	15
Total	100 (100%)	25 (100%)	125

- Never draw individual-level conclusions on the basis of aggregate data
- Beware hidden variables (here Type of Court)

Recap of Key Points so far

1. Tabular and graphical descriptions of data are very useful
2. With qualitative variables (i.e., Nominal/Ordinal) use bar charts and frequency tables
3. With quantitative variables (i.e., Interval or Ratio) use histograms, scatterplots, trend-lines, and grouped frequency distributions
4. Cross-tabulations are useful with two Nominal/Ordinal variables, and so are stacked barcharts
5. Symmetric distributions are easier to work with than are skewed distributions
6. Beware Simpson's Paradox; never infer individual-level relationships from aggregate data

Shifting Gears ... ... now we can try to better understand how certain statistics can be used to capture and convey two features ...

1. what is typical or average or most often seen to occur?
2. how much do things usually differ from what is typical, from this average?

Central Tendency

The Mean

Definition

Central Tendency: A statistical measure that defines the center of a distribution and is most typical, most representative, of the scores that comprise the distribution of the variable of interest

Definition

Mean: The mean is commonly known as the arithmetic average, and is computed by adding up the scores in the distribution and dividing this sum by the sample size

Definition

Sample Mean is denoted by $\bar{x}$ where $\bar{x} = \frac{\Sigma x_i}{n}$ ... add all values of x and divide by the total number of observations with a non-missing value of x

Definition

The Population Mean is denoted by μ where $\mu = \frac{\Sigma x_i}{N}$ ... add all values of x and divide by the total number of observations with a non-missing value of x

The Mean

Example

ID	Salary ($)	ID	Salary ($)
1	2,850	7	2,890
2	2,950	8	3,130
3	3,050	9	2,940
4	2,880	10	3,325
5	2,755	11	2,920
6	2,710	12	2,880

$$\begin{aligned}
\bar{x} &= \frac{\Sigma x_i}{n} \\
&= \frac{x_1 + x_2 + \cdots + x_{12}}{n} \\
&= \frac{2{,}850 + 2{,}950 + \cdots + 2{,}880}{12} \\
&= \frac{35{,}280}{12} \\
&= \$2,940
\end{aligned}$$

Properties of the Mean

1. Changing the value of any observation changes the mean
2. Adding or subtracting a constant k from all observations is equivalent to adding or subtracting the constant k from the original mean
3. Multiplying or dividing a constant k from all observations is equivalent to multiplying or dividing the original mean by the constant k

ID	x	$(x-2)$	$2x$	$\left(\frac{x}{2}\right)$
1	6	4	12	3
2	3	1	6	1.5
3	5	3	10	2.5
4	3	1	6	1.5
5	4	2	8	2
6	5	3	10	2.5
Total	26	14	52	13
Mean	4.33	2.33	8.66	2.16

The Median

Definition

The Median is the middle-value that occurs when the data are arranged in an ascending or descending order
The Median is commonly denoted by Md

1. Arrange the data in either ascending or descending order of the values of x
2. If n is odd, Md is the middle value
3. If n is even, Md is the average of the two middle values

Median Example (n is odd)

Example

ID	Salary ($)	ID	Salary ($)
1	2710	7	2940
2	2755	8	2950
3	2850	9	3050
4	2880	10	3130
5	2890	11	3325
6	2920		

$$Md = \frac{n+1}{2} = 6^{th}$$

$$Md = \$2,920$$

Median Example (n is even)

Example

ID	Salary ($)	ID	Salary ($)
1	2710	7	2920
2	2755	8	2940
3	2850	9	2950
4	2880	10	3050
5	2880	11	3130
6	2890	12	3325

$$Md = \frac{2{,}890+2{,}920}{2}$$

$$Md = \frac{5{,}810}{2} = \$2,905$$

Quartiles

Definition

Quartiles divide the data into four parts and are denoted as Q_1, Q_2, Q_3
Q_1 is the first quartile or the $25^{th} percentile$
Q_2 is the second quartile or the $50^{th} percentile = Md$
Q_3 is the third quartile or the $75^{th} percentile$

Example (n is odd)

1. For Q_1, $i = \left(\frac{p}{100}\right) \times n = \left(\frac{25}{100}\right) \times 11 \approx 3$
2. For Q_3, $i = \left(\frac{p}{100}\right) \times n = \left(\frac{75}{100}\right) \times 11 \approx 9$
3. Therefore, $Q_1 = 3^{rd} = \$2,850$
4. ... and $Q_3 = 9^{th} = \$3,050$

Another Example of Median, Quartiles, etc.

Example (n is odd)

Store owner counts customers who enter the store each hour of the day

$x \sim \{3, 6, 7, 7, 8, 8, 9, 10, 12, 12, 15\}$

1. $\bar{x} = \frac{\Sigma x_i}{n} = 8.18$
2. $Md = 6^{th} = 8$
3. $Mode =$ multiple modes
4. $Q_1 : i = \left(\frac{25}{100}\right) \times 11 = 2.75 \approx 3^{rd} = 7$
5. $Q_3 : i = \left(\frac{75}{100}\right) \times 11 = 8.25 \approx 9^{th} = 12$

Mode

Definition

The Mode is the value with the greatest frequency in the data set

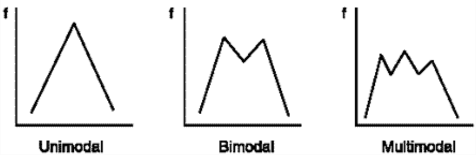
Example

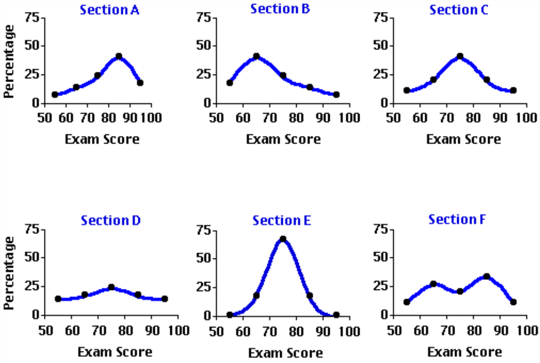
Drink	Freq.
Coke Classic	19
Diet Coke	8
Dr. Pepper	5
Pepsi-Cola	13
Sprite	5
Total	50

Mode = Coke Classic

Note: You could have bimodal or multimodal distributions

Some Distributions ...





Choosing A Measure of Location

1. Mean usually preferred over Median and Mode because
   - uses all observations in the data
   - used in most statistical calculations
   - intuitive
   - However, asymmetric distributions skew the Mean
2. Mode is easy to calculate, and can be used with both qualitative and quantitative data so analysts often gravitate towards it
3. Median is usually preferred when data
   - have extreme scores that lead to a skewed distribution
   - are open-ended (e.g., income with categories of $\leq 25,000$ and/or $\geq 200,000$)
   - have some undetermined values (e.g., time on task with some not completing task)

Measures of Dispersion

Variability ...

1. Say you play a video poker game that costs you $1 to play each time and yet every time you get back 90 cents regardless of whether you win or lose
   Would you play?
2. Say you shift to another game that costs you $1 to play as well but now you get $9 if you win and nothing if you lose.
   Would you play?
   You should ask: What are the odds of winning? Odds are 1 win in 10 tries.
3. Now ask yourself: What is the average earning from each game? Which game are you more likely to play and why?
4. The catch: Identical averages but different variability (0 in first game)
5. Therefore, assessing variability tells us something meaningful

Range

Definition

The Range is the simplest measure of variability, and computed as
$Range = x_{max} - x_{min}$ ... small range implies little variabiity

The Interquartile Range is the difference between the third and the first quartiles;
$IQR = Q_3 - Q1$ ... smaller IQR implies scores do not vary much in the middle 50% of the distribution

Example

1. Range $=$ Max. - Min. $= 627.0 - 315.0 = 312.0$
2. IQR $= Q_3 - Q_1 = 453.0 - 414.0 = 39.0$

Variance and Standard Deviation

Definition

The variance is a measure of variability constructed using all values in a distribution, and its square root is the standard deviation

1. Population Variance: $\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$
2. Population Standard Deviation: $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$
3. Sample Variance: $s^2 = \frac{\sum(x_i - \bar{x})^2}{(n-1)}$
4. Sample Standard Deviation: $s = \sqrt{s^2} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{(n-1)}}$

Calculating Variance and Standard Deviation

1. Let a population $x \sim (1,9,5,8,7)$. Then $\mu = 6$
2. $\Sigma(x-\mu) = 0$
3. $\Sigma(x-\mu)^2 = 40$
4. $\sigma^2 = \frac{\Sigma(x_i-\mu)^2}{N} = \frac{40}{5} = 8$ ... Population Variance
5. $\therefore \sigma = \sqrt{\sigma^2} = \sqrt{8} = 2.83$ ... Population Standard Deviation

1. If these 5 scores were a sample, then the sample mean would be $\bar{x} = 6$
2. $s^2 = \frac{\Sigma(x_i-\bar{x})^2}{(n-1)} = \frac{40}{(5-1)} = \frac{40}{4} = 10$ ... Sample Variance
3. $\therefore s = \sqrt{s^2} = \sqrt{10} = 3.16$ ... Sample Standard Deviation

A Tabular Calculation ...

ID	x	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	2850	-90	8100
2	2950	10	100
3	3050	110	12100
4	2880	-60	3600
5	2755	-185	34225
6	2710	-230	52900
7	2890	-50	2500
8	3130	190	36100
9	2940	0	0
10	3325	385	148225
11	2920	-20	400
12	2880	-60	3600
	Total:	0	301850

$\bar{x} = 2940$

$\Sigma(x_i - \bar{x}) = 0$

$\Sigma(x_i - \bar{x})^2 = 301850$

$s^2 = \frac{301850}{(12-1)} = \$27440.91$

$s = \sqrt{27440.91} = \$165.63$

Why $n-1$ ...

- Calculating s^2 & s requires us to first compute $\bar{x}$
- This calculation leads to an underestimation of variability in the sample and so we make an adjustment

Let $x = \{8, 4, 6\}$

	We know $\mu = 3$		Don't know $\mu = 3$; use $\bar{x}$	
x_i	$(x_i - \mu)$	$(x_i - \mu)^2$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
8	(8 - 3) = 5	25	(8 - 6) = 2	4
4	(4 - 3) = 1	1	(4 - 6) = -2	4
6	(6 - 3) = 3	9	(6 - 6) = 0	0

1. True Variance $= \frac{\Sigma(x_i - \mu)^2}{n} = \frac{35}{3} = 11.66$
2. Estimated Variance (without adjustment) $= \frac{\Sigma(x_i - \bar{x})^2}{n} = \frac{8}{3} = 2.83$
3. Note: We estimated a smaller variance when using $\bar{x}$ so adjustment by $n-1$ will correct this at least somewhat and give $s^2 = 4$

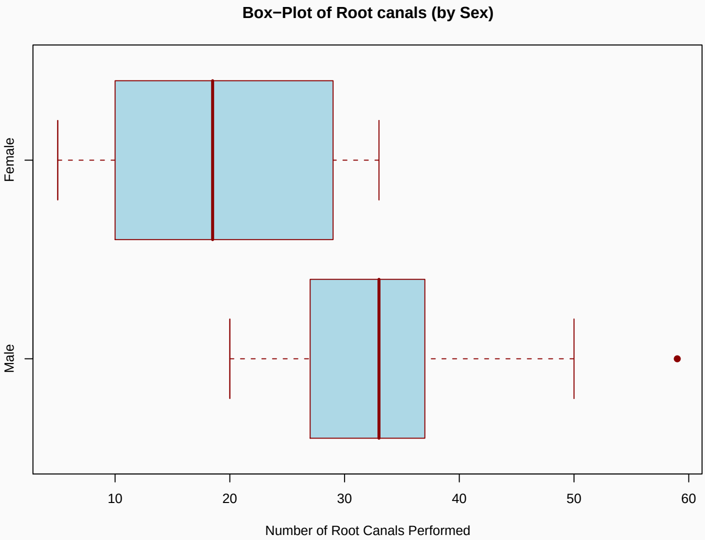
Five-number Summary

- The five-number summary comprises the
  1. Minimum value
  2. Q_1
  3. Md
  4. Q_3, and
  5. the Maximum value
- The distances between these consecutive values tells us something about the center and the shape of the distribution

An Example

Male Dentists	Female Dentists
20	5
25	7
25	10
27	14
28	18
31	19
33	25
34	29
36	31
37	33
44	
50	
59	

Box-Plot of Preceding Data



Another Box-Plot Example