Working with Graphs on Quantitative Data

Dr Tom Ilvento
Department of Food and Resource Economics

Overview

- This lecture looks at strategies for graphing quantitative data
  - Histograms
  - Box Plots (a.k.a, Box and Whisker Plots)
  - Stem and Leaf Plots
  - Scatter Plots
- Excel will do some of these, others can be done by hand or using JMP

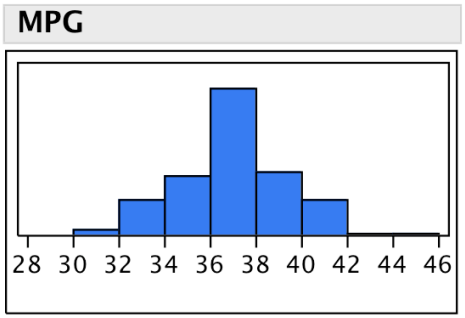
Data Sets used

- Measurement Standard - 100 measurements of a international standard for grams, which weighed just under 10 grams (about the weight of two nickels)
- MPG - the EPA's miles per gallon for 100 sub-compact cars
- Catalog Sales - the amount spent in a year for 1,000 customers for a company that sells electric equipment through catalog sales
- Small data set on the operating costs of several cars
- Data set on the yield of various tomato plants
- State SAT data

Histogram

- As strategy of collapsing the continuous data to show the distribution via a bar chart.
- Good visual depiction of the distribution of a variable, showing shape, modes, skew and outliers
- Can be graphed for both a small and large samples
- Most software programs (including Excel) provide an easy means to construct
- Requires decisions on the number and width of intervals
- Choices made by a user can distort the graph

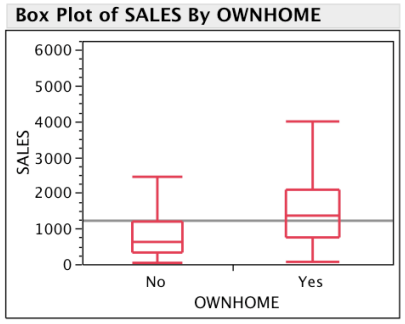
Histogram Examples



Box and Whisker Plots

- Box Plots are based a “Five Number Summary” based on position, including the median and quartiles.
  - minimum
  - 1st Quartile (25th percentile)
  - median (50th percentile)
  - 3rd Quartile (75th percentile)
  - maximum
- Good visual depiction of the distribution of a variable, showing shape, modes, skew and outliers
- Can be graphed for a small or large samples
- There is a fairly uniform approach to construct - no user decisions
- Excellent approach for comparing the distribution of two or more sub-groups

Examples of Box Plots



Stem and Leaf Plot

- A Stem and Leaf plot actually uses the data itself to make the graph - the data are divided into stems and leafs
- Good visual depiction of the distribution of a variable, showing shape, modes, skew and outliers
- Limited to small and medium sized data sets - difficult to produce when the sample size is over 150
- The user (or program) must make some decisions that can influence the shape of the graph

Stem and Leaf Plot

- This approach graphs the data using the data itself
- It provides a graphical picture of a variable
  - Distribution
  - Range
  - Skewness
  - Outliers

Stem and Leaf Plot

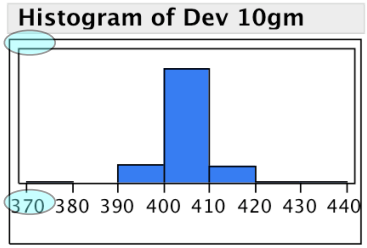
Stem and Leaf of MPG Data

Stem	Leaf	Count
44	9	1
43		
42	1	1
41	002	3
40	0123557	7
39	00345789	8
38	0122345678	10
37	000011122334456677899	21
36	012334455667778888999	20
35	01235667899	11
34	024588	6
33	126899	6
32	5799	4
31	8	1
30	0	1

30|0 represents 30.0

Outliers and Stem and Leaf Plots

- This is a data set of measurements of an international measurement standard
- The standard is measured over many time periods and the measurement is recorded
- The measurement is the deviation below 10 g, multiplied by 10,000
- A measurement of 9.999591 is represented as 409



Stem and Leaf Plot

- Three simple steps!
  1. Sort the data
  2. Choose the Stems
  3. Add the Leaves

Step 1: Sort the Date

- Sort the data from lowest to highest
- It just makes it easier
  - For decisions on choosing stems and leaves
  - For not making an error of missing an observation

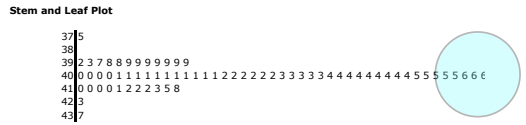
Step 2: Choose the stems

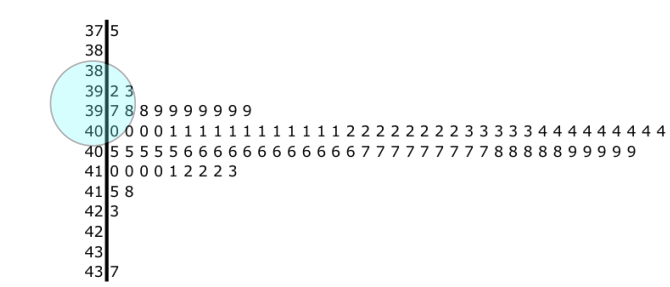
- The Stems are the initial digit in the plot, such as
  - 1 in the number 10;
  - 10 in the number 10.6;
  - 2 in the number 215.
- It is helpful to look at the range of the variable to decide the appropriate stems.
- Stems can be 1, 2 or more digits – for 215 it could be 2 or 21
- **In our example, the stems will be the first two digits (the hundreds and tens digits)**

Step 3: Add the Leaves

- The Leaves, or the following digits. In most cases it will be one digit, but it is possible to use more than one digit
- If doing by hand, make sure the distance between digits are
  - Uniform
  - Large enough to show the separate observations
- **In our example the leaves are the ones digit.**

Stem and Leaf Plots of Measurement Device





Stem and Leaf Plots

- Some statistical packages will create stem and leaf plots for us (JMP, SAS, Minitab, SPSS)
- Each program has unique aspects on how they do it and how it is labeled
- Stem and Leaf plots work best with moderate sized data sets (100 or less). Larger data sets may get unwieldy.

Let's do a Stem and Leaf Plot Together on the Driver's Age Data

- **Sort the data**
- **Choose the Stems** — **10's digit**
- **Add the Leaves** — **one's digit**

16	17	18	20	24	28	34	44	53	73
16	17	18	20	24	28	35	44	56	74
16	17	18	20	24	29	36	45	57	76
16	17	18	21	24	29	37	45	58	76
16	17	18	21	25	30	37	45	65	77
16	17	18	21	26	30	38	45	69	81
16	18	18	22	27	31	38	51	69	86
16	18	18	23	27	31	38	51	70	87
17	18	18	24	27	32	40	51	73	87
17	18	19	24	28	34	42	52	73	88

Driver Age Stem and Leaf

Stem and Leaf

Stem	Leaf	Count
8	6778	4
8	1	1
7	667	3
7	03334	5
6	599	3
6		
5	678	3
5	11123	5
4	5555	4
4	0244	4
3	5677888	7
3	0011244	7
2	5677788899	10
2	00011123444444	14
1	666666667777777788888888888889	30

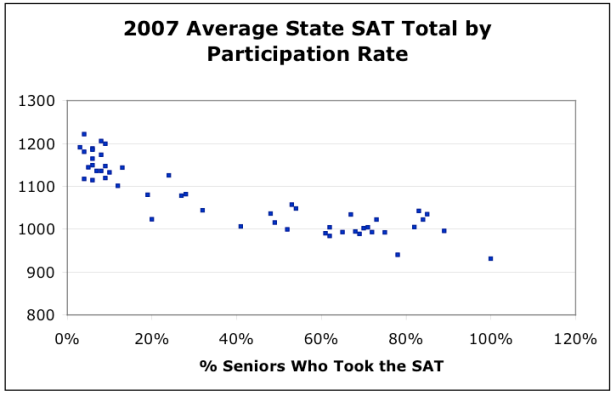
1|6 represents 16

Scatterplots

- Scatter Plots are based on how two continuous variables vary together, particularly over time.
- Most spreadsheet programs provide an easy mechanism to make a scatter plot (also called XY scatterplot).
- In scatter plots we tend to think of the two variables have a function:
  - one of the variables as a dependent variable and label it Y and is on the vertical axis
  - The other variable, labeled X, is the independent variable and is on the horizontal axis
- It forms the basis for correlation and regression

Examples of Scatterplots

SAT Example



Summary

- Graphs are very useful to help us view our data
- They help “tell the story”
- There are graphing techniques for continuous level data – histograms, box plots, stem and leaf, and scatterplots