STATISTICS for MANAGEMENT I
ADM2303

Prof. Brand
Week 7

Administrative Issues

- Quiz (Sunday, March 4th)
  - See course-outline
  - See doc-depot for instructions, time, room, etc.
- A2
  - Soln. uploaded
  - Some common oversights
    - Upload to wrong folder
    - Handwritten solution
    - Solution not in PDF format
    - Submission not registered by way of answer-area
    - Academic integrity neglected
- A3
  - Part-II to be uploaded
  - Part-I available

This Week

- Wrap-up week 6 material
- Macho-chip example
- New topic (returning to data)

Recall SAT (Standardized Academic Test)

- .mu= 500
- .sd= 100
- Area question
  - What is the probability of having a score in excess of 700?

Inverse Problem (1)

- .mu= 500
- .sd= 100
- Inverse question
  - OCGS is giving the SAT a new design. They want to:
    - Keep .sd the same but want to shift .mu such that 1 percent of people get scores greater than 1000
    - *Keep .sd the same but want to shift .mu in order to make sure that the fraction of people getting scores below 450 is 20 percent*

Inverse Problem (2)

- .mu= 500
- .sd= 100
- Inverse question
  - OCGS is giving the SAT a new design. They want to:
    - Keep .mu the same but want to engineer an .sd such that 1 percent of people get scores greater than 1000

Area verses Inverse questions

- X -> z-score -> area [AREA QUESTION]
- Area -> z-score -> X [INVERSE QUESTION]
  - See Cereal Company problem in Sharpe et al
    - Question 2
    - Question 3

Macho Chips Example

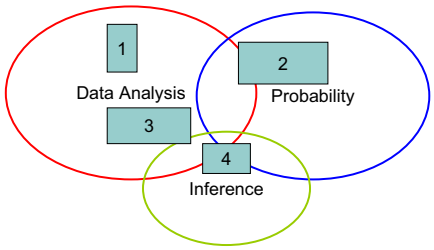
Extend Macho Chips Example

- Consider tabulated summaries (just above part (e)). Now assume that both X and Y are normally distributed, with parameters as tabulated

a) What is the probability that total sales will exceed $350/day

b) What percentage increase in price would be required to assure that the probability of exceeding $350/day was equal to 2.5%.

Birds Eye view of course: Three Phases (Map)



Returning to Data: Recall first week

- Data-types
- Displays for categorical data
  - Bar chart
  - Pie-chart
  - (example, Titanic data)
- As simple as making piles (heights of piles are proportional to counts)

Week 7 Material

- Graphical displays for quantitative data
  - Histograms of continuous data
  - Stem and leaf plots (just in time for the beach)
- Describing histogram (stem and leaf)
  - Qualitative statements
  - Quantitative summaries

EXAMPLE: End-of-year Value ($)

100	128	201	180	154	198	152	178	172	168
183	157	175	108	167	164	160	281	176	184
166	132	210	144	137	183	205	219	163	125
130	160	167	133	171	149	203	189	249	176
140	172	189	117	140	116	180	237	171	218

206	148	138	150	133	154	164	120	136	147
207	87	153	211	227	133	122	145	113	175
136	70	81	153	165	105	60	181	161	147
149	133	104	199	97	152	218	212	120	141
176	135	231	124	131	189	198	172	116	126

TopBox EXAMPLE: Sort Data ($)

100	108	116	117	125	128	130	132	133	137
140	140	144	149	152	154	157	160	160	163
164	166	167	167	168	171	171	172	172	175
176	176	178	180	180	183	183	184	189	189
198	201	203	205	210	218	219	237	249	281

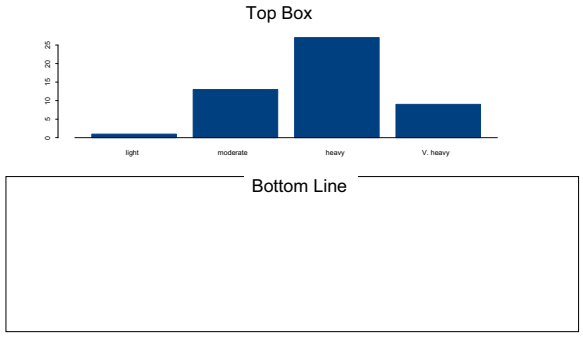
EXAMPLE: SORT + COLOR

100	108	116	117	125	128	130	132	133	137
140	140	144	149	152	154	157	160	160	163
164	166	167	167	168	171	171	172	172	175
176	176	178	180	180	183	183	184	189	189
198	201	203	205	210	218	219	237	249	281

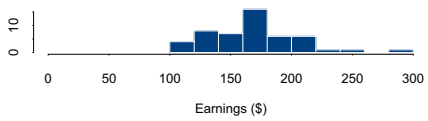
0 to 100	light
101 to 150	moderate
151 to 200	heavy
201 and up	Very Heavy

60	70	81	87	97	104	105	113	116	120
120	122	124	126	131	133	133	133	135	136
136	138	141	145	147	147	148	149	150	152
153	153	154	161	164	165	172	175	176	181
189	198	199	206	207	211	212	218	227	231

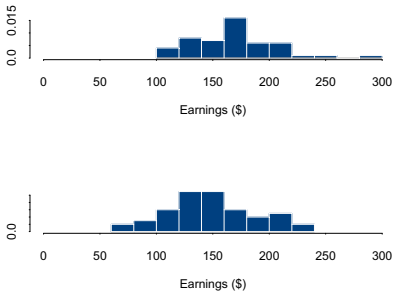
Categorizing Data



EXAMPLE: Histogram (Counts)



EXAMPLE: Histogram (Density)



Histogram options: platform dependent

- In Minitab (for Histogram)
  - Y-Scale options (frequency/percent/density)
    - Frequency = counts
    - Percent = counts/(total counts)
    - Density (ensures that total area =1)

Simple Steps for Getting Frequency Distribution (just the tabulation)

- First, slice up the entire span of values covered by the quantitative variable into equal-width piles called bins.
- The bins and the counts in each bin give the distribution of the quantitative variable.

From DVB

Frequency Distribution: Continuous Data

- Continuous Data: may take on any value in some interval

Example: A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

24, 35, 17, 21, 24, 37, 26, 46, 58, 30,
32, 13, 12, 38, 41, 43, 44, 27, 53, 27

(Temperature is a continuous variable because it could be measured to any degree of precision desired)

Grouping Data by Classes

Sort raw data in ascending order:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

- Find range: 58 - 12 = 46
- Select number of classes: 5 (usually between 5 and 20)
- Compute class width: 10 (46/5 then round off)
- Determine class boundaries:10, 20, 30, 40, 50
- Compute class midpoints: 15, 25, 35, 45, 55
- Count observations & assign to classes

Frequency Distribution Example

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

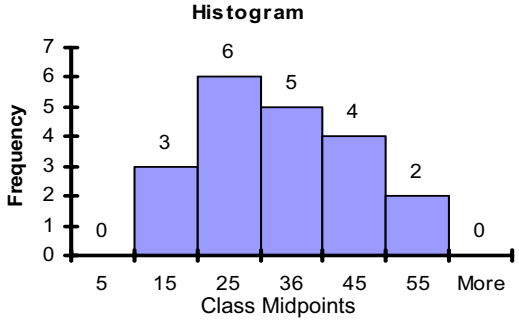
Frequency Distribution

Class	Frequency	Relative Frequency
10 but under 20	3	.15
20 but under 30	6	.30
30 but under 40	5	.25
40 but under 50	4	.20
50 but under 60	2	.10
Total	20	1.00

Histogram Example

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58



No gaps between bars, since continuous data

Choices When Constructing Frequency Distribution or Histogram

- Number of classes (5-20)
- Span of each bin/class (largely determined by number of classes)
- Starting point of first bin
- Lots of guidance for making these choices, but not central to this course

Stem and Leaf Diagram

- A simple way to see distribution details in a data set

METHOD: Separate the sorted data series into leading digits (the **stem**) and the trailing digits (the **leaves**)

Example:

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

- Here, use the 10's digit for the stem unit:
  - 12 is shown as
  - 35 is shown as

Stem	Leaf
1	2
3	5

Example:

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

- Completed Stem-and-leaf diagram:

Stem	Leaves
1	2 3 7
2	1 4 4 6 7 7 8
3	0 2 5 7 8
4	1 3 4 6
5	3 8

Using other stem units

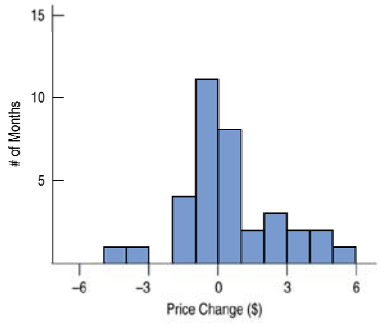
- Using the 100's digit as the stem:
  - Round off the 10's digit to form the leaves
    - 613 would become
    - 776 would become
    - . . .
    - 1224 becomes

Stem	Leaf
6	1
7	8
12	2

Enron Data: Monthly Stock Price Changes

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1997	-$1.44	-0.75	-0.69	-0.88	0.12	0.75	0.81	-1.75	0.69	-0.22	-0.16	0.34
1998	0.78	0.62	2.44	-0.28	2.22	-0.50	2.06	-0.88	-4.50	4.12	1.16	-0.50
1999	3.28	3.34	-1.22	0.47	5.62	-1.59	4.31	1.47	-0.72	-0.38	-3.25	0.03
2000	5.72	21.06	4.50	4.56	-1.25	-1.19	-3.12	8.00	9.31	1.12	-3.19	-17.75
2001	14.38	-1.08	-10.11	-12.11	5.84	-9.37	-4.74	-2.69	-10.61	-5.85	-17.16	-11.59

Enron Example: Comparing Histogram with Stem-and-Leaf



-4	5
-3	1
-2	
-1	7642
-0	99777554321
0	13567888
1	25
2	124
3	33
4	13
5	6

Shape, Center, and Spread

- When describing a distribution, make sure to always tell about three things: shape, center, and spread…

The Shape of the Distribution

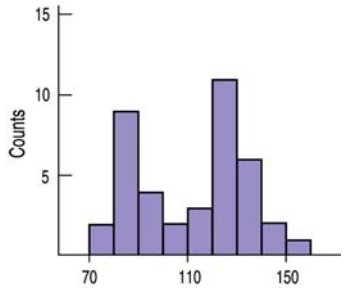
- When talking about the shape of the distribution, make sure to address the following three questions:
  1. Does the histogram have a single, central hump or several separated bumps?
  2. Is the histogram symmetric?
  3. Do any unusual features stick out?

Humps and Bumps

1. Does the histogram have a single, central hump or several separated bumps?
   - Humps in a histogram are called modes.
   - A histogram with one main peak is dubbed unimodal; histograms with two peaks are bimodal; histograms with three or more peaks are called multimodal.

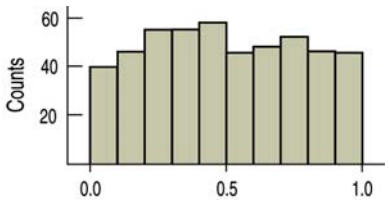
Humps and Bumps (cont.)

- A bimodal histogram has two apparent peaks:



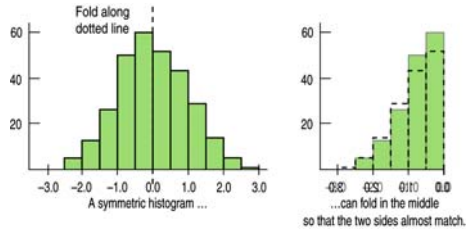
Humps and Bumps (cont.)

- A histogram that doesn't appear to have any mode and in which all the bars are approximately the same height is called uniform:



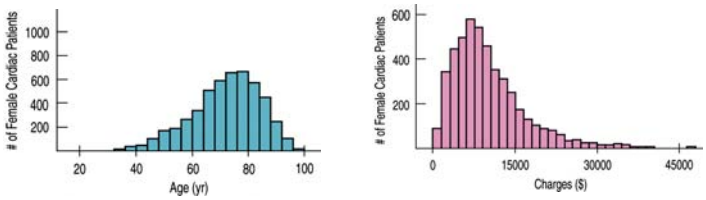
Symmetry

2. Is the histogram symmetric?
   - If you can fold the histogram along a vertical line through the middle and have the edges match pretty closely, the histogram is symmetric.



Symmetry (cont.)

   - The (usually) thinner ends of a distribution are called the tails. If one tail stretches out farther than the other, the histogram is said to be skewed to the side of the longer tail.
   - In the figure below, the histogram on the left is said to be skewed left, while the histogram on the right is said to be skewed right.



Anything Odd?

3. Do any unusual features stick out?
   - Believe it or not, sometimes it's the unusual features that tell us something interesting or exciting about the data.
   - You should always mention any stragglers, or outliers, that stand off away from the body of the distribution.
   - Are there any gaps in the distribution? If so, we might have data from more than one group.

Comparing Distributions

- Often we would like to compare two or more distributions instead of looking at one distribution by itself.
- When making such comparisons it is important that the histograms have been put on the *same scale.* Otherwise, we cannot really compare the two distributions.

What Can Go Wrong?

- Don't make a histogram of a categorical variable—bar charts (or pie charts) should be used for categorical data.
- Choose a scale appropriate to the data.
- Avoid inconsistent scales, either within the display or when comparing two displays.
- Label clearly!

Key Concepts

- Quantitative variables can be displayed using histograms, and/or stem-and-leaf displays. These displays help us to see the distributions of the variables.
- Consider three things when looking at these displays: shape, center, and spread.
- Distributions can be classified as symmetric or skewed (look at how the tails behave with respect to the rest of the distribution).

Key Concepts (cont.)

- A mode is a hump or local high point in the shape of the distribution:
  - unimodal (one mode)
  - bimodal (two modes)
  - multimodal (more than two modes)
  - uniform (relatively flat, no mode)
- Be on the lookout for outliers (extreme values that stand off away from the bulk of the data).

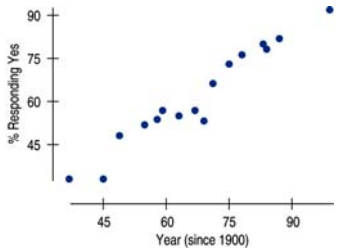
Scatterplots

- Scatterplots may be the most common and most effective display for data.
  - In a scatterplot, you can see patterns, trends, relationships, and even the occasional extraordinary value sitting apart from the others.
- Scatterplots are the best way to check if two quantitative variables are related and the ideal way to picture such associations.

Looking at Scatterplots

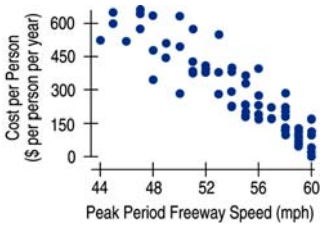
- When looking at scatterplots, we will look for direction, form, and scatter.
- Direction:
  - A pattern that runs from the upper left to the lower right is said to have a negative direction.
  - A trend running the other way has a positive direction.

Looking at Scatterplots (cont.)



- Figure illustrates a positive association between the year since 1900 and the % of people who say they would vote for a female US president.

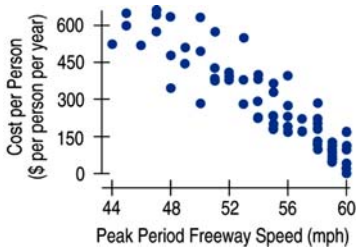
Looking at Scatterplots (cont.)



- Figure illustrates a negative association between peak period freeway speed and cost per person of traffic delays.
- As the peak period freeway speed increases, the cost per person of traffic delays decreases.

Looking at Scatterplots (cont.)

- Form:
  - If there is a straight line (linear) relationship, it will appear as a cloud or swarm of points stretched out in a generally consistent, straight form.
  - Example:



Looking at Scatterplots (cont.)

- Form:
  - If the relationship isn't straight, but curves gently, while still increasing or decreasing steadily,

we can often find ways to make it more nearly straight. Transformations but beyond the scope of this class.

Looking at Scatterplots (cont.)

- Form:
  - If the relationship curves sharply,

the methods of this book cannot really help us.

Looking at Scatterplots (cont.)

- Scatter:
  - At one extreme, the points appear to follow a single, consistent, stream

(whether straight, curved, or bending all over the place).

Looking at Scatterplots (cont.)

- Scatter:
  - At the other extreme, the points appear as a vague cloud with no discernable trend or pattern:

  - Note: we will quantify the amount of scatter soon.

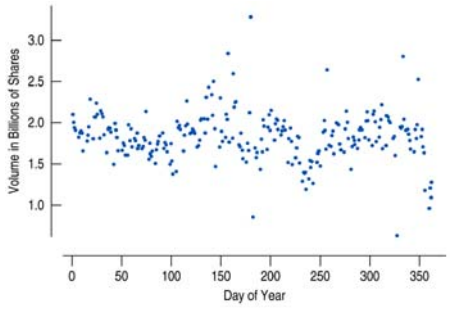
Looking at Scatterplots (cont.)

- Look for the unexpected—often the most interesting thing to see in a scatterplot is the thing you never thought to look for. One example of such a surprise is an outlier standing away from the overall pattern of the scatterplot.

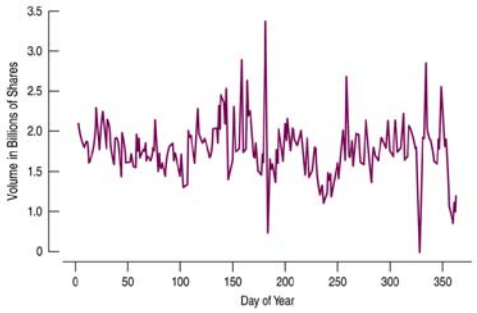
Roles for Variables

- It is useful to determine which of the two quantitative variables goes on the *x*-axis and which on the *y*-axis. By convention, this determination is made based on the roles played by the variables.
- When the roles are clear, the explanatory or predictor variable goes on the *x*-axis, and the response variable goes on the *y*-axis.

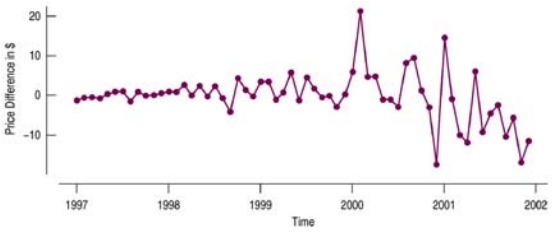
Time Series Plot of daily volume



Time Series: Connecting the Dots



Enron Stock Price changes



Scatter-plots, Association and Causation

- Suppose you collect data for each pair of variables. You want to make a scatter-plot. Choose the explanatory variable. Why?
  - a). T-shirts on sale: price/shirt, number sold
  - b). Scuba diving: depth, pressure
  - c). depth, visibility
  - d). Elem. Sch. Students BW, reading capability
  - h) Age and run-time in marathon

Earnings EXAMPLE Again: TopBox SORTED

100	108	116	117	125	128	130	132	133	137
140	140	144	149	152	154	157	160	160	163
164	166	167	167	168	171	171	172	172	175
176	176	178	180	180	183	183	184	189	189
198	201	203	205	210	218	219	237	249	281

What if you need a few "*bottom line*" quantitative summaries for your data?

Let's begin by considering a ***typical*** value

TYPES OF SUMMARY STATISTICS

- A typical value (location, or central value)
- We might also want to report on "just how typical a value it is."
  - This question refers to the notion of spread or dispersion. (These are non-technical terms.)
- In practice there is no single way to define:
  - Location; and
  - spread

A SIMPLE MEASURE OF LOCATION: MEDIAN

- `Midpoint' of the data
- Sort data from smallest to largest. The median is the value in the middle
- Median exceeds half the values in dataset, and is exceeded by the other half
  - If odd number of datum then median is the $(n/2+1)^{st}$ value
  - If even number of datum then median is chosen as the average of the two mid-values

TopBox Earnings EXAMPLE: DETERMINING THE MEDIAN

MEDIAN
(168+171)/2=$169.5

100	108	116	117	125	12	130	132	133	137
140	140	144	149	152	154	157	160	160	163
164	166	167	167	168	171	171	172	172	175
176	176	178	180	180	183	183	184	189	189
198	201	203	205	210	218	219	237	249	281

Another example:3,5,7,7,38

ANOTHER MEASURE OF LOCATION: THE MEAN

- The mean or average value is obtained by adding all the data values together and then dividing that sum by the number of observations.

Earnings (TopBox) EXAMPLE: CALCULATING THE MEAN

- Sum all 50 values

  100+108+…+281= 8477 $

  N=50

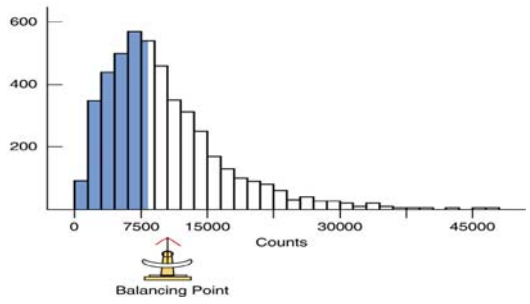
  Mean=(Sum x)/N

  =8477/50

  =$169.5 [compare $169 for median]

$$\bar{x} = \frac{x_1 + x_2 + x_3 \ldots x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

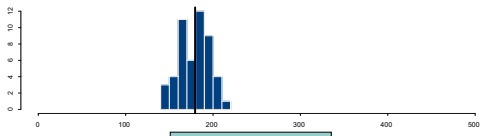
Median versus Mean (xbar)

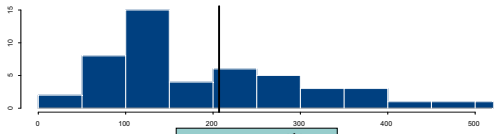


COMPARISON OF MEAN VERSUS MEDIAN

- Median is insensitive to extreme values
- Mean is sensitive to extreme values
- Median and mean can be (dis)similar

HOW TYPICAL?
MEASURES OF SPREAD (DISPERSION)





MEASURE OF SPREAD: IQR

- Inter-Quartile Range (IQR)
- Identified in a similar way as the Median
- `The idea is to split the data into two equal sized groups (upper and lower halfs) and see how far apart the upper and lower groups are.'
- Identify median of
  - lower half of data (Q1), and
  - Upper half of data (Q3)

TopBox Earnings: IQR

Q1=$144

Q3=$184

100	108	116	117	125	128	130	133	133	137
140	140	144	149	152	154	157	160	160	163
164	166	167	167	168	171	171	172	172	175
176	176	178	180	180	183	183	184	189	189
198	201	203	205	210	218	219	237	249	281

IQR = Q3-Q1

= 184-144 = $40

Example: Quiz Scores

	Group					
	1	2	3	4	5	6
Quiz Scores (out of 20)	10	8	0	0	0	4
	10	10	10	8	2	6
	10	10	10	10	10	8
	10	10	10	12	18	14
	10	12	20	20	20	18

Example: Moose versus Tofu

Mooseburgers		McTofu	
Al	$123	Ken	$110
Boris	$136	Latisha	$115
Connie	$144	Maria	$130
Dwight	$150	Nate	$100
Ernie	$110	Otto	$120
Francois	$131	Pablo	$146
Gloria	$140	Quentin	$117
Horace	$160	Rosa	$129
Isaac	$120	Sally	$360
Juan	$130	Ted	$132
		Uta	$107