The UNIVERSITY of VERMONT

Benjamin Fenster
STAT-141 (MacPherson)
21 September 2003

Notes – Basic Data Analysis

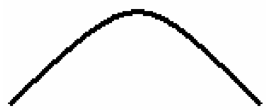
I. Data Types
  a. Quantitative
    i. Observations based on a numerical scale.
    ii. Continuous: Corresponds to an entire interval on the number line.
    iii. Discrete: Set of possible values corresponds to an isolated set of points on the number line.
  b. Qualitative: Everything else
  c. Example
    i. EU = a car
    ii. X = Number of Cylinders = {3, 4, 5, 6, 8, 12} discrete
    iii. Y = ¼ Mile Time = $\{y \mid 8 \le y \le 15\}$ continuous
    iv. Z = Brand = {GM, Ford, …} qualitative
  d. Type of analysis depends largely on the type of data.
  e. Some measurements are intrinsically discrete but we'll still treat them as continuous.
    i. Example: Cost in dollars.
    ii. There are discrete values ($1.01, $1.02, …)
    iii. There are so many, however, that it's better to treat the data as continuous

II. Statistics
  a. What motivates statistics?
    i. We need to answer a question.
    ii. Should I buy a car? Will this drug be effective?
  b. Procedure
    i. Begin with a question
    ii. Design a study (STAT-231 does this in detail)
    iii. Gather Information (STAT-233: Sampling Problems)
    iv. Summarize the Data
    v. (Need some probability at this point)
    vi. Make an inference about the population
    vii. Reach some conclusions
  c. Tools
    i. Need some probability
    ii. Use the computer to help us with each step.

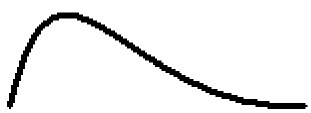
III. Summarizing Data
  a. Start with raw data (we'll assume a proper study was designed, et cetera)
  b. Data Reduction
    i. Too much information isn't useful.
    ii. We need to reduce the amount.
    iii. Order the Data
      1. A simple step that can help quite a bit.
      2. x_i is one measurement from a single EU
      3. x_1 is the first measurement taken, x_2 the second
      4. $x^{(i)}$ is an ordered measure
      5. $x^{(1)}$ is the minimum value, $x^{(2)}$ the next highest
      6. $x^{(1)}$ called the first order statistic, $x^{(2)}$ the second order statistic
      7. Sometimes written $x_{(1)}$ or $x_{\mathbf{1}}$ (subscript in bold)
    iv. Represent Graphically: Techniques discussed later
    v. Do a numerical summary
      1. Find some numbers that "represent" the data
      2. Location
        a. mean, median, etc
        b. Where is the data "located"
      3. Spread: How diverse is the data?

IV. Features of a Data Set

a. Center
   i. What's a typical observation?
   ii. About what point is the data centered?
b. Spread
   i. How much variability?
   ii. Bell Curve



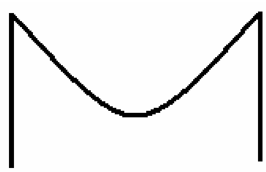
   iii. Skewed to the Right



   iv. Two Populations at Once

   v.

   vi.



c. Shape
d. Outliers: Unusual Observations
e. Gaps
f. Clusters
   i. The example above with two populations has two clusters.
   ii. Might be appropriate to separate into two different groups

V. Graphical Display
a. Histograms are one traditional display
b. Stem-and-Leaf
   i. See the Homes handout
   ii.

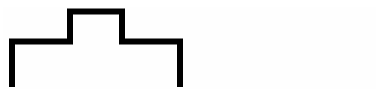
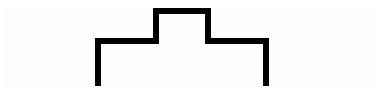
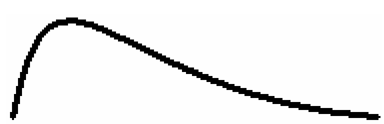
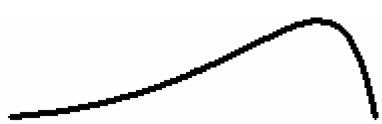
0	9	0	9
1	684	1	468
2	59	2	59
3	5	3	5
4	72	4	27
5		5	
6		6	
7	0	7	0

   iii. Stems on left, represent 10s digit

iv. Leaves on right, represent 1s digit
v. This can be done easily without the computer
vi. Can see min, max easily
vii. Can identify clusters, outliers (red flags: recording errors? mistakes?)
viii. Minitab Report
    1. Includes depths – the numbers on the far left
    2. These are the distance to the nearest end of the dataset.
    3. Standard Format
        a. One line per stem
        b. Like that shown above
    4. Stretched
        a. Two lines per stem
        b. O * O HI (0, 1, 2, 3, 4)
        c. O ● O LO (5, 6, 7, 8, 9)
    5. Squeezed
        a. Five lines per stem
        b. O * (0, 1)
        c. O t (2, 3)
        d. O f (4, 5)
        e. O s (6, 7)
        f. O ● (8, 9)
    6. Guidelines
        a. Pick a number of lines between $\sqrt{n}$ and $2\sqrt{n}$
        b. Truncate, don't round (otherwise it's too hard to refer back to the raw data)
        c. Use one digit leaves
        d. No commas, spaces or decimals
        e. The objective is to simplify, so keep it simple.

c. Frequency Table
    i. See the Voles handout
    ii. Category
        1. Denoted i
        2. "Litter Size" in the Voles example
    iii. Absolute Frequency
        1. Denoted F_i
        2. Number of observations in each category
    iv. Relative Frequency
        1. Denoted RF_i
        2. F_i / n
        3. $RF_3 = 13/170 = 0.0765 = 7.65\%$
    v. Cumulative Relative Frequency
        1. Denoted CRF_i
        2. Sum of all RF_J for $0 \le j \le i$
        3. $CRF_3 = 1/170 + 2/170 + 13/170 = 16/170 = 0.0941$
        4. CRF_N will always be 100%
    vi. Cumulative Distribution Function
        1. CDF
        2. $F(a) = P(x \le a)$
        3. Proportion of elements that are less than or equal to a
        4. Can also be considered the probability that $x \le a$
        5. Voles Example
            a. $F(6) = 0.6353 = P(x \le 6)$
            b. $F(8) = 0.9353 = P(x \le 8)$
    vii. Event Probabilities
        1. $P(x > 6)$

        a. Not in the table!
        b. We do have $P(x \le 6)$
        c. $P(x > 6) = 1 - P(x \le 6) = 0.3647$
        d. “Upper tail” probability or percentage
        e. Watch out for $\ge$ vs $>$ et cetera
    2. $P(3 < x \le 8)$
        a. $P(x \le 8) - P(x < 3)$
        b. $P(x \le 8) - P(x \le 2)$
        c. $F(8) - F(2) = 0.9177$

VI. Numerical Descriptive Measures
- a. Location
    - i.
      
    - ii.
      
    - iii. Two sets of data, identical in shape but positioned differentlky
    - iv. Population Mean
        1. Denoted μ
        2. $\mu = (x_1 + x_2 + \ldots + x_N) / N$
    - v. Sample Mean
        1. Denoted $\bar{x}$.
        2. $\bar{x} = \sum x_i / n = T / n$
        3. T used for “Total”
    - vi. Median
        1. Middle value
        2. “Resistant” (insensitive to the presence of extreme values)
        3. $\sim\mu$ (population), $\sim x$ (sample)
    - vii. Example
        1. Data = {1, 2, 3}
            - a. $\bar{x} = {}^{(1 + 2 + 3)} / {}_{3} = 2$
            - b. $\sim x = x^{(2)} = 2$
            - c. Symmetric distribution
        2. Data {1, 2, 30}
            - a. $\bar{x} = 11$
            - b. $\sim x = 2$ (resistant!)
    - viii. Skewed to Right / Positive Skew
        1. 
        2. $\bar{x} > \sim x$
    - ix. Skewed to Left / Negative Skew
        1. 
        2. $\sim x > \bar{x}$
    - x. There are some exceptions, but generally the data fits this pattern.
    - xi. Trimmed Mean

1. $\bar{x}_{tr}$ = Sum from 2 to n – 1 (eliminate smallest and largest values)
2. $\bar{x}_{tr\ (5\%)}$ Trim 5% off the top and 5% off the bottom.
3. Minitab uses so-called 5% trimmed mean
4. Wants to get as close to 5% as it can if the number of elements doesn't allow exactly 5% to be trimmed.

xii. What to Use

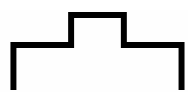
1. Symmetric (no outliers) mean
2. Symmetric (with outliers) trimmed mean
3. Skewed median

xiii. Binary Data

1. Sample n = 100
2. x = {1 approve
0 o/w
3. $\bar{x}$ = (1 + 0 + 0 + 1 + … + 1 + 0) / 100 = (49(0)+ 59(1)) / 100 = 59/100
4. Gives the proportion that approve!
5. Usually designate this p rather than $\bar{x}$ but thesame concept works just as well.

b. Dispersion

i. Example A

1. 
2. $\bar{x}$ = 20
3. R = 5

ii. Example B

1. 
2. $\bar{x}$ = 20
3. R = 15

iii. Range

1. Denoted R
2. Total "width" of the data
3. R = $x^{(n)} – x^{(1)}$

iv. Need a more versatile measure than Range.

v. Variance

1. $\Sigma\ |x_i - \mu|\ /\ N \geq 0$
   a. Almost right.
   b. We can improve still further.
2. Population Variance
   a. $\sigma^2 = \Sigma\ (x_i - \mu)^2\ /\ N$
   b. Always ≥ 0
   c. Has some nice mathematical and statistical properties.
   d. The problem is that the units get squared too.
3. Standard Deviation: $\sigma = \sqrt{(\sigma^2)}$
4. Sample Variance
   a. $s^2 = \Sigma\ (x_i - \bar{x})^2\ /\ N$
   b. We usually divide by (N – 1) to adjust for the tendency to underestimate.
5. Sample Standard Deviation $s = \sqrt{(s^2)}$

vi. Example

1. EU = a bat (the critter, not the stick)
2. n = 11
3. x = distance (cm)
4. x = {62, 23, …, 83}

5. Statistics
    a. $\bar{x} = 48.4$ cm
    b. R = 60 cm
    c. $s^2 = 327 \text{ cm}^2$
    d. s = 18.1 cm

vii. Computational Formula for Variance
1. $s^2 = (\Sigma x^2 - (\Sigma x)^2/n) / (n - 1)$
2. Equivalent mathematically but harder to understand conceptually.
3. Use this when calculating.

c. Quartiles

i. Median cuts data in half.
ii. Quartiles cut data in fourths.
iii. Quartiles = Fourths = Hinges
iv. When n is odd
1. The book says to exclude the median from each half (pg 105)
2. Minitab includes the median in each half.
3. We'll side with minitab.

v. Inter-Quartile Range
1. Denoted IQR
2. $IQR = Q_3 - Q_1$

vi. Five Number Summary
1.

N 25			*Sample Size*
M 13	15593		*Median*
H 7	13685	16457	*Q_1 / Q_3*
1	12784	22934	*Min / Max*

vii. Letter Value Display
1. For larger datasets, may want a more complete summary (1/4, 1/16, …)
2. N Sample Size
3. M Median
4. F Fourth *Also H (Hinge).*
5. E Eighth
6. D 1/16 *John Tukey noticed the pattern with F..E.. and*
7. C 1/32 *decided to continue it down to A, then loop*
8. B 1/64 *around back to Z.*
9. A 1/128
10. Z 1/256

d. BoxPlot

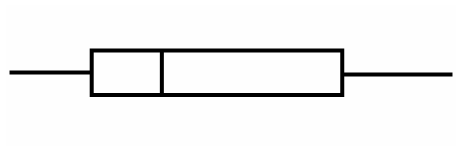
i. A graphical representation using these data.
ii. Need Q_1, Q_2, Q_3, IQR
iii. Inner Fences
1. Boundaries beyond which an observation is considered unusual
2. Lower = $Q_1 - 1.5(IQR)$
3. Upper = $Q_3 + 1.5(IQR)$
4. Bat Example
    a. Lower = 13685 – 1.5(2772) = 9527
    b. Upper = 16457 + 1.5(2772) = 20615
5. These are *mild* outliers

iv. Outer Fences
1. Lower = $Q_1 - 3.0(IQR)$
2. Upper = $Q_3 + 3.0(IQR)$
3. These are *extreme* outliers.

v. Why 1.5 ?

1. With the classic bell curve, we want only 1 in 100 observations to be mild outliers.
2. The value 1.5(IQR) is derived from this.
3. Consider 1.5(IQR) to be one step. Then inner fences are one step away, outer fences are two steps away.

vi.



1. Box width is IQR
2. Center line at Median
3. Whiskers extend to most extreme non-outlier – called adjacents.
4. Minitab marks mild outliers with * and extreme outliers with o
5. Very effective graphic, good to use for final projects.
6. Not good for seeing gaps or clusters.
7. Good for displaying many graphs at the same time.

VII. Explanations for Outliers

a. A mistake!
    i. Does it come from the wrong population?
    ii. Perhaps all funds measured were growth funds but one.

b. Recording Error
    i. The wrong value was recorded.
    ii. Correct it if the correct value can be found. Otherwise remove it.

c. Faulty Measurement Device
    i. The machine taking the measurements may simply be out of calibration.
    ii. This would render all data “strange” in comparison with some other source.

d. A Rare Event
    i. If all other explanations fail, assume the value is legitimate.
    ii. The first thought should always be that it’s some kind of error!

VIII. Z-Score

a. $Z = (x - \mu) / \sigma$

b. (Calculated for a particular ER)

c. Positive = Right of mean, Negative = Left

d. Dividing by σ puts everything on a common scale.

e. A Z score of 1.06 indicates that the measurement is 1.06 standard deviations from the mean.

ERROR: undefinedfilename
OFFENDING COMMAND: </FONT>

STACK: