Political Science 100a/200a

Fall 2001

Descriptive Statistics: summarizing data[1]

Let $Y = (y_1, y_2, y_3, \ldots, y_i, \ldots, y_{n-1}, y_n)$ be a *vector* (i.e., list) of values of a variable.

Examples ...

Defn: A *statistic* is a function that assigns a number to a set of values of a variable Y.

Examples: the average of a list of numbers is a statistic; so is the maximum of the list.

Defn: A *sample* is a subset of the population for which the researcher has data. A *population* is either

1. the set of concrete individuals or units of analysis about which the researcher would like to generalize, or
2. a *process* that produces values of a variable of interest to the researcher. (This is really a "random variable," but "population" is commonly used in this sense, at least conceptually, in social science.)

Examples: public opinion surveys vs. coin-flipping. Dyad-years? congressional elections? cities and crime, etc.

Illustrate with **sample** command in Stata.

Defn: *Descriptive statistics* are techniques for summarizing and describing characteristics of a sample. *Inferential statistics* are procedures for drawing

[1]Notes by James D. Fearon, Dept. of Political Science, Stanford University, October 2001.

inferences about the characteristics of a population (or a social, political, or economic process) from a sample.

If you have only a few observations, then you can (and should) just show your reader all of the data. If you have more than a few observations, you will need to *summarize it*. This can done with numerical summaries (statistics) or graphical summaries.

1. Graphical methods of summarizing information about a variable:
   (a) one-way and two-way scatterplots.
   (b) histograms, relative frequency and frequency. The idea of a "distribution" of values. Manipulating bins in Stata.
   (c) stem and leaf diagrams.
   (d) Cumulative distribution.
2. Statistics (numerical summaries of variables)

   We summarize the distribution of a variable with numbers representing the following (successive) aspects:

   - Where is it *centered*?
   - How *dispersed* or *spread out* is it?
   - Is it *skewed* more to one side or the other?
   - Are the "tails" "fat" or "this"? (This is called "kurtosis.")

   Some of the most important statistics that address these successive questions are called *moments* of the distribution of a variable.

   (a) Measures of central tendency
      i. the *mean* of a variable is just the average of its values. The mean is the *first (centered) moment* or a distribution of a variable.
      The sample mean is typically written
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

where n is the size of the sample. The population mean is typically written

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i,$$

where N is the size of the population.

ii. Rules for $\sum$: $Z = aX + bY + c$ implies $\bar{z} = a\bar{x} + b\bar{y} + c$. But $Z = XY$ does not imply $\bar{z} = \bar{x}\bar{y}$.

(b) the *median* is the value such that half of the other observations are greater and half less (if n even, then median is halfway between the "middle" two observations). Compare $X = (1, 2, 3, 4)$ and $X' = (1, 2, 3, 4, 5)$. And why does median equal mean in these cases?

- Which is "better," mean or median?
  - sensitivity to individual values.
  - the issue of outliers. Examples, msmt errors.
  - the moral of outliers.
- The meaning of the difference between mean and median.
- Why use one rather than other?
  - Often good to use both ...
  - Mean has some nice "statistical properties" ...
- What if your variable is a property, present or absent? What is mean then? What is the median?
- Show: (1) The mean minimizes the *mean squared error* $\frac{1}{n}\sum(x_i - a)^2$. Connection to regression. (2) The median minimizes the *mean absolute error* $\frac{1}{n}\sum|x_i - a|$.

(c) the *mode* is the most common value of a variable, or, with continuous data, the class with most observations (in a histogram, e.g.).

(d) α% trimmed means ...

3. Measures of dispersion

(a) the *range* is $\max(X) - \min(X)$. The *interquartile range* is the difference between the 75th and 25th percentiles.

- The zth percentile of a variable X is the value x_i such that $z\%$ of the values of X are smaller than x_i.
- check range after inputing data ...
- example of why not so helpful often ...

(b) the *variance* is the average of the squared differences between each x_i and the mean of X. This is the *second moment* of a distribution. For the population we typically write

$$var(X) = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2,$$

while for a sample ("sample variance"),

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

- $n - 1$ because otherwise s^2 would be systematically too low and estimate of σ^2; will show why later
- Show:
  i.
  $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \mu^2.$$
  ii. $var(X + c) = var(X)$.
  iii. $var(aX) = a^2 var(X)$.
- units, problems interpreting variance

(c) the *standard deviation* of X is the square root of $var(X)$.

$$\sigma = \sqrt{\frac{1}{N}\sum(x_i - \mu)^2}$$

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

- How interpret?
  Chebyshev's Theorem: For *any* variable X, at least $1-1/k^2$ of the values lie within k s.d.'s of the mean (where k is a number greater than or equal to 1).
  This is not a very tight bound. If histogram of X approximately follows a bell-shaped Normal curve,
  $$y = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$
  then about 68% of the observations will lie with 1 s.d. of mean, and about 95% within 2 s.d.s.
- Graph presidential approval: **graph approve ,bin(2) normal**.
- idea of *standard units*: $z_i = \frac{x_i-\mu}{\sigma}$ is the number of standard deviations that x_i is from the mean.
- The range R should be approximately $4s$, and thus $s = R/4$ is a *crude* estimate of sample s.d. (Show why)

4. Measures of skewness

- Loosely, a distribution/variable is said to be *right skewed* if it has a long right tail, and *left skewed* if it has a long left tail.
- **Defn:** The *kth central moment* of a distribution X is
  $$\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^k.$$
  Note that the $var(X)$ is the second central moment.
- The third central moment indicates skewness: Positive values indicate right skew, and negative values left skew (why?)
- To get a statistic in "standardized" units one typically divides the third central moment by σ^3.

- comments on higher moments, kurtosis ...

5. Box plots – another graphical tool

   - Steps to construct:
     (a) Draw horizontal lines at the median and the 25th and 75th percentiles (values of the variable ranged on the y axis). Connect to make a box.
     (b) Draw vertical lines up to value of most extreme data point that is within 1.5 times the interquartile range of the upper quartile. Likewise a line from the lower quartile.
     (c) Mark points beyond these with asterixes (possible outliers).
   - Indicates central tendency, dispersion, skewness, and possible outliers.
   - illustrate with presidential approval

6. Measures of association between two variables

   (a) The *covariance* of two variables X and Y is the average of the products of the deviations of each variable from its mean. Formally,

   $$cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

   - Note that $cov(X, Y)$ will tend to be positive when there are many cases i such that either both x_i and y_i are greater than their means, or both are less than their means. By contrast, X and Y will tend to "covary negatively" when $x_i > \bar{x} \rightarrow y_i < \bar{y}$ and vice versa. Show with graphs.
   - units $cov(X, Y)$ are the product of units of X and Y, and the number is not very information or descriptive by itself, except for the sign. Thus people typically use a standardized version of covariance ...

   (b) the *correlation coefficient* of two variables X and Y is the covariance of X and Y *when these are expressed in standard units.* Thus, if $X' =$

$(x'_1, x'_2, \ldots, x'_n)$ and $Y' = (y'_1, y'_2, \ldots, y'_n)$ denote X and Y expressed in standard units,

$$\rho(X, Y) = cov(X', Y').$$

This implies

$$\begin{aligned}
\rho(X, Y) \;=\; cov(X', Y') &= \tfrac{1}{n} \sum x'_i y'_i \\
&= \tfrac{1}{n} \sum \left(\tfrac{x_i - \bar{x}}{\sigma_x}\right) \left(\tfrac{y_i - \bar{y}}{\sigma_y}\right) \\
&= \tfrac{cov(X,Y)}{\sigma_x \sigma_y} \\
&= \frac{\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2} \sqrt{\frac{1}{n} \sum (y_i - \bar{y})^2}} \\
&= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}
\end{aligned}$$

About the correlation coefficient:

- $\rho(X, Y)$ is a pure number (no units – show) and $-1 \leq \rho(X, Y) \leq 1$.
- $\rho(X, Y) = \rho(Y, X) = \rho(aX + c, Y)$. (explain in words, noting that $sd(X) = sd(aX + c)$)
- ρ measures how tightly clustered the points in a scatterplot are around an upward or downward sloping line. (note discontinuity at flat or vertical line.) ρ is NOT a reliable measure of slope. Rather, *it is a measure of how reliably one can predict Y if you know X, and vice versa.*
- Important: Suppose there is a causal relationship between a dependent variable Y and an independent variable X, of the form $y_i = a + bx_i + \text{random error}_i$. The *correlation* between X and Y can differ markedly in different samples, depending on the standard deviations of X and Y in a given sample. Three illustrations:
  - i. correlation between **lifeexp** and **ln2gdp** after dropping gdp values one sd from the mean,
  - ii. Predicting graduate school performance from GREs of grad students,

iii. Achen's example.