Descriptive Statistics

Jamie Monogan

University of Georgia

Introduction to Political Methodology
Objectives

By the end of this meeting participants should be able to:

- Distinguish the various measures of central tendency and calculate them.
- Distinguish the various measures of dispersion and calculate them.
- Calculate and interpret quantile and percentile statistics.
Measures of Centrality

- Suppose we have a univariate dataset: \(x_1, x_2, x_3, \ldots, x_n\), which is designated differently when sorted: \(x(1), x(2), x(3), \ldots x(n)\).

- The sample mean is:

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

- The median is:

\[
m = \begin{cases} 
  x_{(n/2+1)} & \text{where } n \text{ is odd} \\
  \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) & \text{where } n \text{ is even}
\end{cases}
\]
The mode is the most common case:

\[ M = x_i \ni \#(x_i) > \#(x_j) \forall i \neq j \]

The midrange is the middle of the range of the data:

\[ \text{mid} = \frac{1}{2}(x(n) - x(1)) \]

The Inter-Quartile Range is the middle 50% of the data:

\[ IQR = [x(n/4), x(3n/4)] \]
Calculating Measures of Centrality

R

library(foreign)
names(nes.2004)
mean(nes.2004$bigbus_therm, na.rm=TRUE) #careful!!!
median(nes.2004$bigbus_therm, na.rm=TRUE)
IQR(nes.2004$bigbus_therm, na.rm=TRUE)

Stata

describe
describe bigbus_therm, detail
Measures of Dispersion

- The sample variance:
  \[ \text{Var}(X) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n - 1} = \frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

- The standard deviation:
  \[ \text{SD}(X) = \sqrt{\text{Var}(X)} = \sqrt{\frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \]

- A very different measure is the median average deviation:
  \[ \text{MAD}(X) = \text{median}(|x_i - \text{median}(x)|) \]
Calculating Measures of Dispersion

**R**

```r
var(nes.2004$bigbus_therm, na.rm=TRUE) # again, careful
sd(nes.2004$bigbus_therm, na.rm=TRUE)
mad(nes.2004$bigbus_therm, na.rm=TRUE)
```

**Stata**

```stata
sum bigbus_therm, detail
egen busMAD=mad(bigbus_therm)
ta busMAD
```

Normally-distributed variables tend to follow the empirical rule, also called the three sigma or 68-95-99.7 rule.
Quantiles and Percentiles

- Quantiles are the relative placement of data values in the sorted list: (0 : 1). (The median is the 0.5 quantile.)

- Percentiles are the same thing scaled: (0 : 100).

- Also: quartiles (4 groups), and quintiles (5 groups).

R

```r
quantile(nes.2004$bigbus_therm, na.rm=TRUE)
quantile(nes.2004$bigbus_therm, na.rm=TRUE, 0.6)
diff(quantile(nes.2004$bigbus_therm,na.rm=TRUE,c(.25,.75)))
```

Stata

```stata
tab bigbus_therm
```
Sometimes it helps to “standardize” the data in order to make generic comparisons.

Basic idea: subtract the mean and divide by the standard deviation for each point:

\[ z_i = \frac{x_i - \bar{x}}{SD(x)} \]
Creating Z-Scores

### R

```r
a <- mean(nes.2004$bigbus_therm, na.rm=TRUE)
b <- sd(nes.2004$bigbus_therm, na.rm=TRUE)
z.bigbus <- (nes.2004$bigbus_therm - a)/b
hist(z.bigbus)

z.bigbus.2 <- scale(nes.2004$bigbus_therm)
hist(z.bigbus.2)
```

### Stata

```stata
egen zBus = std(bigbus_therm)
sum ZBus
hist zBus
```