Stats1 Definitions and Variables

Evan Misshula

2017-09-22

Data Management and Descriptive Statistics

1. Topics for masters statisics B_block

   1.1 Frequency Distribution Presentation, & Central Tendency I
   1.2 Measures of Dispersion
   1.3 Data Distribution and Variance
   1.4 Hypothesis Testing Unit of Analysis
   1.5 Basic Probability, Inference
   1.6 Significance Tests

1. Topics for masters statisics II B_block

   1.1 Chi Square, Expected values & Mean Testing Continued
   1.2 Analysis of Variance (ANOVA)
   1.3 Associations Nominal and Ordinal Data, Bivariate Correlation,
   1.4 Pearson's r and Spearman's Rho
   1.5 Bivariate Regression
   1.6 Multiple Regression

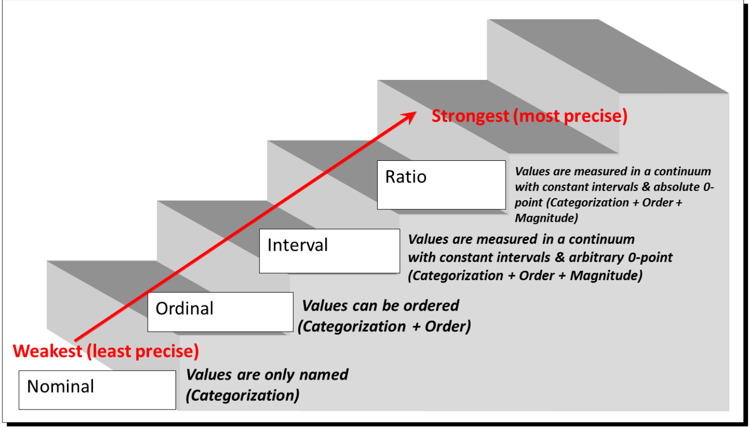
1. Variables B_definition

   - **Variables** are characteristics that change from person to person or object to object in a population of interest
     - Variables can take different values or levels

- They are called variables because they vary between cases
- Each variable has a level of measurement.

1. Levels of Measurement B_block:BMCOL

    - Each variable has a level of measurement.
    - The level of measurement is important because it determines the type of analysis.
    - There are four levels of measurement



1. Categorical, Nominal or Qualitative data B_block
   - The word nominal means names
   - A nominal variable ONLY describes something
   - The only fuction is to label and categorize

- NO INHERENT NUMERIC QUANTITY, NO Ranking of levels or ordering scheme
  - Numbers can be nominal level if there is no quantity associated with them.
  - Categories should be distinct, mututally exclusive, and completely exhaustive

1. Examples B_block

- Sex
- Religious Affiliation
- Race
- Zip code
- State

1. Ordinal Variables B_definition

- Ordinal variables have a ranking
- The root of Ordinal is "Ord" for **order**
- The exact amount of difference is unknown
- rank in most wanted

1. Examples B_block

- rating at a restaurant

- rank in the police department
- letter grade
- Service ratings
  - 1=Poor
  - 2=Fair
  - 3=Good
  - 4=Excellent

1. Interval variable B_definition
   - *Interval Variables* have inherently numeric values
   - We can talk about the difference between two items

1. Interval Variable Gotcha B_block
   - Temperature has difference
   - '0' is arbitrary
   - Does not indicate that there is no temperature

1. Ratio Variables B_block
   - the same as interval but with a *true* 0
   - Number of children
   - Age
   - Prior Arrests

- Commute time

1. Dependent and Independent B_block
    - Dependent Variable is what you are trying to predict
    - Independent is what you are using
    - There are many names for each

1. Synonyms for Dependent B_block
    - outcome
    - response

1. Synonyms for Independent B_block
    - feature
    - explanatory
    - causal

1. Validity B_definition
    - *Validity*: Addresses the question of whether the variable used actually reflects the concept or theory you seek to examine.
    - *Reliability*: A measure is reliable if it is consistent and stable.
        - *stable* is if it remains the same when measured in the same group

- *reliable* is if the same person in different groups will score similarly

1. Univariate descriptive statistics describes one variableB_block

   - Univariate statistics is composed of:
     - Central Tendency
     - Measures of Dispersion
     - Form of the distribution

1. What is a statistic? B_definition:BMCOL
   - A *statistic* is one number that summarizes many numbers
2. Example statistics B_example:BMCOL
   - a batting average
   - a shooting percentage
   - an *average* length of stay for pretrial detention
   - a *median* income for a zip code

1. Goal of Central Tendency B_block
   - We want the single best number that describes **typical** case

- The measure of central tendency you choose depends on the level of measurement of the variable

1. Mode B_definition
   - The *mode* is the most frequently occurring value
   - It is the only MCT appropriate for Nominal Data

1. Median B_definition
   - The *median* is the value in where half the values are greater and half are less

1. Mean B_definition
   - The *mean* is the sum of the values divided by the number of values

1. Calculation example B_block

```
rdata <- c()
rdata <- c(2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 7, 7)
rdata <- c(rdata,7, 7, 8, 8, 8, 9, 9, 10)
table(rdata)
table(rdata)[order(table(rdata),decreasing=TRUE)]
```

```
rdata
 2   3   4   5   6   7   8   9 10
 1   2   1   3   3   4   3   2  1
rdata
 7   5   6   8   3   9   2   4 10
 4   3   3   3   2   2   1   1  1
```

1. Median is the value that splits the distribution B_block
   - into two equal parts
   - often used with distributions that are skewed or have outliers

```
myMedian <- function(x) {
    x1 <- x[order(x,decreasing = F)]
    l <- length(x)
    if(l %% 2 == 0) {
return( .5*(x1[.5*l]+x1[.5*l+1]))
    } else {
return( x1[(l+1)/2])
    }
}
```

1. Do we get the same result for our own median B_block
   - function as the one built into R

```
median(rdata)
myMedian(rdata)
```

```
[1] 7
[1] 7
```

1. Mean is also known as the average B_block
   - Only for Interval or Ratio level data
   - Assumes order and equality of intervals
   - Very sensitive to outliers and skew
   - $\overline{X} = \frac{\sum X}{n}$

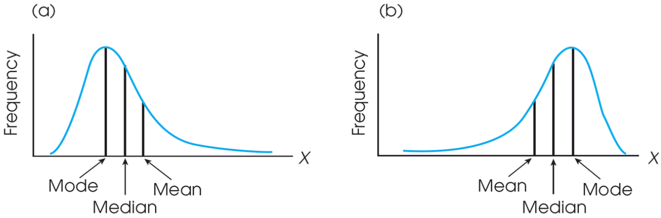
```
  myMean <- function(x) {
      myMean <- sum(x)/length(x)
  return(myMean)
  }
mean(rdata)
myMean(rdata)
```

```
[1] 6.25
[1] 6.25
```

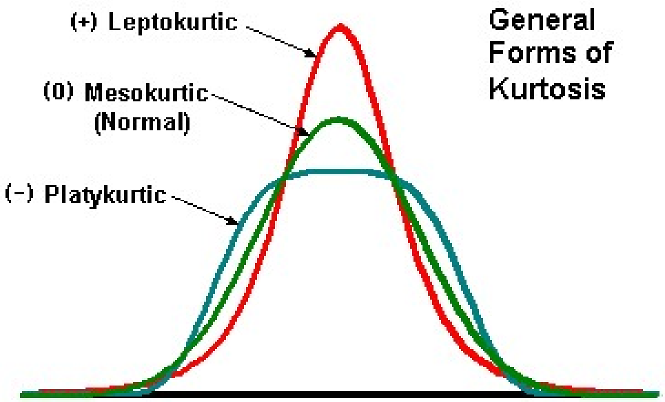
1. Mean, outliers and typical case B_block
   - the mean is sensitive to outliers
   - outliers can be important
   - crashes (stock market)
   - frequent (arrestees)
   - billionaires in income data
   - You need to understand what you are describing/modeling
   - **textbook answer** use the median in presence of skew and outliers

1. Defining skew B_definition
   - A distribution is skew if it is not symmetric
   - If a distribution is skewed it is conventional to use the median not the mean
   - median moves less for a given outlier than the mean

1. Skew examples B_ignoreheading Below are two asymmetrical distributions: (a) positive (right) skew, (b) negative (left) skew.



1. Skew examples B_ignoreheading In the masters program you are also responsible for recognizing different levels of kurtosis



1. Choose the most appropriate measure of B_block
   - central tendency (MCT) for:
     - SEX [sexr]
     - SEV MS ARREST CHARGE [asev1]
     - HOURS PER WEEK WORKED [hrspw]
     - UCR MS ARREST CHARGE [aucr1]
     - NUMBER PRIOR MISD CONV [pmis]

1. Recap so far B_block
   - *level of measurement* limit on the information in a variable
   - *central tendency* describe the typical case
   - *measures of dispersion* tell us how typical the typical case is
   - the measures we should use are dictated by the variable's level of measurement

1. Open the Rikers 1989 Data Set from Blackboard B_block
   - Choose the most appropriate MCT for:
     - SEX [sexr]
     - SEV MS ARREST CHARGE [asev1]
     - HOURS PER WEEK WORKED [hrspw]
     - UCR MS ARREST CHARGE [aucr1]
     - NUMBER PRIOR MISD CONV [pmis]

1. MCT R code B_block

```
library(foreign)
myRikers <- read.spss(file="rikers1989.sav", to.data.fra
head(rikers[1:3,1:4])
```

```
Warning message:
In read.spss(file = "rikers1989.sav", to.data.frame = TR
  rikers1989.sav: Unrecognized record type 7, subtype 18
  caseid                    aucr1    asev1              acd1
1      10 USE/POSS OTHER DRUGS    A MISD                DRU
2      46                ROBBERY C FELONY HARM TO PERS & PR
3      56 USE/POSS OTHER DRUGS    A MISD                DRU
```

1. Two distributions can have the same mean B_block
   - but very different spread of values
   - the amount of variation is very important
   - there can be a little, there can be a lot

1. More names for spread B_block
   - variation, dispersion
   - Variation is important because the typical case can be misleading
   - think of the crash or Bill Gates

1. Central Tendency vs. Variability B_block

- Central Tendency shows typical case
- *Measures of dispersion* show spread of values around the typical case

1. small data examples B_block

```
d1 <- c( 7, 6, 3, 3, 1)
d2 <- c( 3, 4, 4, 5, 4)
d3 <- c( 4, 4, 4, 4, 4)
c(mean(d1),  median(d1),   sd(d1))
c( mean(d2),   median(d2),   sd(d2))
c( mean(d3),   median(d3),   sd(d3))
```

```
[1] 4.00000 3.00000 2.44949
[1] 4.0000000 4.0000000 0.7071068
[1] 4 4 0
```

1. Nominal variables B_block

   - proportion in modal category
   - Index of Qualitative Variation

$$IQV = \frac{K(100^2 - \sum_{i=1}^{K} p_i^2)}{100^2(K-1)}$$

1. Measures of dispersion for ordinal variables B_block
   - proportion in modal category
   - Index of Qualitative Variation

1. Measures of dispersion for interval/ratio variables B_block
   - range
   - variance
   - standard deviation

1. Proportion in modal category formula B_block

$$\frac{Number_{modal}}{Total\ N} * 100$$

1. Proportion in modal category code B_block

```
rdata <- c(1,3,3,3,3,5)
freqTable <- table(rdata)
ordFreqTable <- freqTable[order(freqTable,decreasing = T
propOrdFTable <- prop.table(ordFreqTable)
100*propOrdFTable[1]
```

```
        3
66.66667
```

1. Computing IQV B_block

Fear of Crime	resp
Not Concerned at All	3
A Little Concerned	4
Quite Concerned	6
Very Concerned	20

1. Computing IQV in R B_block

```
myCr <- c(rep(1,3),rep(2,4),rep(3,6), rep(4,20))
myFreq <- table(myCr)
myFreq
```

```
myCr
 1  2  3  4
 3  4  6 20
```

1. continuing the calculation of IQV B_block

```
K <- length(myFreq)
myPropFreq <- prop.table(myFreq)
sqProp <- apply(X=myPropFreq,MARGIN = 1,FUN = function(x
sumSqProp <- sum(sqProp)
IQV <- (K/(K-1))*(1-sumSqProp)
IQV
```

```
[1] 0.7689011
```

1. The style of this B_block
   - is called imperative
   - if we want to make it so we can use it we make it a function

```
IQV <- function(myCr) {
    myFreq <- table(myCr)
    K <- length(myFreq)
    myPropFreq <- prop.table(myFreq)
    sqProp <- apply(X=myPropFreq,MARGIN = 1,
    FUN = function(x){return(x^2)})
    sumSqProp <- sum(sqProp)
    IQV <- (K/(K-1))*(1-sumSqProp)
```

```
    return(IQV)
}
```

1. There are three common measures of variability B_block
   - range
   - variance
   - standard deviation

1. Calculating range B_block
   - $Range = X_{maximum} - X_{minimum}$

```
myData<-c(35,60,80,93,98)
myRange <- max(myData)-min(myData)
myRange
range(myData)
```

```
[1] 63
[1] 35 98
```

1. Advantages of the range B_block:BMCOL
   - easy to compute
   - interval includes all data

2. Disadvantages of the range
    - crude because:
        - it relies on only two points
        - it ignores N-2 points
        - it is very sensitive to outliers

1. should never be used as the sole measure   B_block
    - More examples

```
d1 <- c(10,39,39,40,40,40,40,41,41,47)
d2 <- c(39,39,40,40,40,40,41,41,41,88)
d3 <- c(39,39,40,40,40,40,41,41,41,42)
myRanges<-c(max(d1),max(d2),max(d3))-c(min(d1),min(d2),m
myRanges
```

```
[1] 37 49  3
```

1. Variance and standard deviation   B_block
    - appropriate for Interval/Ratio data
    - not used for nominal or ordinal data
    - there is a one-to-one map from variance to standard deviation

```
 [1] 10 39 39 40 40 40 40 41 41 47
[1] 100.0111
[1] 100.0111
```

1. Formulas for sample variance B_block:BMCOL

   $var(x) = \frac{\sum_{i=1}^{N}(x_i - \bar(x))^2}{N}$

   ```
   d1
   var(d1)
   md1 <- mean(d1)
   sumSq <- unlist(lapply(X=(d1-md1),function(x){return(x^2
   myVar <- sum(sumSq)/(length(d1)-1)
   myVar
   ```

2. Formulas for standard deviation B_block:BMCOL

   $sd(x) = \sqrt(var(x))$

   ```
   sd(d1)
   sqrt(myVar)
   ```

   ```
   [1] 10.00056
   [1] 10.00056
   ```

1. Sample vs. Population Standard Devaiation B_block
   - They both use the difference between the mean ($ μ $ or $\overline{X}$) and each observation (individual X value)
   - The sample uses "n-1" instead of N to adjust for sampling error
   - Sample statistics are used to estimate population parameters
   - Samples tend to have less variation than the populations
   - n-1 is used to adjust for this by inflating a value we know to be artificially low

   - long mathematical derivation **not going to do it**

1. Example of univariate analysis B_block
   - Number of arrests (1,5,7,8,9)

```
arrestData <- c(1,5,7,8,9)
mean(arrestData)
sd(arrestData)
```

1. Insight into standard deviation B_block
   - The reason why the sum of the deviations from the mean is always zero:
     - The mean serves as a balance point for the distribution

- The total distance of individual scores above the mean is exactly equal to the total distance of those below the mean
- The result is the positive and negative numbers cancelling each other out
- The solution is to get rid of the negative signs
- Squaring the deviations makes every number positive

1. Drawbacks of standard deviation B_block:BMCOL
   - It takes into account all values in the distribution
   - Values far from the mean are given extra weight because deviations from the mean are squared
2. Drawbacks of standard deviation B_block:BMCOL
   - Standard deviation and variance are sensitive to outliers (extreme values)

- a display useful in summarizing distribution
- Goes back to Aurther Bowley's work in early 1900's
- Popularized by John Tukey in 1977's *Exploratory Data Analysis*

- class scores

Class Scores:

75	8	36	36	55
55	42	36	50	42
100	83	27	58	55
55	27	83	17	58
82	27	92	50	42

```
myScores<-c(75,8,36,36,55,55,42,36,50,42,100,83,27,58,55,55,
myOrdScores <- myScores[order(myScores)]
stem(myScores,scale=1)
```

```
stem(myScores,scale=2)
```

1. Topics for next time
   - Probability
   - The Gaussian Distribution and the "Normal" Curve
   - Risk of Error