Recitation 2: Introduction to Research Methods for Politics

Dept. of Politics, NYU

POL-850

Spring 2020

Agenda

1. Central tendency: mean, median
2. Spread: variance, standard deviation
3. Correlation

Central tendency: mean, median

Central tendency

1. Mean: central value of a discrete set of numbers;
   sum of the values divided by the number of observations
2. Median: the value in the middle of the distribution that divides the data into two equal-size groups

$$\text{median} = \begin{cases} \text{middle value} & \text{if no. of entries is odd} \\ \dfrac{\text{sum of two middle values}}{2} & \text{if no. of entries is even} \end{cases}$$

Central tendency

1. Calculate mean by hand and check with R:

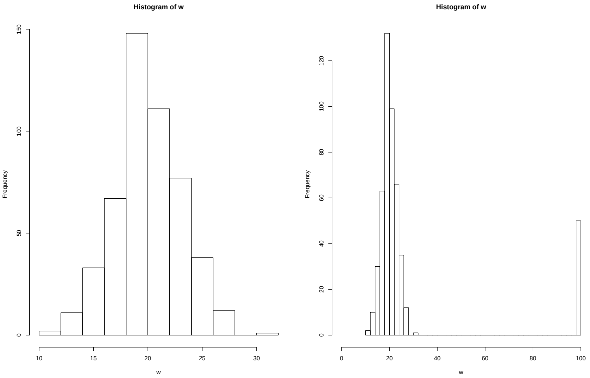
```
set.seed(1)
w <- rnorm(n = 500, mean = 20, sd = 3)  #### generate random sample
hist(w)

#### calculate mean by hand

sum(w)/length(w)
mean(w)   ### use built in mean() function to check
```

2. Calculate median by hand and check with R:

```
#### calculate median by hand
w[order(w)[length(w)/2]]
median(w)  ### use built in median() function to check
```

Central tendency

1. Median more robust to outliers.



2. Left: Mean = 20.06793; Median = 19.88967
3. Right: Mean = 28.0378; Median = 20.30732

Spread: variance, standard deviation

Spread

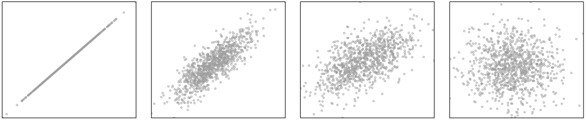
1. NB: In DSS, denominator is n;
2. Usually, according to the sampling theory, we calculate sample or finite population variance and standard deviation with denominator is n-1;
3. Calculate variance and standard deviation by hand and check with R:

```
#### calculate var, sd, corr by hand

x <- rnorm(n = 100, mean = 20, sd = 3)      #### generate random points
y <- 5*x + rnorm(n = 100, mean = 0, sd = 5)

plot(x,y)  ### see correlation

#### var of x, with denominator n-1

var_x <- sum( (x - mean(x))^2 ) / (length(x)-1)
var_x

var(x) ### use R function


#### sd of x, with denominator n-1

sd_x <- sqrt(var_x)
sd_x
```

Correlation: correlation coefficient

Correlation

1. Correlation coefficient: summarizes the direction and strength of the linear association between two variables
2. Ranges from -1 to 1
3. Positive when the two variables tend to move together and negative when they tend to move away from each other



Correlation

1. NB: when you calculate the correlation coefficient, denominators should be consistent with how you calculate standard deviation
2. In dss, denominator is n for sd;
   correlation coefficient also uses denominator n;
3. In R, sd ( ) uses n-1 as the denominator; corr ( ) also uses n-1;
4. $cor(X, Y) = \frac{\sum_1^n z_i^x \times z_i^y}{n}$ where $z_i^x = \frac{x_i - \bar{x}}{sd(x)}$ is the z-scores

Correlation

Calculate correlation coefficient by hand and check with R:

```r
### corr of (x,y), remember the denominator should be the same with sd (or var)

z_x <- (x - mean(x)) / sd_x    ### calculate z score
z_y <- (y - mean(y)) / sd(y)
sum(z_x * z_y) / (length(x) - 1)

cor(x,y)  ### use R function to check
```

Correlation

What if we use n as the denominator (as in dss)

```
##### what if we use denominator n
##### write a function: input: x,y output correlation coeficient; with denominator n

my_corr <- function(x, y){

  var_x <- sum( (x - mean(x))^2 ) / (length(x))
  var_y <- sum( (y - mean(y))^2 ) / (length(y))

  sd_x <- sqrt(var_x)
  sd_y <- sqrt(var_y)

  z_x <- (x - mean(x)) / sd_x
  z_y <- (y - mean(y)) / sd_y

  cor_xy <- sum(z_x * z_y) / (length(x))

  return(cor_xy)
}

my_corr(x,y)  ### our function uses denominator n
cor(x,y)  ### remember, R function uses denominator n-1
```

Application: UK districts.csv (see section 5 R script)