Q-Step: Week 3 Lecture

Descriptive Statistics and Visualization

Spyros Kosmidis

Oxford

January 28, 2022

Roadmap

Previously

- Research Design
- Concepts and Measurement

Today

- Descriptive Statistics
  - Central Tendency
  - Dispersion
- Visualization
  - Actual Example $\rightarrow$ UK GE (2017)

Next Week

- Case Selection

Understanding Data

- Everything that can be entered in a spreadsheet is Data
- Quantifying information helps us analyse complex phenomena
- It allows us to test theoretical expectations (check out week2)
- But, data analysis should not be too complex
- Today, Step 1: Summarize (and Visualize) Data
- But let's go through some definitions before we start looking at numbers

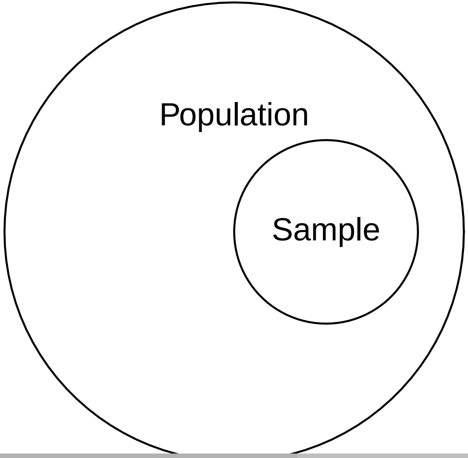
Populations and Samples

- Population$\rightarrow$ a set of cases (someone or something) that we aim to describe or draw inferences about
- Sample $\rightarrow$ a subset of cases drawn from a larger population that we actually use in our analysis

Populations and Samples

- Population→ a set of cases (someone or something) that we aim to describe or draw inferences about

- Sample → a subset of cases drawn from a larger population that we actually use in our analysis



Levels of Measurement

A Typical Definition of Variable

A trait that varies/differs across different units/subjects in a sample or population

Levels of Measurement

A Typical Definition of Variable

A trait that varies/differs across different units/subjects in a sample or population

Levels of Measurement

Levels of Measurement

A Typical Definition of Variable

A trait that varies/differs across different units/subjects in a sample or population

Levels of Measurement

- **Nominal**

Levels of Measurement

A Typical Definition of Variable

A trait that varies/differs across different units/subjects in a sample or population

Levels of Measurement

- **Nominal** (Mutually Exclusive -unordered- Categories, e.g. Gender)

Levels of Measurement

A Typical Definition of Variable

A trait that varies/differs across different units/subjects in a sample or population

Levels of Measurement

- **Nominal** (Mutually Exclusive -unordered- Categories, e.g. Gender)
- **Ordinal**

Levels of Measurement

A Typical Definition of Variable

A trait that varies/differs across different units/subjects in a sample or population

Levels of Measurement

- **Nominal** (Mutually Exclusive -unordered- Categories, e.g. Gender)
- **Ordinal** (Categories go from low to high levels in an ordered fashion, Agree-Disagree, Levels of Education, ordering does not say much about differences between levels)

Levels of Measurement

A Typical Definition of Variable

A trait that varies/differs across different units/subjects in a sample or population

Levels of Measurement

- **Nominal** (Mutually Exclusive -unordered- Categories, e.g. Gender)
- **Ordinal** (Categories go from low to high levels in an ordered fashion, Agree-Disagree, Levels of Education, ordering does not say much about differences between levels)
- **Interval**

Levels of Measurement

A Typical Definition of Variable

A trait that varies/differs across different units/subjects in a sample or population

Levels of Measurement

- **Nominal** (Mutually Exclusive -unordered- Categories, e.g. Gender)
- **Ordinal** (Categories go from low to high levels in an ordered fashion, Agree-Disagree, Levels of Education, ordering does not say much about differences between levels)
- **Interval** (Values are ordered but also linear in terms of their interpretation, e.g. Temperature)

Levels of Measurement

A Typical Definition of Variable

A trait that varies/differs across different units/subjects in a sample or population

Levels of Measurement

- **Nominal** (Mutually Exclusive -unordered- Categories, e.g. Gender)
- **Ordinal** (Categories go from low to high levels in an ordered fashion, Agree-Disagree, Levels of Education, ordering does not say much about differences between levels)
- **Interval** (Values are ordered but also linear in terms of their interpretation, e.g. Temperature)

Measures of Central Tendency and Dispersion

How can we characterize data?

Let's start with a simple example

- Say you are interested in the 2019 election
- And you want to examine the 5 constituencies surrounding the one you voted
- You collect the data (e.g. www.parliament.uk) for the Green Party
- GreenVote=$\{450,880,685,1750,1566\}$
- What is the mean level of Green support in the areas around you?
- Fairly easy task: You add up the #GreenVotes and then divide by the number of observations (i.e. 5)

You might have seen this function like this :

Let's start with a simple example

- Say you are interested in the 2019 election
- And you want to examine the 5 constituencies surrounding the one you voted
- You collect the data (e.g. www.parliament.uk) for the Green Party
- GreenVote={450,880,685,1750,1566}
- What is the mean level of Green support in the areas around you?
- Fairly easy task: You add up the #GreenVotes and then divide by the number of observations (i.e. 5)

You might have seen this function like this : $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

Let's do the calculations

```
sumofx=450+880+685+1750+1566
N=5
mu=sumofx/N
mu
```

```
## [1] 1066.2
```

1066.2 is the average number of votes the Greens got in 5 random constituencies during the 2019 election.

Let's do the calculations

```
sumofx=450+880+685+1750+1566
N=5
mu=sumofx/N
mu
```

```
## [1] 1066.2
```

1066.2 is the average number of votes the Greens got in 5 random constituencies during the 2019 election.

Let's proceed with some simple examples that will help you understand Central Tendency and Dispersion

Central Tendency, A Toy Example, Pt1

- If X=$\{4,5,5,5,6,7,15,3\}$, the mean is

Central Tendency, A Toy Example, Pt1

- If X={4,5,5,5,6,7,15,3}, the mean is

```
X=c(4,5,5,5,6,7,15,3)
mean(X)
```

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

Central Tendency, A Toy Example, Pt1

- If X={4,5,5,5,6,7,15,3}, the mean is

```
X=c(4,5,5,5,6,7,15,3)
mean(X)
```

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

- the median is

```
median(X)
```

```
## [1] 5
```

Central Tendency, A Toy Example, Pt1

- If X={4,5,5,5,6,7,15,3}, the mean is

```
X=c(4,5,5,5,6,7,15,3)
mean(X)
```

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

- the median is

```
median(X)
```

```
## [1] 5
```

Can you think of cases where the median is preferred to the mean?

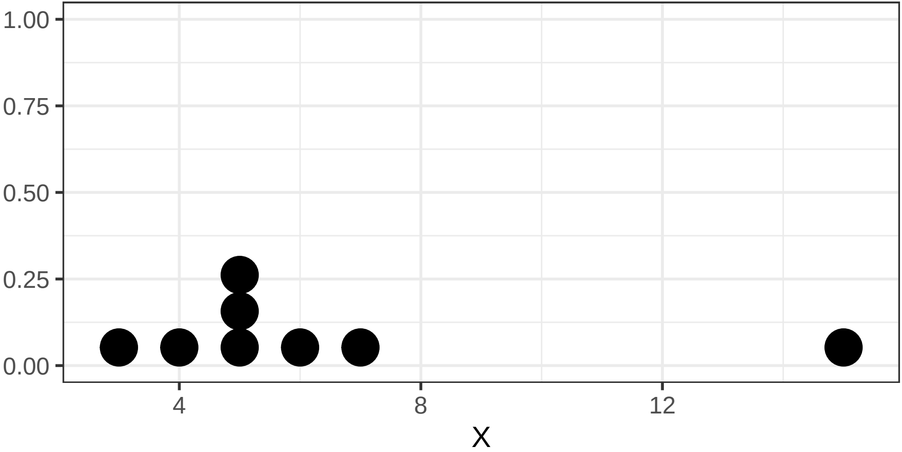
A Toy Example, Pt2

- Very often we might be interested in the values that appear more often than others. The most frequent value is called the **mode**

A Toy Example, Pt2

- Very often we might be interested in the values that appear more often than others. The most frequent value is called the **mode**

```
qplot(X, geom="dotplot", binwidth=0.6)+theme_bw()
```



Data Dispersion Measures

- Range: maximum value minus minimum value of the variable in a data set
- Variance (σ^2): sum of the squared deviations from the mean divided by n-1
- Standard deviation (σ): square root of the variance

Data Dispersion: Variance

- By inspecting deviations from the mean (i.e. the ‘typical observation’), we are able to see how disperced the data is
- These deviations are both negative and positive, in order not to arrive at a variance of zero we square the individual distances
- Here is the formula:

Data Dispersion: Variance

- By inspecting deviations from the mean (i.e. the 'typical observation'), we are able to see how disperced the data is
- These deviations are both negative and positive, in order not to arrive at a variance of zero we square the individual distances
- Here is the formula:

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

Data Dispersion: Variance

- By inspecting deviations from the mean (i.e. the 'typical observation'), we are able to see how disperced the data is
- These deviations are both negative and positive, in order not to arrive at a variance of zero we square the individual distances
- Here is the formula:

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

```
X=c(4,5,5,5,6,7,15,3)
sqdiff=(4-mean(X))^2+(5-mean(X))^2+(5-mean(X))^2+
  (5-mean(X))^2+(6-mean(X))^2+(7-mean(X))^2+
  (15-mean(X))^2+(3-mean(X))^2
variance=sqdiff/(length(X)-1)
variance
```

```
## [1] 13.92857
```

Data Dispersion: Standard Deviation

Data Dispersion: Standard Deviation

- The standard deviation is the most common way to measure deviation from the mean and is simply the square root of the variance
- Best way to think of it is as a kind of rough typical distance of an observation to the mean
- Very similar to variance, in fact, only a square root away:

Data Dispersion: Standard Deviation

- The standard deviation is the most common way to measure deviation from the mean and is simply the square root of the variance
- Best way to think of it is as a kind of rough typical distance of an observation to the mean
- Very similar to variance, in fact, only a square root away:

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

Data Dispersion: Standard Deviation

- The standard deviation is the most common way to measure deviation from the mean and is simply the square root of the variance
- Best way to think of it is as a kind of rough typical distance of an observation to the mean
- Very similar to variance, in fact, only a square root away:

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

- The standard deviation is crucial for inferential statistics. It is key to understand the statistical significance of a parameter estimate (be patient for a couple of weeks)

Recall the previous example

Recall the previous example

```
X=c(4,5,5,5,6,7,15,3)
var(X) #Same as before when we did it by hand!
```

```
## [1] 13.92857
```

Recall the previous example

```r
X=c(4,5,5,5,6,7,15,3)
var(X) #Same as before when we did it by hand!
```

```
## [1] 13.92857
```

```r
sd(X)
```

```
## [1] 3.7321
```

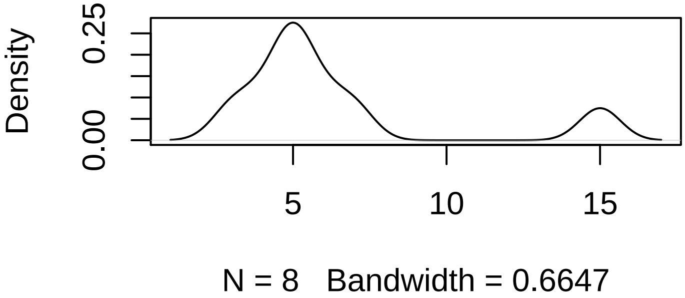
Recall the previous example

```
X=c(4,5,5,5,6,7,15,3)
var(X) #Same as before when we did it by hand!
```

```
## [1] 13.92857
```

```
sd(X)
```

```
## [1] 3.7321
```

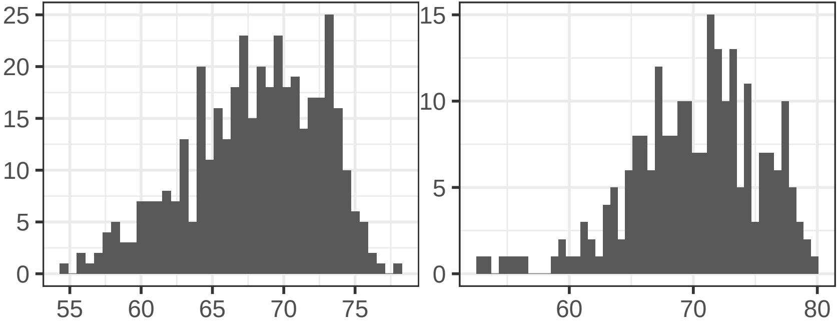


Now let's do a more interesting example

Q: What was the mean level of turnout in 2017 in Brexit and Remain constituencies?

```
brseat<-read.csv("brseat.csv")
```

Let's plot constituency turnout for Brexit and Remain constituencies:



What do you observe (if anything)?

Turnout in Brexit and Remain Areas (2017)

```
summary(brseat$Turnout[brseat$Year1==2017])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   53.02   65.42   69.16   68.75   72.39   79.52
```

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==0])
```

```
## [1] 67.96286
```

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==1])
```

```
## [1] 70.13509
```

```
var(brseat$Turnout[brseat$Year1==2017 & brseat$remain==0])
```

```
## [1] 21.28935
```

```
var(brseat$Turnout[brseat$Year1==2017 & brseat$remain==1])
```

```
## [1] 27.68395
```

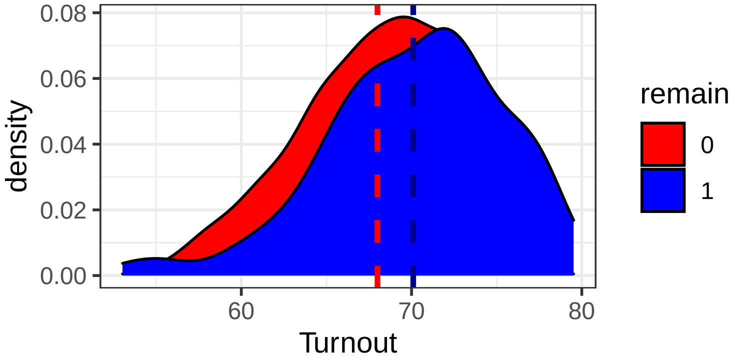
Turnout in Brexit and Remain Areas (2017)

```
median(brseat$Turnout[brseat$Year1==2017 & brseat$remain==0])
```

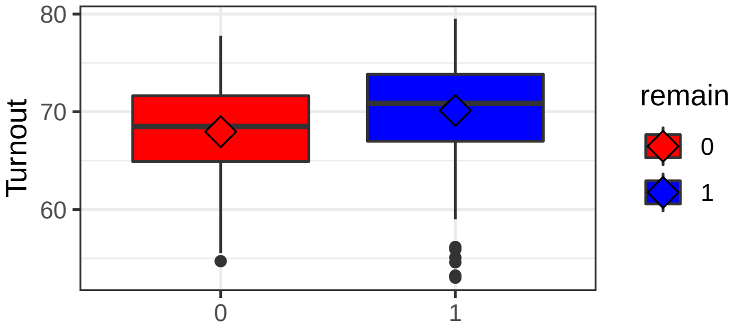
```
## [1] 68.50045
```

```
median(brseat$Turnout[brseat$Year1==2017 & brseat$remain==1])
```

```
## [1] 70.85917
```



As a boxplot



- **box**: 25 and 75th percentile
- **line**: 50th percentile aka median
- **diamond**: mean
- **circles**: some outliers

Turnout in Leave and remain Areas

Turnout in Leave and remain Areas

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==0])
```

```
## [1] 67.96286
```

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==1])
```

```
## [1] 70.13509
```

Turnout in Leave and remain Areas

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==0])
```

```
## [1] 67.96286
```

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==1])
```

```
## [1] 70.13509
```

Why is turnout higher in Remain areas?

Turnout in Leave and remain Areas

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==0])
```

```
## [1] 67.96286
```

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==1])
```

```
## [1] 70.13509
```

Why is turnout higher in Remain areas? Is it the referendum?

Turnout in Leave and remain Areas

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==0])
```

```
## [1] 67.96286
```

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==1])
```

```
## [1] 70.13509
```

Why is turnout higher in Remain areas? Is it the referendum?

```
mean(brseat$Turnout[brseat$Year1==2015 & brseat$remain==0])
```

```
## [1] 65.28172
```

```
mean(brseat$Turnout[brseat$Year1==2015 & brseat$remain==1])
```

```
## [1] 68.3166
```

Turnout in Leave and remain Areas

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==0])
```

```
## [1] 67.96286
```

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==1])
```

```
## [1] 70.13509
```

Why is turnout higher in Remain areas? Is it the referendum?

```
mean(brseat$Turnout[brseat$Year1==2015 & brseat$remain==0])
```

```
## [1] 65.28172
```

```
mean(brseat$Turnout[brseat$Year1==2015 & brseat$remain==1])
```

```
## [1] 68.3166
```

Unlikely to have influenced remain and leave areas differently.

Turnout in Leave and remain Areas

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==0])
```

```
## [1] 67.96286
```

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==1])
```

```
## [1] 70.13509
```

Why is turnout higher in Remain areas? Is it the referendum?

```
mean(brseat$Turnout[brseat$Year1==2015 & brseat$remain==0])
```

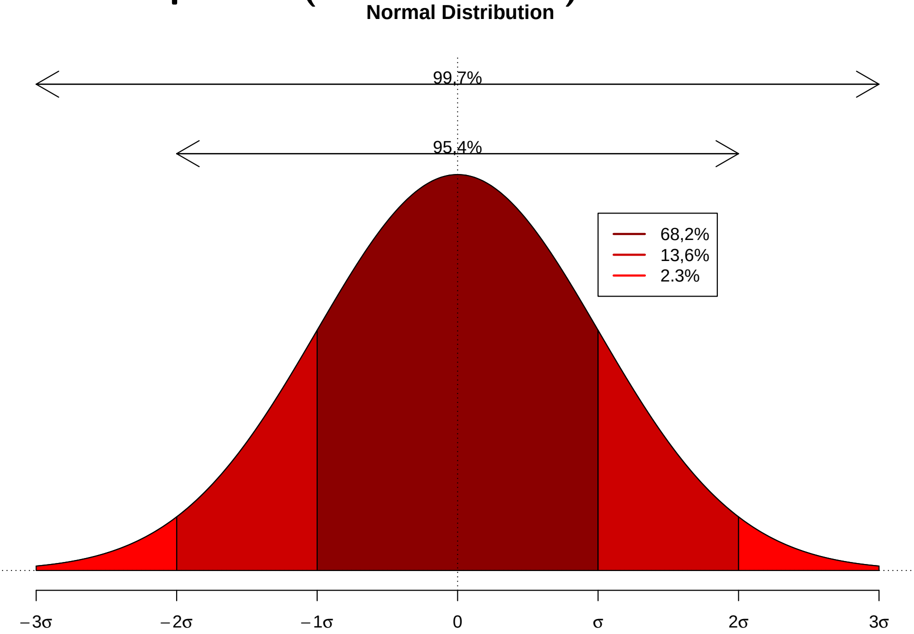
```
## [1] 65.28172
```

```
mean(brseat$Turnout[brseat$Year1==2015 & brseat$remain==1])
```

```
## [1] 68.3166
```

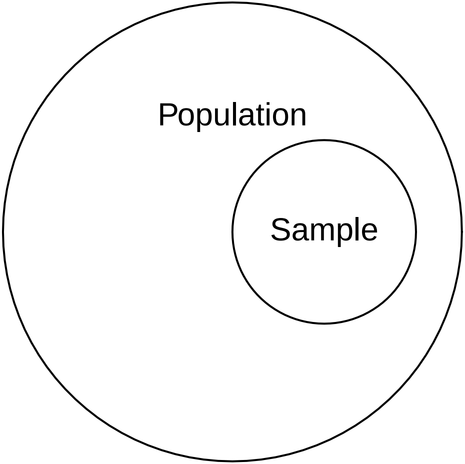
Unlikely to have influenced remain and leave areas differently. The effect is uniform, yet hard to establish causality!

Some Properties (for the Future)



Next Week: Case Selection

How do the cases you choose affect the conclusions you draw?



Thank you!