Statistical Reasoning

Week 4

Sciences Po - Louis de Charsonville

Spring 2018

Week 4 : Distributions

Research Paper

Distributions and graphs

Measures of central tendency

- Mean
- Median
- Mode

Measures of Variability

- Range(s)
- Standard deviation

Normal distribution

Research Paper

Research Paper

Timeline

Research Proposal	**Today**
1^{st} draft	**6 March**
2^{nd} draft	**10 April**
Final draft	**24 April**

Submission's Rules

- A **word** document (following template on the Google Drive).
- A **do-file** showing *all* commands in Stata with comments in green.

The Word document

- Provisional paper title
- Introduction stating and accounting the research question
- Brief theory section describing your hypotheses
  - Describe how you think the independent variable you chose are supposed to influence the dependent variable (better if you have a few references).
- Brief description of the dataset
  - Objectives of the survey, date, data collecting methods, sampling, etc.
- Description of the dependent and independent variables as they exist unmodified in the original data
  - Names, codes, values, what they measure, missing values
- Description of all variable renamings, recodings, how missing values have been managed ;
- Univariate statistics on all variables with 1/2 sentence(s) describing their distribution.

Exporting results from Stata into Word

- Tables : select `copy table`, paste in Excel, edit, paste in Word.
  - Add footnotes.
- Graphs : save in `.tif` format, insert as a picture in Word
- *More details :* section 13.4 in the **Stata Guide**.

Distributions and graphs

Distributions

- A distribution is a collection of data, or scores, on a variable.
- Scores are usually arranged in order from smallest to largest.

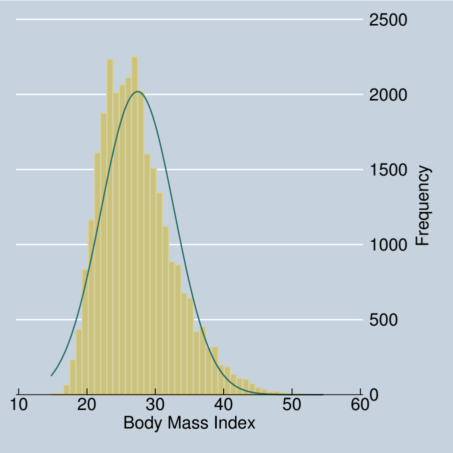


Figure – Distribution of BMI

Example - Gun control in the US

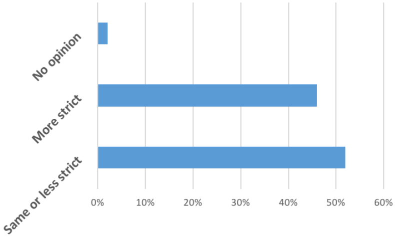
Would you like to see gun laws in the US made more strict, less strict, or remain as they are ?

- the same or less strict : 52%
- more strict : 46%
- no opinion : 2%

Example - Gun control in the US

Would you like to see gun laws in the US made more strict, less strict, or remain as they are ?

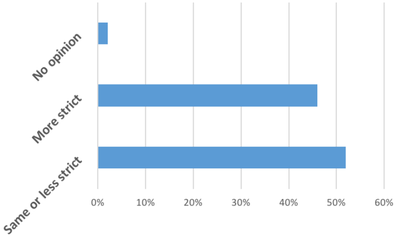
- the same or less strict : 52%
- more strict : 46%
- no opinion : 2%

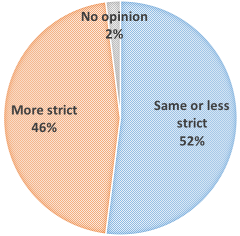


Example - Gun control in the US

Would you like to see gun laws in the US made more strict, less strict, or remain as they are ?

- the same or less strict : 52%
- more strict : 46%
- no opinion : 2%





Distribution of Qualitative Variables

Frequency distributions

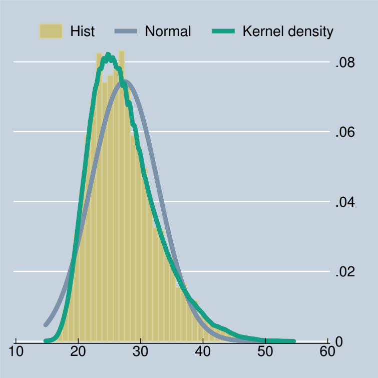
- Provide the number of observations in each category and/or the corresponding percentage
  - Be careful : percentages should sum to 100%
  - How have you dealt with *missing values* ?
- **Cumulative frequencies or percentages :**provide the number/percentage of observations below or equal to a given value or category (only with *ordinal data*)
- **Stata** : these statistics are obtained with `tab` or `fre`, and can be visualized using **bar graphs** and **histograms**

Distribution of Quanlitative Variables

- Categorize a quantitative variables.
  - Example : earnings in `nhis9711`
- Compute measures of central tendency and variability.
- Plot the probability density function (*kernel density*)

Distribution of Quanlitative Variables

- Categorize a quantitative variables.
  - Example : earnings in `nhis9711`
- Compute measures of central tendency and variability.
- Plot the probability density function (*kernel density*)



Measures of central tendency

Measures of central tendency

- Collection of scores of a variable : **distribution**
- How spread out the scores are ?
- What is the most common score ?
- etc.

One set of distribution characteristics that research are interested in is central tendency :

- mean
- median
- mode

Stata

- use `sum` or `tabstat`
- *primarily appropriate for quantitative variables*

Mean

- Arithmetic average of a distribution of scores :

$$\bar{x} = \sum_{i=1}^{N} \omega_i x_i \,, \text{ with } \omega_i \text{ weight of obs } i$$

- Most commonly used
- Denoted μ for the *population mean* and $\bar{x}$ for the *sample mean*

Weaknesses

- Sensitive to extreme values (*outliers*)
- A distribution may have very few scores near the mean

Mean

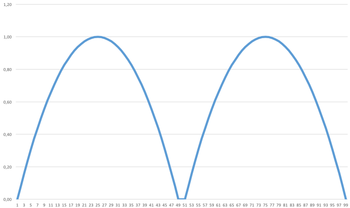
- Arithmetic average of a distribution of scores :

$$\bar{x} = \sum_{i=1}^{N} \omega_i x_i \,, \text{ with } \omega_i \text{ weight of obs } i$$

- Most commonly used
- Denoted μ for the *population mean* and $\bar{x}$ for the *sample mean*

Weaknesses

- Sensitive to extreme values (*outliers*)
- A distribution may have very few scores near the mean



$$X_1 = \{2,3,5,6\}$$

$$X_2 = \{0,3,5,8\}$$

X_1, X_2 have the same mean.

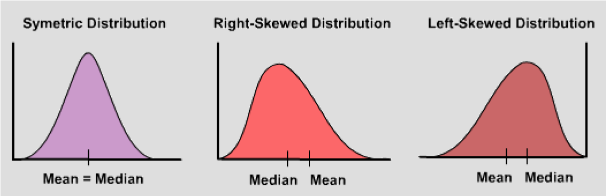
Median

- The score in the distribution that marks the 50th percentile
- 50% of the scores in the distribution fall above the median and 50% fall below it.
- Not sensitive to outliers.
- Comparing the mean and the median gives an idea whether the distribution is skewed or not.

Median

- The score in the distribution that marks the 50th percentile
- 50% of the scores in the distribution fall above the median and 50% fall below it.
- Not sensitive to outliers.
- Comparing the mean and the median gives an idea whether the distribution is skewed or not.

Skewed or not ?

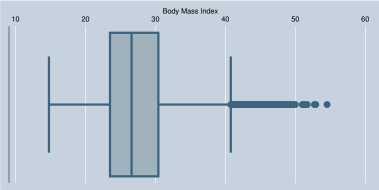


Beyond the median

- **Quartiles** : divide the sample into 4 equal parts
- **Deciles** : divide the sample into 10 equal parts
- **Percentiles** : divide the sample into 100 equals parts.

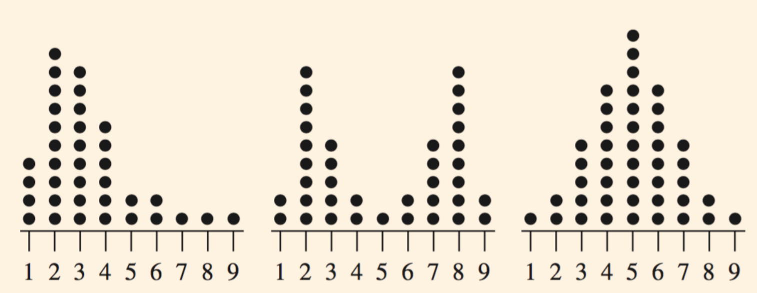
Stata

- use `summarize` (with options `details`)
- or draw a boxplot with `graph hbox`



Mode

- The **most frequent value** in the sample
- A series of values can be unimodal (one mode), bimodal(two modes) or multimodal (several modes).
- Not used a lot.



Example

$$X = \{86, 90, 95, 100, 100, 110, 110, 115, 120\}$$

- Mean ?
- Median ?
- Unimodal ? Bimodal ?

Measures of Variability

Measures of Variability

Old saying

"If your head is in the freezer and your feet are in the oven, on average you're comfortable."

Measures of Variability

Old saying

"If your head is in the freezer and your feet are in the oven, on average you're comfortable."

- Measures of central tendency do not inform us on the dispersion of scores in the distribution

Measures of dispersions

- Range
- Variance
- Standard deviation (most informative and widely used)

Range

- Range = difference between the largest score and the smallest score.

$$Range = Max - Min$$

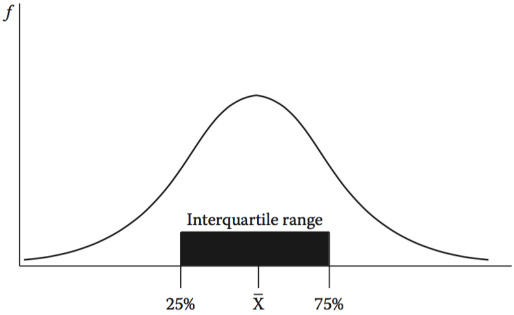
Range

- Range = difference between the largest score and the smallest score.

$$Range = Max - Min$$

- Another common measure : **Interquartile range (IQR)** :

$$IQR = Q_3 - Q_1$$



Standard deviation

- **Deviation** : refers to the distance between an individual score and in the average score
- **Standard** : means *average*
- Standard deviation is the average distance between individual observation and the mean of the distribution.

Formula

Population :

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

Estimate based on a *sample :*

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

Shape

The shape of the distribution refers to how the observations are distributed around the mean

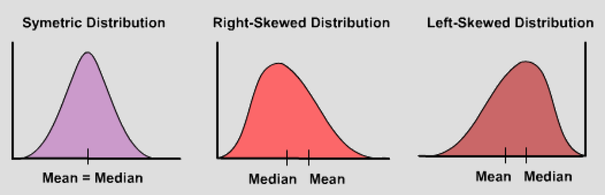
- symmetrically distributed ?
- Are the widely spread around the mean ? (Outliers ?)

Describing the shape :

- Skewness (asymmetry)
- Kurtosis (flatness)

Skewness

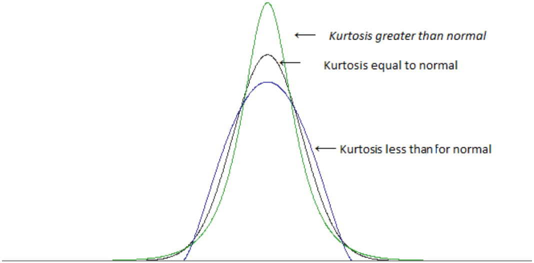
- **Right-skewed** (*positive skew*) : outliers pull the mean upwards (a few very high values), graphically the mean is pulled to the right, the right-hand tail is longer. Most observations are clustered at the lower end.
- **Left-skewed** (*negative skew*) : outliers pull the mean downwards (a few very low values), graphically the mean is pulled to the left, the left-hand tail is longer. Most observations are clustered at the higher end.



Kurtosis

The shape of a distribution of scores in terms of its flatness or peakedness (compared to the normal distribution)

- A normal distribution has a kurtosis of 3.
- **Leptokurtic** : a higher peak and thinner tails (than the normal curve, $kurtosis > 3$)
- **Platykurtic** : a lower peak and thicker tails (than the normal curve, $kurtosis < 3$)



Normal distribution

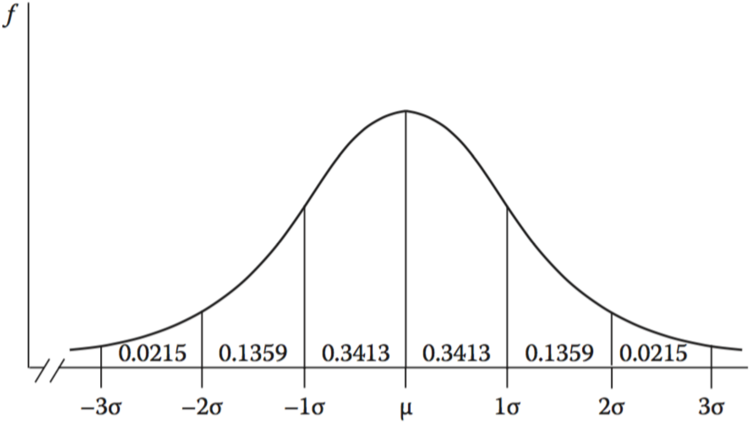
Normal distribution $\mathcal{N}(\mu, \sigma^2)$

- **Normal distribution** : extremely important to statistics
- often referred as the **bell curve**

Properties

- **symmetric** and **unimodal**
- **mean = median = mode**
- $\mathcal{N}(0,1)$ **standard normal distribution**

Normal curve



Normality assessment

Visual assessment

- Distributions : `hist, normal, kdensity, gr (h)box`
- Diagnostics : `symplot, qnorm, g(ladder)`

Formal assessment

- Use `su x, d` to assess the symmetry ($skewness \sim 0$) and flatness ($kurtosis \sim 3$) of a variable.
- Use `tabstat x y, s(skew kurt) c(s)` to compare a variable with its transformation (often to log-units)

PRACTICE