Research Methods for Social Sciences:

Working with numbers: Basic Tools

Alex Teytelboym & Earth Sivakul

@t8el

Hilary 2013

Outline

Qualitative data

- Descriptive statistics are the tabular, graphical, and numerical methods used to summarize data
- Summarizing qualitative data
  - Frequency Distribution
  - Relative Frequency Distribution
  - Bar Graph
  - Pie Chart

Qualitative data

Example

Village Health Service

- Clients of an International Health Service were asked to rate the quality of the services as being **excellent**, **above average**, **average**, **below average**, or **poor**.
- The ratings provided by a sample of 20 clients are: Below Average, Above Average, Above Average, Average, Above Average, Average, Above Average, Average, Above Average, Below Average, Poor, Excellent, Above Average, Average, Above Average, Above Average, Below Average, Poor, Above Average, Average.

Frequency distribution

Example

Village Health Service

Rating	**Frequency**
Poor	2
Below average	3
Average	5
Above average	9
Excellent	1
Total	**20**

Quantitative data

- Summarizing quantitative data
  - Frequency Distribution
  - Relative Frequency Distributions
  - Cumulative Distributions
  - Histogram
- A frequency distribution is a listing of intervals of possible values for a variable, together with a tabulation of the number of observations in each interval. It is formed by dividing possible values into intervals and counting observations in each interval. Intervals must be exhaustive and mutually exclusive. The number of intervals must be chosen carefully (not too many, not too few).

Frequency distribution

Example

Life expectancy at birth using World Bank 2009 data

Excel spreadsheet

Let's have a look at how you can do frequency distributions and histograms in Excel.

Cross-tabs and scatter diagrams

- Crosstabulation and a scatter diagram are two methods for summarizing the data for two variables simultaneously.
- A crosstabulation is a tabular summary of data for two variables.

Cross-tab

Example

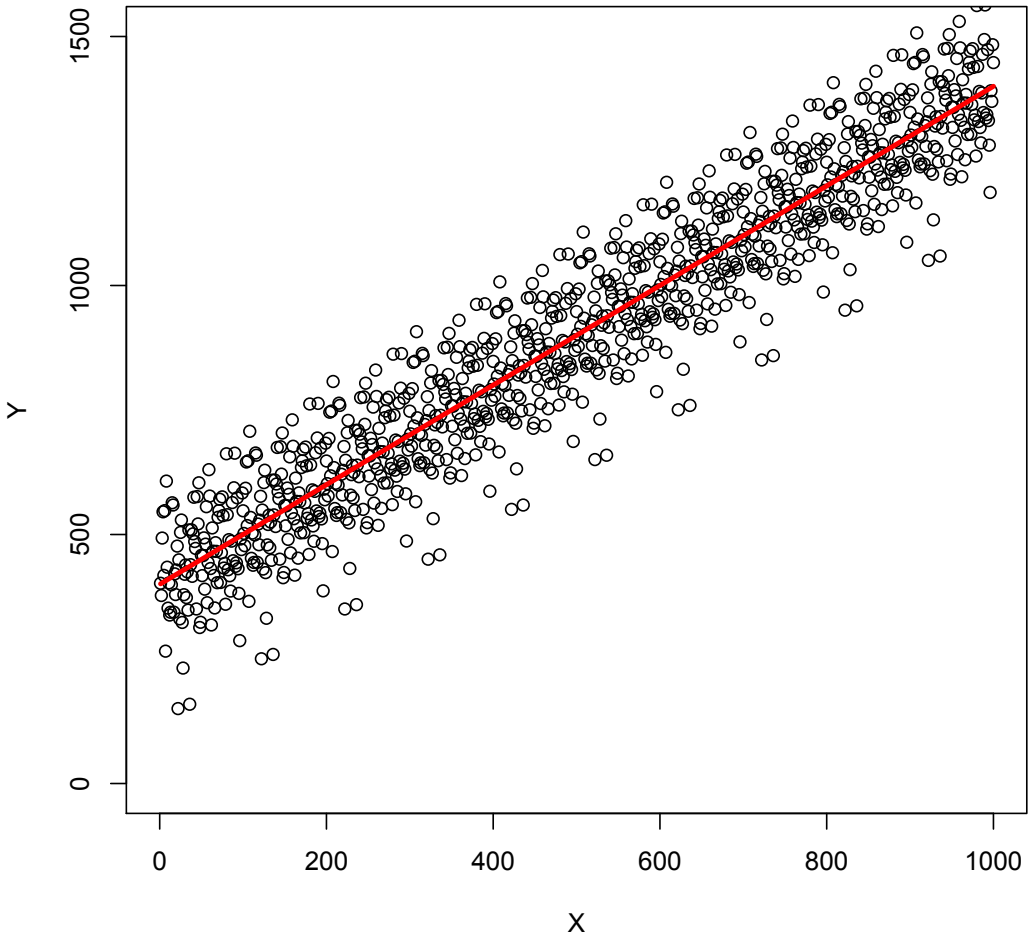
SOS Services NGO

The number of projects that SOS services had for each type of funding source and value for the past two years is shown below.

	Funding source				
Value range	**Gov**	**Private**	**IO**	**NGO**	**Total**
$\leq \$99,000$	18	6	19	12	55
$> \$99,000$	12	14	16	3	45
Total	30	20	35	15	100

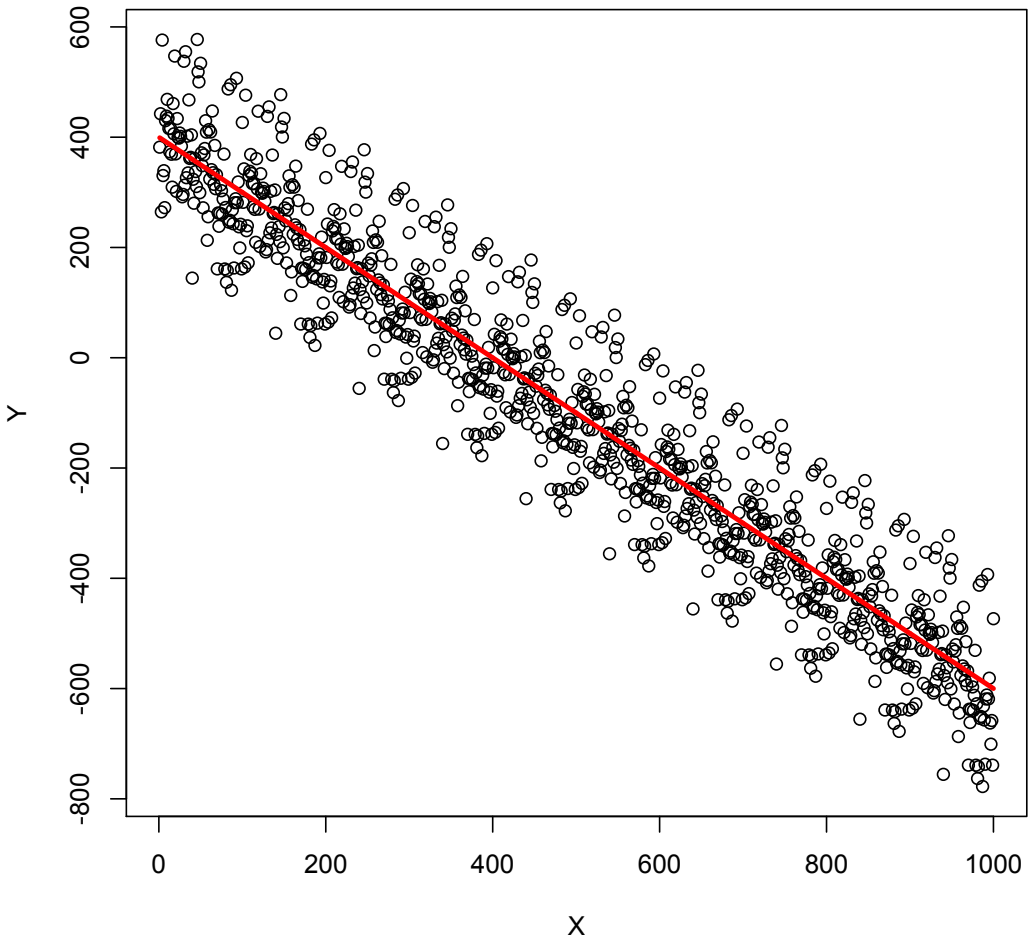
Scatter diagram

Positive relationship



Scatter diagram

Negative relationship



Mean

- The mean of a data set is the average of all the data values.
- The sample mean $\overline{X}$ is the point estimator of the population mean μ.

Sample mean

Sample mean is calculated as:

$$\overline{X} = \frac{\sum_{i=1}^{i=n} X_i}{n}$$

where $\sum_{i=1}^{i=n} X_i$ is the sum of the values of n observations and n is the size of the sample.

Population mean

Population mean is calculated as:

$$\mu = \frac{\sum_{i=1}^{i=N} X_i}{N}$$

where $\sum_{i=1}^{i=N} X_i$ is the sum of the values of N observations and N is the size of the population.

Sample mean

Example

Wages

See the excel spreadsheet.

Seventy workers were randomly sampled in a firm:

$$\overline{X} = \frac{\sum_{i=1}^{i=n} X_i}{n} = \frac{37682}{70} = 538.31$$

Median and mode

- The median of a data set is the value in the middle when the data items are arranged in ascending order.
  - Whenever a data set has extreme values, the median is the preferred measure of central location. Eg. Income in two countries.
- The mode of a data set is the value that occurs with greatest frequency.
  - The mode doesn't need to be unique

Measures of Variability

- Variance
- Standard deviation

Variance

- Variance is a measure of variability that utilizes all the data. It is based on the difference between the value of each observation (X_i) and the mean ($\overline{X}$ for a sample, μ for a population).
- The variance is the average of the squared differences between each data value and the mean.
- For a sample the variance is computed as follows:

$$s^2 = \frac{\sum_{i=1}^{i=n}(X_i - \overline{X})^2}{n-1}$$

- For a population the variance is computed as follows:

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

Standard deviation

- The standard deviation of a data set is the positive square root of the variance.
- For a sample the standard deviation is computed as follows:

$$s = \sqrt{s^2}$$

- For a population the standard deviation is computed as follows:

$$\sigma = \sqrt{\sigma^2}$$

Applications

- Mean and median household income of US
  - Why are the differences?
- Standard deviation of household income
  - What does the pattern mean?

Probability

- Probability of an outcome ('success'): proportion of times that outcome would occur in a long sequence of repeated observations.

$$Prob(z) = \frac{number\ of\ successes}{total\ n\ of\ possible\ outcomes} = \frac{s}{n} = p(z)$$

- Probability of obtaining head in flip of coin is $\frac{1}{2}$
- Probability of a single roll of a diee $p(rolling\ a\ four) = 1/6 = 0.1667$
- Probabilities can be expressed as proportions (values between 0 and 1) or as percentages (%).

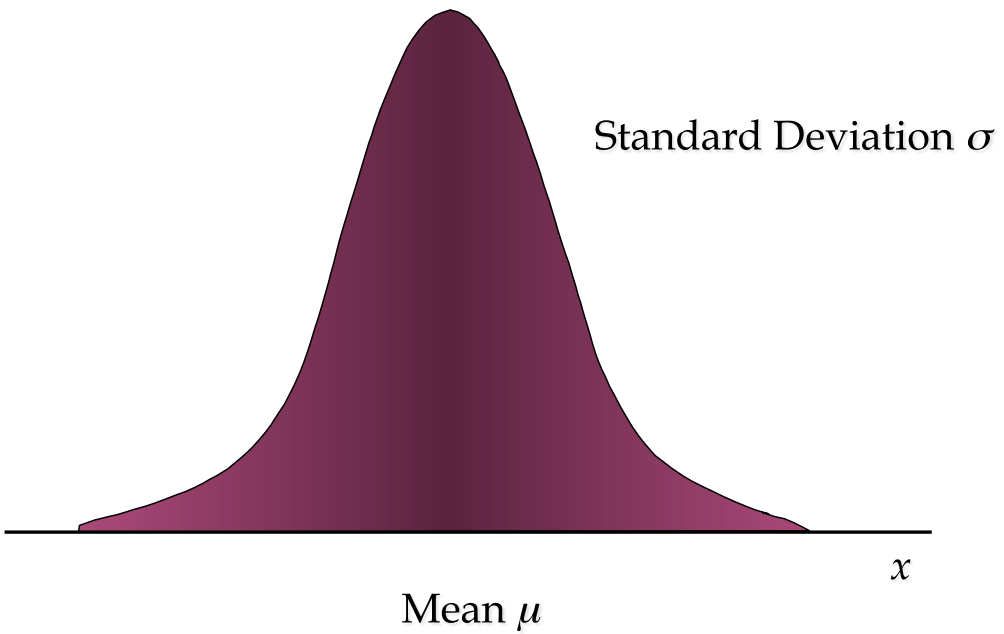
Probability Distributions

- These sound scary but they are just lists of the possible outcomes and their probabilities.
- Discrete Probability Distributions, examples of which include:
  - Binomial Distribution
  - Poisson Distribution
- Continuous Probability Distributions, examples of which include:
  - Normal Distribution
  - Exponential Distribution

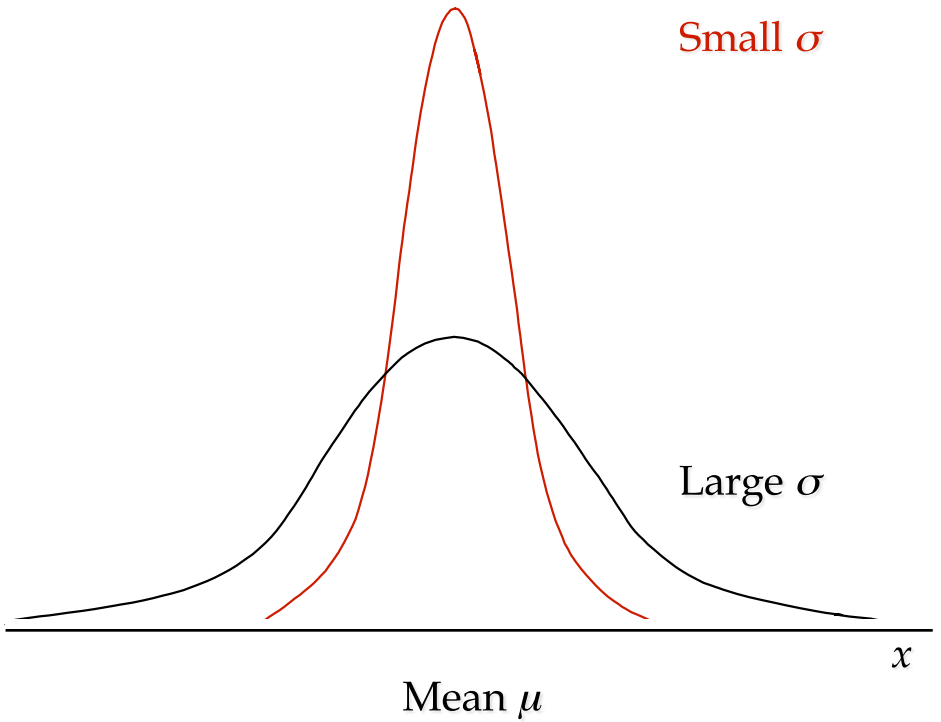
Normal distribution

- The distribution is symmetric; its skewness measure is zero.
- The entire family of normal probability distributions is defined by its mean μ and its standard deviation σ.
- Examples: exam results, height, weight, firm size.
- The highest point on the normal curve is at the mean, which is also the median and mode.
- The mean can be any numerical value: negative, zero, or positive.
- The standard deviation determines the width of the curve: larger values result in wider, flatter curves. Probabilities for the normal random variable are given by areas under the curve. The total area under the curve is 1 (.5 to the left of the mean and .5 to the right).

Normal distribution



Normal distribution



Normal distribution I

- A random variable having a normal distribution with a mean of 0 and a standard deviation of 1 is said to have a standard normal probability distribution.
- Empirical normal distribution can converted into standard normal distribution.
- Converting to the Standard Normal Distribution (i.e from $x \sim N(\mu, \sigma)$ to $z \sim N(0, 1)$):

$$z = \frac{x - \mu}{\sigma}$$

- z-score indicates the distance to the mean in terms of units of σ. Eg., z-score of $+1.00$ indicates that the original score lies one σ unit above (right to) the mean.

Normal distribution

- The conversion allows us to:
  - Use the normal distribution table
  - To find out probability (area) from z-score
  - To find z-score for certain tail probabilities

Normal distribution

- You can the find the table of z-scores and probabilities for a standard normal distribution at the back of most statistics textbooks.
- Or you can google it...
- For example:
  http://www.math.unb.ca/~knight/utility/NormTble.htm

Normal distribution I

Example

Village households income

Monthly income (x) of households in a village is distributed normally with a mean (μ) of \$72 and a standard deviation (σ) of \$8. What percentage of households have monthly income between \$70 – 79?

Step 1: Convert x (i.e. $x_1 = 70$ and $x_2 = 79$) to the standard normal distribution z.

$z_1 = (x_1 - \mu)/\sigma = (70 - 72)/8 = -0.25$

$z_2 = (x_2 - \mu)/\sigma = (79 - 72)/8 = 0.88$

Normal distribution I

Example

Village households income

Monthly income (x) of households in a village is distributed normally with a mean (μ) of \$72 and a standard deviation (σ) of \$8. What percentage of households have monthly income between \$70 – 79?

Step 2: Find the area under the standard normal curve to the left of $z = -0.25$ and $z = 0.88$

Area right to 0.25 is 0.4013. So Area left to -.25 is 0.4013.

Area left to 0.88 is (1-0.1894) = 0.8106

Normal distribution I

Example

Village households income

Monthly income (x) of households in a village is distributed normally with a mean (μ) of \$72 and a standard deviation (σ) of \$8. What percentage of households have monthly income between \$70 – 79?

Step 3: The area between -0.25 and 0.88 =0.8106-0.4013 =**0.4093**

. . . **which is proportion of households that have monthly income between \$70 – \$79**

Normal distribution I

Example

Village households income

Monthly income (x) of households in a village is distributed normally with a mean (μ) of \$72 and a standard deviation (σ) of \$8. What is the probability that a household selected at random will have a monthly income less than 61?

Step 1: Convert x to the standard normal distribution.

$z = (x - \mu)/\sigma = (61 - 72)/8 = -1.38$

Normal distribution I

Example

Village households income

Monthly income (x) of households in a village is distributed normally with a mean (μ) of \$72 and a standard deviation (σ) of \$8. What is the probability that a household selected at random will have a monthly income less than 61?

Step 2: Find the area under the standard normal curve to the left of $z = -1.38$

It equals the area to the right of $z = 1.38$.

So the area/probability is 0.0838.

So the probability of selecting a household with a income less than 61 is about 8%.

Normal distribution I

Example

Village households income

Monthly income (x) of households in a village is distributed normally with a mean (μ) of \$72 and a standard deviation (σ) of \$8. What is the probability that a household selected at random will have a monthly income more than 80?

Step 1: $z = (x - \mu)/\sigma = (80 - 72)/8 = 1.00$

Normal distribution I

Example

Village households income

Monthly income (x) of households in a village is distributed normally with a mean (μ) of \$72 and a standard deviation (σ) of \$8. What is the probability that a household selected at random will have a monthly income more than 80?

Step 2: Find the area under the standard normal curve to the right of $z = 1.00$

The area/probability is 0.1587

So the probability of selecting a household with a income more than 80 is about 16%.

Central limit theorem I

- For random sampling, as sample size grows, the sampling distribution of $\overline{X}$ approaches a normal distribution, **regardless of the shape of the population distribution**.
- The distribution of a mean tends to be Normal, even when the distribution from which the average is computed is decidedly non-Normal.
- What does it mean for the distribution of the sample mean to the Normal? Let's fix a sample size, say $n = 30$. We pick 30 people randomly from the population and record their sample mean. Then we pick another 30 people and record their sample mean. If you do this loads of times, the distribution of the sample mean:
  - will be normal AND
  - it will have the same mean as the population distribution AND
  - it will have a variance equal to the variance of the population distribution divided by the sample size.

Central limit theorem II

- So the bigger your sample size is the smaller is the variance of the distribution of the sample mean!
- But if we know what distribution ONE sample came from, we only need ONE sample to be able to make inference about the mean of the population.
  - We should be able to say how likely it is that we will have got within a certain range of the mean - just like we did in the village incomes example.
- Thus, the Central Limit theorem is the foundation for many statistical procedures because the distribution of the phenomenon under study does not have to be Normal because its average will be.
- How large is sample size high enough? $n = 25 - 30$ generally sufficient for sampling distribution to be close to normal; smaller samples adequate if population distribution is already bell-shaped.
  - If you sample just one person at a time, the sampling distribution of the mean will be the same as the population distribution - not very useful!

Central limit theorem III

- But if you sample the whole population, the sample distribution of the mean will have variance zero - think why!

Question 1

The Current Population Survey of about 60,000 households in the US in 1992 indicated that 10.3% of whites, 31.0% of blacks, and 26.7% of Hispanics in the US have annual income below the poverty level.
a. Are these numbers statistics or parameters? Explain.
b. Using a certain method we could conclude that the percentage of all black households in the US having income below the poverty level is at least 30% but no greater than 32%. What type of statistical method does this illustrate – descriptive or inferential?
c. Regarding the number of countries and years included in the Survey, is it cross-sectional data, time series data or none of them? Explain.

Solution to Question 1

The Current Population Survey of about 60,000 households in the US in 1992 indicated that 10.3% of whites, 31.0% of blacks, and 26.7% of Hispanics in the US have annual income below the poverty level.
a. Are these numbers statistics or parameters? Explain.

Answer

Percentage of households below the annual poverty line is quantitative, continuous variable.

CPS: is a survey, not a census. Population of interest: total population of the US.

CPS is a sample hence Y is a **statistic**.

If it were the population, then Y is a **parameter**.

Solution to Question 1

The Current Population Survey of about 60,000 households in the US in 1992 indicated that 10.3% of whites, 31.0% of blacks, and 26.7% of Hispanics in the US have annual income below the poverty level.
b. Using a certain method we could conclude that the percentage of all black households in the US having income below the poverty level is at least 30% but no greater than 32%. What type of statistical method does this illustrate – descriptive or inferential?

Answer

From the statistic (% of poor HH), we are drawing an inference about what happens in the total population. The only information we have is that in our sample there is a certain proportion of HH below the poverty line. Using statistic techniques, we infer - with a certain error - the percentage of poor HH in the total population of interest. Considering that the sample design is appropriate, as we increase the sample (# of HH) the error will be smaller.

Solution to Question 1

The Current Population Survey of about 60,000 households in the US in 1992 indicated that 10.3% of whites, 31.0% of blacks, and 26.7% of Hispanics in the US have annual income below the poverty level.
c. Regarding the number of countries and years included in the Survey, is it cross-sectional data, time series data or none of them? Explain.

Answer

Number of countries: 1 (US) and number of years: 1 (1992).
It is not cross-sectional or time series data. Data are collected at the same or approximately the same time (synchronic) and in only one country. However, if we compare the CPS from different years and we contrast the different states, it would be cross-sectional and time-series data.

Question 2

Question 2. Since statistics on AIDS are not fully reliable, the Council of a city in South Africa conducts a pilot study where 25 respondents are asked: "During the past month, how many people have you known personally that have AIDS? Below is reported the number of people with AIDS that the 25 respondents met personally in the last month. You are asked to summarize the data and to report it to the Council.
We answer parts a) to d), f) and g) of this question in Stata.

Solution to Question 2

e) Compute mean and median from the frequency distribution you constructed in point a. Compare the grouped with the ungrouped measures.

Answer

Middles of each interval: $\{1, 4, 7, 10, 13, 16\}$. To calculate the mean for grouped data:

$$\bar{Y}_G = \sum_{i=1}^{i=n} f_i(Y)Y = 0.2 \times 1 + 0.28 \times 4 + 0.28 \times 7 + 0.12 \times 10 + 0.08 \times 13 + 0.04 \times 16$$
$$= 5.92$$

Median for grouped data: interval 6 to 8, the middle value of which is 7.

Solution to Question 2

h) Suppose another researcher had just interviewed 50 women and found they had met an average of 4 people with AIDS. What is the mean across both samples (not considering the 26 th value from g)?

Weighted mean:

$$\bar{Y} = \frac{n_1 \times \bar{Y}_1 + n_2 \times \bar{Y}_2}{n_1 + n_2} = \frac{25 \times 6 + 50 \times 4}{25 + 50} = 4.67$$

Question 3

Table 1 shows the homicide rate in selected countries of South America.
We answer parts a), b), d) and f) of this question in Stata.

Solution to Question 3

Table 1 shows the homicide rate in selected countries of South America. c) In Table 1, countries are also classified according to their "level of violence". Thus, countries with a homicide rate of 10 or lower are classified as "low level of violence", 11 to 20 as "intermediate level of violence" and 21 and over as "high level of violence". - What type of variables are "homicide rate" and "level of violence"? - What are the advantages and limitations of the new variable "level of violence"? Can you calculate the mean, the median, or mode for these data? If so, do so and interpret.

Answer

One way to summarize information is to reduce the number of values a variable has. An example of this strategy is to group values in intervals and after that to label them. By doing so, the variable becomes qualitative. Thus, while "homicide rate" is a quantitative continuous variable, "level of violence" is a qualitative ordinal variable.

Solution to Question 3

Table 1 shows the homicide rate in selected countries of South America.

Answer (cont.)

- Advantages: information is reduced. Therefore, instead of having multiple values of the variable, you obtain only 3 categories. This makes easier to interpret results, to classify the units of analysis and to present results in general.
- Limitations: if you were only given the variable level of violence, you will be unable to calculate most statistical measures.
  Mean and median are not possible for categorical variables, although median can be possible for an ordinal variable. In this case, we can observe that 50% of the cases have low level of violence.
  Mode = Low level of violence, with 50% of the total sample.

Solution to Question 3

Table 1 shows the homicide rate in selected countries of South America.
e) Table 3 shows the Human Development Index (HDI) for selected countries. Then, countries with a HDI from 1.000 to 0.800 are classified as of "high development" and countries from 0.799 to 0.500 as "medium development". What type of variable is the HDI? And level of development?

Answer

While HDI is a continuous quantitative variable, level of development is an ordinal qualitative variable.