Social Media and Political Participation

Lab 2

pablo.barbera@nyu.edu

January 7, 2014

Today

- Concepts in statistical analysis
- Working with single variables (univariate analysis)
- Studying the relationship between variables (bivariate analysis)
- Visualizing statistical relationships
- In-class exercise: analysis of a dataset

Concepts in statistical analysis

Important concepts in statistical analysis

When we perform a statistical analysis, we are usually interested in examining a list of **variables** that belong to a number of **units**:

1. **Variables** are properties that can be measured or counted
2. **Units** are the objects we are analyzing

For example:

- Individuals (units) and height, weight, age... (variables)
- Countries (units) and size, regime type, location... (variables)
- Social media accounts (units) and number of followers, posts, type of social media site... (variables)

Types of variables

Four types:

1. Continuous: height, geographic coordinates...
2. Counts: number of likes, age in years...
3. Ordinal: academic grades, clothing sizes...
4. Categorical: type of post, gender...

Difference is important because it implies different types of statistical analyses.

Univariate analysis

Univariate analysis for continuous variables

When a variable is continuous or a count, we can summarize it with the following measures:

- Mean (average), the sum of all its values divided by the number of values in the variable.
- Median, the middle value of a variable
- Minimum and maximum values of a variable
- Quantiles, the values that divide the variable in equal intervals

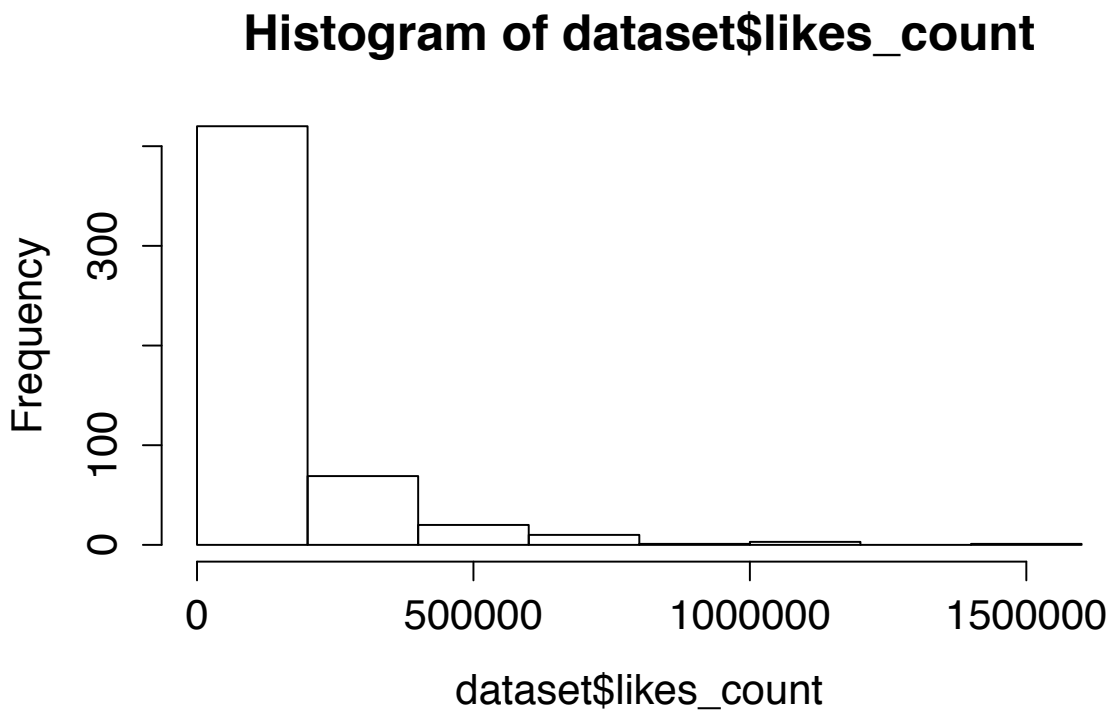
The function to compute all of these in R is `summary`.

```
# computing summary statistics for number of likes
> summary(dataset$likes_count)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0   51910   92150  145100  171800 1572000
```

Univariate analysis for continuous variables

When a variable is continuous or a count, we can use an **histogram** to study its distribution.

```
# generating an histogram in R
> hist(dataset$likes_count)
```



Univariate analysis for categorical variables

When a variable is categorical, instead we use **frequency tables** to look at the distribution of the different values.

```
# computing frequency table for month of year
> table(dataset$month)

 1  2  3  4  5  6  7  8  9 10 11 12
50 46 52 41 26 48 48 46 28 45 44 50
```

We can also easily compute proportions with `prop.table`

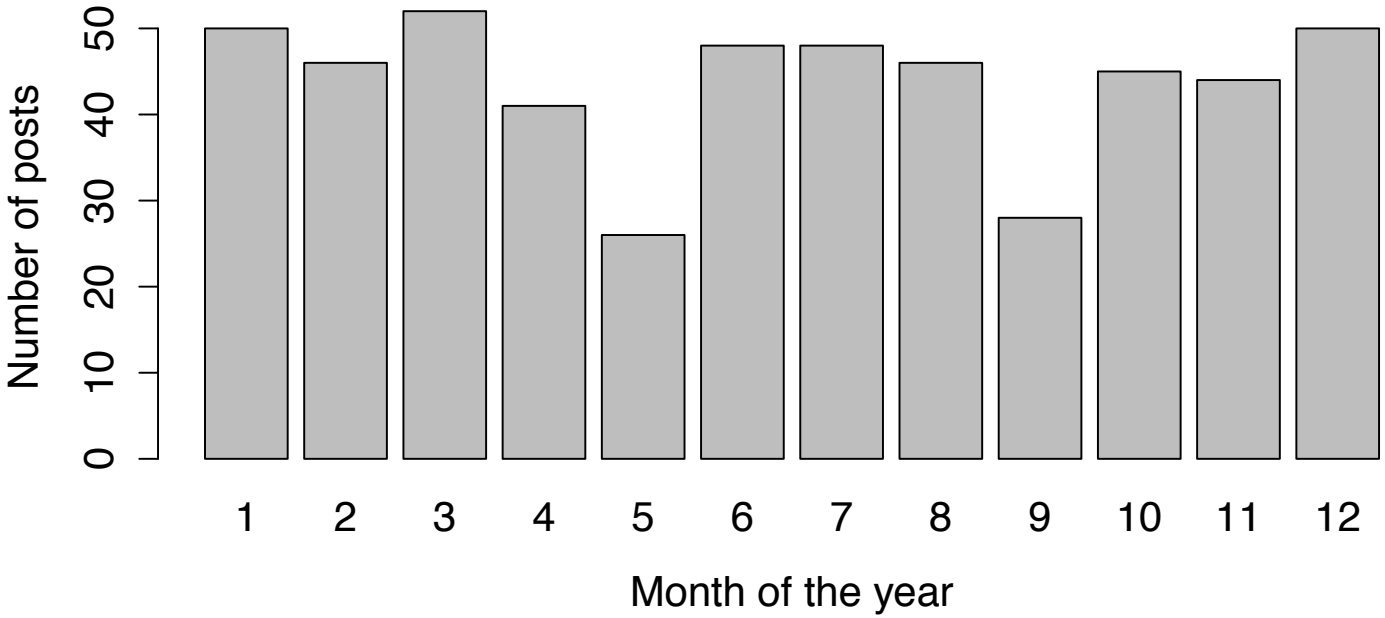
```
# creating proportion table for type of post
> prop.table(table(dataset$month))

   1    2    3    4    5    6    7    8    9   10   11   12
0.10 0.09 0.10 0.08 0.05 0.09 0.09 0.09 0.05 0.09 0.08 0.10
```

Univariate analysis for categorical variables

A frequency table can also be visualized in a bar chart:

```
# generating a bar chart in R
> barplot(table(dataset$type))
```



Note that the `barplot` command is applied to the frequency table, already computed with `table`.

Univariate analysis with R

The R script `lab2_univariate_analysis.R` shows how to:

- Open and summarize a dataset
- Continuous variables are examined using the summary functions
- Categorical variables are examined with frequency tables
- Visualization of variables: histograms and bar charts.

Bivariate analysis

Bivariate analysis

Often we're interested in learning about the relationship between two variables. For example:

- Are men more likely to wear dark clothes?
- Do Facebook posts that receive many likes also receive more comments or shares?
- In what month of the year did a page receive more likes?

Note that each of these questions corresponds to a different combination of categorical and continuous variables. Let's now turn to each possible case.

Two categorical variables

When two variables are categorical, we use **contingency tables** to examine their relationship.

```
# creating contingency table for month and post type
> table(dataset$type, dataset$month)

         1  2  3  4  5  6  7  8  9 10 11 12
link     0  0  0  1  1  0  0  0  0  0  0 16
photo   48 46 52 37 25 48 47 46 28 45 43 33
status   0  0  0  2  0  0  1  0  0  0  1  0
video    2  0  0  1  0  0  0  0  0  0  0  1
```

One categorical, one continuous

When two variables are of different types, we **aggregate** the continuous variable over each different value of the categorical one.

```
# computing mean number of likes for each month
> aggregate(dataset$likes_count,
+           by=list(month=dataset$month),
+           FUN=mean)
   month         x
1      1 188840.04
2      2 183459.02
3      3 146685.04
4      4 122619.34
5      5  96129.65
6      6  98936.50
...          ...
```

Here I'm computing the mean, but it could be any other statistic (sum, minimum, maximum...)

Two continuous variables

To measure the association between two continuous variables, we can compute the correlation coefficient.

It takes values between -1 (negative association) and $+1$ (positive association). A value of 0 implies no association whatsoever.

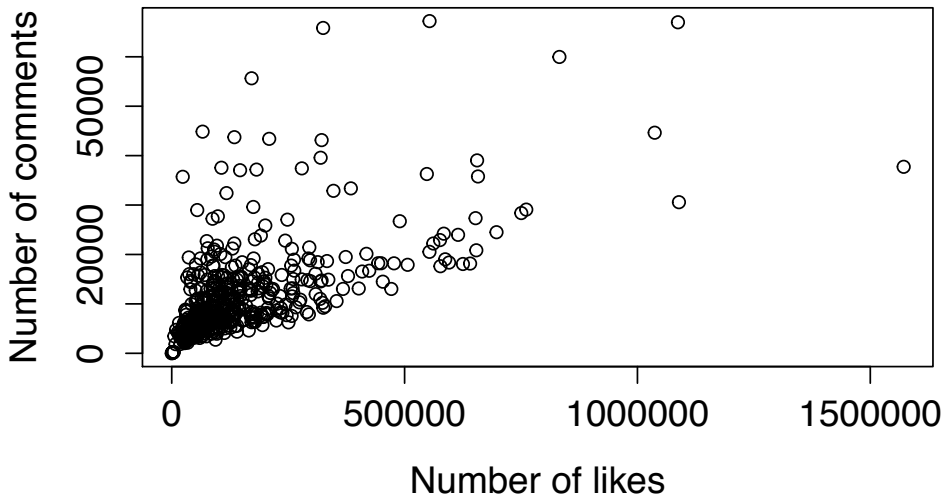
```
# do posts that get more likes also receive more comments?
> cor(dataset$likes_count, dataset$comments_count)
[1] 0.6258727
```

A positive value, close to 1, means that high values of the first variable usually appear associated to high values of the second variable.

Two continuous variables

The relationship between two variables can be visualized with a scatter plot, where each dot represents one observation:

```
# scatter plot comparing number of likes and number of comments
> plot(x=dataset$likes_count, y=dataset$comments_count,
+      xlab="Number of likes", ylab="Number of comments")
```



Note the use of `xlab` and `ylab` options to add titles to each axis.

Univariate analysis with R

The R script `lab2_bivariate_analysis.R` shows how to:

- Create a contingency table
- Aggregate a continuous variable over values of a categorical variable
- Compute a correlation coefficient
- Display a scatter plot of two variables

In-class exercise

In-class exercise: statistical analysis

Create your own R script (with comments) that:

1. Opens the dataset `lab1_obama_data.csv`
2. Runs different commands that help you answer the following questions:
   1. What type of post (photo, status, link, video) was the most common on Barack Obama's Facebook page in 2013?
   2. Do posts that contain photos receive more likes on average than links?
   3. Are more liked posts also more shared? Answer this question computing a correlation coefficient and creating a scatter plot of these two variables

Optional Create a plot that displays the total number of comments on the page each month. Do you notice unusual? Why?