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

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

Approximate timing of Lab 2

- Concepts in statistical analysis = 15 minutes
- Univariate analysis = 35 min (15 minutes for slides, 20 minutes working with R)
- BREAK (10 minutes)
- Bivariate analysis = 40 minutes (15 minutes for slides, 25 minutes working with R)
- In-class exercise = 20 minutes

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)

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)

This is the basic terminology of statistics.
Ask students to see if they can come up with more examples.
A few other possible examples:

- speed/color and cars
- square feet/number of windows and apartments
- number of pages/genre of books
- vote share / ideology of a party

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.

Types of variables

Four types:
- Continuous: height, geographic coordinates...
- Counts: number of likes, age in years...
- Ordinal: academic grades, clothing sizes...
- Categorical: type of post, gender...

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

Definitions:

- Continuous: anything that can be given a real (decimal) number
- Counts: anything that can only be given integers (non decimal) numbers
- Ordinal: discrete values that are ordered (one is higher than the other)
- Categorical: discrete values that are not ordered (no hierarchy)

Ask students to see if they can come up with more examples.
A few other possible examples:

- Continuous: speed, ideology, square feet...
- Counts: number of retweets/followers, number of votes...
- Ordinal: agreement with a statement, income categories...
- Categorical: colors, vote choice...

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 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
```

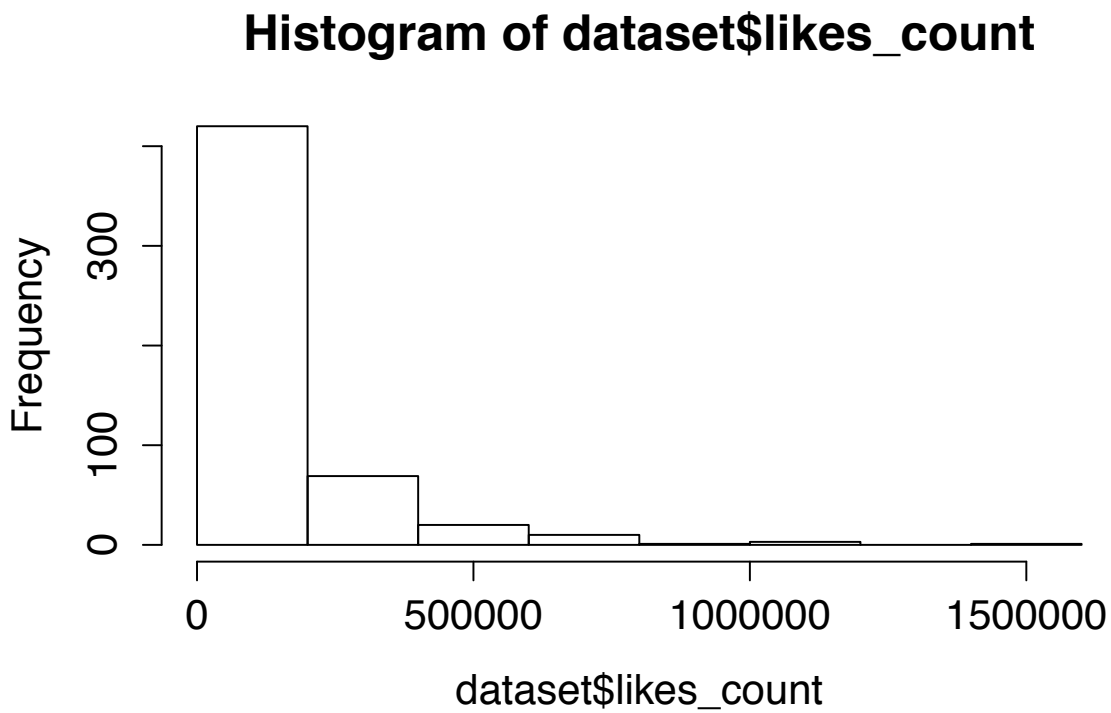
We start with univariate analysis: when there's only one variable and we want to "describe" or summarize its main properties.
There are two types of statistics that we can compute, where a statistic is a number that measures the variable in some way:

1. Measures of central tendency: mean and median. Use example below to explain difference: median implies 50% of posts have less than 92K likes; and 50% have more. Mean implies that if you aggregate all likes and divide by number of posts, that's what you get.
2. Measures of dispersion: min, max, quantiles. Example to explain quantiles: if we have 100 numbers, from 1 to 100, then quantile 1 is 25, and quantile 3 is 75. Also use results from likes count to explain.

Univariate analysis for continuous variables

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

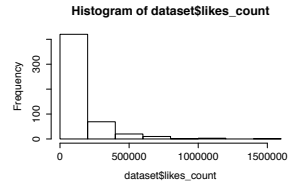
```
# generating an histogram in R
> hist(dataset$likes_count)
```



Univariate analysis for 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)
```



A graphic sometimes is worth more than a thousand words. That's why plots are usually more informative than tables or numbers. Depending on the type of variable we have, we will need to prepare different types of plots.
For a count (continuous?) variable like the number of likes, a histogram is what we need.
Each bar represents the number of posts that have a like count within that interval (e.g. first interval is 0 to 250K). So x axis is number of likes, in intervals, and y axis is number of posts that fall in that interval.

In this case, we see that most posts get a relatively small number of likes, and there are a few that get many ($> 1.5M$). Those are "outliers".

Univariate analysis for categorical variables

When a variable is categorical or ordinal, 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

When a variable is categorical or ordinal, 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
```

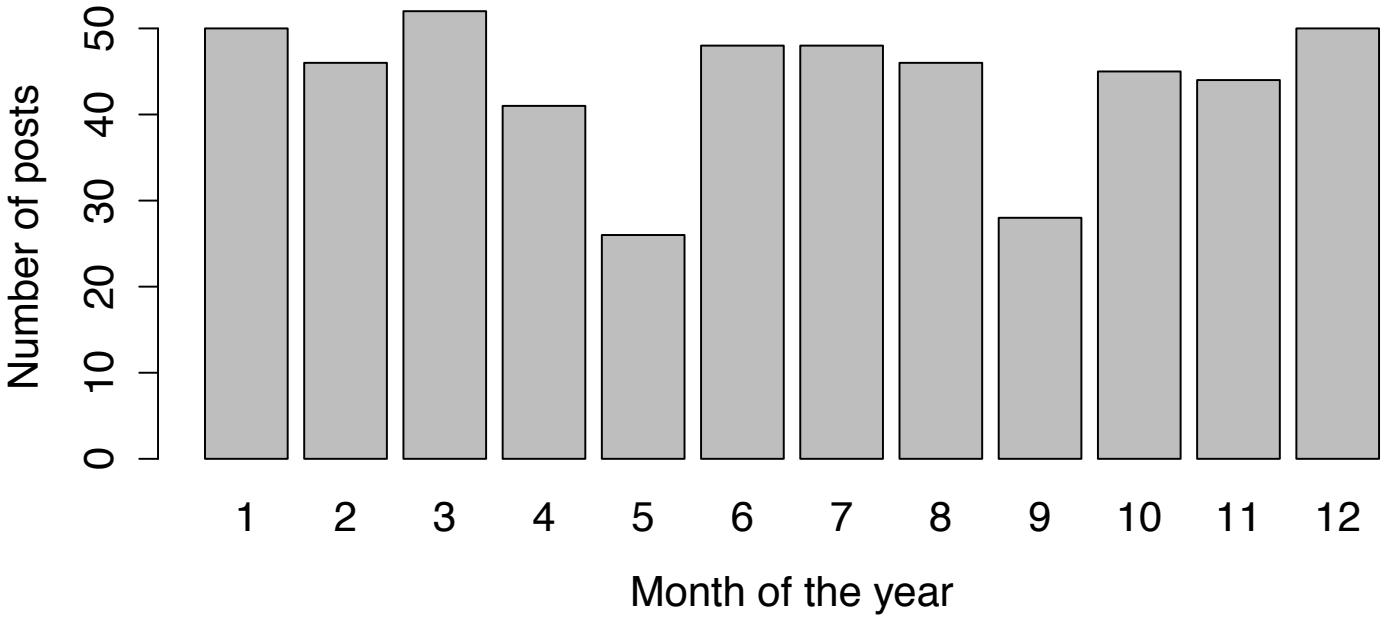
When we have a categorical variable, computing the mean or median is meaningless (what is the mean of a month variable, for example?). That's why we use frequency tables, that measure the proportion of each category (or the number of times each possible value appears).
The example here is the number of posts per month (so first row is month number, second row is number of posts that were sent that month).

We can compute either the count (first example) or the proportions (second example). Note that in the second example we need to use the `prop.table` function, which takes the frequency table as argument.

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))
```

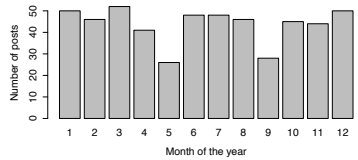


Note that the `barplot` command is applied to the frequency table, already computed with `table`.

Univariate analysis 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`.

As before, often it's better to prepare a graph with the distribution of the data. When we have a categorical variable, the equivalent of the histogram is a bar chart. Here, the x axis indicates each possible value of the categorical variable, and the y axis is the number of times each value is used.
For example, we learn that Obama didn't update his Facebook account as often in May and September.

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.

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.

Bivariate analysis literally means "of two variables". We're interested in bivariate analysis because it helps us understand the world. In contrast with univariate analysis, which is usually descriptive, when we consider two variables we're examining how one affects the other. Under certain conditions, we can also talk about "causal" relationships.
Ask students for more examples:

- Do young people tend to vote more frequently?
- Is the intake of aspirins associated with alleviation of the symptoms of the flu?
- Do Twitter users with more followers receive more retweets too?

Often, we have some ideas about what the answer can be, and we can refer to those as "hypotheses".

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
```

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
```

The first case is when we have two categorical variables (in this case, type of Facebook post, and month of the year).
A contingency table (also called cross tabs) is a two-way table where the rows indicate values of the first variable and the columns values of the other variable. Each individual cell corresponds to a count of the observations that fall within that combination of categories.
For example, 48 posts in January were photos, 2 were videos, and 0 were links or statuses.

The R command is table too, with the row variable first and the column variable second.

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...)

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...)

This is a bit more complicated. The idea is that we summarize the continuous variable for each possible value of the categorical variable.
In this example, we compute the mean number of likes for posts sent in each month of the year.
The first number indicates that posts from January received 188K likes on average.

And of course, depending on the situation, we might be interested in computing other statistics (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.

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.

The more frequent case, however, is when we have two continuous variables. There are different ways of summarizing this relationship, and here I will focus on one that is very commonly use: the correlation coefficient.

A correlation is a measure of the linear relationship between two variables: to what extent do high values of one variable appear at the same time as high values of another variable for the same individual?

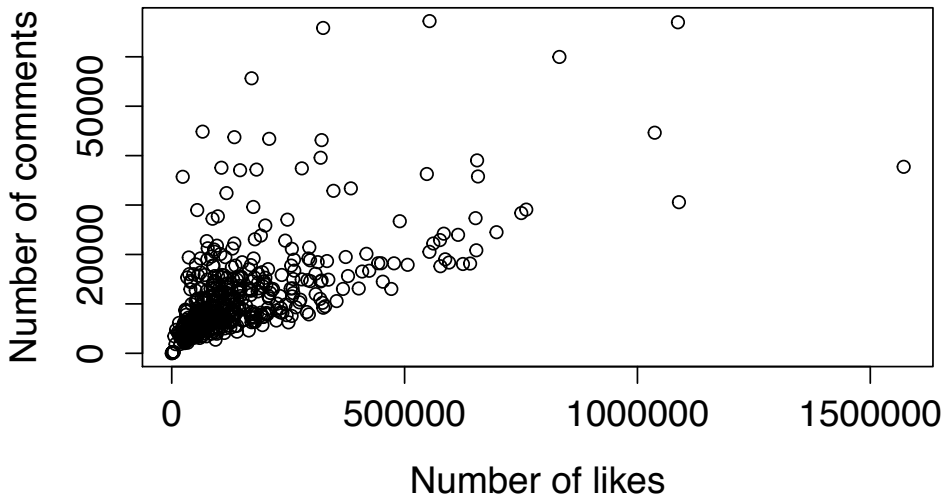
Formula (just in case): $r = \frac{1}{n-1} \Sigma \left(\frac{x-\bar{x}}{s_x} \right) \left(\frac{y-\bar{y}}{s_y} \right)$

This coefficient takes value between -1 and $+1$. Explain following slide.

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.

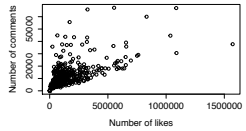
As we discussed earlier, a good graphic can be more informative than a number. To examine the relationship between two variables, we can use a scatter plot, where each dot represents one observation (one individual, one post...), and its position depends on the values of the two variables.
The horizontal axis is referred to as x-axis by convention; the vertical axis is the y axis.

We can add labels to the axes with the xlab and ylab options.

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.