Stat-202-2015-XD-W3-Mon (Pg 8)

What is the difference between a variable and a random variable

Hint a "random variable" is not a kind of "variable". (separate concepts) (related though)

On the first day of class (and on page 1 of the text) we defined variable as

"a characteristic of a case". ← object of study in a data set

This is meant a measured characteristic of a case recorded as data.

A variable concerns what is measured about a real phenomenon.

A random variable is a function on the set of outcomes of a random phenomenon

Whereas the variable concerns just the things that were measured, a random variable concerns all the things that are possible, both concern outcomes of a random phenomenon

Variable - in practice / what actually happened

Random Variable - in theory / everything that can happen and how likely it is.

There is a relationship between variable and random variable.

The variable is a characteristic of a case
cases might be people
variable might be height of person

For a random variable
the <u>random phenomenon</u> is
"pick a person a random."
the <u>sample space</u> (outcomes)
is the set of 7 bil people on Earth
The random variable is the function
that assigns to each outcome (person)
the number (his or her height)

Notice that this height random variable
is discrete (7 bil values) but with
so many values, people usually think about
it as continuous, and approximate the discrete
random variable with a continuous one,
one where all heights are possible
When we talk about the distribution
of a variable or a density curve
of a variable we mean the distribution/
density curve of the associated random variable.

Foreshadow Chapter 5

In Chapter 5 (Sampling distributions) we will consider a random phenomenon

"Pick 100 cases out of the whole population, (eg 7 bil people)"

Sample space is all possible sets of 100 cases (all possible samples)

Random variable is the function that assigns to each sample the mean of the height of the 100 cases in the sample

(sample mean)

Obviously it is not always 100 - could be 250 or 500, etc, called "n" (sample size)

We want to study the distribution (sampling distribution) of this random variable (sample mean) in particular how it depends on n.

What's the big deal? We want to infer something about the population based on the sample. The theory of sampling distributions will give us the tools to put this on solid theoretical basis.

Stat 202-2015S W6-Tues (Pg 1)

Chapter 1: graphical and numerical methods to describe a single variable

Chapter 2: graphical and numerical methods to describe relationships between pairs of variables

Example ① Score on first exam, Score on second exam } relationship?

② Size of a coffee at starbucks, price of a coffee at starbucks } relationship?

③ stress of students, lack of sleep of students } relationship?

Two variables measured on the same cases are said to be associated if knowing the values of one tells you something about the values of the other that you would not know without this information

In the coffee example one variable perfectly predicted the other

In the other examples the variables don't perfectly predict each other — all are associated

Two variables can be quantitative or categorical (or both one of each)

When considering relationships between variables ask yourself

1) Am I simply trying to describe relationship
or 2) Am I trying to show that one variable explains or causes changes in the other.

A response variable measures an outcome of a study (think dependent variable)
An explanatory variable explains or causes change in the response variable (think independent variable)

We don't use independent/dependent variables in statistics because independent means something different (related to not associated)

Stress and Sleep - not clear which causes/explains what

Other times it is clear.

Some Statistical techniques require a distinction between explanatory/response variables
Others don't.

Pg 5

The most common way to display a relationship between two quantitative variables is with a scatterplot

A scatterplot shows the relation between two quantitative variables measured on the same cases

One variable x-axis
Other variable y-axis

Each case appears as a single point (x,y)

Load data set Oasis

Examining Scatter plots

- Look for the overall pattern
- Look for striking deviations from pattern
- Describe pattern in terms of
  - form
  - direction
  - strength
  - Outliers (an individual that falls outside overall pattern

Direction

Two variables are positively associated when above average values of one tend to accompany above average values of the other and below average values tend to occur together also

Negatively associated when above average values of one occur with the below average values of the other and vice versa

Strength

Determined by how closely points follow a clear form

Form

- Could be linear
- Or some other function

- Sometimes to see data better we use a transformation (eg log transformation)

Stat 202 2015XD - W3 - Monday
Stat 202 - 2015S W6 - Friday (pg 7)

Review - Chapter 2 - Relationships between pairs of variables

Associated - two variables measured on same cases are associated if knowing values of one tells you something about values of other.

dep Response Variable - measures outcome of a study
indep Explanatory variable - explains or causes change in response variable

Scatterplot

Oasis - strong positive
Bioclocks - weak positive
Fidget - moderate negative

Form - eg line, also could be something else. A curve for example, Sine wave. We will consider only lines

Strength - How closely do data follow form?

Direction Pos/neg depending on whether above average values of one occur with above average values of other (pos) or below average values of other (neg)

Pg 2

Correlation

A number which quantifies the direction and strength of the linear relationship between two variables.

↓

To the extent that the form is a line this is useful. Not necessarily as useful if form is something else.

Direction: { positive association
{ r is pos
{ negative association
{ r is negative

Strength: { strong association
{ r is close to ±1
{ weak association
{ r is close to 0

Defined

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

$$s_x = \sqrt{\frac{1}{n-1} \sum (x - \bar{x})^2}$$

Standardized x × Standardized y

z-score