Why is my evil lecturer forcing me to learn statistics?

Self-test answers

- Based on what you have read in this section, what qualities do you think a scientific theory should have?

A good theory should do the following:

1. Explain the existing data.
2. Explain a range of related observations.
3. Allow statements to be made about the state of the world.
4. Allow predictions about the future.
5. Have implications.

- What is the difference between reliability and validity?

Validity is whether an instrument measures what it was designed to measure, whereas reliability is the ability of the instrument to produce the same results under the same conditions.

- Why is randomization important?

It is important because it rules out confounding variables (factors that could influence the outcome variable other than the factor in which you're interested). For example, with groups of people, random allocation of people to groups should mean that factors such as intelligence, age, gender and so on are roughly equal in each group and so will not systematically affect the results of the experiment.

- Compute the mean but excluding the score of 252.

First, we first add up all of the scores:

$$\sum_{i=1}^{n} x_i = 22 + 40 + 53 + 57 + 93 + 98 + 103 + 108 + 116 + 121$$
$$= 811$$

We then divide by the number of scores (in this case 11):

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{811}{10} = 81.1$$

The mean is 81.1 friends.

- Twenty-one heavy smokers were put on a treadmill at the fastest setting. The time in seconds was measured until they fell off from exhaustion: 18, 16, 18, 24, 23, 22, 22, 23, 26, 29, 32, 34, 34, 36, 36, 43, 42, 49, 46, 46, 57. Compute the mode, median, upper and lower quartiles, range and interquartile range.

First, let's arrange the scores in ascending order: 16, 18, 18, 22, 22, 23, 23, 24, 26, 29, 32, 34, 34, 36, 36, 42, 43, 46, 46, 49, 57.

The mode: The scores with frequencies in brackets are: 16 (1), 18 (2), 22 (2), 23 (2), 24 (1), 26 (1), 29 (1), 32 (1), 34 (2), 36 (2), 42 (1), 43 (1), 46 (2), 49 (1), 57 (1). Therefore, there are several modes because 18, 22, 23, 34, 36 and 46 seconds all have frequencies of 2, and 2 is the largest frequency. These data are multimodal (and the mode is, therefore, not particularly helpful to us).

The median: The median will be the $\frac{1}{2}(n+1)\text{th}$ score. There are 21 scores, so this will be 22/2 = 11. The 11th score in our ordered list is 32 seconds.

The mean: The mean is 32.19 seconds:

$$\begin{aligned}\bar{X} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ &= \frac{16 + 18 + 18 + 22 + 22 + \ldots + 46 + 46 + 49 + 57}{21} \\ &= \frac{676}{21} \\ &= 32.19\end{aligned}$$

The lower quartile: This is the median of the lower half of scores. If we split the data at 32 (not including this score), there are 10 scores below this value. The median of 10 = 11/2 = 5.5th score. Therefore, we take the average of the 5th score and the 6th score. The 5th score is 22, and the 6th is 23; the lower quartile is therefore 22.5 seconds.

The upper quartile: This is the median of the upper half of scores. If we split the data at 32 (not including this score), there are 10 scores above this value. The median of 10 = 11/2 = 5.5th score above the median. Therefore, we take the average of the 5th score above the median and the 6th score above the median. The 5th score above the median is 42 and the 6th is 43; the upper quartile is therefore 42.5 seconds.

The range: This is the highest score (57) minus the lowest (18), i.e. 39 seconds.

The interquartile range: This is the difference between the upper and lower quartile: 42.5−22.5 = 20.

- Assuming the same mean and standard deviation for the Beachy Head example above, what's the probability that someone who threw themselves off of Beachy Head was 30 or younger?

As in the example, we know that the mean of the suicide scores was 36, and the standard deviation 13. First we convert our value to a *z*-score: the 30 becomes (30−36)/13 = −0.46. We want the area below this value (because 30 is below the mean), but this value is not tabulated in the Appendix. However, because the distribution is symmetrical, we could instead ignore the minus sign and look up this value in the column labelled 'Smaller Portion' (i.e. the area above the value 0.46). You should find that the probability is .32276, or, put another way, a 32.28% chance that a suicide victim would be 30 years old or younger. By looking at the column labelled 'Bigger Portion' we can also see the probability that a suicide victim was aged 30 or more! This probability is .67724, or there's a 67.72% chance that a suicide victim was older than 30 years old!

Smart Alex's solutions

Task 1

What are (broadly speaking) the five stages of the research process? ①

1. Generating a research question: through an initial observation (hopefully backed up by some data).
2. Generate a theory to explain your initial observation.

3. Generate hypotheses: break your theory down into a set of testable predictions.
4. Collect data to test the theory: decide on what variables you need to measure to test your predictions and how best to measure or manipulate those variables.
5. Analyse the data: look at the data visually and by fitting a statistical model to see if it supports your predictions (and therefore your theory). At this point you should return to your theory and revise it if necessary.

Task 2

What is the fundamental difference between experimental and correlational research? ①

- In a word, *causality*. In experimental research we manipulate a variable (predictor, independent variable) to see what effect it has on another variable (outcome, dependent variable). This manipulation, if done properly, allows us to compare situations where the causal factor is present to situations where it is absent. Therefore, if there are differences between these situations, we can attribute cause to the variable that we manipulated. In correlational research, we measure things that naturally occur and so we cannot attribute cause, but instead look at natural covariation between variables.

Task 3

What is the level of measurement of the following variables? ①

- The number of downloads of different band's songs on iTunes:
    - This is a discrete ratio measure. It is discrete because you can download only whole songs, and it is ratio because it has a true value of 0 (no downloads at all).
- The names of the bands downloaded.
    - This is a nominal variable. Bands can be identified by their name, but the names have no meaningful order. That fact that Norwegian black metal band 1349 called themselves 1349 does not make them better than British boy-band has-beens 911; the fact that 911 were a bunch of talentless idiots does, though.
- The position in the iTunes download chart.
    - This is an ordinal variable. We know that the band at number 1 sold more than the band at number 2 or 3 (and so on) but we don't know how many more downloads they had. So, this variable tells us the order of magnitude of downloads, but doesn't tell us how many downloads there actually were.
- The money earned by the bands from the downloads.
    - This variable is continuous and ratio. It is continuous because money (pounds, dollars, euros or whatever) can be broken down into very small amounts (you can earn fractions of euros even though there may not be an actual coin to represent these fractions).
- The weight of drugs bought by the band with their royalties.
    - This variable is continuous and ratio. If the drummer buys 100 g of cocaine and the singer buys 1 kg, then the singer has 10 times as much.
- The type of drugs bought by the band with their royalties.
    - This variable is categorical and nominal: the name of the drug tells us something meaningful (crack, cannabis, amphetamine, etc.) but has no meaningful order.
- The phone numbers that the bands obtained because of their fame.
    - This variable is categorical and nominal too: the phone numbers have no meaningful order; they might as well be letters. A bigger phone number did not mean that it was given by a better person.
- The gender of the people giving the bands their phone numbers.
    - This variable is categorical and binary: the people dishing out their phone numbers could fall into one of only two categories (male or female).
- The instruments played by the band members.

  - This variable is categorical and nominal too: the instruments have no meaningful order but their names tell us something useful (guitar, bass, drums, etc.).
- The time, they had spent learning to play their instruments.
  - This is a continuous and ratio variable. The amount of time could be split into infinitely small divisions (nanoseconds even) and there is a meaningful true zero (0 time spent learning your instrument means that, like 911, you can't play at all).

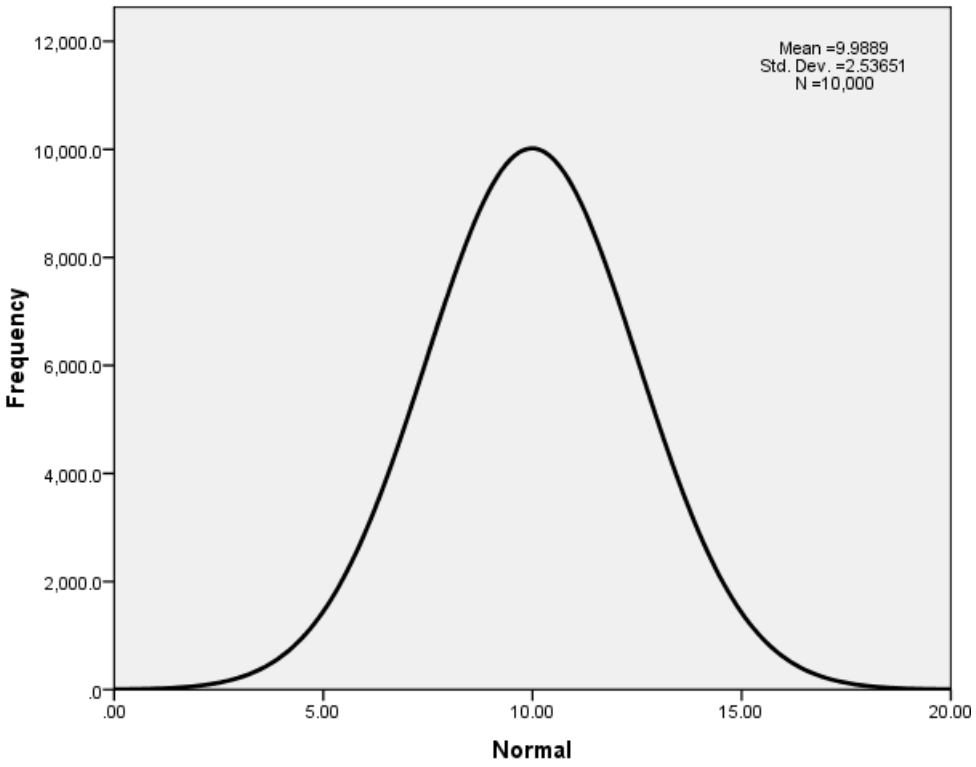
Task 4

Say I own 857 CDs. My friend has written a computer program that uses a webcam to scan my shelves in my house where I keep my CDs and measure how many I have. His program says that I have 863 CDs. Define measurement error. What is the measurement error in my friends CD counting device?

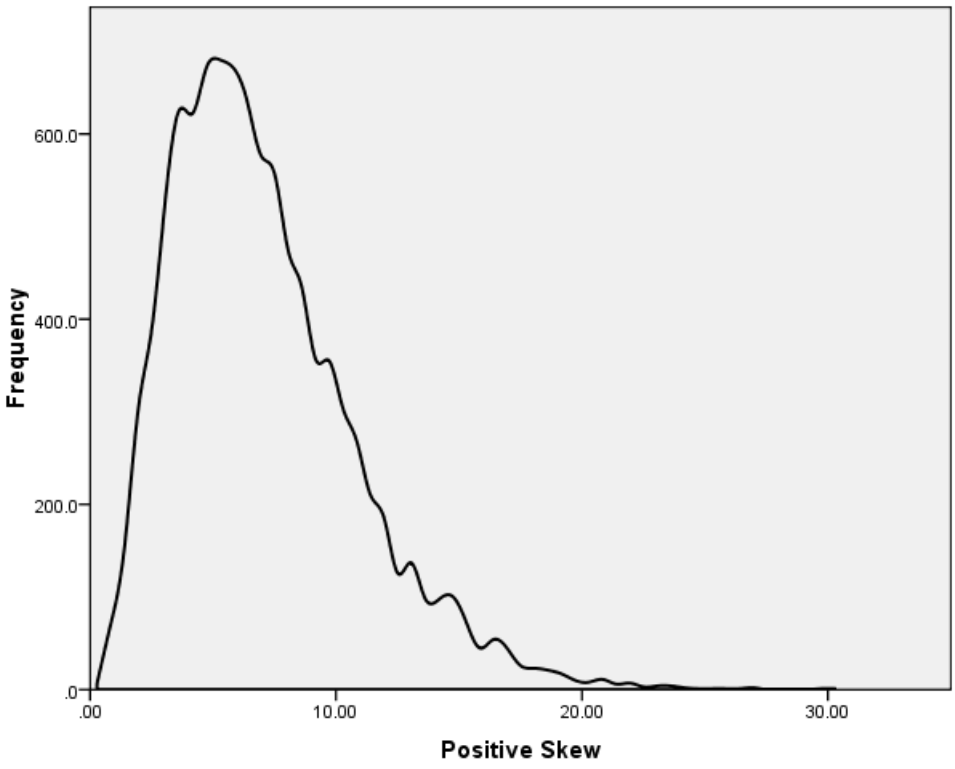
- Measurement error is the difference between the true value of something and the numbers used to represent that value. In this trivial example, the measurement error is 6 CDs. In this example we know the true value of what we're measuring; usually we don't have this information so we have to estimate this error rather than knowing its actual value.

Task 5

Sketch the shape of a normal distribution, a positively skewed distribution and a negatively skewed distribution.
Normal:



Positive skew:



Negative skew: