Unit 3 Statistics

Task 1 Data - collocations

Complete the list of collocations with the noun ´data´ with other examples.

ADJ.

E.g. We have collected the ***raw*** *data and are about to begin analysing it.*

QUANT.

E.g. One vital ***item*** *of data was missing from the table.*

VERB + DATA

E.g. They are not allowed to ***hold*** *data on people's private finances.*

DATA + VERB

E.g. This data ***reflects*** *the magnitude of the problem.*

DATA + NOUN **acquisition, capture, collection | entry, input | storage | access, retrieval | analysis, handling, management, manipulation, processing | exchange, interchange, transfer, transmission | protection, security | source | archive, bank** (also **databank**), **base** (also **database**), **file | system**

PREP. **in the ~** *We have found some very interesting things in the data.* | **~ about** *Data about patients is only released with their permission.* | **~ for** *We have no data for southern Spain.* | **~ from** *My aim is to synthesize data from all the surveys.* | **~ on** *data on the effects of pollution*

PHRASES **the acquisition/handling/storage, etc. of data, a source of data**

(http://oxforddictionary.so8848.com/search?word=data)

Task 2 "Lies, damned lies and statistics"

A) Try to decide what is suspicious about the following cases:

- We hear a prediction that life expectancy will reach 150 years in the next century, based on simple extrapolation from increases over the past 100 years.

- In 1999, Sally Clark, a young British lawyer, was tried, convicted, and given a life sentence for murdering her two baby sons. Her first child died in 1996, aged 11 weeks, and her second died in 1998, aged 8 weeks. The verdict depended on what has become a byword for the misunderstanding and misuse of statistics, when the paediatrician Sir Roy Meadow in his role as expert witness for the prosecution, claimed that the chance of two children dying from cot death was 1 in 73 million. He obtained this figure by simply multiplying together the chance for the two deaths separately.

B) Data

Data are the raw material on which the discipline of statistics is built as well as the raw material from which individual statistics themselves are calculated. These data are typically numbers. In fact, however, data are more than merely numbers. To be useful, the numbers must be associated with some (1) ___________. For example, we need to know what the measurements are measurements *of*, and just *what* has been counted. To produce valid and accurate results when we (2) _________ our statistical analysis, we also need to know something about how the values have been (3) _________ . Did everyone we asked give answers to a questionnaire, or did only some people answer? If only some answered, are they properly (4) ____________ of the population of people we wish to make a statement about, or is the sample distorted in some way? We also need to know if a measuring instrument is (5)___________ and whether the data are up (6) _____ ________ . There is an infinite number of such questions which could be asked, and we need to be alert for any which could influence the conclusions we (7) ___________ .

C) Measures of central tendency

Measures of central tendency are numbers that describe what is average or typical of the distribution.

These measures include the mean, median and mode and standard deviation.

Match each definition with the type of average:

a) It is the value such that half the numbers in the data set are larger and half are smaller
b) It is the value taken most frequently in a sample
c) It is the value found by adding all the numbers up and dividing by how many there are

Which of the statistics would you choose when talking about average salaries?

Which of the statistics would be suitable when presenting the number of children per family?

D) Complete the text with phrases below:

a) a very ´representative´ value of the set
b) from individual values in a set of numbers
c) of the general size of the values in the data
d) than the mean and median
e) can be misleading

Dispersion

Averages, such as the mean and the median, provide single numerical summaries of collections of numerical values. They are useful because they give an indication (1) … . However, single summary values (2) … . In particular, single values might deviate substantially (3) … . To illustrate, suppose that we have a set of a million and one numbers, taking the values 0,1,2,3,4,…, 1,000,000. Both the mean and the median of this set of values are 500,000.
But, it is readily apparent that this is not (4) … . At the extremes, one value in the set is half a million larger and one value is half a million smaller (5) … .

E) Standard deviation

Standard deviation is a number that indicates how much each of the values in the distribution deviates from the mean (or centre) of the distribution. If the data points are all close to the mean, then the standard deviation is close to ______. If many data points are far from the mean, then the standard deviation is ______ . If all the data values are equal, then the standard deviation is __________ .

Here is how to calculate standard deviation. Based on the description, write the mathematical formula for standard deviation.

- Calculate the mean (M)
- Subtract the mean from each subject´s score (X)
- Square the answer
- Sum the squared scores (Σ)
- Divide by the number of participants minus 1
- Take the square root of the answer

F) Read the text about normal distribution and identify one piece of information that is false:

A normal distribution of data means that most of the examples in a set of data are close to the ´average´, while a few examples are at one extreme or the other.

Normal distribution graphs have these characteristics:

- The curve has a single peak
- It is bell-shaped
- The mean (average) lies at the centre of the distribution and the distribution is symmetrical around the mean
- The two ´tails´ of the distribution cross the x axis

G) Skewness

Measures of dispersion tell us how much the individual values deviate from each other. But they do not tell us in what way they deviate. In particular, they do not tell us if the larger deviations tend to be for the larger values or the smaller values in the data set. E.g. if there are five company employees, four of whom earned about $10,000 per year and one earned around ten times that, the measure of dispersion (the standard deviation, for example) would tell us that the values were quite widely spread out, but would not tell us that one of the values was much larger than the others. Indeed, the standard deviation for the five values $90,000, $89,999, $89,998, $889,997 and $1 is exactly the same as for the original five values. What is different is that the anomalous value (the $1value) is now very small instead of very large. To detect this difference, we need another statistic to summarize the data, one which picks up on and measures the asymmetry in the distribution of values. One kind of asymmetry in distribution of values is called *skewness.* Our original employee salary example, with one anomalously large value of $99,999, is right skewed because the distribution has a long ´tail´ stretching out to the single very large value. This distribution has many smaller values and very few larger values. In contrast, the distribution of values in which $1 is the anomaly, is left skewed, because the bulk of the values bunch together and there is a long tail stretching down to the single very small value.

Draw a right and a left skewed curve.

Task 3 Assignment

Performance analyst

Amy is a performance analyst for a professional football club. One of her key job roles is to analyse football matches to see how well the team and players are performing from a statistics perspective.

´Professional football is such big business now that football teams are looking for as much detail as they can about the performance of the team as a whole, and of the individual players, so that they can start to see who is performing well, and who isn´t. This has lots of influence over team selections as the statistics that I produce are often good indicators of whether or not a player is tired and needs a rest, or even if they look like they´re not putting in enough effort. I can also give players detailed breakdowns of what they have done during the game. For example, how many passes they have completed, how far they have run, how many shots they´ve had on target. I keep a record of this for them over the course of the season and we can use this to help develop the players.

As well as being able to analyse players and our own team, I analyse lots of games of opposition teams if we have a big game coming up (for example, I might look at their star striker and be able to give the staff a breakdown of where they prefer to shoot from and how they score most of their goals) which the manager and the coaching staff find really useful – and sometimes they might even get me to compare the performance statistics of a couple of players that they´re interested in buying.

One of the key problems I face is that sometimes the playing and coaching staff don´t understand some of the statistic that I present them with, so I do need to spend some time with them to explain what is going on.´

What type of statistics do you think would be useful to use in this type of sporting environment?

How do you think you could make the statistics easier to understand?

How much do you think players would appreciate this type of feedback?