Everything you ever wanted to know about statistics (well, sort of)

Self-test answers

- We came across some data about the number of friends that 11 people had on Facebook (22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 252). We calculated the mean for these data as 96.64. Now calculate the sums of squares, variance and standard deviation.

To calculate the sum of squares, take the mean from each value, then square this difference. Finally, add up these squared values:

Value	Mean	Difference	Difference squared
22	96.64	-74.64	5570.59
40	96.64	-56.64	3207.68
53	96.64	-43.64	1904.13
57	96.64	-39.64	1571.04
93	96.64	-3.64	13.22
98	96.64	1.36	1.86
103	96.64	6.36	40.50
108	96.64	11.36	129.13
116	96.64	19.36	374.95
121	96.64	24.36	593.59
252	96.64	155.36	24137.86
		Sum =	**37544.55**

So, the sum of squared errors is a massive 37,544.55.

The variance is the sum of squared errors divided by the degrees of freedom ($N - 1$). There are 11 scores and so the degrees of freedom are 10. The variance is, therefore, 37,544.55/10 = 3754.45.

Finally, the standard deviation is the square root of the variance: $\sqrt{3754.45}$ = 61.27.

- Calculate these values again but excluding the outlier (252).

To calculate the sum of squares, take the mean from each value (note that it has changed because the outlier is excluded), then square this difference. Finally, add up these squared values:

Value	Mean	Difference	Difference squared
22	81.10	-59.10	3492.81
40	81.10	-41.10	1689.21
53	81.10	-28.10	789.61
57	81.10	-24.10	580.81
93	81.10	11.90	141.61
98	81.10	16.90	285.61
103	81.10	21.90	479.61
108	81.10	26.90	723.61
116	81.10	34.90	1218.01
121	81.10	39.90	1592.01
		Sum =	10992.90

So, the sum of squared errors is 10,992.90.

The variance is the sum of squared errors divided by the degrees of freedom ($N - 1$). There are 10 scores and so the degrees of freedom are 9. The variance is, therefore, 10,992.90/9 = 1221.43.

Finally, the standard deviation is the square root of the variance: $\sqrt{1221.43}$ = 34.95.

Note, then, that like the mean itself the standard deviation is hugely influenced by outliers: the removal of this one value has halved the standard deviation!

- Compute the range but excluding the score of 252.

The range is the lowest score (22) subtracted from the highest score (now 121), which gives us 121–22 = 99.

- We came across some data about the number of friends that 11 people had on Facebook. We calculated the mean for these data as 96.64 and standard deviation as 61.27. Calculate a 95% confidence interval for this mean.

First we need to calculate the standard error,

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N}} = \frac{61.27}{\sqrt{11}} = 18.47$$

The sample is small, so to calculate the confidence interval we need to find the appropriate value of t. First we need to calculate the degrees of freedom, $N - 1$. With 11 data points, the degrees of freedom are 10. For a 95% confidence interval we can look up the value in the column labelled 'Two-Tailed Test', '0.05' in the table of critical values of the t-distribution (Appendix). The corresponding value is 2.23.

The confidence intervals are, therefore:

lower boundary of confidence interval = $\bar{X} - (2.23 \times SE)$ = 96.64 – (2.23 × 18.47) = 55.63

upper boundary of confidence interval = $\bar{X} - (2.23 \times SE)$ = 96.64 + (2.23 × 18.47) = 137.65

- Recalculate the confidence interval assuming that the sample size was 56.

First we need to calculate the new standard error,

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N}} = \frac{61.27}{\sqrt{56}} = 8.19$$

The sample is big now, so to calculate the confidence interval we can use the critical value of z for a 95% confidence interval (i.e. 1.96). The confidence intervals are, therefore:

lower boundary of confidence interval = $\overline{X} - (1.96 \times SE)$ = 96.64 – (1.96 × 8.19) = 80.59

upper boundary of confidence interval = $\overline{X} - (1.96 \times SE)$ = 96.64 + (1.96 × 8.19) = 112.69

- What are the null and alternative hypotheses for the following questions:

1. 'Is there a relationship between the amount of gibberish that people speak and the amount of vodka jelly they've eaten?'
    - Null hypothesis: There will be no relationship between the amount of gibberish that people speak and the amount of vodka jelly they've eaten.
    - Alternative hypothesis: There will be a relationship between the amount of gibberish that people speak and the amount of vodka jelly they've eaten.
2. 'Is the mean amount of chocolate eaten higher when writing statistics books then when not?'
    - Null hypothesis: There will be no difference in the mean amount of chocolate eaten when writing statistics textbooks compared to when not writing them.
    - Alternative hypothesis: The mean amount of chocolate eaten when writing statistics textbooks will be higher than when not writing them.

Smart Alex's solutions

Task 1

- Why do we use samples?

We are usually interested in populations, but because we cannot collect data from every human being (or whatever) in the population, we collect data from a small subset of the population (known as a sample) and use these data to infer things about the population as a whole.

Task 2

- What is the mean and how do we tell if it's representative of our data?

The mean is a simple statistical model of the centre of a distribution of scores: a hypothetical estimate of the 'typical' score. We use the variance, or standard deviation, to tell us whether it is representative of our data. The standard deviation is a measure of how much error there is associated with the mean: a small standard deviation indicates that the mean is a good representation of our data.

Task 3

- What's the difference between the standard deviation and the standard error?

The standard deviation tells us how much observations in our sample differ from the mean value within our sample. The standard error tells us not about how the sample mean represents the sample itself, but how well the sample mean represents the population mean. The standard error is the standard deviation of the sampling distribution of a statistic. For a given statistic (e.g. the mean) it tells us how much variability there is in this statistic across samples from the same population. Large

values, therefore, indicate that a statistic from a given sample may not be an accurate reflection of the population from which the sample came.

Task 4

- In Chapter 1 we used an example of the time taken for 21 heavy smokers to fall off of a treadmill at the fastest setting (18, 16, 18, 24, 23, 22, 22, 23, 26, 29, 32, 34, 34, 36, 36, 43, 42, 49, 46, 46, 57). Calculate the sums of squares, variance, standard deviation and standard error of these data.

To calculate the sum of squares, take the mean from each value, then square this difference. Finally, add up these squared values:

Value	Mean	Difference	Difference squared
18	32.19	-14.19	201.37
16	32.19	-16.19	262.13
18	32.19	-14.19	201.37
24	32.19	-8.19	67.08
23	32.19	-9.19	84.46
22	32.19	-10.19	103.85
22	32.19	-10.19	103.85
23	32.19	-9.19	84.46
26	32.19	-6.19	38.32
29	32.19	-3.19	10.18
32	32.19	-0.19	0.04
34	32.19	1.81	3.27
34	32.19	1.81	3.27
36	32.19	3.81	14.51
36	32.19	3.81	14.51
43	32.19	10.81	116.85
42	32.19	9.81	96.23
49	32.19	16.81	282.56
46	32.19	13.81	190.70
46	32.19	13.81	190.70
57	32.19	24.81	615.51
		Sum =	**2685.24**

So, the sum of squared errors is a massive 2685.24.

The variance is the sum of squared errors divided by the degrees of freedom ($N - 1$). There are 21 scores and so the degrees of freedom are 20. The variance is, therefore, 2685.24/20 = 134.26.

The standard deviation is the square root of the variance: $\sqrt{134.26}$ = 11.59.

The standard error will be: $\dfrac{s}{\sqrt{N}} = \dfrac{11.59}{\sqrt{21}} = 2.53$

The sample is small, so to calculate the confidence interval we need to find the appropriate value of t. First we need to calculate the degrees of freedom, $N - 1$. With 21 data points, the degrees of freedom are 20. For a 95% confidence interval we can look up the value in the column labelled 'Two-Tailed Test', '.05' in the table of critical values of the t-distribution (Appendix). The corresponding value is 2.09. The confidence intervals are, therefore:

lower boundary of confidence interval = $\overline{X} - (2.09 \times SE)$ = 32.19 – (2.09 × 2.53) = 26.90

upper boundary of confidence interval = $\overline{X} - (2.09 \times SE)$ = 32.19 + (2.90 × 2.53) = 37.48

Task 5

- What do the sum of squares, variance and standard deviation represent? How do they differ?

All of these measures tell us something about how well the mean fits the observed sample data. Large values (relative to the scale of measurement) suggest the mean is a poor fit to the observed scores, and small values suggest a good fit. They are also, therefore, measures of dispersion, with large values indicating a spread-out distribution of scores and small values showing a more tightly packed distribution. These measures all represent the same thing, but differ in how they express it. The sum of squared errors is a 'total' and is, therefore, affected by the number of data points. The variance is the 'average' variability, but with units squared. The standard deviation is the average variation, but converted back to the original units of measurement. As such, the size of the standard deviation can be compared to the mean (because they are in the same units of measurement).

Task 6

- What is a test statistic and what does it tell us?

A test statistic is a statistic for which we know how frequently different values occur. The observed value of such a statistic is typically used to test hypotheses, or to establish whether a model is a reasonable representation of what's happening in the population.

Task 7

- What are Type I and Type II errors?

A Type I error occurs when we believe that there is a genuine effect in our population when, in fact, there isn't. A Type II error occurs when we believe that there is no effect in the population when, in reality, there is.

Task 8

- What is an effect size and how is it measured?

An effect size is an objective and standardized measure of the magnitude of an observed effect. Measures include Cohen's d, the odds ratio and Pearson's correlations coefficient, r.

Task 9

- What is statistical power?

Power is the ability of a test to detect an effect of a particular size (a value of .8 is a good level to aim for).