Exercise Set 8 - Solution

1 Comparing analysis methods for earth nitrogen content

Both sets (X and Y) are paired since they are based on repetitive measurements on the same earth samples. The zero hypothesis H_0 states that both methods are identical, so that the average difference is zero.

First we can compute the difference between the results for each earth sample. This gives us a list of 15 difference values. It is then possible to find the mean and the unbiased estimation for the standard deviation.

$$d_i = y_i - x_i \quad \text{for all } i \qquad \bar{D} = 0.1067 \qquad s_D^2 = 0.04112 \qquad s_D = 0.2038$$

Since the variance wasn't known and was estimated, we use the Student t-test.

$$t = \frac{\bar{D}}{s_D/\sqrt{n}} = 2.038$$

The number of degrees of freedom is $N_D - 1 = 14$.

We are just checking for difference/deviation, not specifically if M2 gives (for example) larger values. So we have a two-sided test. Hence, the probability is found via $2 * CDF - 1 = P$, so we would have to look for where the CDF of the t-distribution reaches 98.75%.

This value cannot directly be looked up, but we know that the corresponding t-score will be close to $qt_{14}(99\%) = 2.624$, the and larger than $qt_{14}(97.5\%) = 2.145$.

The t-score we have is in fact lower than both values so the null hypothesis cannot be rejected (and it could not even be rejected at the 95% confidence level, where 2.145 would be the cut-off). This means one cannot claim "the two measurement methods are significantly different in the mean value they find", which is what the company would have liked to do. One cannot say that method 2 is less *accurate* than method one.

Could one say that method 2 is less *precise* than method 1? If we would have looked at the unbiased estimator for the std. dev. of each data set individually, we would have found $s_X = 0.575$ and $s_Y = 0.6033$. These values are much larger than s_D so the variation comes primarily from the samples, not from the measurement device. This is a bad starting point for such a comparison. Nevertheless s_Y is a little bit larger than s_X, so could one argue that the latter method is less precise? The difference turns out not to be significant, but proving this would require a separate test!.

2 Novel diet for a healthy lifestyle - revisited

This time both sets aren't paired, so we have to estimate the combined variance.

$$\bar{X}_1 = 93.0 \qquad s_1^2 = 48.18 \qquad \bar{X}_2 = 103.6 \qquad s_2^2 = 73.17$$

No information about the sample variance is given, so we cannot assume them to be equal and have to do a Welch test. Both series have $n_1 = n_2 = 12$ datapoints.

First we determine the degrees of freedom by calculating:

$$a = (\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2 = 102.16 \qquad b = \frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1} = 4.85 \qquad a/b = 21.1$$

The closest integer to $\frac{a}{b}$ is 21, this is the degree of freedom we have to use.

At this stage we could ask the question if a 2-sample t-test would also be a decent approximation. The degree of freedom we find is very close to the degree of freedom of the 2-sample t-test, $(n_1-1)+(n_2-1) = 22$. The standard deviations are somewhat different however. Importantly, as we do not have any information about the variances being equal, we stick to the Welch test.

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = -3.33$$

As we want to know if they are **different** (not, for example, if 1 is larger than 2, only if they are different), we do a 2-sided test. The value is big and negative, so lets focus on the critical region at low values of T. We compute the critical value as the 1% quantile of the t_{21} distribution, for example in R using "qt(0.01,21)", and find -2.51. As our T-value is lower than the critical value, it is very unlikely that both diets are equivalent. We reject the null hypothesis.

3 Are the dice fair?

If the dice were fair, the expected frequencies can be computed using the Binomial distribution $\mathcal{B}(n,p)$, with $n = 4$ and $p = 0.5$ the probability to get an even result. The expected frequency to get k even dice is $200 \cdot \binom{n}{k} p^k (1-p)^{n-k} = 12.5 \cdot \binom{n}{k}$

Number of even dice	x_i	0	1	2	3	4
Observed absolute frequencies	n_i	10	41	70	57	22
Expected absolute frequencies	$p_i * 200$	12.5	50	75	50	12.5

Using both the observed and the expected frequencies, we find than χ^2:

$$\chi^2 = \sum_{i=0}^{4} \frac{(n_i - p_i * 200)^2}{p_i * 200} = 10.653$$

The number of degrees of freedom is $\nu = 5 - 1$. χ^2 is bigger than the quantile $q\chi_4^2(95\%) = 9.488$, we can reject the hypothesis that the dice are fair with 95% of confidence. (Note that for the chi^2 there is no one-sided/two-sided distinction, as we always look at a squared deviation).

We could check if we can be even more confident in our rejection. Since $\chi^2 < q\chi_4^2(97.5\%) = 11.14$, it is however not possible to reject the hypothesis with 97.5% of confidence.

4 Are politicians living the same life as the "average" person?

The first null hypothesis to test is that FFA politicians' income are the same as the average in the population. It is tested against the hypothesis that the revenues of FFA politicians are bigger.

We use the χ^2 test as before. The expected frequencies, when considering that the null hypothesis is true, are the percentage of the population with a given income times 100, the number of FFA in this survey.

FFA Income	$>$ 200 kCHF	100 to 200 kCHF	75 to 100 kCHF	$<$ 75 kCHF
Observed frequencies	50	25	15	10
Expected frequencies	8.33	16.67	50	25

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = 246.1$$

This is clearly bigger than $q\chi_3^2(99\%) = 11.34$, so we can reject the null hypothesis with 99% of confidence.

The same things are done with FS politicians.

FS Income	$>$ 200 kCHF	100 to 200 kCHF	75 to 100 kCHF	$<$ 75 kCHF
Observed frequencies	30	40	45	45
Expected frequencies	13.33	26.67	80	40

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = 43.4$$

This is bigger than $q\chi_3^2(99\%)$ too, so the same conclusion follows.

Finally the income of the politicians from both parties is compared. The null hypothesis (both income are the same) is tested against the hypothesis that FFA's revenues are bigger.