Practical "Hypothesis testing"
P2

Read the paper provided for the discussion (Du Prel JB et al., 2009. Confidence Interval or P-Value? *Dtsch Arztebl Int* 106:335-339). In groups of four or five, discuss the following questions:

1. Consider an example from Gardner and Altman (1986) where samples of 100 diabetic and 100 non-diabetic men of a certain age are compared in a study. A difference of 6.0 mm Hg is found between their mean systolic blood pressures, and the standard error of this difference between sample means is 2.5 mm Hg.

(a) A two independent samples two-tailed t-test is used and the associated p-value is 0.02. Based on the p-value, would you say this might be due to a real difference, or that it might be due to chance? Does the p-value provide information about the exact size and direction of the difference?

Suggested answer:
The observed difference might be due to chance. The probability that this difference, in standard errors, is greater than what was found, is of 2$\%$ under the null hypothesis of no difference. The smaller the p-value, the less likely it is that the difference can be assigned to chance, and so the less plausible is the null hypothesis of no difference. The p-value does not provide information about the exact size nor about the direction of the difference.

(b) The best approximation to the population mean difference is provided by the difference of the sample means, 6.0. How precise is this? Explain why this can be shown by a confidence interval.

Suggested answer:
This mean does not give any information about how exact it is. It lacks a measure of precision. Variability of systolic blood pressure in the sample, as well as the sample size, play a role there. A confidence interval gives a range of values considered plausible for the population. It depends on the standard error of the difference of the sample means, and hence (see Session 1) on the standard deviation and the sample size, as well as on the degree of "confidence" we wish to have.

(c) Looking at the confidence interval shown in Figure 1, explain, as in 1a), the difference in mean systolic blood pressure. Is it possible that this interval does not contain the true mean difference?

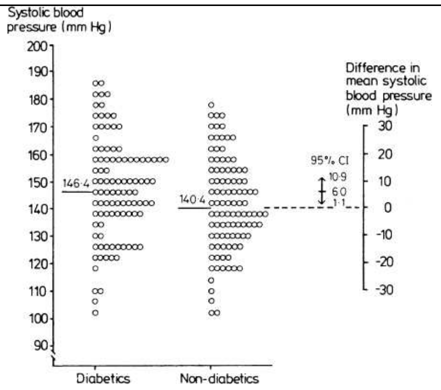

Figure 1: Systolic blood pressure in 100 diabetics and 100 non-diabetics with mean levels of 146.4 and 140.4 mm Hg, respectively. The difference between the sample means of 6.0 mm Hg is shown to the right together with the 95% confidence interval from 1.1 to 10.9 mm Hg.

Suggested answer:

The observed difference might not be due to a true difference in blood pressure, but simply due to chance. It could be that the true difference between population means lies outside the confidence interval in Figure 1, though it is more likely that it lies near the point estimate (6.0).

(d) If a systolic blood pressure difference of at least 4 mm is defined as clinically relevant, to which of the four examples in Figure 2 of Du Prel et al. (2009) does this example correspond?

Suggested answer:

Statistically significant and clinically relevant (b).

(e) Would it be equally helpful to report confidence intervals for the mean systolic blood pressure of diabetics and non-diabetics separately instead of the confidence interval for the mean difference?

Suggested answer:

The main question of the study concerns the difference in systolic blood pressure, so separate confidence intervals for the mean of each of the two groups will not necessarily provide the precision or statistical significance of the difference.

2. Suppose that the same analysis is done using samples half the size, that however have the same observed mean difference (6.0) and standard deviation (do not confound with standard error) of the difference in sample means as in 1a. Similarly to Figure 1, means and confidence interval are displayed in Figure 2.

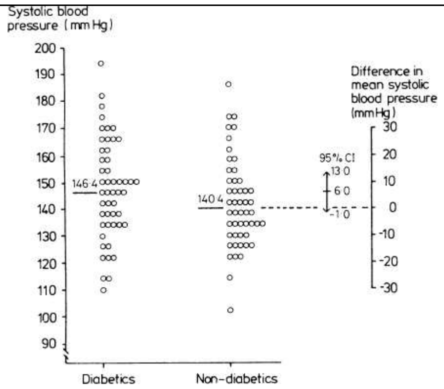

Figure 2: Systolic blood pressure in 50 diabetics and 50 non-diabetics with mean levels of 146.4 and 140.4 mm Hg, respectively. The difference between the sample means of 6.0 mm Hg is shown to the right together with the 95% confidence interval from -1.0 to 13.0 mm Hg.

(a) Explain why the confidence interval is now wider. Does this affect your conclusion about the difference in means?

Suggested answer:
The difference in sample means is the same, as well as the standard deviation of the difference, but the sample size decreases and we thus lose precision. We saw during Session 1 that if we have one sample, the confidence interval for the mean is defined as $[\bar{x} - c * SEM, \bar{x} + c * SEM]$, with some c > 0 and $\bar{x}$ being the sample mean, and SEM being the standard error of the sample mean. We also saw that the standard error of the sample mean is inversely proportional to the square root of the sample size. We can extend this to two samples, as in the present example, so reducing the sample size leads to a wider confidence interval.

(b) As before, if a systolic blood pressure difference of at least 4 mm is defined as clinically relevant, to which of the four examples in Figure 2 of Du Prel et al. (2009) does this example correspond?

Suggested answer:
Clinically relevant, not statistically significant (d).

(c) Does a p-value $>$ 0.05 imply that there is no difference in mean systolic blood pressure between diabetics and non-diabetics? What is the risk of simply distinguishing between significant and non-significant?

Suggested answer:
We can say that there is not enough evidence to reject the null hypothesis that there is no difference between diabetics and non-diabetics with regard to their mean systolic blood pressure. Using a strict convention such as rejecting if $p <$ 0.05 and not rejecting otherwise does not take into account questions such as the sample size nor clinical or biological relevance. Besides, a p-value of 0.04 is not so different from a p-value of 0.06, but would lead to opposite conclusions about significance if we simply used the dichotomy significant/non-significant at a 5$\%$ level.

(d) Explain what "publication bias" is.

Suggested answer:

Workers, readers and journals ignore findings that are potentially clinically useful only because they are not statistically significant. A study has shown that high-impact journals prefer to publish significant results (Easterbrook et al. 1991). This can distort reviews and meta-analyses.

3. Researchers plan to conduct a study on the quality of life (QoL) after breast cancer surgery to see whether women who undergo breast conserving surgery are more satisfied than women who undergo mastectomy. It is known that women after mastectomy score their QoL on average as 15 and it is expected that women after breast conserving surgery will score their QoL on average as 25. The standard deviation in both groups is about 13. How many women should be included in each group, assuming both groups have similar number of women, to reach the power level of 90$\%$?

   Suggested answer:
   We have to sample case where we test H_0 : *QoL after mastectomy is the same as QoL after breast conserving surgery* vs H_1 : *QoL after mastectomy is different than QoL after breast conserving surgery*. Moreover, $\bar{x}_1 = 25$, $\bar{x}_2 = 15$, $\sigma_{\bar{x}_1} = \sigma_{\bar{x}_2} = 13$. Using $\alpha = 5\%$ and equal sample size in both groups, $k = 1$, we can calculate:

$$n_1 = (13^2 + 13^2)\left(\frac{1.96 + 1.28}{25 - 15}\right)^2 = 35.48$$

   Thus, 36 women per group, 72 women in total should be included in the study.