Biostats VHM 801 Course
Fall 2016, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to home assignment II

The solution contains the level of detail required for a 100% mark (although some leverage is applied for minor points), except for the last question where results are shown and discussed for all three variables instead of one. The data are from Strain et al. (1994), *J. Amer. Med. Assoc.* **272**, 1043–1048, and also form the basis of a story in the EESEE supplement to the IPS textbook.

1. Clinical dependence

The 26 participants were selected as volunteers meeting certain enrolment criteria, one of them being that they believed they were psychologically or physically dependent on caffeine. The inference can (at best, without considering further biases) be considered representative for a population defined by these two properties (volunteers, dependence on caffeine). It seems natural to assume that the outcome of the clinical assessment for the 26 persons could be considered as a binomial setting, if no further characteristics are used to distinguish the persons. Therefore, if X denotes the number of clinically dependent persons, we would assume $X \sim \text{Bin}(26, p)$, where p is the population probability of a clinical dependence. Because the sample with $X_{\text{obs}} = 15$ does not meet the conditions for the classical method for confidence intervals, we use instead the "plus four" method, as follows.

$$\hat{p} = X_{\text{obs}}/26 = 15/26 = 0.577,\ \text{SE}(\hat{p}) = \sqrt{0.577\,(1-0.577)/26} = 0.0969,$$
$$\tilde{p} = (X_{\text{obs}} + 2)/(26+4) = 17/30 = 0.567,\ \text{SE}(\tilde{p}) = \sqrt{0.567\,(1-0.567)/30} = 0.0905,$$
$$95\%\ \text{CI} : \tilde{p} \pm z^*\text{SE}(\tilde{p}) = 0.567 \pm 1.96 \cdot 0.0905 = 0.567 \pm 0.177 = (0.389, 0.744),$$

using all decimals in the calculations. We are 95% confident that this interval will cover the population value. The "exact" CI based on the binomial distribution (from software) is similar: $(0.369, 0.766)$.

2. Double-blind study

A double-blind study is one where neither the subjects nor the experimentor knows the true allocation to treatment groups. For the present study it means that it was not known in which treatment period the subjects received the caffeine and no-caffeine capsules. The blinding of the subjects was primarily done to avoid any placebo effects, whereas the blinding of the experimentators was done to avoid any bias in recording of evidence of withdrawal symptoms. It is possible that the presence of research staff with knowledge of the true treatment group could affect answers to questions or any general discussion of the participants' experiences of the treatment periods.

3. Randomization

Randomization is a safeguard against any unexpected (and undesired) effects arising from a systematic assignment of treatments. For example, if all subjects received the no-caffeine pill first, the caffeine effect would be inseparable from any period effect. Randomization was also necessary to ensure that the experiment was truly double-blind.

The randomization involves a random assignment of each subject to receive either the caffeine or the no-caffeine treatment first. This could be achieved by flipping a coin, drawing red and black cards from a deck, using a table of random digits, or drawing random numbers between 0 and 1 (where e.g. numbers ≤ 0.5 would imply the subject to receive the caffeine treatment first, and numbers > 0.5 to receive the no-caffeine treatment first). In any case equal probability of the two possibilities should be ensured.

4. Study periods one week apart

The two study periods could not be on consecutive days because there might be a residual effect of one or both of the treatments (in particular, the no-caffeine treatment) for some days after the treatment period. If such a residual effect affects the results of the next treatment period, we talk about a carry-over effect. The periods were held one week apart to make the study periods as equal as possible. Different weekdays might involve different habits and activities for the subjects; this would in particular be true if one day fell during the week and another day during the weekend.

5. Probability of a withdrawal symptom for vigour

Before any probabilities can be computed one needs to determine the distribution in question. The scores for vigour are quantitative records on a scale ranging at least from 1 to 30 units. The range of the scale should be wide enough to approximate the distribution by a continuous probability distribution, and the obvious choice is a normal distribution. This assumption cannot be checked from the present data but the researchers had access to sufficient data to assess the normal distribution. Without that information our use of a normal distribution is inevitably somewhat pragmatic. The actual procedure of determining extreme values (corresponding to subjects with withdrawal symptoms) at a cut-off of the mean minus two times the standard deviation appears to be derived from the normal distribution (it would be less intuitive for a skewed distribution). It appears that the vigour scores have been computed as a sum of scores across relevant questions in the questionnaire. By the central limit theorem it is plausible that such cumulative scores would be approximately normally distributed.

If we assume the vigour scores to be approximately normally distributed in the population of college students, the probability that a randomly selected person's score (Y) is below the mean (μ) minus two times the standard deviation (σ) is calculated as follows:

$$\mathrm{P}(Y < \mu - 2\sigma) = \mathrm{P}\left(\frac{Y-\mu}{\sigma} < \frac{\mu - 2\sigma - \mu}{\sigma}\right) = \mathrm{P}(Z < -2) = 0.023,$$

where $Z \sim \mathrm{N}(0,1)$. In view of the uncertainty involved in the approximation by a normal distribution it seems fair to say that the probability is approximately 2.5%. Instead of carrying out the above calculation one could refer to the $68-95-99.7\%$ rule for the normal distribution.

6. Probability of withdrawal symptoms among subjects

For the presence or absence of a withdrawal symptom for vigour among 11 persons from the general population we assume a binomial setting. It seems fair to assume independence between subjects and equal probabilities to experience vigour below the cut-off. If Y denotes the count of subjects with a withdrawal symptom we therefore assume $Y \sim \mathrm{Bin}(11, p)$, where p is the probability for one person. Question 5 gave the approximate value $p \approx 0.025$. The desired probability is now

$$\mathrm{P}(Y \geq 5) = \mathrm{P}(Y = 5) + \mathrm{P}(Y = 6) + \ldots P(Y = 11) = 0.0000040 \quad (0.0000026 \text{ for } p = 0.023).$$

The probability is very small, and it is therefore very unlikely that the observed 5 out of 11 subjects with a withdrawal symptom in the caffeine period could have happened by chance alone. We conclude that the probability of a withdrawal symptom among the subjects was larger in the caffeine period than in the general population. This could perhaps be attributed to the subjects being non-representative of the population the cut-off for withdrawal symptoms was based on (i.e. college students) but it also seems more plausible that it is a result of the lack of caffeine. *Note*: the probability computed is the P-value for a statistical test of $\mathrm{H}_0 : p = 0.025$ against the alternative $\mathrm{H}_a : p > 0.025$. We would therefore reject H_0 in favour of H_a, and could say there was significant evidence of a larger proportion of participants with withdrawal symptoms than in the general population of college students.

7. Comparison of motor skill scores in caffeine and caffeine-free periods

Motor skill scores are quantitative on a scale ranging roughly from 200 to somewhat about 400. Our first idea is to analyze the data using the normal distribution. The statistical design is two paired samples. For the differences of scores between caffeine and caffeine-free periods, say $D_1, \ldots, D_{11}$, we assume a normal distribution $\mathrm{N}(\mu, \sigma)$. We want to test the null hypothesis $\mathrm{H}_0 : \mu = 0$ against the two-sided alternative $\mathrm{H}_a : \mu \neq 0$ specified in the question; it seems perhaps plausible that there is no obvious direction of the impact of caffeine withdrawal. The estimates and calculations are summarized in the table below.

	Estimation			Test		
Outcome	$\hat{\mu} = \bar{D}$	$\hat{\sigma} = s$	95% CI for μ	t	H_a	P-value
motor skill score	20.2	48.7	(-12.6, 52.9)	1.37	$\mu \neq 0$	0.20

The results show a higher score in the caffeine period; the difference is about 20 beats per minute. The t-test is however totally non-significant, so the data show no evidence of a real difference between caffeine-free and caffeine periods. The differences do not follow a normal distribution too well (right skewness (1.19), some possible outliers, and $P = 0.064$ on the A-D normality test) so some caution is needed with the estimates and confidence intervals. Despite these (minor) deviations from normality one would think the conclusion remains the same. *Note*: because non-parametric methods were not covered in the course prior to the home assignment, no additional analysis using non-parametric methods was expected.

8. Comparison of fatigue, vigour and depression scores in caffeine and no-caffeine periods

The statistical design is the same as for the motor skill scores analyzed above, and we use the same notation, statistical model and steps of the analysis. The question motivates one-sided alternative hypotheses for each of the outcomes, given in the table below together with the relevant calculations.

	Estimation			Test		
Outcome	$\hat{\mu} = \bar{D}$	$\hat{\sigma} = s$	95% CI for μ	t	H_a	P-value
fatigue score	-9.09	14.20	(-18.63, 0.45)	-2.12	$\mu < 0$	0.030
vigour score	9.73	12.97	(1.01, 18.44)	2.49	$\mu > 0$	0.016
depression score	-7.36	6.92	(-12.01, -2.72)	-3.53	$\mu < 0$	0.003

For all three outcomes, the data show a difference between the two periods in the same direction as anticipated. The differences are of the order of magnitude of 7-10 units, but without knowing anything about how the scales are constructed it is difficult to assess the biological significance. All P-values are significant at significance level 0.05, only weakly so for fatigue and vigour scores, but quite strongly for depression scores. The normality assumption seems quite ok with at most weak skewness, reasonably straight normal probability plots (not shown) and P-values for the Anderson-Darling normality test well above 0.05 (the lowest value is 0.254, for the depression scores). Therefore we will in our conclusion will reject the null hypotheses and claim evidence of the directional alternatives. The indication of all types of caffeine withdrawal symptoms are strong enough for us to claim that it is unlikely to have occurred by chance alone. So the study supports the presence of such symptoms, at least in the specifically selected study population.