Biostats VHM 801 Course
Fall 2023, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to Mid-term Exam, October 2023

The solution is more detailed than required for a 100% score, by including answers to all six subquestions and multiple points for discussion in **(f)**. The data are from Swain JF et al. (1990), Comparison of the effects of oat bran and low-fiber wheat on serum lipoprotein levels and blood pressure, *New England Journal of Medicine* **322**, 147–152.

Question 1

Part (a)

We let X denote the total cholesterol level for a person on baseline diet, and assume $X \sim \mathrm{N}(186, 31)$. Then the total cholesterol level is borderline high or high if it exceeds 200, and

$$P(X > 200) = P\left(\frac{X - 186}{31} > \frac{200 - 186}{31}\right) = P(Z > 0.452) \approx P(Z > 0.45) = 0.3264 \approx 0.33,$$

using Table B of PSLS. According to this calculation, about one third of the people have borderline high (or higher) total cholesterol levels. A similar calculation for the LDL cholesterol level $Y \sim \mathrm{N}(115, 23)$ yields $P(Y > 130) = P(Z > 0.652) \approx 0.26$, where we computed the z-score as: $z = (130-115)/23 = 0.652$. The estimated proportion of people with borderline high (or higher) LDL levels is a bit lower than for the total cholesterol levels. We could have concluded that already after seeing that the z-score was higher than for total cholesterol (0.452), without computing the actual probability.

Part (b)

Counting the number of subjects (X) out of 20 with a high total cholesterol level, would correspond to a binomial setting. We therefore assume $X \sim B(20, 0.04)$. The following calculations are all based on this binomial distribution, which is not directly included in the Stevens (S) textbook table.

- expected number of persons: $\mathrm{E}X = np = 20 \cdot 0.04 = 0.8$,
- $P(X = 0) = (1 - 0.04)^{20} = 0.442$,
- $P(X > 1) = 1 - P(X \leq 1) = 1 - P(X = 0) - P(X = 1) = 1 - 0.442 - 0.368 = 0.190$, using the calculation

$$P(X = 1) = \binom{20}{1} 0.04^1 \, (1 - 0.04)^{19} = 20 \cdot 0.04 \cdot (1 - 0.04)^{19} = 0.368.$$

  Without this calculation, we could get lower and upper bounds for $P(X = 1)$ from the entries for the $B(20, 0.01)$ and $B(20, 0.05)$ distributions in the S table, leading to: $0.017 \leq P(X > 1) \leq 0.264$.

Part (c)

Denote by $X_1, \ldots, X_{20}$ the total cholesterol levels for the 20 subjects at their baseline levels. We assume this to be a simple random sample (or the $X_1, \ldots, X_{20}$ to be i.i.d.) from a normal distribution $\mathrm{N}(\mu, \sigma)$. We don't have any descriptive information about the values based on which we can evaluate the assumption. For a 95% confidence interval we use $t^* = 2.093$ from a t-distribution with 19 df,

$$95\%\ \mathrm{CI}: \ \bar{X} \pm t^* s/\sqrt{n} = 186 \pm 2.093 \cdot 31/\sqrt{20} = 186 \pm 14.5 = (171.5, 200.5).$$

This confidence interval is given directly in the Minitab listing. Note that we use the t-distribution because the standard deviation is not known. As 200 is (barely) included in the confidence interval, there is no statistical evidence against $\mathrm{H}_0 : \mu = 200$ at the 5% significance level, even if the P-value for testing H_0 against $\mathrm{H}_a : \mu \neq 200$ must be close to 0.05. We can therefore say that the data indicate the (population) mean cholesterol level to be lower than 200, but the data are not strong enough to rule out a level of 200 (or higher) at the 5% significance level.

Part (d)

As the cholesterol levels at baseline and with high fiber supplement were obtained on the same subjects, we have two paired samples. Again assuming normal distributions, the natural statistical model is therefore that $(X_1, Y_1), \ldots, (X_{20}, Y_{20})$ are independent pairs with $X_i \sim \mathrm{N}(\mu_b, \sigma_b)$ (the baseline values from **(c)**) and with $Y_i \sim \mathrm{N}(\mu_h, \sigma_h)$ (the high fiber values). The mean change in cholesterol levels is $\mu_b - \mu_h$ which we estimate by the difference between the sample means,

$$\hat{\mu}_b - \hat{\mu}_h = \bar{X} - \bar{Y} = 186 - 172 = 14.$$

In order to carry out statistical inference for $\mu_b - \mu_h$, we need to form the differences $D_i = X_i - Y_i$ and in particular compute their estimated standard deviation (s_D). If we had a value of s_D, we could do a one-sample t-test for the hypothesis $\mathrm{H}_0 : \mu_D = 0$ with 19 degrees of freedom using the formula $t = (\bar{X} - \bar{Y})/(s_D/\sqrt{20})$. If we had the data values, we could form the differences and use statistical software to test the hypothesis. In both cases we would need to think about the alternative hypothesis; arguments could be made in favour of both one-sided and two-sided alternatives.

Part (e)

The number of subjects (X) out of 20 that guessed correctly the identity of the two fiber diets would correspond to a binomial setting. We therefore assume $X \sim B(20, p)$. Our observation was 17 out of 20 correct guesses ($X_{\mathrm{obs}} = 17$). This to have happened by chance alone corresponds to $\mathrm{H}_0 : p = 0.5$. We want to test this hypothesis against a two-sided alternative $\mathrm{H}_a : p \neq 0.5$. The choice between one-sided and two-sided alternatives is not obvious here, but it does seem possible that people could be systematically wrong in their guesses (and hence have $p < 0.5$). We carry out this test as an exact test in the binomial distribution, using the S table for $B(20, 0.5)$,

$$P = 2 \cdot P(X \geq 17) = 2 \cdot (0.001 + 0 + 0 + 0) = 0.002.$$

The table has no entries for 18, 19 and 20, meaning that these probabilities are less than 0.0005, and they were set to 0 above. The P-value is very small, so we reject the null hypothesis. It could not have happened by chance alone that 17 out of 20 guessed the diets correctly. (*Note:* The test could also have been carried out by a classical, normal approximation z-test: $z = 3.13$ and $P = 2 \cdot 0.0009 = 0.0018 \approx 0.002$ from Table B.)

Part (f)

Several critical points can be raised. First, the wording of the conclusion to say that "oat bran has little cholesterol-lowering effect" seems unfortunate when the study results did in fact show a reduction in both cholesterol levels between baseline and the high fiber diet. Even if the reduction had been non-significant (by the analysis discussed in **(d)**), that does not necessarily translate into "little effect".

Second, the study itself can also be criticized. The sample size is small, and the study subjects do not seem to represent any natural reference population, nor were they randomly selected. It could be speculated that the baseline diets of employees at a hospital, in particular among dieticians, does not adequately represent the diets in the population. So the stated conclusion seems bold considering the power of the study (the low sample size) and the representativity of the study.