STATISTICAL INFERENCE (MS-C1620). 8.1.–17.4.2025. Aalto University. University Lecturer Pekka Pere.

Suggested solutions to the questions of exam 17.4.2024

1.

a) The lower and upper bounds of the classical 95 % confidence interval for the proportion of green cubes chosen are

$$\frac{y}{n} \pm 1.960 \sqrt{\frac{\frac{y}{n}\left(1-\frac{y}{n}\right)}{n}}.$$

Above 1.960 is the 0.975th quantile of the Standard normal distribution (`qnorm(0.975)`).) In the present application the bounds are

$$\frac{17}{20} - 1.960 \sqrt{\frac{\frac{17}{20}\left(1-\frac{17}{20}\right)}{20}} = 0.694.$$

and

$$\frac{17}{20} + 1.960 \sqrt{\frac{\frac{17}{20}\left(1-\frac{17}{20}\right)}{20}} = 1.006.$$

The bounds can be calculated with R as follows:

```
n <- 20
p <- 17/n
q <- 1-p
p-1.960*sqrt(p*q/n)
p+1.960*sqrt(p*q/n)
```

The bounds of the plus four confidence interval are

$$\frac{y+2}{n+4} \pm 1.960 \sqrt{\frac{\frac{y+2}{n+4}\left(1-\frac{y+2}{n+4}\right)}{n+4}}.$$

For the present application the bounds are

$$\frac{17+2}{20+4} - 1.960 \sqrt{\frac{\frac{17+2}{20+4}\left(1-\frac{17+2}{20+4}\right)}{20+4}} = 0.629.$$

and

$$\frac{17+2}{20+4} + 1.960 \sqrt{\frac{\frac{17+2}{20+4}\left(1-\frac{17+2}{20+4}\right)}{20+4}} = 0.954.$$

The bounds can be calculated with R as follows:

```
n <- 24
p <- 19/n
q <- 1-p
p-1.960*sqrt(p*q/n)
p+1.960*sqrt(p*q/n)
```

The plus four interval is a bit wider:

$$0.325 \approx 0.9541441 - 0.6291892 > 1.0064906 - 0.6935094 \approx 0.313.$$

The Wald interval is not meaningful as it overshoots the theoretical maximum $\pi = 1$. The plus four interval does not and is meaningful. The plus four interval is more meaningful also in the sense that it is more trustworthy in small samples.

b) The rule of three gives the 95% confidence interval:

$$\left[0, \frac{3}{301}\right] \approx [0, 0.01].$$

2.

a) Two evaluations have been made of a single person. This is a case of matched samples so a pairwise t-test is appropriate. It is typically a desired test because of its greater power compared to the corresponding t-test for independent samples.

The test statistic of the pairwise t-test is

$$\frac{\hat{\mu}_D - \mu_{D_0}}{s_D/\sqrt{n}}.$$

Here $\hat{\mu}_D = \sum_{j=1}^{n} D_j/n$, D_js are pairwise differences of the observations, $s_D^2 = \sum_{j=1}^{n}(D_j - \hat{\mu}_D)^2/(n-1)$, n is the sample size, and μ_{D_0} is μ according to the null hypothesis. In the present circumstance $\mu_{D0} = 0$, $\hat{\mu}_D = -4.97508$, $s_D = 221.5834$, and $n = 1\,244$. The test statistic takes value -0.7919049:

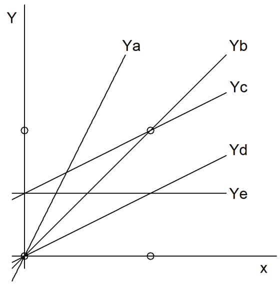
$$\frac{-4.97508}{221.5834/\sqrt{1244}} = -0.7919049.$$

b) The pairwise test statistic follows the t(n-1)-distribution (t-distribution with $n-1$ degrees of freedom) under the null hypothesis.

c) The quantiles of the t(n-1)-distribution and the Standard Normal distribution essentially agree for large n. Here $n = 1244$ is large so the quantiles of the Standard Normal distribution can be used for testing. The 0.9995th quantile of the Standard Normal distribution is 3.290 so the critical values of the test statistic are -3.290 and 3.290. The test statistic falls inside this range ($-0.7919049 < -3.290$) so the null hypothesis is not rejected at significance level 0.001.

d) The normality assumption is not necessary for testing here. The sample is large enough for the Central limit theorem to assure that the sample mean of the differences follows a Normal distribution.

3.



Correct answers: E, E, B, D, and A. The vertical distance between points is marked by $2a$ in the explanations below.

a) E: Line Ye describes best the average of observations Y.

b) E: Intuitively: The regressand has no tendency to increase as the regressor increases. More formally: The other lines yield larger squares of residuals which the method of least squares aims to minimise. Formally: Line Ye yields residual sum of squares (RSS) $a^2+a^2+a^2+a^2 = 4a^2$. All other lines result in a larger RSS. E.g. line Yb yields RSS $(2a)^2 + 0^2 + 0^2 + (2a)^2 = 8a^2 > 4a^2$.

c) B: $\mathsf{R}^2 = 1 - \mathsf{RSS}/\mathsf{SYY}$. Here $\mathsf{RSS} = \mathsf{SYY}$ so $\mathsf{R}^2 = 0$.

d) D: The regression line is Ya, Yb, or Yd. The observations on the left do not affect the RSS. Obviously, line Ya is associated with the largest squared residuals so the answer must be line Yb or line Yd. With line Ye the impact of the right-most observations to the RSS is $0^2 + (2a)^2 = 4a^2$. With line Yd the impact is $a^2 + a^2 = 2a^2 < 4a^2$. Line Yd minimises the RSS.

e) A: SYY equals $a^2+a^2+a^2+a^2 = 4a^2$. The RSS is $(2a)^2+0^2+a^2+a^2 = 6a^2$. $\mathsf{R}^2 = 1 - \mathsf{JNS}/\mathsf{KNS} = 1 - 6a^2/4a^2 = -1/2 < 0$!

The result in item e) is possible because there is no intercept in the model. A regression line through the origin may explain the data more poorly than the intercept alone or a horizontal regression line.

4.

a) The degrees of freedom are determined as follows: There are $n = 45 + 54 + 50 = 149$ observations, and $m = 3$ group means are estimated. The null distribution of the F test statistic is $\mathsf{F}(m-1, n-m)$. In the present context the degrees of freedom are $m - 1 = 3 - 1 = 2$ and $n - m = 149 - 3 = 146$.

b) The null hypothesis is that the group means equal in the corresponding populations (a population is e.g. students or persons in general who are given a menu with calorie information before the food item). The null hypothesis means that it does not matter whether the calories are stated before or after a food item or not at all.

c) p-value is the probability under the null hypothesis of obtaining a value of the test statistic which deviates as much as the observed test statistic or more from the theoretical null-value towards more extreme non-null values. In the present case, the null distribution of the test statistic is the $\mathsf{F}(2, 146)$ distribution. It takes nonnegative values only, and the extreme values are large values. In the present case the p-value is the probability of obtaining 3.60 or a larger value of the test statistic from the $\mathsf{F}(2, 146)$ distribution. Thus the probability of the right tail of this distribution, starting from 3.60, should be calculated. (R command `1-pf(3.60,2,146)` would do the trick.)

d) The p-value is 0.03 or less than 0.05. The null hypothesis is hence rejected at significance level 0.05. At least one of the three group means differs from the rest in the respective populations. It matters how, or if at all, calories are reported in the menus.

One is tempted to infer that students order food with less calories if the calories are printed to the left of (before) the food item in the menu. However, for one to be able to draw this conclusion a new test focusing on this statement should be carried out next.

The researchers carried out another experiment with students from Israel. The point of this experiment was that the menus of Israeli students are written in Hebrew which is read from right to left. A calorie warning for these students should hence be more effective if it is printed to the right of a food item!