Biostats VHM 801 Course
Fall 2019, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to Midterm Exam, October 2019

The solution is more detailed than required for a 100% score, by including more detail and general discussion in the answers than could be managed within the time constraints of the exam. The data and context of the question are taken from the EESEE (Encyclopedia of Electronic Statistical Examples and Exercises) story "Kicking a helium-filled football". In the marking, scores were adjusted to an expectation of less than five completed subquestions for a full mark.

Question 1

Subquestion a)

Both studies can be considered as experiments. The first study included two experimental settings: kicks with and against the wind; this increased the scope of the study (i.e., the settings, or population, the results could potentially be representative for). Characteristics of the player are not known; these would have been helpful to determine the scope of the results. In the second study, the use of a novice player and the restriction to a windless day limits the scope of the results. The results might not be representative for any broader population.

The sample size (number of replications within each experimental setting) was fairly low in the first study (4 kicks), and much larger in the second study (39 kicks). We have no information about the logistics of the experimental design in Study 1, but the trials in the second study can be thought of as blocks (each block consisting of one kick with each of the footballs). This seems appropriate for such a long series as 39 kicks, because the player may either improve or get tired over time.

There is limited information on whether or how randomization (of the order of the kicks with the two footballs) might have taken place. In the second study, and considering the trials as blocks, the two kicks within each trial should have been in random order (it appears they were just alternated, so not actually randomized).

In the first study, the player knew the identity of the two footballs; that is, the study was not blinded. Lack of blinding increases the risk of bias. In the second study, the player was blinded to the identity of the footballs, but possibly not to the two footballs being used in an alternating sequence.

In summary, the results of the two studies hardly can be said to contradict each other because they were done under different conditions (e.g., windless day vs. with and against the wind). Our confidence in the results of the second study is probably larger because of its better design and larger replication.

Subquestion b)

The question is how to interpret the intervals, e.g. "plus or minus 0.7 yards" for the air-filled ball. The most natural, perhaps "naive", interpretation is as the range within which most of the air-filled balls travelled; that would be a range within the distribution of actual distances. The statement could also be interpreted as being about the precision of the averages; in that case, we might think it to be a confidence interval. In both cases, the lack of indication of a confidence or probability for the range makes it hard to interpret precisely. Alternative and less formal interpretations could be as an indicator of spread (standard deviation) or as the precision of mean (standard error of the mean).

Looking at the data, it is clear that the interval is too narrow to be a (meaningful) range within the distribution of distances. For example, the IQR for the air-filled football is $29 - 23 = 6$ yards. Nor

is it a 95% confidence interval for the mean. In fact, the value 0.7 yards was most likely obtained by (incorrectly) rounding off the standard error of the mean, whereas the value 1 yard is indeed the correctly rounded standard error of the mean for the helium-filled footballs. As a rough approximation (ignoring the uncertainty from estimating the standard deviation, and ignoring any problems with assuming a normal distribution), an interval with a margin of error of one standard error will give a confidence level around 68% (the 68-95-99.7 rule for an approximate normal distribution of the mean).

One way of improving the statement is to let it reflect an approximate 95% confidence interval. Using $z^* = 1.96$ or $t^* = 2.042$ for a t-distribution with 30 degrees of freedom (Table C of PSLS does not include t^* for 38 df) gives a margin of error of approximately 1.5. The statement could be written: "the average distance travelled by the air-filled ball was 26.0 yards with a margin of error of 1.5 yards, 19 times out of 20". Other ways of improving the statement are possible as well, depending on the desired interpretation.

Subquestion c)
We let X denote the distance of a kick with the helium-filled ball, and assume $X \sim \mathrm{N}(26.385, 6.214)$. Then

$$P(X < 15) = P\left(\frac{X - 26.385}{6.214} < \frac{15 - 26.385}{6.214}\right) = P(Z < -1.832) \approx 0.0336 \approx 0.034,$$

using Table B of PSLS (the exact value is 0.03346). (It may be more precise to compute $P(X < 14.5)$, due to the round-off of the distances; that value equals 0.028.) The calculation tells us that it is not very unlikely to observe a single kick below 15 yards. Given that the dataset comprises 39 kicks, we would expect $1 - 2$ kicks below 15 yards (among 100 observations we would expect 3.4 kicks below 15 yards). The actual number of helium-filled football kicks below 15 yards is 4, so that seems a bit too many compared with what would be expected from a normal distribution. Still, the conclusion does not seem clear.

Some problems with the calculation above are (i) the reliance on the normal distribution and on the parameters estimated from data including any suspect observations (in particular, the standard deviation could be too large), and (ii) how to properly account for the number of observations. For problem (i) it would be natural to assess the normality of the distribution by a normal probability plot, possibly both with and without the suspect observations. If the distribution differs substantially from a normal distribution, other methods will be needed. A different approach is the determination of "suspected outliers" by the rules used for the boxplot. Here we would get the lower bound for suspected outliers as: $Q_1 - 1.5 \cdot IQR = 24 - 1.5 \cdot 6 = 15$, meaning that all observations below 15 would be labelled as "suspected outliers". The figure gives the impression that some of these observations are really quite far below the others, but from the data listing it seems that the biggest gap in the distribution is between 16 and values above (the closest being 22).

The problem (ii) of how to incorporate the number of observations also exists with the "suspected outlier" method, and needs to be dealt with in other ways; one possibility is to use procedures for multiple testing (to be discussed in Session 10), another is to use tests specifically designed to detect single or multiple outliers (these may however also be based on an assumed normal distribution).

Subquestion d)
The distances of the 39 kicks with the air- and helium-filled footballs should be considered as two paired samples, with the trials forming the pairs or blocks, as discussed above. If $X_1, \ldots, X_{39}$ and $Y_1, \ldots, Y_{39}$ denote the distances of the air- and helium-filled footballs, respectively, we let $D_i = Y_i - X_i$

for $i = 1, \ldots, 39$. The estimated mean difference is $\bar{Y} - \bar{X} = \bar{D} = 0.46$. The distribution of the differences looks roughly symmetrical with no strong outliers (despite our discussion from **c)**), so inference based on assuming a normal distribution $D_i \sim \mathrm{N}(\mu_d, \sigma_d)$ is probably acceptable. If we were to calculate a 90% confidence interval manually, we would need to use $t^* = 1.697$ for a t-distribution with 30 degrees of freedom (because again the value for 38 df is not available in Table C of PSLS), giving

$$90\%\text{ CI: } : \bar{D} \pm t^* s_D/\sqrt{39} = 0.46 \pm 1.697 \cdot 6.87/\sqrt{39} = 0.46 \pm 1.87 = (-1.41, 2.33).$$

The software listings give the exact confidence interval as $(-1.39, 2.32)$. It includes 0 and ranges well into both positive and negative values, and it seems any difference in mean distance between the two footballs is likely to be fairly small (certainly smaller than the results reported by Study 1). For formal statistical inference we would have to set up our statistical hypotheses, and because of the focus on helium-filled footballs travelling further the most natural choice would be:

$$\mathrm{H}_0 : \mu_d = 0 \quad \text{versus} \quad \mathrm{H}_a : \mu_d > 0.$$

With a two-sided alternative hypothesis, the fact that the 90% CI includes zero would tell us that $P > 0.10$. The P-value for a one-sided alternative hypothesis is half of that for the two-sided alternative (if the estimate is in the direction of the alternative, which is the case here), so we may conclude thet $P > 0.10/2 = 0.05$. Thus we have no (or not sufficient) evidence to say that the kicks with the helium-filled football are longer than those with the air-filled football.

Subquestion e)

When the focus shifts from the actual distances of the 39 pairs of kicks to whether the helium-filled football travels longer than the air-filled football in each pair, the data essentially reduce to binary indicators, or signs. The procedure therefore corresponds to a sign test. Among the 39 pairs, the distances are equal in two pairs. This probably does not mean that the actual distances were identical but after round-off they appear so. We will remove those observations and concentrate on the 37 pairs with a "winner". We let X denote the count of pairs where the helium-filled football travelled longest, and assume $X \sim B(37, p)$. The statistical analysis proceeds in the following steps:

- estimation: $\hat{p} = X/n = 20/37 = 0.54$,
- hypotheses: $\mathrm{H}_0 : p = 0.5$, $\mathrm{H}_a : p > 0.5$,
- test: the classical z-test meets its conditions for use, because when $p = 0.5$ the expected number of events is large (i.e., $= 37 \cdot 0.5 = 18.5 > 10$, and the same for non-events), so we calculate:

$$\begin{aligned} z &= (\hat{p} - p_0)/\sqrt{p_0 \cdot (1 - p_0)/n} = (0.54 - 0.5)/\sqrt{0.5 \cdot 0.5/37} = 0.49, \\ P &= \mathrm{P}(Z > 0.49) = P(Z < -0.49) = 0.312. \end{aligned}$$

The high P-value leads us to a similar conclusion as in **d)**: there is no evidence to indicate that the helium-filled football has a higher probability than 0.5 of travelling longer than the air-filled football. Thus we have no support from the data to say that the helium-filled football will most likely travel further than the air-filled football.

Note: It is possible to use the normal approximation to the binomial distribution to get an approximate P-value for the exact test in the binomial distribution, but this refinement is not quite necessary in an exam situation (the value obtained from the normal approximation is 0.37, whereas the exact value in $B(37, 0.5)$ is 0.371).