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

Solution to Midterm Exam, October 2016

The solution is more detailed than required for a 100% score, by including several options in different questions and including answers to all five parts of **d)**. Also the discussion throughout is more verbose than could be reasonably be managed within the time constraints of the exam. The data are from Gumpertz et al. (1997), *J. Agric. Biol. Environ. Statist.* **2**, 131−156.

Question 1

Subquestion a)

The study was a field trial, but because no treatments beyond standard practices were imposed on the two fields it would most naturally be described as observational. The recorded variables could be described as follows

- *mois*: quantitative and continuous,
- *dis*: binary (and categorical),
- *col*: a count, therefore quantitative and discrete.

Subquestion b)

For each of the fields, the values of soil moisture constitute a random sample. It seems natural to assume the observations to be i.i.d. (independent, with the same distribution). The descriptive statistics show that the distribution is hardly a normal distribution; both samples have moderate right-skewness, and the normality tests are both significant at $P < 0.05$. The box-plots do not show any suspected outliers, although the data from field B include a couple of large values. Overall, this leaves us with two options for analysis, both of which are reasonable when justified appropriately.

Option 1: analysis by normal distribution procedures. Although the distribution is not normal (as discussed), this choice can be justified by the robustness of t-procedures based on the normal distribution (with $n > 15$, some skewness and outliers can be tolerated). The inference will somewhat approximate because we don't know how much the results will be affected by the non-normality. The parameter of interest is the mean (μ).
Option 2: analysis by non-parametric procedures, to avoid any normal distribution assumptions. Because the major "problem" with the distribution is a right-skewness, we should not use procedures that assume symmetry of the distribution (i.e., Wilcoxon's signed rank test). The parameter of interest is the median (m).

The question of interest is whether the distribution systematically exceeds 10(%); this we should interpret as either the mean or median exceeding 10. Therefore the natural statistical test procedure is to test $\mathrm{H}_0 : \mu = 10$ (or $\mathrm{H}_0 : m = 10$) versus $\mathrm{H}_a : \mu > 10$ ($\mathrm{H}_a : m > 10$). The confidence intervals for μ and m in Minitab's (Graphical) Summary Report can only be used directly to test against two-sided alternatives. We could manually compute a t-test for the mean ($t = 2.49,\ \mathrm{df} = 19,\ P = 0.011$) or a sign test for the median ($X = 11,\ P = 0.41$), or we can rely on assessments from the confidence intervals. A summary of results is given in the table on the next page.

From the 95% CI for the median for field A, we cannot assess the P-value for the test against the one-sided H_a; we only know that the P-value against a two-sided H_a is > 0.05. For field B, the tests against the one-sided H_a all have P-values much greater than 0.5 because the estimate is actually lower than 10. In conclusion, there is evidence in favour of $\mu > 10$ in field A, but there is no evidence for $m > 10$, and in field B the data directly contradict the idea that the mean or median would be > 10.

Parameter	Statistic	Field A	Field B
mean	estimate	11.15	8.27
	95% CI	(10.18,12.11)	(7.31,9.23)
	P for one-sided H_a	< 0.025 (0.011)	> 0.975
median	estimate	10.23	7.70
	95% CI	(9.70,11.92)	(6.95,8.70)
	P for one-sided Ha	? (0.41)	> 0.975

Subquestion c)

The samples from the two fields should be considered as independent samples. The statistical analysis divides across the same lines as in **b)**, depending on whether normality is assumed or not. If yes, an (approximate) two-sample t-test is appropriate. If not, the Mann-Whitney-Wilcoxon test can be used, and because the two distributions seem pretty similar in shape it seems reasonable to assume them to have the same shape, so that the inference can be interpreted in terms of the medians. All results are given in the Minitab listings, and are summarized in the table below. The alternative hypothesis should by default be two-sided.

Parameter	Estimated difference	95% CI	Test statistic	P-value
mean	$\hat{\mu}_A - \hat{\mu}_B = 2.87$	$(1.559, 4.189)$	$t = 4.43$	< 0.0005
median	$\hat{m}_A - \hat{m}_B = 2.825$	(1,740, 3.920)	$W = 555$	0.0001

From both analyses, we conclude that field A has significantly higher soil moisture levels than field B, expressed as either higher mean or higher median soil level.

Subquestion d.i)

If we (in one of the fields) denote by X the number among all the quadrats with disease, we would naturally assume a binomial distribution: $X \sim \text{Bin}(400, p)$. The main difficulty with this assumption is the assumed independence (see the published paper for discussion), but we have no other choice here. For the confidence intervals we can use the classical (normal approximation) method because both the numbers of diseased and non-diseased quadrats easily exceed 15.

Field	Sample proportion	95% confidence interval
A	$\hat{p} = 54/400 = 0.135$	$0.135 \pm 1.96\sqrt{.135 * .865/400} = 0.135 \pm 0.033 = (0.102, 0.168)$
B	$\hat{p} = 61/400 = 0.1525$	$0.1525 \pm 1.96\sqrt{.1525 * .8475/400} = 0.1525 \pm 0.035 = (0.117, 0.188)$

It is seen that the two confidence intervals clearly overlap, whereby each estimate is even inside the other interval. This would lead one to conclude that there does not seem to be any major difference in disease proportion between the two fields. We can carry out a statistical test to confirm this impression, but that was not considered as part of the exam.

Subquestion d.ii)

Counting the quadrats (from which the 20 soil samples were taken) with disease would correspond to a binomial setting. We therefore assume $X \sim \text{Bin}(20, p)$ with $p = 0.12$. This binomial distribution is not included in our statistical table, so we need a manual calculation,

$$P(X = 0) = (1 - p)^{20} = (1 - 0.12)^{20} = 0.88^{20} = 0.078.$$

So it is not very likely, but not impossible, that all quadrats would be disease-free. The expected number of quadrats with disease equals: $np = 20 \cdot 0.12 = 2.4$. If the proportion of diseased quadrats was larger than 0.12, the expected number of quadrats with disease will be larger, and the probability of all quadrats being disease-free will be lower.

Subquestion d.iii)

Let now $X \sim \text{Bin}(5, p)$ denote the count of colonized leaves in a quadrat. As described, a valid estimate for p is the proportion of colonized leaves among the $400 \cdot 5 = 2000$ leaves sampled in the field. We compute this as follows,

$$\hat{p} = (87 \cdot 1 + 31 \cdot 2 + 21 \cdot 3 + 16 \cdot 4 + 18 \cdot 5)/(5 \cdot 400) = 0.183.$$

We could also compute (and interpret) this value as the mean count of X across the 400 quadrats, divided by 5. We next compare the observed proportions of the counts $0-5$ among the 400 quadrats (e.g. $227/400 = 0.568$ for a count of 0) with those expected from $\text{Bin}(5, p)$, with either $p = \hat{p}$ or $p = 0.20$, a rounded-off value of $\hat{p}$ allowing us to use a binomial distribution table (Table 1 in Stevens).

count	0	1	2	3	4	5
observed prop.	.568	.218	.078	.053	.040	.045
$\text{Bin}(5, 0.183)$	.364	.408	.183	.049	.005	.000
$\text{Bin}(5, 0.20)$	.328	.410	.205	.051	.006	.000

The agreement between the observed proportions and the binomial probabilities is poor. The observed proportions at zero and for large counts are way too large, and they are conversely much too low for counts 1 and 2. We conclude that the binomial distribution does not fit well to these data. In fact the distribution is more dispersed than a binomial distribution; this phenomenon is called *overdispersion* (a topic well beyond the scope of VHM 801).

Subquestion d.iv)

The observed proportions for field B were already computed in **d.iii)**, but here we need to combine the counts $3-5$, and compute the same quantities for field A also (with all decimals included):

observed prop.	0	1	2	3+
field A	.7400	.0725	.0450	.1425
field B	.5675	.2175	.0775	.1375

If we denote by X_A and X_B the counts of colonized leaves in samples from fields A and B, we have

$$\begin{aligned} P(X_B > X_A) &= P(X_A = 0, X_B > 0) + P(X_A = 1, X_B > 1) + P(X_A = 2, X_B > 2) \\ &= 0.74 \cdot (1 - 0.5675) + 0.0725 \cdot (0.0775 + 0.1375) + 0.0450 \cdot 0.1375 = 0.342 \end{aligned}$$

Note that the calculation assumes X_A and X_B to be independent. The chance that a randomly sampled quadrat from field B has more colonized leaves than a quadrat from field A is about $1/3$.

Subquestion d.v)

The population the quadrats within a field are representative for, is samples taken from the same field under similar circumstances, for example in areas of the field not covered by the lattice or at different times. One could also think of other fields that are entirely similar to the actual fields but that is perhaps mostly hypothetical because it would be very difficult to find other fields that are similar enough to the ones sampled.

If different treatments were applied to entire fields A and B, the experimental unit for the treatment would be the field whereas the measurement unit would be the quadrat. If there was only one field per experimental unit, it would be impossible to establish statistical significance between treatments because replication is needed for statistical inference. The multiple quadrats within a field do not capture the between-field variation and do therefore not represent true replicates. Thus the answer to the question is no.