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

Mid-term exam, 28 October 2016

All aids are allowed, except a computer-like device (including tablets and smartphones) and personal assistance. The exam consists of one question with four subquestions (labeled by letters **a)-d)** with weights as indicated) that should all be answered to achieve the maximal 15 points. The subquestions can be answered independently of each other. The mid-term exam accounts for 15% of the course mark; however, every student may choose to waive the result of the mid-term exam. The duration of the mid-term exam is 1 hour.

Generally, **statistical models and methods should be specified**. If you realize that you need more information than is provided to carry out the analysis, specify what information you need, how you would obtain it using Minitab or other statistical software, and how you would use it.

Question 1 (*15 points*)

The pathogen *Phytophthora capsici* causes bell peppers to wilt and die. Because bell peppers are an important commercial crop, this disease has been the subject of much agricultural research. It is thought that the pathogen may spread with surface water. We consider data from a study carried out in the 1990s at two naturally infested bell pepper fields, labeled here simply as A and B. The fields were cultivated and prepared by the growers, using standard cultural practices. Three variables were recorded from the fields: soil moisture (or soil water content, expressed as a percentage; *mois*), presence of the pathogen indicated by lesions on any of the plants in a plot (no/yes; *dis*), and the number of leaves (out of 5) that showed a colonization by the pathogen in an assay performed on leaf material (*col*).

Records of soil moisture were obtained at 20 randomly selected locations in each of the two fields. The Minitab listings show the values obtained together with some descriptive statistics and selected statistical analyses. More information about the data for *dis* and *col* is given before subquestion **d)**.

a) (*2 points*) Describe the study type (i.e., experimental versus observational), and characterize the variable type for each of the 3 variables measured (in each field).

b) (*4 points*) A question of interest was whether the soil moisture (*mois*) in the fields systematically exceeded 10%. Carry out a statistical analysis to assess and draw conclusions about this question, for both fields A and B. Include in your answer the relevant information about the distributions of *mois* to justify your choice of analytical approach; a full descriptive analysis is not required. Your analysis should, for each field, include an estimate representing the center of the distribution, and an assessment of the question based on statistical inference.

c) (*3 points*) Perform also a statistical analysis to estimate and compare the soil moisture levels in the two fields. Include also here a description and justification of your chosen approach, and draw conclusions.

The observations of disease presence and leaf colonization were taken on quadrats (or plots) into which the fields were divided. Each field was a square lattice set up with 20 rows and 20 columns, and thus a total of 400 quadrats with 2 to 3 bell pepper plants per quadrat. In both fields, the quadrats could be labeled as $(1,1),(1,2),\ldots,(1,20),(2,1),(2,2),\ldots,(20,20)$, where the first number corresponds to rows and the second number to columns. The fields each had their own lattice layout, as described. The numbers of quadrats in fields A and B with the different observed values of *dis* and *col* are shown in the tables below.

Variable	no	yes
dis_A	346	54
dis_B	339	61

Variable	0	1	2	3	4	5
col_A	296	29	18	21	14	22
col_B	227	87	31	21	16	18

d) (*6 points*) For this question, you should answer **three of the five parts** *i)-v)* below, each with the same score. It is allowed (but not recommended) to answer additional parts, in which case your score for **d)** will be for the best 3 parts among those answered.

i) (*2 points*) Compute 95% confidence intervals for the probability of a quadrat in field A to contain diseased plants, and carry out a similar calculation for field B. Interpret your results, and use them for an informal comparison between the disease occurrence in the two fields. Note that a formal comparison based on statistical inference between the disease levels in the two fields is not requested.

ii) (*2 points*) There could be a concern that the 20 soil moisture samples taken from quadrats in each of the fields would not include any quadrats where disease was present. If the proportion of diseased quadrats was assumed to be at least 12%, do you think the researchers should be concerned that all 20 quadrats sampled would be disease-free? Make suitable assumptions, and compute the probability for this to happen, as well as the expected number of disease-free quadrats among those sampled.

iii) (*2 points*) Because the number of colonized leaves is a count out of 5 leaves, it seems natural to assume a binomial distribution. Considering field B only, estimate the probability p of a leaf being colonized as the proportion of leaves in the data that showed colonization. Use this estimated probability to assess whether the binomial distribution seems to give a good description of the data, by comparing (some of) the observed proportions for leaf numbers $0-5$ with the corresponding expected proportions from the binomial distribution. (*Hint:* You are allowed to round off the estimated p to avoid calculating probabilities in the binomial distribution by hand.)

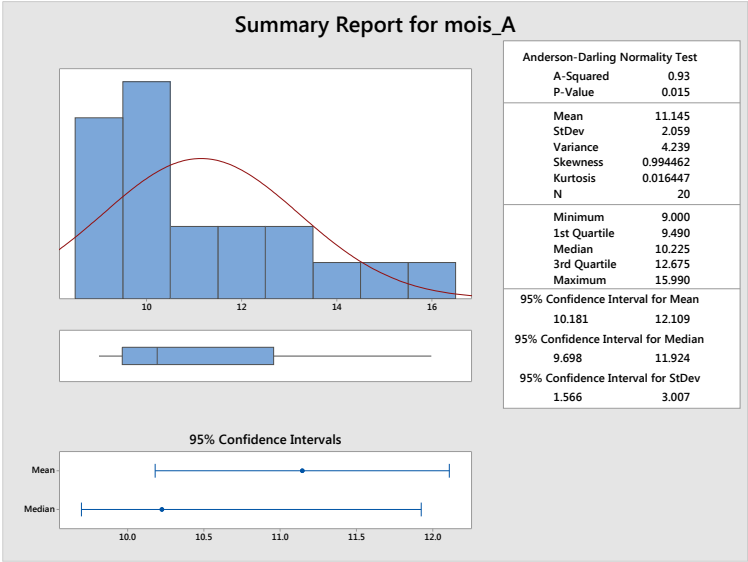
iv) (*2 points*) Without using any specific distribution, compute the sample proportions of quadrats with 0, 1, 2, and 3 or more (i.e., 3, 4, or 5) leaves colonized, in each of the fields A and B. Use these proportions to calculate the probability that a randomly selected quadrat from field B has a higher number (without distinguishing between 3, 4 and 5) of colonized leaves than a randomly selected quadrat from field A.

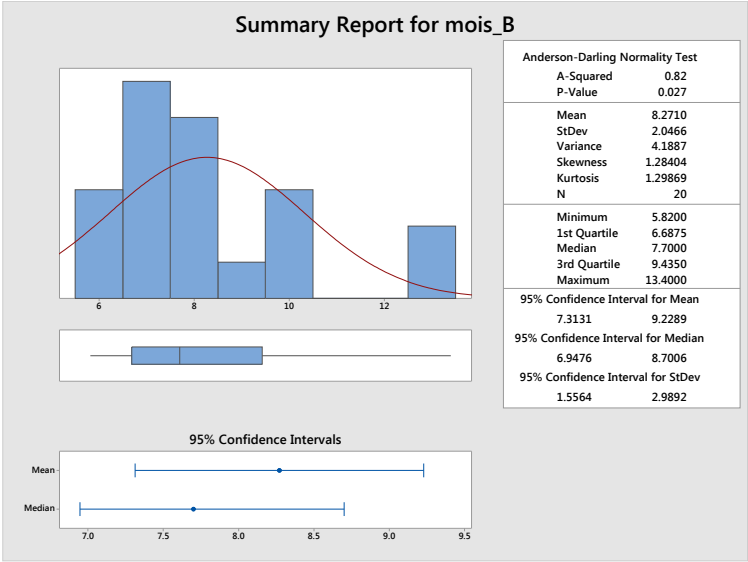
v) (*2 points*) Describe the population(s) the statistical inference in the previous questions for the data from the two fields could be representative for. Additionally, if different treatments against the pathogen had been employed on (entire) fields A and B, could the disease data from the quadrats have been used to determine whether a significant difference between the two treatments existed? Explain your reasoning.

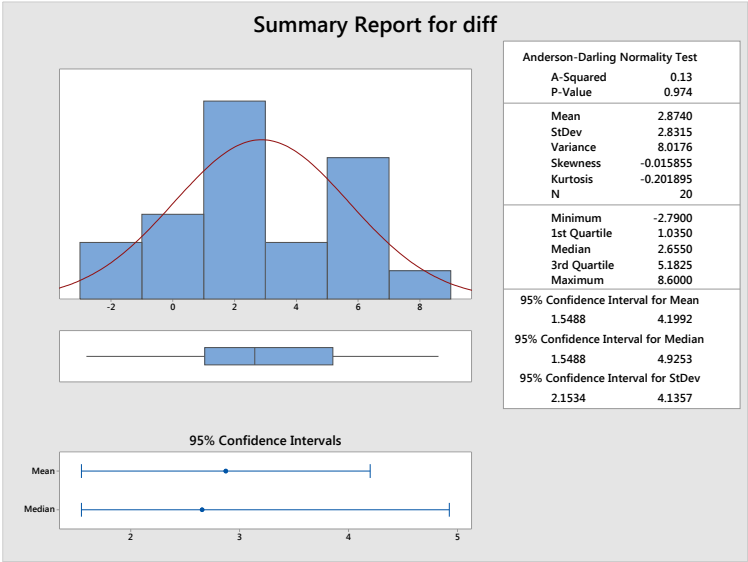
Minitab (version 17) listings for all questions

```
MTB > GSummary  'mois_A' 'mois_B'.
Summary Report for mois_A
Summary Report for mois_B

MTB > Let 'diff' = 'mois_A'-'mois_B'
MTB > Print 'mois_A' 'mois_B' 'diff'.
Data Display

Row  mois_A   mois_B    diff
  1    9.67    7.93   1.740
  2   12.03    6.83   5.200
  3   14.58    8.41   6.170
  4   11.31    8.79   2.520
  5    9.00    7.47   1.530
  6   15.99    7.39   8.600
  7    9.79    9.65   0.140
  8    9.35    7.74   1.610
  9    9.25    8.38   0.870
 10   10.30    7.33   2.970
 11    9.89    7.66   2.230
 12    9.43    6.64   2.790
 13   11.35    5.82   5.530
 14    9.07   10.48  -1.410
 15   13.37   13.40  -0.030
 16   10.15    6.10   4.050
 17   11.58    6.45   5.130
 18   12.89    6.52   6.370
 19   14.07    9.81   4.260
 20    9.83   12.62  -2.790

MTB > GSummary  'diff'.
Summary Report for diff
```

(continues on the next page)







```
MTB > WTest 0.0 'diff';
SUBC>   Alternative 0.
Wilcoxon Signed Rank Test: diff

Test of median = 0.000000 versus median not = 0.000000

          N for   Wilcoxon         Estimated
       N   Test  Statistic      P     Median
diff  20     20      195.0  0.001      2.888


MTB > TwoSample 'mois_A' 'mois_B';
SUBC>   Confidence 95.0;
SUBC>   Test 0.0;
SUBC>   Alternative 0.
Two-Sample T-Test and CI: mois_A, mois_B

Two-sample T for mois_A vs mois_B

         N   Mean  StDev  SE Mean
mois_A  20  11.15   2.06     0.46
mois_B  20   8.27   2.05     0.46

Difference = mu (mois_A) - mu (mois_B)
Estimate for difference:  2.874
95% CI for difference:  (1.559, 4.189)
T-Test of difference = 0 (vs not =): T-Value = 4.43  P-Value = 0.000  DF = 37


MTB > Mann-Whitney 95.0 'mois_A' 'mois_B';
SUBC>   Alternative 0.
Mann-Whitney Test and CI: mois_A, mois_B

         N  Median
mois_A  20  10.225
mois_B  20   7.700

Point estimate for eta1 - eta2 is 2.825
95.0 Percent CI for eta1 - eta2 is (1.740,3.920)
W = 555.0
Test of eta1 = eta2 vs eta1 not = eta2 is significant at 0.0001
```