Applied Statistics

MS Comprehensive Examination

April 21, 2018

Instructions:

Please answer all questions.

Record answers on the pages provided.

Points are as noted.

You may use books, notes, and a calculator for this exam.

You have three hours.

1. (20 points) We wish to study and compare two dichotomous populations, A and B. Assume that each population is large. For simplicity let's regard the characteristic under study as the answer to a Yes/No question. Denote the proportion saying "Yes" in population A by p_A and similarly for population B. This question explores a few sample size issues for this study. In this question, please derive your sample size answers from basic principles. Do NOT just plug the numbers given into a formula.

   a. Say that we need to know the sample size required to estimate p_A to within a margin of error M = .05 with 90% confidence. What sample size is required?

   b. Say that preliminary work has given us guesses for these parameters of p_{A0} =.70 and p_{B0} =.60. Using this information, how many fewer observations would be required to complete the task in part a?

   c. Using the preliminary information, if we choose equal sample sizes for each group, what is the sample size required for each group to have a margin of error for estimating the difference $p_A - p_B$ with 95% confidence equal to M = 0.05?

2. (20 points) Do twins have the same IQ between older and younger (one was born first)? To study this issue, a random sample of 7 pairs of twins was taken and their full scale IQ were recorded. Below, the results are summarized in a table:

Older twin	96	89	102	104	129	98	91
Younger twin	89	87	103	96	125	101	96

   a. Use Wilcoxon signed rank test (see two different tables at the end) at level $\alpha = 0.10$ to decide whether there is a significant difference between twins IQ. Find the exact p-value and use it to make your decision.

   b. Graph the pertinent data so that you can make an argument as to whether or not it is reasonable to answer this question using parametric methods and the appropriate t-test.

   c. No matter what your answer to part b, do the appropriate t-test (see t-table at the end). Again use $\alpha = 0.10$. You should i) find an approximate p-value OR ii) find the critical value to make your decision.

3. (20 points) Following are some statistics from data from a normal population. In this problem, we will use this data to test H_0: $\sigma^2 = 16$ versus Ha: $\sigma^2 > 16$ with level of significance $\alpha = .05$.

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Data	30	6.227	5.167	-3.002	3.256	5.059	8.695	22.666

Do this two ways:

a. the usual way using the exact chi-square distribution. Give your test statistic, the critical value, your decision, and bounds on the P-value provided by the attached chi-square table.

b. using the appropriate normal approximation to the chi-square (see standard normal table at the end). Give your test statistic, the critical value, your decision, and the P-value provided by the attached normal table. To use the normal approximation, recall these facts regarding the chi-square distribution:

i) If W is chi-square with degrees of freedom ν, then $E(W) = \nu$ and $Var(W) = 2\nu$ and
ii) If ν is large, then W is approximately normal.

c. Comment on the similarities and differences between your answers in parts a and b. In particular, discuss whether or not the normal approximation is justified and the impact of this on your answer to part b.

4. (40 points) In 1965, data on the connection between radioactive waste exposure and cancer mortality was published. The data was collected from 9 counties that were located near an Atomic Energy Commission facility in Hanford, Washington.

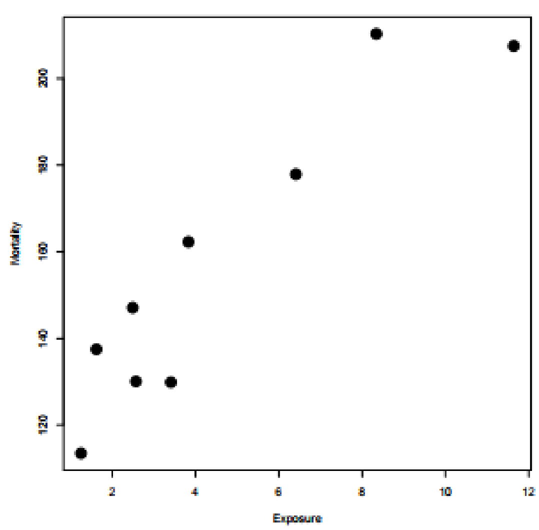
The data given the index of exposure and the cancer mortality rate during 1959-1964 for the nine counties affected. Higher index of exposure values represent higher levels of contamination.

Variable Description:		
	County	Name of county
	Exposure	Index of exposure
	Mortality	Cancer mortality per 100,000 man-years*

The data is as follows:

	County	Exposure	Mortality
1	Umatilla	2.49	147.1
2	Morrow	2.57	130.1
3	Gilliam	3.41	129.9
4	Sherman	1.25	113.5
5	Wasco	1.62	137.5
6	HoodRiver	3.83	162.3
7	Portland	11.64	207.5
8	Columbia	6.41	177.9
9	Clatsop	8.34	210.3

The scatterplot:



* This is a measure of cancer rate per 100,000 people for a certain amount of time

Here is the numerical summary of Exposure and Mortality:

```
> sd(Mortality)
[1] 34.79135
> mean(Mortality)
[1] 157.3444
> sd(Exposure)
[1] 3.491192
> mean(Exposure)
[1] 4.617778
```

Output from fitting the simple linear regression for predicting Mortality from Exposure is shown below:

```
> lm.out=lm(Mortality~Exposure)
```

```
> summary(lm.out)

Call:
lm(formula = Mortality ~ Exposure)

Residuals:
    Min      1Q  Median      3Q     Max
-16.295 -12.755   4.011   9.398  18.594

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  114.716      8.046  14.258 1.98e-06 ***
Exposure       9.231      1.419   6.507 0.000332 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 14.01 on 7 degrees of freedom
Multiple R-Squared: 0.8581,Adjusted R-squared: 0.8378
F-statistic: 42.34 on 1 and 7 DF,  p-value: 0.0003321
```

```
> anova(lm.out)
Analysis of Variance Table

Response: Mortality
          Df Sum Sq Mean Sq F value    Pr(>F)
Exposure   1 8309.6  8309.6         0.0003321 ***
Residuals  7 1373.9   196.3
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

a. (3 points) What is the expected mortality rate for a county with an exposure index of 3?
b. (3 points) For part (a), R reports the following two intervals: (130.1, 154.7) and (107.1, 177.7). Which one is the 95% confidence interval for the mean, and which is the 95% prediction interval for a new observation? How could you tell?
c. (3 points) Interpret the estimated slope of the fitted model.
d. (3 points) What is the correlation between Mortality and Exposure?
e. (3 points) Is there a significant linear relationship between Mortality and Exposure? Provide a null hypothesis, a test statistic, p-value, and conclusion.

f. (3 points) What is the estimated variance of the observations at a conditional mean?
g. (4 points) What is the Total Sums of Squares for this data?
h. (4 points) What is the F-value in the ANOVA output?
i. (4 points)What is the relationship between the t-value of Exposure and F-value in ANOVA? Why is that?
j. (10 points) Derive the formula for intercept and slope using least square technique.

Chisq table:

v	α					
	0.100	0.050	0.025	0.010	0.005	0.001
1	2.7055	3.8415	5.0239	6.6349	7.8794	10.8276
2	4.6052	5.9915	7.3778	9.2103	10.5966	13.8155
3	6.2514	7.8147	9.3484	11.3449	12.8382	16.2662
4	7.7794	9.4877	11.1433	13.2767	14.8603	18.4668
5	9.2364	11.0705	12.8325	15.0863	16.7496	20.5150
6	10.6446	12.5916	14.4494	16.8119	18.5476	22.4577
7	12.0170	14.0671	16.0128	18.4753	20.2777	24.3219
8	13.3616	15.5073	17.5345	20.0902	21.9550	26.1245
9	14.6837	16.9190	19.0228	21.6660	23.5894	27.8772
10	15.9872	18.3070	20.4832	23.2093	25.1882	29.5883
11	17.2750	19.6751	21.9200	24.7250	26.7568	31.2641
12	18.5493	21.0261	23.3367	26.2170	28.2995	32.9095
13	19.8119	22.3620	24.7356	27.6882	29.8195	34.5282
14	21.0641	23.6848	26.1189	29.1412	31.3193	36.1233
15	22.3071	24.9958	27.4884	30.5779	32.8013	37.6973
16	23.5418	26.2962	28.8454	31.9999	34.2672	39.2524
17	24.7690	27.5871	30.1910	33.4087	35.7185	40.7902
18	25.9894	28.8693	31.5264	34.8053	37.1565	42.3124
19	27.2036	30.1435	32.8523	36.1909	38.5823	43.8202
20	28.4120	31.4104	34.1696	37.5662	39.9968	45.3147
21	29.6151	32.6706	35.4789	38.9322	41.4011	46.7970
22	30.8133	33.9244	36.7807	40.2894	42.7957	48.2679
23	32.0069	35.1725	38.0756	41.6384	44.1813	49.7282
24	33.1962	36.4150	39.3641	42.9798	45.5585	51.1786
25	34.3816	37.6525	40.6465	44.3141	46.9279	52.6197
26	35.5632	38.8851	41.9232	45.6417	48.2899	54.0520
27	36.7412	40.1133	43.1945	46.9629	49.6449	55.4760
28	37.9159	41.3371	44.4608	48.2782	50.9934	56.8923
29	39.0875	42.5570	45.7223	49.5879	52.3356	58.3012
30	40.2560	43.7730	46.9792	50.8922	53.6720	59.7031
31	41.4217	44.9853	48.2319	52.1914	55.0027	61.0983
63	77.7454	82.5287	86.8296	92.0100	95.6493	103.4424
127	147.8048	154.3015	160.0858	166.9874	171.7961	181.9930
255	284.3359	293.2478	301.1250	310.4574	316.9194	330.5197
511	552.3739	564.6961	575.5298	588.2978	597.0978	615.5149
1023	1081.3794	1098.5208	1113.5334	1131.1587	1143.2653	1168.4972