Spring 2017

STA 580: Applied Statistical / Biostatistical Methods

MAT 387: Introduction Applied Statistical / Biostatistical Methods

Final **Name:______________________**

For full credit, show all of your work and use appropriate notation. Do not simply write the final numerical answer. No credit for correct final answer without a valid argument. Show your work graphically in all relevant questions. Please be organized and neat.

Provide a short report for each question listed below. Be sure that your report includes all the appropriate numerical and graphical summaries of the data, as well as appropriate justification for any inferential procedures that you choose to use. In addition, carefully state the conclusions of your analysis. **Make sure to include all your SAS and/or R codes and output as an appendix to your reports**. All hypothesis testing problems should specify the null and alternative hypotheses and report the p-value of the data.

1. An educator believes that a new reading curriculum will help elementary school students improve some aspects of their reading ability. She arranges for a third-grade class of 21 students to take part in the new curriculum for an eight-week period. A control classroom of 23 third-graders follows the standard curriculum. At the end of the eight weeks, all students are given a Degree of Reading Power (DRP) test, which measures aspects of reading that the treatment is designed to improve "**DRPscores.txt**".

Test the hypothesis that the treatment group performed better than the control group on the test. State your conclusions.

Treat	24
Treat	56
Treat	43
Treat	59
Treat	58
Treat	52
Treat	71
Treat	62
Treat	43
Treat	54
Treat	49
Treat	57
Treat	61
Treat	33
Treat	44
Treat	46
Treat	67
Treat	43
Treat	49
Treat	57
Treat	53
Control	42
Control	46
Control	43
Control	10
Control	55
Control	17
Control	26
Control	60
Control	62
Control	53
Control	37
Control	42
Control	33
Control	37
Control	41

Control 42
Control 19
Control 55
Control 54
Control 28
Control 20
Control 48
Control 85

2. A company is rated as acceptable in quality control if more than 90% of units produced at its facilities are found to be defect-free, and it is rated as excellent in quality control if more than 95% are defect-free. Suppose that a random sample of 500 units is selected and tested for defects, and that 18 units are found to have defects.

a. Does this data show at the 5% level of significance that the company is acceptable?

b. Does it show that the company is excellent? Construct a 95% confidence interval for proportion of defect-free units.

c. What sample size should a reliability engineer use to estimate this proportion to within 2% with 95% confidence if it is assumed that the proportion of units that are defect-free is at least 90%?

3. A large corporation requires that its employees attend a 1-day sexual harassment seminar. The Director of Human Resources of this corporation would like to determine whether or not the information presented in this seminar is retained over a long period of time. To this end, a random sample of 40 employees is selected from recently hired employees who are scheduled to take this seminar. Each of the employees in this sample completes a test of knowledge concerning sexual harassment and related legal issues immediately after the seminar, and then takes a similar test 6 months later. The scores are contained in the file "harass.txt".

```
Employee     Test1 Test2
1      72    36
2      92    91
3      93    83
4      87    89
5      90    74
6      84    95
7      99    107
8      87    94
9      85    93
10     88    55
11     88    118
12     85    124
13     92    113
14     87    100
15     83    38
16     82    62
17     66    3
18     71    71
19     83    88
20     85    70
21     64    55
22     81    74
23     92    122
24     81    102
```

25	99	64
26	76	76
27	77	65
28	55	30
29	66	66
30	82	103
31	80	63
32	84	87
33	88	61
34	75	51
35	70	79
36	87	94
37	84	62
38	94	109
39	72	64
40	95	52

Does that data indicate at the 5% level of significance that the mean score has changed after 6 months? Construct a 95% confidence interval for the difference between the mean scores.

4. The goal for the mean time to resolve software problems by the software support group of a large corporation is 24 hours. Suppose that 60 software problem items are randomly selected from all such items over the past quarter, and the mean time to resolve the problems was 22.4 hours with a standard deviation of 9.6 hours.

a. Does this data show at the 10% level of significance that the mean resolution time is less than 24 hours?

b. Construct a 90% confidence interval for the mean resolution time.

c. What sample size would be required to estimate the mean time to within 0.5 hours with 90% confidence if it is assumed that the standard deviation will be no more than 10 hours?

5. A study of industries in North Texas compared the experience of entry-level managers in telecommunication companies with the experience of entry-level managers in software-services companies. Suppose that a random sample of size 20 entry-level managers was selected separately from each group of companies and the experience of each manager was obtained. The data summary are:

$$\textbf{Telecom}\ \ (\textbf{Sample 1}):\ \bar{x}_1 = 4.7\ \ ,\ s_1 = 1.75$$

$$\textbf{Software}\ (\textbf{Sample 2}):\ \bar{x}_2 = 6.0\ \ ,\ s_1 = 0.75$$

a. Does this data show that there is a difference at the 5% level of significance?

b. Construct a 90% confidence interval for the difference between the means.

6. A large corporation would like to determine if employee job satisfaction will improve if it includes profit sharing based on quality scores for its factory workers. To answer this question, a pilot program was begun at one of its factories. A random sample of 30 workers from this factory was selected and, separately, a random sample of 30 workers was selected from another of its factories that did not implement this program. Prior to the start of the program each worker in these samples was given a test of job satisfaction as part of their normal review process. This test was then administered to the same employees six months after the start of the new program. Use 5% level of significance for the following questions. The data are contained in the file "**Pilot.txt**".

Factory	**Before**	**After**
Pilot	55	60
Pilot	106	111
Pilot	64	58
Pilot	66	82
Pilot	62	68
Pilot	87	90
Pilot	71	79
Pilot	85	88
Pilot	105	99
Pilot	103	104
Pilot	56	62
Pilot	89	98
Pilot	50	54
Pilot	108	119
Pilot	67	75
Pilot	46	50
Pilot	102	107
Pilot	61	69
Pilot	90	89
Pilot	55	59
Pilot	82	85
Pilot	55	62
Pilot	95	98
Pilot	64	67
Pilot	96	96
Pilot	74	69
Pilot	90	91
Pilot	65	72
Pilot	34	40
Pilot	92	98
NonPilot	87	91

NonPilot	48	45
NonPilot	86	88
NonPilot	64	66
NonPilot	76	72
NonPilot	56	58
NonPilot	96	94
NonPilot	103	104
NonPilot	74	77
NonPilot	122	123
NonPilot	96	99
NonPilot	117	118
NonPilot	65	59
NonPilot	71	61
NonPilot	77	73
NonPilot	54	48
NonPilot	58	57
NonPilot	74	76
NonPilot	77	76
NonPilot	62	64
NonPilot	47	50
NonPilot	87	88
NonPilot	71	72
NonPilot	57	58
NonPilot	31	31
NonPilot	88	91
NonPilot	106	107
NonPilot	72	72
NonPilot	75	75
NonPilot	84	87

a. Is there a difference between the mean satisfaction scores of these two factories before the pilot program is started?

b. Let **SatisImprov** be defined as $\mathbf{SatisImprov} = \mathbf{After} - \mathbf{Before}$.

i. Is there a difference between the means of **SatisImprov** at these factories?

ii. Construct a 95% confidence interval for **SatisImprov** at the pilot factory.

7. To develop which muscles need to be subjected to conditioning program in order to improve one's performance on the flat serve used in tennis, the study "An Electromyographic -Cinematrographic Analysis of the Tennis Serve" was conducted by the Department of Health, Physical Education and Recreation at the Virginia Polytechnic Institute and State University in 1978. Five different muscles

1: anterior deltoid
2: pectorial major
3: posterior deltoid
4: middle deltoid
5: triceps

were tested on each of three subjects, and the experiment was carried out three times for each treatment combination. The electrographic data, recorded during the serve, are given in the following table. Data file "electromyographic.txt".

	Muscle				
Subject	**1**	**2**	**3**	**4**	**5**
1	32	5	58	10	19
	59	1.5	61	10	20
	38	2	66	14	23
2	63	10	64	45	43
	60	9	78	61	61
	50	7	78	71	42
3	43	41	26	63	61
	54	43	29	46	85
	47	42	23	55	95

Use $\alpha = 0.01$ level of significance to test the hypothesis that

a. Different subjects have equal electromyographic measurements.

b. Different muscles have no effect on electromyographic measurements.

c. Subjects and type of muscle do not interact.

8. (**only for graduate students**) A quality control engineer studied the relationship between years of experience of a system control engineer on the capacity of the engineer to complete within a given time a complex control design including the debugging of all computer programs and control devices. A group of 25 engineers having a wide difference in experience (measured in months of experience) were given the same control design project. The results of the study are given in the following table with $y = 1$ if the project was successfully completed in the allocated time and $y = 0$ if the project was not successfully completed. Data file "ExperienceCompetingTask.txt".

Months of Experience (**experience**)	Project Success $(y = 0,1)$ (**completing the task**)
2	0
4	0
5	0
6	0
7	0
8	1
8	1
9	0
10	0
10	0
11	1
12	1
13	0
15	1
16	1
17	0
19	1
20	1
22	0
23	1
24	1
27	1
30	0
31	1
32	1

a. Determine whether experience is associated with the probability of completing the task.

b. Compute the probability of successfully completing the task for an engineer having 24 months of experience. Place a 95% confidence interval on your estimate.

9. (**only for graduate students**) **Geriatric study**. A researcher in geriatrics designed a prospective study to investigate the effects of two interventions on the frequency of falls. One hundred subjects were randomly assigned to one of the two interventions: education only $(X_1 = 0)$ and education plus aerobic exercise training $(X_1 = 1)$. Subjects were at least **65** years of age and in reasonably good health. Three variables considered to be important as control variables were gender $(X_2 :\ 0 = \text{female}\ ;\ 1 = \text{male})$, a balance index (X_3), and a strength index (X_4). The higher the balance index, the more stable is the subject: and the higher the strength index, the stronger is the subject. The subject kept a diary recording the number of falls (Y) during the six months of the study. The data are given in the following table. Data file "GeriatricStudy.txt".

▨	**Dependent (or Response) variable**	**Independent Variables (or Predictors)**			
Subject i	**Number of Falls** Y	**Intervention** X_1	**Gender** X_2	**Balance Index** X_3	**Strength Index** X_4
1	1	1	0	45	70
2	1	1	0	62	66
3	2	1	1	43	64
4	0	1	1	76	48
5	2	1	0	51	72
6	1	1	1	73	39
7	0	1	1	40	54
8	0	1	0	66	37
9	2	1	1	80	81
10	2	1	1	56	60
11	2	1	1	59	64
12	3	1	1	81	44
13	2	1	1	28	68
14	3	1	1	76	66
15	1	1	0	37	46
16	2	1	1	45	34
17	4	1	0	80	55
18	3	1	0	41	78
19	1	1	0	32	64
20	2	1	1	48	77
21	3	1	0	81	58
22	1	1	1	43	60
23	1	1	0	66	57
24	0	1	1	52	53
25	0	1	1	63	40
26	2	1	0	33	64
27	0	1	0	74	77
28	0	1	1	73	52
29	3	1	0	79	89
30	3	1	0	56	54
31	2	1	1	47	73
32	1	1	0	38	71

33	1	1	1	28	73
34	5	1	1	51	75
35	1	1	1	40	70
36	0	1	1	83	83
37	4	1	1	43	29
38	1	1	1	40	59
39	0	1	1	19	61
40	1	1	0	63	51
41	1	1	1	33	64
42	4	1	1	13	56
43	1	1	1	62	52
44	1	1	1	53	34
45	0	1	1	42	55
46	3	1	0	92	82
47	0	1	1	54	75
48	4	1	1	56	87
49	1	1	0	28	63
50	0	1	0	39	50
51	4	0	0	74	59
52	11	0	0	45	87
53	3	0	1	59	65
54	2	0	0	48	59
55	6	0	0	63	90
56	3	0	1	56	43
57	3	0	1	63	58
58	6	0	0	75	69
59	3	0	1	49	66
60	3	0	0	55	68
61	5	0	0	90	58
62	7	0	1	98	49
63	7	0	1	53	53
64	4	0	0	36	30
65	4	0	0	24	42
66	4	0	0	33	55
67	9	0	0	50	57
68	7	0	0	76	80
69	9	0	1	89	58
70	7	0	1	65	62
71	3	0	0	19	76
72	3	0	1	33	62
73	2	0	1	34	55
74	8	0	0	73	58
75	4	0	1	39	71
76	2	0	1	54	57
77	7	0	1	66	59
78	3	0	0	51	65
79	3	0	1	28	49
80	2	0	1	36	51
81	3	0	0	13	18
82	1	0	1	54	48
83	4	0	1	52	84
84	4	0	0	47	83
85	5	0	0	66	50
86	4	0	0	31	71
87	4	0	0	46	73
88	4	0	0	68	72

89	3	0	0	80	65
90	3	0	1	67	44
91	1	0	0	13	67
92	3	0	0	25	89
93	10	0	1	43	70
94	9	0	0	73	60
95	7	0	0	43	62
96	5	0	1	76	46
97	2	0	1	33	55
98	4	0	0	69	48
99	4	0	1	50	52
100	2	0	0	37	56

a. Fit the Poisson regression model with the response function

$$\mu = E(Y \mid X_1, X_2, X_3, X_4) = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4}$$

$$\Updownarrow$$

$$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

State the estimated regression coefficients, their estimated standard deviations. and the estimated response function.

b. Obtain the deviance residuals and present them in an index plot. Do there appear to be any outlying cases?

c. Assuming that the fitted model is appropriate, use the likelihood ratio test to determine whether gender (X_2) can be dropped from the model: control α at 0.05. State the full and reduced models, decision rule, and conclusion. What is the P-value of the test?

d. For the fitted model containing only X_1, X_3 and X_4 in first-order terms, obtain an approximate 95% confidence interval for β_1. Interpret your confidence interval. Does aerobic exercise reduce the frequency of falls when controlling for balance and strength?

e. Fit the **four-simple** Poisson regression models with one independent variable, one X , at a time. Interpret the results.