

Statistics 7—Practice Final Examination

1. The following is a list of some statistical techniques and methods that you have learned about in this course.

| | | |
|---------------------------|----------------------------|----------------------------|
| • Frequency table. | • Histograms. | • Stem and leaf plots. |
| • Boxplots. | • Quantile plots. | • Scatterplots. |
| • One-sample z -test. | • z confidence interval. | • t confidence interval. |
| • One-sample t -test. | • Two-sample t -test. | • Sign test. |
| • Matched pairs t -test | • ANOVA (one-way). | • Regression. |

For each of the situations described below, select the technique that you believe is the most applicable.

If it is a statistical hypothesis test, state the null and alternative hypotheses. (State them in terms specific to the example, e.g. “mean for women equals mean for men”, rather than in general terms like “ $\mu_0 = \mu_1$ ”.) Do **not** go into details of the computations required.

If it is a graphical or descriptive technique, say in a few words what you would be looking for.

If it is a confidence interval, no further explanation is required here.

- A researcher is interested in how annual personal income varies between different regions of the country. She has samples from four regions and would like a graphical summary.
 - The researcher is also interested in whether mean annual income is different for Asian Americans and African Americans in New England. She has data from a SRS of each population.
 - An investigator is interested in the success of a job training program for current welfare recipients. If fewer than 30% of participants in the program are able to find work within three months, the program will be discontinued.
 - A researcher is interested in the effect of diet on the speed that mice can complete a particular task. He chooses two mice from each of several litters and each is assigned to one of two diet groups.
 - The researcher now believes that different breeds of mice respond differently to the two diets. He repeats the above experiment but now with three different breeds of mice.
 - A high school principal is interested in how well she can predict the number of days that her students miss school as a function of their GPA.
2. Researchers investigated whether a certain strain of bacteria is responsible for heart disease. Two groups of patients are studied; the first group consists of people who have been diagnosed with heart disease, the second group is completely healthy. The scientists detected the bacteria in 72 of 90 patients with heart disease, and only in one of 24 patients without heart disease.
- Is this study an experiment, survey, or observational study? Explain briefly.
 - Is the study prospective or retrospective? Explain briefly.
 - What is the treatment? What is the response?
 - Do the results of the study prove that the bacteria cause heart disease? Explain briefly.
 - Suppose that the 90 patients with heart disease are a simple random sample of all heart disease patients. Compute a 90% confidence interval for the fraction of all heart disease patients who have the bacteria.

3. A demographer is interested in U.S. population and where urban areas are growing and shirking in the U.S. Using a random number table, she selects five cities west and five cities east of the Mississippi from the 100 biggest cities in the U.S. (according to the 1990 census). This is what she finds.

| Populations in 1000s of people | | | | | |
|--------------------------------|------|------|-------------|------|------|
| | East | | | West | |
| City | 1990 | 1980 | City | 1990 | 1980 |
| Chicago | 2783 | 3005 | Lubbock | 186 | 173 |
| Raleigh | 208 | 150 | El Paso | 515 | 426 |
| Detroit | 1027 | 1203 | Fremont | 173 | 132 |
| Tampa | 280 | 272 | Bakersfield | 175 | 106 |
| Columbus | 632 | 565 | Tulsa | 367 | 360 |

Knowing that the population tends to be quite skewed, the wise demographer takes the log of the data before she conducts analysis. She computes the following summary statistics on the log scale.

| | 1990 | 1980 | Difference (1990-1980) |
|------|-------------------|-------------------|------------------------|
| EAST | $\bar{x} = 6.457$ | $\bar{x} = 6.411$ | $\bar{x} = 0.047$ |
| | $s = 1.041$ | $s = 1.186$ | $s = 0.187$ |
| WEST | $\bar{x} = 5.539$ | $\bar{x} = 5.328$ | $\bar{x} = 0.211$ |
| | $s = 0.505$ | $s = 0.614$ | $s = 0.190$ |

The demographer has several questions she would like to ask.

- Test the null hypothesis that large cities in the east on average did not grow between 1980 and 1990. What do you conclude?
 - Construct a 95% confidence interval for the difference in average population between large cities in the east and the west in 1990. What do you conclude?
 - Test the hypothesis with significance level $\alpha = 0.5$, that cities in the west are growing faster than cities in east. What do you conclude?
 - Summarize your conclusions in words that can be understood by someone with no statistical training. What specific advice concerning the design of this study would you give to the political scientist?
4. In this problem we use data from a sample of size 80 of Harvard undergraduate men to investigate the question of whether Harvard men educated at private secondary schools are as likely as Harvard men educated in public schools to prefer boxers over briefs.
- Is this an experiment, survey, or observational study. Explain briefly.
 - Suppose 40% of Harvard undergraduate men wear briefs. Let X be the number of men in the sample that wear briefs. What is the distribution, mean, and variance of X ? Compute $\Pr(30 < X \leq 40)$.

Stata output from the actual study appears below.

| | Private | Public | Total |
|--------|---------|--------|-------|
| boxers | 21 | 28 | 49 |
| briefs | 6 | 25 | 31 |
| Total | 27 | 53 | 80 |

- (c) Consider the population of Harvard undergraduate men who attended private secondary schools. Construct a 97% confidence interval for the proportion of this population who wear boxers.
- (d) Now consider the population of Harvard undergraduate men who attended public secondary schools. Test the hypothesis that half of these men wear boxers. Report a p-value and your conclusion.

5. In this problem we investigate the affect of four diets on blood coagulation time in lab rats. Several rats were randomly assigned to one of four diet groups for several days and then the time required for blood coagulation timed (in seconds). A summary of the data and an ANOVA table are given below.

```
. tabulate diet, summarize(time)
```

| diet | Summary of time | | |
|-------|-----------------|-----------|-------|
| | Mean | Std. Dev. | Freq. |
| 1 | 61 | 1.8257419 | 4 |
| 2 | 66 | 2.8284271 | 6 |
| 3 | 68 | 1.6733201 | 6 |
| 4 | 61 | 2.6186147 | 8 |
| Total | 64 | 3.8448158 | 24 |

```
. oneway time diet
```

| Source | Analysis of Variance | | | F | Prob > F |
|----------------|----------------------|----|---------|------|----------|
| | SS | df | MS | | |
| Between groups | 112 | 3 | 37.3333 | 3.27 | 0.0000 |
| Within groups | 228 | 20 | 11.4 | | |
| Total | 340 | 23 | 14.7826 | | |

Bartlett's test for equal variances: $\chi^2(3) = 1.6680$ Prob> $\chi^2 = 0.644$

- (a) State null and alternative hypotheses that the ANOVA table is constructed to test. Perform the hypothesis test and state your conclusion.
- (b) Estimate the mean difference between Diet 1 and Diet 2, in the time required for blood coagulation.
- (c) Compute a 95% CI for your estimate and interpret the interval for someone not trained in statistics.

6. In this problem we investigate the rate of increase of freshmen minority enrollments at Michigan State University between 1981 and 1990. The data are plotted on the following page; the predictor variable year is the actual year minus 1980 (that is 1981 is coded as 1, 1982 as 2, etc.). For the purpose of this question we assume independence between years. Some computer output (with missing values) appears on the next page.

- (a) Use the regression line to predict minority enrollment in 1991.
- (b) What fraction of the variation in the minority enrollments can be accounted for by year?
- (c) What is the conditional standard deviation of minority enrollments given year?
- (d) Use the standard deviation you computed in part (c) to approximate the probability that 1991 minority enrollments will be greater than 4800.
- (e) What is your estimate for the rate of increase in minority enrollments? Is there evidence that minority enrollments are increasing? Give a p-value and an confidence interval.
- (f) Use the following three plots to comment briefly on whether linear regression is appropriate for these data. Explain briefly what each plot tells you if it is relevant to this question.

```
. summarize
```

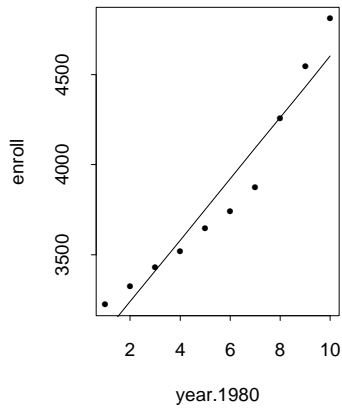
| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-----|--------|-----------|------|------|
| enroll | 10 | 3835.9 | 535.4359 | 3224 | 4810 |
| year | 10 | 5.5 | 3.02765 | 1 | 10 |

```
. regress enroll year1980
```

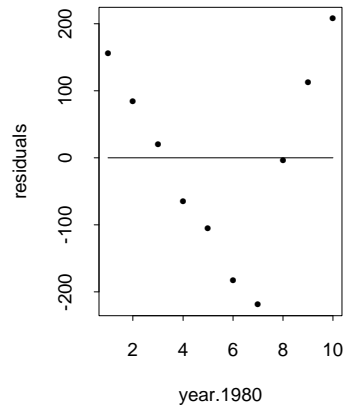
| Source | SS | df | MS | Number of obs = | 10 |
|----------|------------|----|------------|-----------------|--------|
| Model | 2396676.15 | 1 | 2396676.15 | F(1, 8) = | 104.46 |
| Residual | 183548.752 | 8 | 22943.5939 | Prob > F = | 0.0000 |
| | | | | R-squared = | 0.9289 |
| | | | | Adj R-squared = | 0.9200 |
| Total | 2580224.90 | 9 | 286691.656 | Root MSE = | 151.47 |

| enroll | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|--------|----------|-----------|--------|-------|----------------------|
| year | 170.4424 | 16.67646 | XX.XXX | X.XXX | XXX.XXXX XXX.XXX |
| _cons | 2898.467 | 103.4747 | XX.XXX | X.XXX | XXXX.XXX XXXX.XX |

#1: data and regression line



#2



#3

