

# Stat 401 B/xm - Fall 2001 - Exam 1

## Please put your name on the back of the last page

This exam has two parts: a series of short answer questions and a longer problem. Each short answer is worth 10 points. The long problem is worth 40 points. You have two hours to do the exam.

Please write your answers in the enclosed spaces. A piece of ruled paper is attached for your answers to the long problem. If you need more room, continue on the back of the page.

The data set and computer output for the longer problem are attached. I have included SAS code and output and JMP printouts. WARNING: I tried to anticipate **any** output someone might need. There is more output than you need.

**Oncampus students:** If you want something you don't see, please ask for it. If you want something and aren't sure how to interpret or find it in the output, please ask.

**Offcampus students:** Please call or e-mail me if you have any questions.

The last page is a formula sheet. If you want a formula and you don't see it, please ask.

1. Many environmentalists are concerned about amphibian decline. In many places around the world, frogs are less abundant than they were a few years ago. Consider the following study of the green frog in central Iowa. Some researchers randomly selected some ponds in central Iowa. In 1987, they visited each pond and counted the number of frogs. They revisited the same ponds in 1990 and again counted the number of frogs. The researchers wanted to estimate the change in mean abundance from 1987 to 1990. One pond had a very large number of frogs in both years.

a) Circle the most appropriate statistical method to test the hypothesis of no change in mean abundance from 1987 to 1990:

t-test

Wilcoxon rank sum test

paired t-test

Explain your choice.

b) The appropriate test gave a p-value of 0.34. Consider the following statement from the results: "The mean abundance in 1990 is the same as that in 1987."

Is this statement appropriate? Why or why not?

2. The same ponds were sampled again in 1995. The difference in mean abundance for one rare species was 26 animals. A randomization test of the difference between 1987 and 1995 was used to evaluate this result. The 130 values in the randomization distribution are summarized in the enclosed stem-and-leaf diagram.

Legend: 2 6 represents a difference of 26 animals.

Using this information, what is the 2 sided p-value for a test of the hypothesis that the true difference is 0?

Describe, in words, an appropriate conclusion.

3. A research group is studying the prevalence of a particular disease on pig farms in Iowa. In each of three years, the researchers drew a random sample of farms and calculated the prevalence of the disease at each farm. These observations were averaged to estimate the mean prevalence in Iowa. The resources available for the study increased, so the sample size (N) got larger during the study. Here are the data (N=sample size,  $\bar{x}$  = sample average, s = sample standard deviation):

Year	N	$\bar{x}$	s
1	9	47%	15%
2	25	52%	20%
3	36	48%	36%

The mean prevalence was most precisely estimated in which year? Explain your answer.

4. In a study of insect response to pesticides, two strains of fruitfly were bred. One was resistant to DDT; the other was susceptible to DDT. Fecundity (number of eggs laid) was measured on 25 randomly selected females from each strain. Summary statistics for each group are:

Group	N	mean	s.d.
Resistant	25	25.26	7.77
Susceptible	25	23.63	9.77

- a) Compute a pooled estimate of the standard error of the difference.
- b) How many degrees of freedom does this have?
- c) Compute a 95% confidence interval for the difference in means. The 0.975 percentile of the appropriate t distribution is 2.01.
- d) If your advisor wanted a hypothesis test (of difference = 0) instead of a confidence interval, what can you say about the p-value **without** doing the test calculations?
5. Some friends of mine are studying a fungal disease of strawberries. One recent experiment examined the effect of temperature on the germination of fungal spores. Strawberry plants, growing in pots, were randomly assigned to one of 6 temperatures (25°, 27.5°, 30°, 32.5°, 35°, or 37.5°). Five plants were used for each temperature (total of 30 plants). A suspension of fungal spores was sprayed on each plant, then the plants were put in growth chambers set at the appropriate temperatures. The plants were removed from the growth chambers after 48 hours. Eight leaf disks (small circular areas of leaf) were removed from each plant. The percent of germinated spores was calculated for each leaf disk (total of 240 observations, 40 at each temperature).
- a) What is the experimental unit for this study? Why?
- b) What is the observational unit for this study? Why?
- c) Given your answers to a) and b), do you expect a problem with the assumption of independence? Why or why not?

6. Shown to the left are a stem-and-leaf and a box plot of the mercury concentration in fish from 120 lakes in Maine. Each of the 120 observations represents the mean concentration in fish from one of the lakes. On the stem and leaf plot, 25 0 means that there is an observation with 2.50 ppm of mercury. The median concentration is 0.41 ppm (parts per million). The mean concentration is 0.48 ppm. Eating mercury is not a good thing, so lots of folks are concerned about how much mercury they will ingest if they eat fish they have caught.

a) If my uncle fishes at one (and only one) lake in Maine, is it more appropriate to tell him the median concentration or the mean concentration?

Why?

b) What, if any, features of this distribution violate the usual t-test assumptions?

7. **40 points** The following data were collected as part of a study on antibody responses in diabetic mice. The question for this part of the study was 'Does (or by how much) does insulin change the amount of serum albumen'? 36 diabetic mice were used in this part of the study. 18 were randomly assigned to receive an injection of insulin; the other 18 mice (a control group) received an injection of saline solution. The data and a lot of SAS and JMP output are on the following pages.

The JMP output are on three pages. The analyses and descriptions on the first and second pages are appropriate if the data are two samples. The first page gives analyses of the original data (untransformed). The second page gives analyses of the log transformed values. The third page of analyses and descriptions are appropriate if the data are paired.

A packet of SAS output is included, if you want to see this. This information includes:

	SAS Page numbers:
Two printouts of the data	1, 10
Descriptive statistics for each group	2 - 5
Side by side box plots	6
2 Sample t test on raw data	7
2 Sample t test on log transf.	8
Wilcoxon rank sum test	9
Descriptive statistics on paired difference	11 - 12
Paired t-test and Wilcoxon signed rank test	11

Your goal is to answer the investigator's question, 'Does (or by how much) does insulin change the amount of serum albumen'?

There is a piece of ruled paper on the next page for your answers. I have more if you need it.

- a) What is (are) the most appropriate methods to answer the investigator's question.
- b) Please explain why you are choosing a particular method.
- c) Please summarize your conclusions in few sentences.

## Formulae

I have tried to anticipate formulae you might need, but I have not copied all formulae. If you think you need something else, don't assume you're wrong. Just ask for the formula.

$$\begin{aligned} \bar{X} &= \sum_{i=1}^N X_i / N \\ \text{Var } X &= \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2 \\ s_x &= \sqrt{\text{Var } X} \\ \text{s.e. } \bar{X} &= \frac{s_x}{\sqrt{N}} \\ \text{pooled s.d. } s_p &= \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}} \\ \text{pooled s.e. of } \bar{X}_1 - \bar{X}_2 &= s_p \sqrt{\frac{1}{N_1} + \frac{1}{N_2}} \\ \text{s.e. without pooling: s.e. of } \bar{X}_1 - \bar{X}_2 \text{ } s_W &= \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}} \\ \text{T ratio: } T &= \frac{\text{estimate} - \text{parameter}}{\text{s.e. of estimate}} \\ \text{d.f. } W &= \frac{s_W^4}{\text{s.e.}_1^4 / (N_1 - 1) + \text{s.e.}_2^4 / (N_2 - 1)} \\ \text{mean of Wilcoxon rank sum statistic} &= N_1 \bar{R} \\ \text{s.d. of Wilcoxon rank sum statistic} &= s_R \sqrt{\frac{N_1 N_2}{N_1 + N_2}} \end{aligned}$$