

# Case-based Social Statistics II

CSSS 322

Spring 2002

## Solutions to Midterm Examination

Exam: Tuesday, May 7, 10:30am - 12:00pm

Professor: Mark S. Handcock

Name: Ronald Aylmer Fisher

1. Please write your name in the above space.
2. **You need to do all 4 questions.** All questions are of equal value (but not necessarily of equal difficulty). attempted.
3. Do not turn the page until so instructed. (You will have 90 minutes to work after the examination has been discussed with you.)
4. You may use your crib sheet. Otherwise this is a closed book examination.
5. If you do not have enough room for your work in the place provided, use the back of a nearby page. (However, be sure to mark clearly which problem the material on the back of any page refers to.) If you pull the pages apart, sign all pages.
6. Answers should unambiguously state, in words, the approach taken. You should show your work so that partial credit can be given. Poorly described solutions will be penalized. unsupported answers
7. Good luck!

Question	Subject	Points available	Points earned
1	Fair pay for fair work	33	
2	More fair pay for fair work	24	
3	Gender Discrimination and Salaries	28	
4	Mortgage Discrimination	15	
Total		100	

**Question 1) Fair pay for fair work (33 points)**

The Federal minimum wage law says that all adult workers should be paid at least \$5.15 per hour. To allow some flexibility to temporary workers in the fruit picking industry in California, the local law states that the average wage paid to workers in a farm should be \$5.50 per hour. Some workers can be paid less and some workers can be paid more, but the average must be \$5.50.

To test compliance, the State randomly samples 10 workers from a large farm. They find the average wage to be \$5.25 per hour and the standard deviation to be \$0.08 per hour.

**a) (6 points)**

Construct a 95% confidence interval for the mean wage paid to workers on this farm.

**Solution:** The 95% confidence interval is

$$\begin{aligned}\bar{X} \pm t_{\alpha/2}(n-1) \times \frac{s}{\sqrt{n}} \\ 5.25 \pm 2.262 \times \frac{0.08}{\sqrt{10}} \\ 5.25 \pm 0.057 \\ \text{that is } (5.19, 5.31) \text{ dollars}\end{aligned}$$

**b) (8 points)**

Construct and interpret a 95% *prediction* interval for the wage paid to a randomly chosen worker on this farm.

**Solution:** To get the 95% prediction interval for a single worker's wage, we first have to get  $\hat{\sigma}_X$ :

$$\hat{\sigma}_X = s\sqrt{1 + \frac{1}{n}} = 0.08\sqrt{1 + \frac{1}{10}} = 0.0839.$$

The 95% prediction interval for the wage is then

$$\begin{aligned}\bar{X} \pm t_{\alpha/2}(n-1) \times \hat{\sigma}_X \\ 5.25 \pm 2.262 \times 0.0839 \\ 5.25 \pm 0.190 \\ \text{that is } (5.06, 5.44) \text{ dollars}\end{aligned}$$

We interpret this as: We can be 95% confident that the wage paid to a randomly chosen worker on this farm is between \$5.06 and \$5.44.

**c) (4 points)**

The farm claims that it is paying its workers fairly, and it is unclear if they are paying higher or lower wages. Translate this claim into a null and alternative hypothesis about the wages expressed in symbols. Be sure to define the symbols you use.

**Solution:** Let  $\mu$  be the mean wage paid to all workers. The claim is that the farm is paying its workers fairly, that is, at \$5.50 per hour. Formally,

$$H_0 : \mu = 5.50$$

$$H_1 : \mu \neq 5.50$$

Note that the alternative hypothesis is two sided.

d) (4 points)

Test the null hypothesis in c) based on the sample of wages. Use the 95% level of confidence. What is your conclusion?

**Solution:** From a), the 95% confidence interval is from \$5.19 to \$5.31. We reject the null hypothesis as \$5.50 falls outside the confidence interval. On this basis there is sufficient evidence, at the 5% significance level, to reject the null hypothesis that the farm is paying its workers fairly.

e) (4 points)

The State is concerned that the farm is under-paying its workers. It claims that any deviation from the overall rate will be toward lower rates and that higher rates are not possible.

Translate this claim into a null and alternative hypothesis about the wages that can be tested with the above data. Be sure to define the symbols you use.

**Solution:** The State claims that the alternative hypothesis should only cover deviations from the null in the direction of lower wages than \$5.50 per hour. Formally,

$$H_0 : \mu = 5.50$$

$$H_1 : \mu < 5.50$$

Note that the alternative hypothesis is one sided, reflecting the “guilty verdict” of the theory. The null hypothesis can also be chosen to be  $H_0 : \mu \geq 5.50$ .

f) (7 points)

Test the null hypothesis in e) based on the sample of wages. Use the 95% level of confidence. What is your conclusion?

**Solution:** As this is a one-sided hypothesis test with  $\alpha = 0.05$ , the 95% confidence interval are all values above

$$\bar{X}_{men} - t_{\alpha}(n-1) \times \frac{s}{\sqrt{n}}.$$

This lower bounds is

$$5.25 - 1.833 \times \frac{0.08}{\sqrt{10}}$$

$$5.25 - 0.046$$

that is values less than \$5.30

Note that the confidence interval, corresponds to a one-sided test rather than a two-sided test. We use  $t_{0.05}(9) = 1.833$ . We reject the null hypothesis as \$5.50 falls outside the confidence interval and the confidence interval falls completely in the region specified by the alternative hypothesis. On this basis there is sufficient evidence, at the 5% significance level, to reject the null hypothesis that the farm’s claim of fair wages is true, at the 95% level of confidence. Thus we find in favor of the State’s hypothesis.

**Question 2) More fair pay for fair work (24 points)**

**a) (4 points)**

Briefly describe the meaning of Type I error in the hypothesis test of part c) of the previous question.

**Solution:** A Type I error is to decide that the null hypothesis is false when in fact it is true. In this case that means to decide that the farmers are under-paying their workers when in fact they are not and the average is \$5.50.

**b) (5 points)**

Briefly describe the meaning of Type II error in this case?

**Solution:** A Type II error is to decide that the null hypothesis is true when in fact it is false. In this case that means to decide that the farmers are paying their workers fairly when in fact they are not and the average is below \$5.50.

**c) (5 points)**

Given the confidence interval in part a) of the previous question, is the probability of Type II error large or small. Briefly say why.

**Solution:** Our best estimate of  $\mu$  is  $\bar{X} = 5.25$ . If this were the actual value then we will not reject the null hypothesis if the confidence interval includes 5.50. From a), \$5.50 is far outside the interval, so it will happen quite rarely. Hence the probability of Type II error is quite small in this case.

**d) (8 points)**

Calculate the  $p$ -value of the hypothesis test in part c) of the previous question.

**Solution:** At the  $\alpha$  error level, we will reject the null if 5.50 is outside the confidence interval for  $\mu$ . The two-sided intervals are:

95%	\$5.19 to \$5.31
99%	\$5.17 to \$5.33

It is clear that the confidence level will need to be very large (that is, greater than 99.9999%) to just include \$5.50. Hence the  $p$ -value is less than 0.000.

**e) (3 points)**

Which of the two hypothesis tests considered in the previous question do you think is more appropriate? Explain briefly.

**Solution:** An argument can be made for both tests. The decision rests on whether it is reasonable to exclude average values of wages above \$5.50. If it is the one sided test is more appropriate. However, even with the State's claim it seems reasonable to consider the likelihood that the farm is being more than fair. Hence the two-sided test is more appropriate.

### Question 3) Gender Discrimination and Salaries (28 points)

In establishing gender discrimination in salaries by a corporation, one of the most important forms of evidence is a statistical record of salaries paid in the past to men and women. Suppose in a particular firm that the salaries paid for all managers are:

	Gender	
	Women	Men
sample size	$n_{women} = 15$	$n_{men} = 22$
sample mean	$\bar{X}_{women} = \$24,467$	$\bar{X}_{men} = \$33,095$
sample standard deviation	$s_{women} = \$2,806$	$s_{men} = \$4,189$

#### a) (8 points)

Construct a 99% confidence interval for the mean difference in wages between men and women.

**Solution:** We first should calculate the standard deviation of the difference in means. There are two options. If we think the two standard deviations are different we can use the two-sample formula:

$$\begin{aligned} s_{women-men} &= \sqrt{\frac{s_{men}^2}{n_{men}} + \frac{s_{women}^2}{n_{women}}} \\ &= \sqrt{\frac{4.189^2}{22} + \frac{2.806^2}{15}} \\ &= 1.150 \end{aligned}$$

Alternatively we can use the two-sampled pooled formula that acts as if the two variances are the same:

$$\begin{aligned} s_{pooled} &= \sqrt{\frac{s_{men}^2 \times n_{men} + s_{women}^2 \times n_{women}}{n_{men} + n_{women} - 2}} \\ &= \sqrt{\frac{4.189^2 \times 22 + 2.806^2 \times 15}{22 + 15 - 2}} \\ &= 3.795 \end{aligned}$$

The estimate of the variance of the difference is then:

$$\begin{aligned} s_{diff} &= s_{pooled} \sqrt{\frac{1}{n_{men}} + \frac{1}{n_{women}}} \\ &= 1.271 \end{aligned}$$

These numbers are close, even though the two sample values appear to be different. Hence we prefer the direct method, rather than the pooled method here. The 99% confidence interval is

$$\begin{aligned} &\bar{X}_{men} - \bar{X}_{women} \pm t_{\alpha/2}(n_{men} + n_{women} - 2) \times s_{men-women} \\ &33.095 - 24.467 \pm 2.72 \times 1.150 \\ &8.628 \pm 3.128 \\ &\text{that is } (5.500, 11.750) \text{ dollars} \end{aligned}$$

**b) (4 points)**

The corporation claims that they treat men and women equally.

Translate this claim into a null and alternative hypothesis about women's and men's wages expressed in symbols. Be sure to define the symbols you use.

**Solution:** Let  $\mu_{women}$  be the mean salary for women and  $\mu_{men}$  be the mean salary for men. The claim is that the population mean difference is zero. Formally,

$$H_0 : \mu_{women} = \mu_{men}$$

$$H_1 : \mu_{women} \neq \mu_{men}$$

Note that the alternative hypothesis is two sided. We could consider a one-sided test (with the alternative specifying that  $\mu_{women} < \mu_{men}$ ) but this is not necessarily true.

**c) (4 points)**

Test the null hypothesis in b) based on the above data. Use the  $\alpha = 0.01$  significance level. What is your conclusion?

**Solution:** From a), the 98% confidence interval for the difference between the mean salaries is from \$5,204 to \$12,052. We reject the null hypothesis as 0 falls outside the confidence interval. On this basis there is sufficient evidence, at the 1% significance level, to reject the null hypothesis that the two salaries coincide.

**d) (9 points)**

The State wants to collect more evidence. To do so it looks the workers that work together at the same job within the corporation. They randomly select 20 pairs of men and women and find the following results:

	Women	Men
sample size	$n = 20$	
sample mean	$\bar{X}_{women} = \$25,837$	$\bar{X}_{men} = \$30,456$
sample standard deviation	$s_{women-men} = \$3,286$	

The standard deviation is for the differences between women's and men's wages in each pair.

Test the null hypothesis of equal average wages using this new sample. Use the 95% level of confidence. What is your conclusion?

**Solution:** This is a paired data situation. The null and alternative hypotheses are the same as before:

$$H_0 : \mu_{women} = \mu_{men}$$

$$H_1 : \mu_{women} \neq \mu_{men}$$

Let  $\mu = \mu_{men} - \mu_{women}$ . The natural estimate of  $\mu$  is  $\bar{X} = \bar{X}_{men} - \bar{X}_{women}$  and the natural estimate of the variance of  $\bar{X}$  is the sample variance  $s_{women-men}$ . The 95% confidence interval is then

$$\bar{X}_{men} - \bar{X}_{women} \pm t_{\alpha/2}(n - 1) \times \frac{s_{women-men}}{\sqrt{n}}$$

$$30.456 - 25.837 \pm 2.093 \times \frac{3.286}{\sqrt{20}}$$

$$4.619 \pm 1.538$$

that is (3.081, 6.157) thousand dollars

We use  $t_{0.025}(19) = 2.093$ . We reject the null hypothesis as 0 falls outside the confidence interval. On this basis there is sufficient evidence, at the 5% significance level, to reject the null hypothesis that the mean salaries are the same, at the 95% level of confidence.

e) (3 points)

Which of the two hypothesis tests considered in this question do you think is a better test of the claim? Explain briefly.

**Solution:** Both are valid tests and are dictated by the source of evidence available. In the first case the evidence take the form of independent samples. In the second it is a set of 20 matched pairs of salaries. The matched information has the advantage of having an overall reduced variance due to the removal of the overall variation in salaries. This results in a variance of 3.286 in the paired case compared to 3.785 in the unpaired case. In this sense the paired test will be more powerful for a given confidence level.

#### Question 4) Mortgage Discrimination (16 points)

A large North-Western Bank can choose to offer home loans to King County residents that apply and qualify under broad income guidelines for the typical home. The bank investigates the credit worthiness of each applicant before it decides which of the applicants will be offered loans. In 1999 the bank offered home loans to 3190 out of 6600 applicants who were Seattle residents, and to 660 out of 2420 applicants who were not from the city of Seattle, but were still in King County. Assume all applicants were equally qualified for the loan.

Location	Number of Applicants	Number Offered	Number Denied
Non-Seattleites	2420	660	1760
Seattleites	6600	3190	3410
Totals	9020	3850	5170

One measure of the likelihood of receiving a loan is the offer rate, defined as the percentage of qualified applicants that were offered loans.

In this question we will explore the relative offer rates for Seattle residents only.

##### a) (5 points)

What is the offer rate for Seattleites? The City rules say that the bank should offer 25% of Seattle applicants a loan. We wish to conduct a hypothesis test of this based on the above data.

State the null and alternative hypotheses in symbols and words. Be sure to identifying all symbols that you use.

**Solution:** The offer rate for Seattleites is  $3190/6600 = 0.4833$ . The null hypothesis is

$$H_0 : \pi = 0.25$$

against the alternative hypothesis

$$H_1 : \pi \neq 0.25$$

Here  $\pi$  is the proportion of all applicants offered a loan. This is a two-sided alternative. Note that implicitly the alternative claimed is that the offer rate is less than 25% but it would be too much to presume it here. Thus the one-sided alternative is not acceptable, i.e.,

$$H_1 : \pi < 0.25$$

The null hypothesis states that the proportion offers is 25%. The alternative hypothesis states that the proportion of offers is not 25%.

##### b) (6 points)

Construct a confidence interval appropriate for this test. Use the 95% level of confidence.

**Solution:** The point estimate of  $\pi$  is  $f = 3190/6600 = 0.4833$ . The standard error of the estimate is  $S_f = \sqrt{f(1-f)}/\sqrt{n} = \sqrt{0.4833(1-0.4833)}/\sqrt{6600} = 0.00615$ . The confidence interval is then:

$$f \pm t_{\alpha/2}(n-1) \times s_f$$

For an  $\alpha = 0.05$  the interval is:

$$\begin{aligned} &0.4833 \pm 1.96 \times 0.00615 \\ &0.4833 \pm 0.0121 \\ &=(47.1\%, 49.5\%) \end{aligned}$$

Note that as  $n > 40$ , we approximate the  $t$ -multiplier by that with infinite degrees of freedom, 1.96.

**c) (4 points)**

Test the hypothesis using the 95% level of confidence. What is your conclusion?

**Solution:** As the hypothesized value of 25% does not fall inside the confidence interval we reject the null hypothesis at the 95% confidence level.