

Solutions to (modified) practice exam 4

Statistics 224 Practice exam 4 FINAL Your Name _____

Friday 12/21/07

Professor Michael Iltis (Lecture 2)

Discussion section (circle yours) :

section: 321 (3:30 pm M)

322 (2:25 pm M)

323 (4:35 pm M)

Problem	max points	points scored
1		
2		
3		
4		
5		
6		
Total	120	

Do all 6 problems.

Rules :

1. No notes allowed
2. Standard hand calculator allowed
3. Numerical answers without supporting work (or rationale) may receive no points
4. Failure to follow rules may result in lost points

1. In an effort to determine the most effective way to teach safety principles to a group of employees at Weedco, four different methods were used. A sample of 20 employees were randomly assigned to one of the four groups. The first group was given programmed instruction booklets and worked through the course at their own pace. The second group attended lectures. The third group watched television presentations, and a fourth was divided into small discussion groups. At the end of the session, a test was given to the four groups. A high score of 10 was possible. The results were :

TEST GRADES

Programmed instruction	Lecture	Group TV	discussion
6	8	7	5
5	7	9	5
6	8	6	6
5	8	8	6
6	8	9	5

The following is an Analysis of Variance Mini-tab software output with missing information:
ANALYSIS OF VARIANCE ON GRADES

SOURCE	DF	SS	MS	F
TREAT	<u>3</u>	26.550	8.850	<u>14.16</u>
ERROR	<u>16</u>	<u>10</u>	.625	
TOTAL	<u>19</u>	36.550		

a) Complete the missing values : The total number of observations is $N=20$ so the total degrees of freedom is $N - 1 = 19$. The df for treatment is $k-1 = 3$ where k = the number of treatments or populations being compared. This gives the df for error = $19-3 = 16$ since the total is the sum of the other two. Same is true for sum of squares for error which must be 10 since the total SS is the sum of the other two is 36.550. $MSE = 10/16 = .625$ is just SSE / df for error. F is just the ratio $MS(tr)/MSE$ of MS for treatment over MS for error $8.85/.625 = 14.16$

b) Test at the .05 level that there is no difference among the four means.

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$$

$$H_a: \mu_i \neq \mu_j \text{ for some } i \neq j$$

Reject H_0 at $\alpha = .05$ if $F > F_{.05}(3,16) = 3.24$ for 3 numerator and 16 denominator degrees of freedom

Decision : reject H_0 since $F = 14.16 > 3.24$

2. W&A Beverages has observed the following Overhead Costs associated with Gallons of Output over the past twelve months. Additional columns, their sums and relevant formulas are included to aid in answering questions

Month	x = Thousands of Gallons of Output	y = Overhead cost (in \$1000)	x^2	y^2	xy
January	18	45.6	324	2079.36	820.8
February	44	62.4	1936	3893.76	2745.6
⋮	⋮	⋮	⋮	⋮	⋮
December	40	60.0	1600	3600.00	2400
TOTAL	408	684	14900	39376.00	23600

Additional information

$$S_{xx} = \sum (x_i - \bar{x})^2 = 14900 - \frac{408^2}{12} = 1028, \quad S_{yy} = SST = 39376 - \frac{684^2}{12} = 388$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = 23600 - \frac{(408)(684)}{12} = 344, \quad b = \hat{\beta} = \frac{S_{xy}}{S_{xx}}, \quad a = \hat{\alpha} = \bar{Y} - b\bar{X}$$

$$SSR = b^2 S_{xx}, \quad SST = SSR + SSE, \quad MSE = \frac{SSE}{n-2} = s_e^2$$

100(1- α)% confidence interval for β : $b \pm t_{\alpha/2, n-2} \frac{s_e}{\sqrt{S_{xx}}}$. 100(1- α)% prediction

interval for an individual Y when $x = x_0$: $\hat{y}(x_0) \pm t_{\alpha/2, n-2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$

a) An estimate of the variable overhead cost (in \$1000) per thousand gallons of output is $b = S_{xy}/S_{xx} = 344/1028 = .33463$ (this is the slope since we want the change in y over (per) the unit change in x (in \$1000 units). The key hint is the word "per")

b) An estimate of the monthly fixed cost (in \$1000) is

$$a = \bar{Y} - b\bar{X} = (684/12) - (.33463)(408/12) = 57 - .33463(34) = 45.6226 \quad (\text{fixed "set up" cost})$$

c) The fraction of the total sum of squares error from the mean monthly overhead cost explained by the gallons produced variable and the least square line is

$$r^2 = SSR/SST = S_{xy}^2 / (S_{xx} S_{yy}) = 344^2 / (1028 \times 388) = .29668$$

d) A 95% confidence interval for the true variable overhead cost estimated in (a) is

The CI for true slope β is (with $t_{.025, 10} = 2.228$, $s_e = \sqrt{(388 - 344^2/1028)/10} = 5.22386$)

$$b \pm t_{\alpha/2, n-2} \frac{s_e}{\sqrt{S_{xx}}} = .33463 \pm 2.228(5.22386)/\sqrt{1028} = .33463 \pm .36300 = [-.02837, .69763]$$

e) If 30 thousand gallons of output are planned for next month, what would you predict for the overhead cost (to the nearest \$10) Answer : the fitted value (on the estimated line when $x_0 = 30$) is $\hat{y}(30) = 45.6226 + .33463(30) = 55.6615$

f) A 95% prediction interval (in \$) for the prediction of overhead cost in the previous problem is (to the nearest \$100) Answer: 95% P.I. (for $x_0 = 30$) by above formula is

$$55.6615 \pm 2.228(5.22386) \sqrt{1 + 1/12 + (30 - 34)^2/1028} = 55.6615 \pm 12.20072 = [43.46078, 67.86222]$$

3. Let

X = amount of raw material added to a chemical process (in gms)

Y = amount of usable final product (in gms)

Suppose the true regression equation

$$E[Y|x] = -1 + 2.5x$$

is known to hold for $6 \leq x \leq 14$ with error standard deviation $\sigma = .1$ gm.

a) What is the expected usable final product when 10 gms of raw material is added to the chemical process ?

$$-1 + 2.5(10) = -1 + 25 = 24 \text{ gms}$$

b) What could go wrong if we try to estimate the expected value of usable final product when 15 gms of raw material are added to the process?

We could try to use the regression model but we have no guarantee that it works since the value 15 falls outside of the observed interval of x values $6 \leq x \leq 14$ for which we know the model holds.

c) By how many gms do we expect usable final product to increase for each additional gm of raw material added ?

This is the increase in y when increase in x is 1 gm which is the slope of the estimated line (times 1 gm) so by the regression line given, $\Delta y = 2.5$ gms (where slope = 2.5)

d) What is the probability that the amount of usable final product would exceed 24.2 gms if 10 gms of raw material is added ?

The mean value was found in part a) to be 24 and the regression model assumes the Y -values are normally distributed about this mean with S.D. $\sigma = .1$ so standardizing Y , we want

$$P(Y > 24.2 | x_0 = 10) = P\left(Z = \frac{Y - 24}{.1} > \frac{24.2 - 24}{.1} = 2\right) = P(Z > 2) = P(Z < -2) = .0228$$

from the standard Z -table

4. A physical anthropology study of the ability of individuals to walk in a straight line reported that in a sample of $n = 20$ randomly selected healthy men their cadence (which is the number of strides per second) data had a sample mean of

$$\bar{x} = .9255 \quad \text{and standard deviation} \quad s = .0809$$

A normal probability plot yielded substantial support to the assumption that the population distribution of cadence is (approximately) normal.

a) (8 points) Calculate and **interpret** a 95% confidence interval for population mean cadence. Your interpretation should answer the question of whether you know or in what sense the actual population mean cadence lies in the interval you found.

From the t-table $t_{.025,19} = 2.093$ is the critical value with $20-1 = 19$ degrees of freedom

$$\text{The C.I. is } \bar{x} \pm t_{.025,19} s / \sqrt{n} = .9255 \pm 2.093(.0809) / \sqrt{20} = .9255 \pm .03786 = [.887638, .96336]$$

The interpretation is that if we were to do similar experiments (with sample size $n=20$) many times, 95% of the time we would be right to say that the (fixed) population mean μ lies in the (random) C.I. . But for a particular interval either μ does or it doesn't lie in the interval and we don't know which holds.

4. b) (6 points) Calculate a 95% prediction interval for a future single value X_{21} :

$$\text{The P.I. is } \bar{x} \pm t_{.025,19} s \sqrt{1+1/n} = .9255 \pm 2.093(.0809) \sqrt{1.05} = .9255 \pm .1735 = [.7520, 1.0990]$$

The interpretation is similar to that for the 95% C.I.

c) (6 points) Calculate a 95% confidence interval for the true population standard deviation σ of men's cadence. From the definition of chi-squared

and from the chi-squared table we find $\chi^2_{.975,19} = 8.907 < \chi^2 = \frac{(n-1)s^2}{\sigma^2} < \chi^2_{.025,19} = 32.852$

Solving for σ gives $\sqrt{\frac{(n-1)s^2}{32.952}} = .06758 < \sigma < \sqrt{\frac{(n-1)s^2}{8.907}} = .118157$ as the 95% C.I.

5. A family that owns two automobiles is selected at random. Let

A = older auto is American

B = the newer auto is American

Suppose these probabilities are known : $P(A)=.8, P(B)=.5$ and $P(A \cap B)=0.4$.

Determine :

a) Draw a Venn diagram for the problem, then find the probability that at least one auto is American.

In the Venn diagram .4 would go inside $A \cap B$ and .4 in the other part of A lying outside of B . The other part of B outside of A would have a .1 in it. The complementary region outside of both A and B would have a .1 in it so that the sum of all the numbers adds to the total probability 1.

The event "at least one auto is American" is the event $A \cup B$.

From the general addition formula for probabilities we have

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = .8 + .5 - .4 = .9$$

b) The probability that neither auto is American

The event "neither auto is American" is the event $\bar{A} \cap \bar{B} = \overline{A \cup B}$
(the equivalence follows by De Morgan's relation)

so
$$P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - .9 = .1$$

c) The probability that the newer auto is American given that the older auto is American.

$$P(B | A) = \frac{P(A \cap B)}{P(A)} = \frac{.4}{.8} = .5$$

d) Are the events A and B independent?

Yes since $P(B | A) = P(B) = .5$ or equivalently since $P(A \cap B) = P(A) \cdot P(B) = .8 \cdot .5 = .4$

6. Customers at a gas station select either regular (A) or premium (B) or diesel fuel (C) assume that successive customers make independent choices with $P(A) = .5$, $P(B) = .4$, $P(C) = .1$. Let X_1, X_2 , and X_3 be the number of customers who select regular, premium or diesel fuel respectively (these will depend on the total number of customers under consideration).

a) Among the next 10 customers, how many ways are there for exactly 3 to purchase regular fuel, 5 to purchase premium and 2 to purchase diesel if order does not matter ? Hint : note that there are 10 choose 3 ways for 3 customers to purchase regular and of the remaining 7 customers there are 7 choose 5 ways for 5 to purchase premium (so the remaining 2 purchase diesel). This is the multinomial coefficient

$$\frac{10!}{3!5!2!} = \binom{10}{3} \binom{7}{5} = 2520$$

b) Among the next 10 customers, what is the joint probability $P(X_1=3, X_2=5, X_3=2)$ that exactly 3 will purchase regular fuel, 5 will purchase premium and 2 diesel ? Hint: All selections are made independently. What kind of joint distribution is this ?

This is the multinomial probability

$$P(X_1=3, X_2=5, X_3=2) = \frac{10!}{3!5!2!} (.5)^3 (.4)^5 (.1)^2 = .032256$$

c) Among the next 25 customers, what are the mean and variance of the number who select premium fuel ? What kind of random variable is this number and with what parameters ? Hint : either a customer does or doesn't select premium. The probability that premium is not selected is $1 - P(B) = .6$. Explain your reasoning.

The number who select premium is a binomial random variable with parameters $n=25$ and $p = .4$. Hence the mean is $np = 10$ and the variance is $np(1-p) = 6$.

d) If from a finite population of 100 one gallon containers of fuel of which 40 contain premium fuel, we randomly select 25 of these containers, what kind of random variable is the number that contain premium fuel. and with what parameters and what is the expected number of containers that contain premium fuel?

This is now a hypergeometric random variable X (the finite population analogue of a binomial random variable) with parameters the total population size $N=100$, the number $m=40$ which contain premium fuel, and sample size $n=25$. The mean number is $np = 10$ again, where p is the proportion $40/100 = .4$ containing premium. To derive this last expected value, we can write $X = X_1 + X_2 + \dots + X_{25}$ where X_i is 1 if the i^{th} container selected holds premium fuel and 0 else. Unlike Bernoulli trials, these X_i 's are dependent, however the expected value of a sum is the sum of the expected values $E[X] = E[X_1] + E[X_2] + \dots + E[X_{25}] = np = 10$ (whether dependent or independent r.v.'s) same as for the binomial r.v. since the sample was random and so each $E[X_i] = 0 \cdot (1-p) + 1 \cdot p = p = 40/100 = .4$ is the same (each equally likely, 10 with fuel).