

Forty four people took the exam. The range of scores was from 27 to 100 with the mean=72.5 and the standard deviation=17.7. The median was 77.

1. i) Let X be a normal(20,16) random variable. Find $P(X \leq 24)$.

ii) Let X_1, \dots, X_{50} be iid each normal(20,16). In such a sample what is the expected number that will be less than or equal to 24?

iii) In part ii) find the probability that at most 40 of the sample will be less than or equal to 24.

Solution:

i)

$$P(X \leq 24) = P\left(\frac{X - 20}{4} \leq \frac{24 - 20}{4}\right) = P(N(0, 1) \leq 1) = 0.84$$

ii) The number of values in the sample which are less than or equal to 24 is a binomial(50,.84) random variable. Let W denote this random variable. So the answer is

$$E(W) = 50 \times 0.84 = 42$$

iii)

$$P(W \leq 40) \doteq P(W \leq 40.5) = P\left(N(0, 1) \leq \frac{40.5 - 42}{\sqrt{50 \times 0.84 \times 0.16}}\right) = 0.281$$

■

2. Suppose a urn contains six balls with two of them labeled zero, one labeled one and three labeled two. Let X be the value of a ball selected at random from the urn.

i) Find $E(X)$.

ii) If $Y = 3X^2 + 5$ then find $E(Y)$.

Solution:

i)

$$E(X) = \sum_x xp(x) = 0 \times \frac{2}{6} + 1 \times \frac{1}{6} + 2 \times \frac{3}{6} = \frac{7}{6}$$

ii)

$$E(X^2) = \sum_x x^2p(x) = 0 \times \frac{2}{6} + 1 \times \frac{1}{6} + 4 \times \frac{3}{6} = \frac{13}{6}$$

and so

$$E(Y) = 3 \times \frac{13}{6} + 5$$

■

3. Exhibit a discrete random variable which has a variance of 8.5. That is list its set of possible values along with the corresponding probabilities.

Solution:

The easiest example is to take a two point distribution which is symmetric about zero. So let X take on the values a and $-a$ each with probability $1/2$. Since for all a we have $E(X) = 0$ we need to find the value of a for which $E(X^2) = 8.5$. But

$$E(X^2) = (-a)^2 \times .5 + a^2 \times .5 = a^2$$

which means $a = \sqrt{8.5}$.

■

4. A random sample of university students were asked if they went to a gym at least twice a week and if they worked at a job at least 10 hours a week. The results are given in the table. At level $\alpha = 0.01$ test whether the time spent at the gym is independent of having a job.

	gym ≥ 2	gym < 2
work ≥ 10	21	39
work < 10	19	21

Solution:

Under the hypothesis of independence the expected number in the first cell is $E = 100(60/100)(40/100)$
Hence

$$\sum \frac{(O - E)^2}{E} = \frac{(21 - 24)^2}{24} + \frac{(39 - 36)^2}{36} + \frac{(19 - 16)^2}{16} + \frac{(21 - 24)^2}{24} < 6.635 = \chi_{1,0.01}^2$$

and we accept the hypothesis of independence at level $\alpha = 0.01$.

■

5. In an attempt to develop a model that could be used to predict the level of crime in different states demographic statistics were collected for 47 US states for a recent year. The variables are given just below.

- Rt: Crime rate: number of offenses reported to police per million population
- Age: The number of males of age 14-24 per 1000 population
- Ed: Mean number of years of schooling x 10 for persons of age 25 or older
- U2: Unemployment rate of urban males per 1000 of age 35-39
- X: The number of families per 1000 earning below 1/2 the median income
- Ex0: Per capita expenditure on police by state and local government

The model

$$Rt = \beta_0 + \beta_{Age}Age + \beta_{Ed}Ed + \beta_{U2}U2 + \beta_X X + \beta_{Ex0}Ex0 + Z$$

was fit to the data. The results are given below along with some anova tables.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-524.3743	95.1156	-5.513	2.13e-06	***
Age	1.0198	0.3532	2.887	0.006175	**
Ed	2.0308	0.4742	4.283	0.000109	***
U2	0.9136	0.4341	2.105	0.041496	*
X	0.6349	0.1468	4.324	9.56e-05	***
Ex0	1.2331	0.1416	8.706	7.26e-11	***

Residual standard error: 21.3 on 41 degrees of freedom

Multiple R-Squared: 0.7296

F-statistic: 22.13 on 5 and 41 DF, p-value: 1.105e-10

Analysis of Variance Tables

Response: Rt

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Age	1	551	551	1.2140	0.2769718	
Ed	1	7260	7260	15.9994	0.0002586	***
U2	1	7363	7363	16.2263	0.0002373	***
X	1	638	638	1.4064	0.2424900	
Ex0	1	34394	34394	75.8007	7.26e-11	***
Residuals	41	18604	454			

Response: Rt

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Ex0	1	32533	32533	71.6985	1.525e-10	***
X	1	7398	7398	16.3046	0.0002304	***
U2	1	3	3	0.0056	0.9407905	
Ed	1	6489	6489	14.3011	0.0004980	***
Age	1	3783	3783	8.3368	0.0061751	**
Residuals	41	18604	454			

Response: Rt

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X	1	2205	2205	4.8602	0.033152	*
Ed	1	5770	5770	12.7158	0.000939	***
Ex0	1	37827	37827	83.3650	1.975e-11	***
Age	1	2394	2394	5.2762	0.026799	*
U2	1	2010	2010	4.4295	0.041496	*
Residuals	41	18604	454			

Use this output to answer the following questions. If you are unable to answer a part of question because of missing output explain what additional information is needed.

i) In the full model at level $\alpha = 0.05$ test the hypothesis $H : \beta_{Age} = \beta_{U2} = 0$ against $K : \text{at least one of them is not } 0$.

ii) In the full model at level $\alpha = 0.05$ test the hypothesis $H : \beta_{Ed} = 2$ against $K : \beta_{Ed} > 2$.

iii) Note that in the first table labeled Coefficients U2 as a p -value of 0.0415 while in the second of the anova tables its p -value is 0.941. Briefly explain why this is possible.

iv) In class it was explained that good regression models can be used for prediction. It was also pointed out that often people attach a “story” to the model which allows them to claim that that the independent variables in their model explain the behavior of the dependent variable. Can the model given here be used for prediction? Is there a good “story” that goes with the model? Briefly justify your answer.

Solution:

i) Using the third anova table we find

$$\frac{(2394 + 2010)/2}{454} = 4.85 > 3.23 = f_{2,41;0.05}$$

so we reject H .

ii) Since

$$\frac{2.0308 - 2}{.4742} = 0.065 < 1.683 = t_{41;.05}$$

we fail to reject H .

iii) Note the first p -value is testing how important U2 is the full model while the second p -value measures how much additional explanation is gain from adding U2 to the model which just contains Ex0 and X. This just reflects the fact that U2 becomes more important once Ed and Age are added to the model.

iv) R-square = 0.7296, which is not bad for data of this type. All the variables are significant at least at the 0.05 level. So it appears that this model could be useful for prediction. Note however that coefficients for the five independent variables are all positive. This means that the model will predict an increase in the level of crime when any of these variables increase. Standard wisdom might accept this for Age, U2 and X. But do we believe that an increase in Ed or Ex0 would cause crime to increase?

■

6. An experiment was carried out to study the effectiveness of three methods of teaching reading. In a large school Sixty-six students were randomly divided into three groups of 22 students and one method was assigned to each group. At the end of the term each student was given a test to measure their reading level. Their score on the test, which was denoted by post1, was the response variable. An one way analysis of variance was run on this data and the results are given below.

```
groups<-gl(3,22,66)
> anova(lm(post1~groups))
```

Analysis of Variance Table

```
Response: post1
          Df Sum Sq Mean Sq
groups    2  108.12   54.06
Residuals 63  640.50   10.17
```

The means for the three methods are:

```
[1] 6.681818 [2] 9.772727 [3] 7.772727
```

Each student was also given the reading test before the term. Their scores on this pre-test were denote by pre1. The dummy variables

```
d1<-c(rep(1,22),rep(0,44))
d2<-c(rep(0,22),rep(1,22),rep(0,22))
```

were created and the model

$$post1 = \beta_0 + \beta_1 pre1 + \beta_2 d1 + \beta_3 (d1 \times pre1) + \beta_4 d2 + \beta_5 (d2 \times pre1) + Z \quad (1)$$

was fit to the data. Two anova tables for this model were also found. Finally a plot of post1 against pre1 was also constructed.

```
exp.lm<-lm(post1~pre1 + d1 + d1*pre1 + d2 + d2*pre1)
```

Residuals:

Min	1Q	Median	3Q	Max
-5.283	-1.900	0.413	1.648	4.417

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.2377	1.5255	-0.156	0.877
pre1	0.8768	0.1572	5.576	6.19e-07 ***
d1	2.0233	2.4572	0.823	0.414
d2	3.3302	2.4885	1.338	0.186
pre1:d1	-0.4105	0.2366	-1.735	0.088 .
pre1:d2	-0.1900	0.2506	-0.758	0.451

Residual standard error: 2.408 on 60 degrees of freedom
Multiple R-Squared: 0.5352,
F-statistic: 13.82 on 5 and 60 DF, p-value: 5.596e-09

Analysis of Variance Table

Response: post1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
pre1	1	239.74	239.74	41.3386	2.330e-08 ***
d1	1	115.81	115.81	19.9691	3.550e-05 ***
d2	1	27.64	27.64	4.7655	0.03296 *
pre1:d1	1	14.13	14.13	2.4357	0.12386
pre1:d2	1	3.34	3.34	0.5751	0.45122
Residuals	60	347.97	5.80		

Response: post1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
d1	1	64.12	64.12	11.0563	0.001511 **
d2	1	44.00	44.00	7.5869	0.007771 **
pre1	1	275.07	275.07	47.4299	3.899e-09 ***
d1:pre1	1	14.13	14.13	2.4357	0.123859
d2:pre1	1	3.34	3.34	0.5751	0.451224
Residuals	60	347.97	5.80		

Use this output to answer the following questions. If you are unable to answer a part of question because of missing output explain what additional information is needed.

i) At level $\alpha = .05$ test the hypothesis that there is no differences between the average effectiveness of three methods against the alternative that there are differences.

ii) If μ_1 is the true average or mean response under method 1 and μ_2 is the true average for method 2 find a 95% confidence interval for $\mu_1 - \mu_2$.

iii) In the full model given by 1 test $H : \beta_1 = \beta_2 = 0$ against the alternative K at least one of them is not zero at level $\alpha = 0$.

iv) How can one predict a potential student's post1 score before they are taught by one of the three methods? Does this experiment help one decide which method to use with such a student? Briefly justify your answer.

Solution:

i) Using the information from the anova table we find

$$\frac{108.12/2}{640.50/63} = 5.32 > 3.14 = f_{.05,2,63}$$

and so we reject H .

ii) Since $t_{63,.025} = 1.998$ and the residual mean Sq is just our pooled estimate of the variance the interval is

$$6.68 - 9.77 \pm (1.998)\sqrt{(2 \times 10.17)/22}$$

iii) Cannot do. You need an anova table where these two variables come last.

iv) Part i) tells us that the second method is best. But the output from model(1) also tells us that pre1, a student's score on pre-test, is also a very good predictor of their final score. The two anova tables and the plot strongly suggest that β_3 and β_5 are not needed in the model. So our prediction for a student taught with method two and pre-test score pre1 would be

$$-0.2377 + 0.8768pre1 + 3.3302 \times 1$$

Of course we should recalculate the parameter estimates with β_3 and β_5 omitted from the model.

■

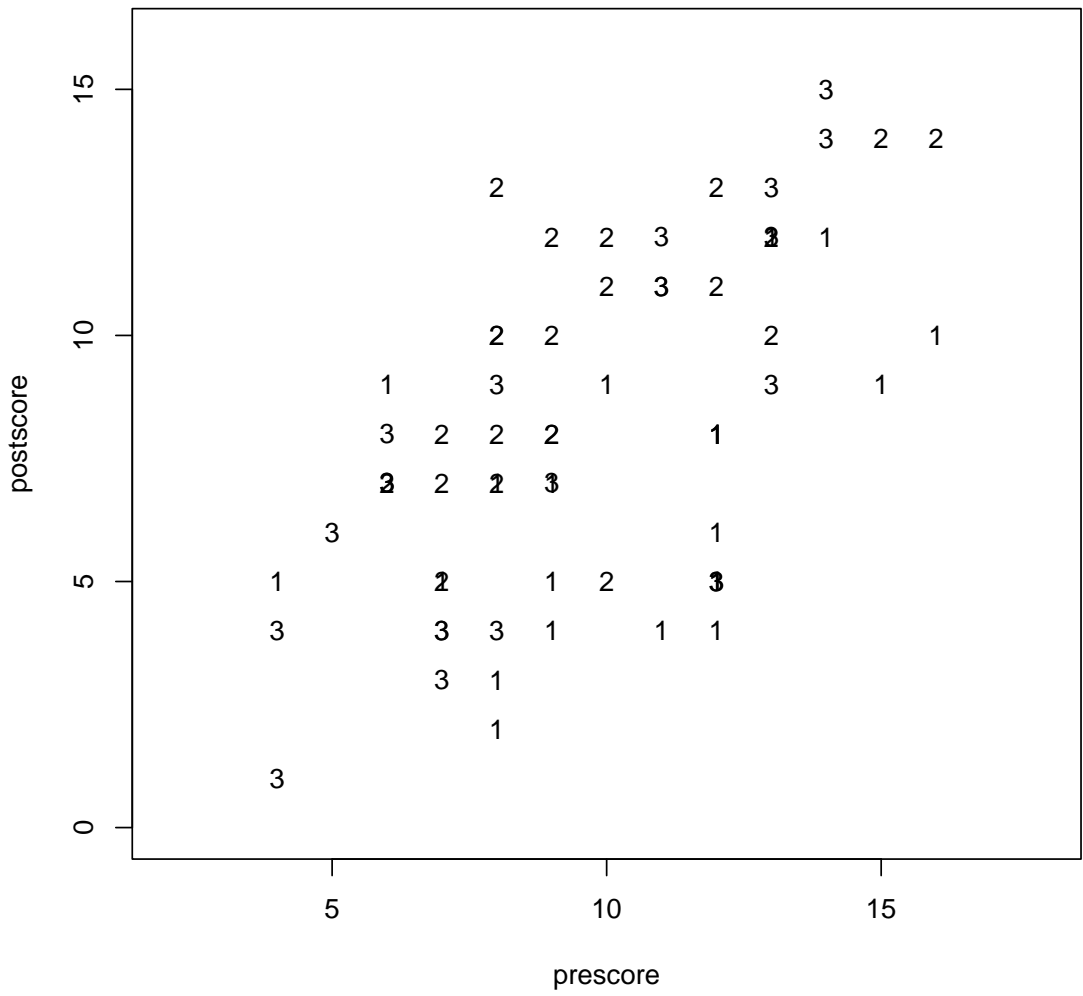


Figure 1: For each of the three methods the plot of students post training scores, post1, against their pre training scores, pre1. The 22 students who received method one are denoted by 1's, the second group by 2's and the third by 3's.