Stat 135, Fall 2006 A. Adhikari

HOMEWORK 9 SOLUTIONS

1. The boxplots indicate that the means are different, and also that the variances look different. So I'll use `oneway.test` without assuming equal variances. Read the `dogs` dataset into R, and then:

> dogs ← t(dogs)
> dogs1 ← c(dogs[,1], dogs[,2], dogs[,3])
> groups ← c(rep(1, 10), rep(2, 10), rep(3, 10))
> dogs ← cbind(dogs1, groups)
> oneway.test(dogs1 ~ groups, data=dogs)
> kruskal.test(dogs1 ~ groups, data=dogs)

The F-test of "all three underlying means equal" versus "the underlying means are not all equal" results a p-value of just around 5.5%. The non-parametric test of "all three underlying distributions are the same" versus "the distributions are not all the same" has a p-value of about 5.9%, so the nonparametric result is consistent with the parametric one. **Don't** be tied like a slave to the 5% cutoff. These p-values are on the small side and it's hard to have much faith in the null hypothesis of no difference.

If you did the anova assuming equal variances, you'll get a p-value of about 1%. Since I don't see any *a priori* reason to assume equal variances, and the plots seem to indicate unequal variances, I prefer to use the other two tests.

2. The boxplots indicate unequal means and unequal variances; the sample sizes are large enough that we should be able to pick up the differences. The simplest way to use Bonferroni's method is to construct three simultaneous 95%-confidence intervals for the three population means. (No, there's nothing saying you have to do 95% intervals. Do 99% intervals if you like. Just say what you're doing.)

So you will make three 98.33% intervals, one for each mean. In two case you'll be using the $t_5 2$ and in the remaining case the $t_2 7$. The appropriate values of t are respectively about 2.474 and 2.552 respectively. So the confidence interval for the mean of C57 is $231.6038 \pm 2.474 \times 67.53584/\sqrt{53}$ which is $(208.65, 254.55)$. The other two are $(30.08, 78.64)$ and $(95.75, 164.67)$. The three intervals are completely distinct; none overlaps either of the others. I conclude that the three means are different.

3. The boxplots indicate possibly unequal variances and possibly equal means. With such small samples it will be very hard to detect differences between the underlying means. Indeed, the one-way anova (assuming unequal underlying SDs) gives a p-value of over 60%, supporting the null hypothesis of equal underlying means. The nonparametric test has a p-value of 34%, supporting the hypothesis of the same underlying distribution. Even though the SDs appear unequal, we have such small samples that any estimates are going to have very large standard errors. So the hypothesis of "same distribution" is not rejected.

4. It is important to remember that a random variable in standard units has expectation 0, variance 1, and expected square 1:

$$E(X^*) = \frac{E(X) - E(X)}{SD(X)} = 0 \qquad Var(X^*) = \frac{Var(X)}{(SD(X))^2} = 1 \quad \Rightarrow \quad E(X^{*2}) = 1$$

It is also important to note that $R(X, Y) = E(X^*Y^*)$ by definition.

a) The first equality is true because multiplication is commutative. Next, note that Since $E(X^*) = E(Y^*) = 0$ and $SD(X^*) = SD(Y^*) = 1$, the definition says that $R(X^*, Y^*) = E(X^*Y^*) = R(X, Y)$.

b) Using the hint,

$$E[(X^* + Y^*)^2] \geq 0 \quad \Rightarrow \quad E[X^{*2} + 2X^*Y^* + Y^{*2}] = 1 + 2\rho + 1 \geq 0 \quad \Rightarrow \quad \rho \geq -1$$

$$E[(X^* - Y^*)^2] \geq 0 \quad \Rightarrow \quad E[X^{*2} - 2X^*Y^* + Y^{*2}] = 1 - 2\rho + 1 \geq 0 \quad \Rightarrow \quad \rho \leq 1$$

5. Use the fact that $\sigma^2 = \frac{n-1}{n}s^2$.

$$r = \frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{s_x\sqrt{\frac{n-1}{n}}}\right)\left(\frac{y_i - \bar{y}}{s_y\sqrt{\frac{n-1}{n}}}\right) = \frac{1}{n}\frac{n}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

6a. By definition, $r = Cov(x, y)/(\sigma_x\sigma_y)$. So $Cov(x, y) = r\sigma_x\sigma_y$. In class we derived a formula for the slope in terms of the covariance, and so:

$$\hat{b} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{Cov(x, y)}{\sigma_x^2} = \frac{r\sigma_x\sigma_y}{\sigma_x^2} = r\frac{\sigma_y}{\sigma_x} = r\frac{s_y}{s_x}$$

b) The formula $\hat{a} = \bar{y} - \hat{b}\bar{x}$ for the intercept was derived in class.

The equation of the regression line is $\hat{y} = \bar{y} - \hat{b}\bar{x} + \hat{b}x$. If you plug in $\bar{x}$ for x then you find $\hat{y} = \bar{y}$ and so the line goes through the point $(\bar{x}, \bar{y})$.

c) Look at the equation of the line in the previous part:

$$\hat{y} - \bar{y} = \hat{b}(x_i - \bar{x}) \quad \Rightarrow \quad \hat{y} - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x_i - \bar{x}) \quad \Rightarrow \quad \left(\frac{\hat{y} - \bar{y}}{\sigma_y}\right) = r\left(\frac{x - \bar{x}}{\sigma_x}\right)$$

d) If a student is 1.5 SDs above average on the midterm, then the student's midterm score in standard units is 1.5. Therefore by **c** the predicted final score in standard units will be $1.5r$. For midterm and final scores r will be a number strictly between 0 and 1 (because you expect the association to be positive but not perfectly linear). So the predicted final score will be less than 1.5 SDs above average.

Similarly if the student's midterm score in standard units is -1.5, then the predicted final score will be $-1.5r$ in standard units, that is, fewer than 1.5 SDs below the mean.

7a)

$$\frac{1}{n}\sum\hat{y}_i = \frac{1}{n}\sum(\hat{a} + \hat{b}x_i) = \hat{a} + \hat{b}\bar{x} = \bar{y} - \hat{b}\bar{x} + \hat{b}\bar{x} = \bar{y}$$

b) By the definition of variance,

$$\sigma_{\hat{y}}^2 = \frac{1}{n}\sum(\hat{y}_i - \bar{\hat{y}})^2 = \frac{1}{n}\sum(\hat{y}_i - \bar{y})^2$$

by the result of **a**. So

$$\sigma_{\hat{y}}^2 = \frac{1}{n}\sum(\hat{a} + \hat{b}x_i - \bar{y})^2 = \frac{1}{n}\sum(\bar{y} - \hat{b}\bar{x} + \hat{b}x_i - \bar{y})^2 = \frac{1}{n}\sum(\hat{b}(x_i - \bar{x}))^2 = \hat{b}^2\sigma_x^2 = r^2\sigma_y^2$$

When $r = 0$ the regression line is flat; its equation is $\hat{y} = \bar{y}$. So fitted values are all the same and hence their variance is zero.

When $r = 1$ then the points all fall exactly on a line, which must be the same as the regression line. Hence the fitted values are the same as the observed values of y. So the variances are identical.

8a.

$$\bar{\hat{e}} = \frac{1}{n}\sum(\hat{y}_i - y_i) = \bar{\hat{y}} - \bar{y} = \bar{y} - \bar{y} = 0$$

b) Because $\bar{\hat{e}} = 0$,

$$\sigma_{\hat{e}}^2 = \frac{1}{n}\sum \hat{e_i}^2 = \frac{1}{n}\sum(\hat{y}_i - y_i)^2 = \frac{1}{n}\sum(\bar{y} - \hat{b}\bar{x} + \hat{b}x_i - y_i)^2 = \frac{1}{n}\sum[\hat{b}(x_i - \bar{x}) - (y_i - \bar{y})]^2$$

Expand the square to see that the expression becomes

$$\hat{b}^2\sigma_x^2 - 2\hat{b}r\sigma_x\sigma_y + \sigma_y^2 = \sigma_y^2(r^2 - 2r^2 + 1) = (1 - r^2)\sigma_y^2$$

When $r = 0$ we have the constant prediction $\bar{y}$, no matter what the value of x. The residuals are then just the deviations of the data (y) from their average ($\bar{y}$) and therefore the mean squared residual is just the variance σ_y^2.

When $r = 1$ then all the points lie on the regression line and the residuals are all zero. Hence their variance is also zero.

9a. Add the results of **7b** and **8b**:

$$\sigma_{\hat{e}}^2 + \sigma_{\hat{y}}^2 = (1 - r^2)\sigma_y^2 + r^2\sigma_y^2 = \sigma_y^2$$

b) This is the same equality as in **a**. Just multiply each element in **a** by $1/n$.

10a. This is just **9b** restated.

$$(y_i - \bar{y})^2 = [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 = (y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2$$

Sum both sides and notice that the result of **9b** implies that the cross-product term must be 0.

b) "Uncorrelated" means "correlation = 0" which is equivalent to "covariance = 0".

Part **a** shows that the residuals are uncorrelated with the fitted values. Since the fitted values are a linear function of x, the residuals must also be uncorrelated with x.

In order to prove this by algebra it's enough to show that $Cov(x, \hat{e}) = 0$, as follows.

$$Cov(x, \hat{e}) = \frac{1}{n}\sum(x_i - \bar{x})(\hat{e}_i - \bar{\hat{e}}) = \frac{1}{n}\sum(x_i - \bar{x})(\hat{y}_i - y_i) = \frac{1}{\hat{b}n}\sum(\hat{b}x_i - \hat{b}\bar{x})(\hat{y}_i - y_i)$$

By a calculation that should now be familiar, the first term in the product is $\hat{y}_i - \bar{y}$. And therefore

$$Cov(x, \hat{e}) = \frac{1}{\hat{b}n}\sum(\hat{y}_i - \bar{y})(\hat{y}_i - y_i) = 0$$

by the result of **a**.