Statistics and Bioinformatics
Problem Set 13
Due in class Tuesday, January 11, 2011

For the following questions, assume we're referring to the true or population covariance.

1) The covariance between two variables is a measure of how changes in one variable ___________ (cause, are related to) changes in the other variable.

2) The covariance between x and y is _________ (equal to, greater than, less than) the covariance between y and x.

3) The covariance between a variable x and a constant is equal to _______ (one, zero, the variance of x).

4) The covariance between a variable x and itself is equal to ________ (1, zero, the variance of x).

5) The covariance between a variable x and x multiplied by a constant a is equal to ___________ (a, a^2) times __________ (var(x), sd(x)).

6) Write an equation for the covariance between x and y in terms of the mean x and the mean y.

$$Cov(x,y) = E(x-\bar{x})(y-\bar{y})$$

7) What is the maximum possible covariance between x and y?

$S_x S_y$

8) What is the minimum possible covariance between x and y?

$-S_x S_y$

9) If two variables are independent, then their covariance is ________ (zero, 1, -1).

10) If the covariance between two variables is 0, then the two variables are _________ (independent, can't tell for sure).

11) If the points plotted in a scatterplot between two variables fall on a straight line, then their covariance is equal to the _________ (product, sum) of their ________ (standard deviations, variances).

12) The covariance between x and ax, where a is a constant, is equal to a Var(x).

13) The covariance between x and ay, where a is a constant and x and y are any variables, is equal to a Cov(x, y).

14) The correlation coefficient between x and itself is equal to ________ (1, 0, -1).

15) If the points plotted in a scatterplot between two variables fall on a straight line, then the correlation coefficient between the two variables is equal to 1 (1, 0, 1).

16) The correlation coefficient between x and y is equal to the covariance between x and y ________ (divided by, multiplied by) the product of the ________ (standard deviations, variances) of x and y.

17) The maximum possible value of the correlation coefficient is 1.

18) The minimum possible value of the correlation coefficient is -1.

19) If two variables are independent, their correlation coefficient is 0.

20) If the correlation coefficient of two variables is zero, then we _________ (can, cannot) conclude that they are independent.

21) If the correlation coefficient of two variables is close to 1, then we ________ (can, cannot) conclude that changes in one variable cause changes in the other variable.

22) The best fit line relating two variables is the line that ________ (minimizes, maximizes) the residual _________ (sum of squares, covariance, correlation).

23) The residual sum of squares is the sum of the square deviations between each point and the best-fit line along the ________ (x, y) axis.

24) The residual degrees of freedom in a linear regression is equal to the total number of data points ________ (plus, minus) ________ (1, 2, 3, 4, 5).

25) The residual mean squares is equal to the residual ________ (SS, df) divided by the residual ________ (SS, df).

26) The residual sum of squares in a linear regression is 234, and the number of data points is 223. What is the residual mean squares? $MS_{res} = \frac{234}{221} = 1.059$

27) The line sum of squares in a linear regression is 45. What is the line mean squares? $M_{line} = 45$

28) For the previous two questions, calculate the F ratio and its tail probability. Is the regression significant? $F = 45/1.059 = 42.49$ $p = 1 - pf(42.49, 1, 221)$ $= 4.8 \times 10^{-10}$

yes, Significant

29) The sum of squares for the line in a linear regression between x and y is the sum of square deviations between each __________ (observed, predicted) y value and the mean of the ________ (x, y) values.

30) The degrees of freedom for the line in a linear regression between x and y is ________ (1, 2, 3, 4, 5).

31) The total degrees of freedom [redacted] in a linear regression between x and y is the total number of points ________ (minus, plus) ________ (1, 2, 3, 4, 5).

32) The slope of the best fit line on a scatter diagram relating two variables x (on the horizontal axis) and y (on the vertical axis) is equal to the ratio of ________ (covariance between x and y, variance in x, variance in y) to ________ (covariance between x and y, variance in x, variance in y).

33) For each of the following, calculate the slope of the best fit line in a plot of x against y, and the correlation coefficient of x and y.

a) The covariance between x and y is 8. The variance in x is 2. The variance in y is 4.

$\text{Slope} = \frac{8}{2} = 4$

$r = \frac{8}{\sqrt{2 \cdot 4}} = \frac{8}{\sqrt{8}} > 1 \therefore$ impossible!

b) The covariance between x and y is 132. The variance in x is 23. The variance in y is 30.

$$\text{Slope} = 132/23 = 5.7$$

$$r = 132/\sqrt{23 \cdot 30} > 1 \quad \therefore \text{ impossible!}$$

c) The covariance between x and y is 25. The variance in x is 25. The variance in y is 25.

$$\text{Slope} = 25/25 = 1$$

$$r = 25/\sqrt{25^2} = 1$$

d) The covariance between x and y is 10. The variance in x is 25. The variance in y is 4.

$$\text{Slope} = 10/25 = 0.4$$

$$r = 10/\sqrt{100} = 1$$

34) For the following examples, present an ANOVA table that tests the significance of the linear regression, and state how much variation in y is explained by the line.

a) The line sum of squares is 4.4, and the residual sum of squares is 2541. The number of data points is 100.

$$df_{res} = 98 \Rightarrow MS_{res} = \frac{2541}{98} = 25.9$$

$$df_{line} = 1 \Rightarrow MS_{line} = 4.4$$

$$F = \frac{MS_{line}}{MS_{res}} = \frac{4.4}{25.9} = 0.170$$

$$p = 1 - pf(0.17, 1, 98) = 0.68$$

$$R^2 = SS_{line}/SS_{tot} = 4.4/(2541+4.4) = 0.0017$$

	df	SS	MS	F	p
line	1	4.4	4.4	0.17	0.68
resid	98	2541	25.9		

b) The line sum of squares is 5.1, and the residual sum of squares is 231. The number of data points is 93.

$df_{res} = 91$, $df_{line} = 1$

$F = \frac{5.1}{2.54} = 2.01$

$MS_{res} = \frac{231}{91} = 2.54$

$p = 1 - pf(2.01, 1, 91) = 0.15$

$MS_{line} = 5.1$

$R^2 = 5.1 / (231 + 5.1) = 0.02$

	df	SS	MS	F	p
line	1	5.1	5.1	2.01	0.15
resid	91	231	2.54		

Present your results for the following problem in a Word file emailed to the assistant tosaric@unizd.hr.

35) For the following data points, perform a linear regression. Calculate the mean x, the mean y, the covariance of x and y, the variance of x and y, the correlation coefficient, the sum of squares and mean squares of the line and residuals, the coefficient of determination, and present the ANOVA table. State whether the regression is significant, and state how much of the variation in y is explained. Include in your document a scatter plot of x against y, with the best-fit line drawn through the plotted points.

a) x: number of cricket chirps per second: 20, 16, 19.8, 18.4, 17.1, 15.5, 14.7, 17.1, 15.4, 16.2, 15, 17.2, 16, 17, 14.4.
y: air temperature: 31.4, 22.0, 34.1, 29.1, 27.0, 24.0, 20.9, 27.8, 20.8, 28.5, 28.1, 27.0, 28.6, 24.6

.

Let's say the number of chirps per second was observed to be 19.0. What do you predict the air temperature to be?