STA 2101 Assignment 3[1]

The questions on this assignment are not to be handed in. They are practice for Quiz Three on Friday October 4th. There is a posted formula sheet that will be provided with the quiz. The linear algebra questions are more review.

1. Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution with parameter θ.

   (a) Find the limiting distribution of

   $$Z_n = 2\sqrt{n}\left(\sin^{-1}\sqrt{\overline{X}_n} - \sin^{-1}\sqrt{\theta}\right).$$

   Hint: $\frac{d}{dx}\sin^{-1}(x) = \frac{1}{\sqrt{1-x^2}}$. The measurements are in radians, not degrees.

   (b) In a coffee taste test, 100 coffee drinkers tasted coffee made with two different blends of coffee beans, the old standard blend and a new blend. We will adopt a Bernoulli model for these data, with θ denoting the probability that a customer will prefer the new blend. Suppose 60 out of 100 consumers preferred the new blend of coffee beans. Using your answer to the first part of this question, test $H_0 : \theta = \frac{1}{2}$ using a variance-stabilized test statistic. Give the value of the test statistic (a number), and state whether you reject H_0 at the usual $\alpha = 0.05$ significance level. In plain, non-statistical language, what do you conclude? This is a statement about preference for types of coffee, and of course you will draw a directional conclusion if possible.

   (c) If the probability of an event is p, the *odds* of the event is (are?) defined as $p/(1-p)$. Suppose again that $X_1, \ldots, X_n$ are a random sample from a Bernoulli distribution with parameter θ. In this case the *log odds* of $X_i = 1$ would be estimated by

   $$Y_n = \log\frac{\overline{X}_n}{1-\overline{X}_n}.$$

   Naturally, that's the natural log. Find the approximate large-sample distribution (that is, the asymptotic distribution) of Y_n. It's normal, of course. Your job is to give the approximate (that is, asymptotic) mean and variance of Y_n.

   (d) Again using the Taste Test data, give a 95% confidence interval for the log odds of preferring the new brand. Your answer is a pair of numbers.

2. The label on the peanut butter jar says peanuts, partially hydrogenated peanut oil, salt and sugar. But we all know there is other stuff in there too. There is very good reason to assume that the number of rat hairs in a jar of peanut butter has a Poisson distribution with mean λ, because it's easy to justify a Poisson process model for how the hairs get into the jars. There is a government standard that says the true expected number of rat hairs in a 500g jar may not exceed 8 A sample of thirty 500g jars yields $\overline{X} = 9.2$.

   (a) State the model for this problem.

   (b) What is the parameter space Θ?

   (c) State the null hypothesis in symbols. Because nothing will happen if the number of rat hairs is significantly *less* than the standard, I think we need a one-sided test here.

   (d) Find a variance-stabilizing transformation for the Poisson distribution. You may use the fact that a Poisson has expected value and variance both equal to the parameter λ.

   (e) Using your variance-stabilizing transformation, derive a test statistic that has an approximate standard normal distribution under H_0.

   (f) Calculate your test statistic for these data. Do you reject the null hypothesis one-sided at $\alpha = 0.05$? Answer Yes or No.

   (g) In plain, non-statistical language, what do you conclude? Your answer is something about peanut butter and rat hairs.

3. If the $p \times 1$ random vector $\mathbf{x}$ has variance-covariance matrix $\boldsymbol{\Sigma}$ and $\mathbf{A}$ is an $m \times p$ matrix of constants, prove that the variance-covariance matrix of $\mathbf{Ax}$ is $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top$. Start with the definition of a variance-covariance matrix:

$$cov(\mathbf{Z}) = E(\mathbf{Z} - \boldsymbol{\mu}_z)(\mathbf{Z} - \boldsymbol{\mu}_z)^\top.$$

4. If the $p \times 1$ random vector $\mathbf{x}$ has mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, show $\boldsymbol{\Sigma} = E(\mathbf{x}\mathbf{x}^\top) - \boldsymbol{\mu}\boldsymbol{\mu}^\top$.

5. Let the $p \times 1$ random vector $\mathbf{x}$ have mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, and let $\mathbf{c}$ be a $p \times 1$ vector of constants. Find $cov(\mathbf{x} + \mathbf{c})$. Show your work.

6. Let the $p \times 1$ random vector $\mathbf{x}$ have mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$; let $\mathbf{A}$ be a $q \times p$ matrix of constants and let $\mathbf{B}$ be an $r \times p$ matrix of constants. Derive a nice simple formula for $cov(\mathbf{Ax}, \mathbf{Bx})$.

7. Let $\mathbf{x}$ be a $p \times 1$ random vector with mean $\boldsymbol{\mu}_x$ and variance-covariance matrix $\boldsymbol{\Sigma}_x$, and let $\mathbf{y}$ be a $q \times 1$ random vector with mean $\boldsymbol{\mu}_y$ and variance-covariance matrix $\boldsymbol{\Sigma}_y$. Let $\boldsymbol{\Sigma}_{xy}$ denote the $p \times q$ matrix $cov(\mathbf{x}, \mathbf{y}) = E\left((\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{y} - \boldsymbol{\mu}_y)^\top\right)$.

   (a) What is the (i, j) element of $\boldsymbol{\Sigma}_{xy}$? You don't need to show any work; just write down the answer.

   (b) Find an expression for $cov(\mathbf{x} + \mathbf{y})$ in terms of $\boldsymbol{\Sigma}_x$, $\boldsymbol{\Sigma}_y$ and $\boldsymbol{\Sigma}_{xy}$. Show your work.

   (c) Simplify further for the special case where $Cov(X_i, Y_j) = 0$ for all i and j.

   (d) Let $\mathbf{c}$ be a $p \times 1$ vector of constants and $\mathbf{d}$ be a $q \times 1$ vector of constants. Find $cov(\mathbf{x} + \mathbf{c}, \mathbf{y} + \mathbf{d})$. Show your work.

8. Let $\mathbf{x} = (X_1, X_2, X_3)^\top$ be multivariate normal with

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 0 \\ 6 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

   Let $Y_1 = X_1 + X_2$ and $Y_2 = X_2 + X_3$. Find the joint distribution of Y_1 and Y_2.

9. Let X_1 be Normal(μ_1, σ_1^2), and X_2 be Normal(μ_2, σ_2^2), independent of X_1. What is the joint distribution of $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$? What is required for Y_1 and Y_2 to be independent? Hint: Use matrices.

10. Show that if $\mathbf{w} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ positive definite, $Y = (\mathbf{w} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{w} - \boldsymbol{\mu})$ has a chi-squared distribution with p degrees of freedom.

11. You know that if $\mathbf{w} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{A}\mathbf{w} + \mathbf{c} \sim N_r(\mathbf{A}\boldsymbol{\mu} + \mathbf{c}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$. Use this result to obtain the distribution of the sample mean under normal random sampling. That is, let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. Find the distribution of $\overline{X}$. You might want to use $\mathbf{1}$ to represent an $n \times 1$ column vector of ones.

12. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$.

   (a) Show $Cov(\overline{X}, (X_j - \overline{X})) = 0$ for $j = 1, \ldots, n$.

   (b) Why does this imply that if $X_1, \ldots, X_n$ are normal, $\overline{X}$ and S^2 are independent?

13. Consider the usual multiple regression model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is an $n \times p$ matrix of known constants with linearly independent columns, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants, and $\boldsymbol{\epsilon}$ is multivariate normal with mean zero and covariance matrix $\sigma^2\mathbf{I}_n$. The constant $\sigma^2 > 0$ is unknown. We have $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$, $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ and $\mathbf{e} = (\mathbf{y} - \widehat{\mathbf{y}})$

    (a) Show $\mathbf{X}^\top\mathbf{e} = \mathbf{0}$.

    (b) If the model has an intercept, why does this last result show that the sum of residuals equals zero?

    (c) Let $\mathbf{1}$ denote an $n \times 1$ column of ones, and let $\overline{\mathbf{x}} = \frac{1}{n}\mathbf{X}^\top\mathbf{1}$.

        i. What are the dimensions of the matrix $\overline{\mathbf{x}}$?

        ii. If the model has an intercept, what is the first element of $\overline{\mathbf{x}}$?

        iii. What is the second element of $\overline{\mathbf{x}}$?

    (d) We are interested in the predicted value of y (height of the least-squares regression plane) when all the explanatory variables are set to their sample mean values. Express this quantity in terms of $\overline{\mathbf{x}}$ and $\widehat{\boldsymbol{\beta}}$.

    (e) Assuming the model has an intercept, simplify your answer to the last question. What do you get?

14. High School History classes from across Ontario are randomly assigned to either a discovery-oriented or a memory-oriented curriculum in Canadian history. At the end of the year, the students are given a standardized test and the median score of each class is recorded. Please consider a regression model with these variables:

    X_1 Equals 1 if the class uses the discovery-oriented curriculum, and equals 0 if the class uses the memory-oriented curriculum.

    X_2 Average parents' education for the classroom.

    X_3 Average family income for the classroom.

    X_4 Number of university History courses taken by the teacher.

    X_5 Teacher's final cumulative university grade point average.

    Y Class median score on the standardized history test.

    The full regression model (as opposed to the reduced models for various null hypotheses) implies

    $$E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5.$$

    For each question below, please give

    - The null hypothesis in terms of β values.
    - $E[Y|X]$ for the reduced model you would use to answer the question. Don't re-number the variables.

(a) If you allow for parents' education and income and for teacher's university background, does curriculum type affect test scores? (And why is it okay to use the word "affect?")

(b) Controlling for parents' education and income and for curriculum type, is teacher's university background (two variables) related to their students' test performance?

(c) Correcting for teacher's university background and for curriculum type, are parents' education and family income (considered simultaneously) related to students' test performance?

(d) Taking curriculum type, teacher's university background and parents' education into consideration, is parents' income related to students' test performance?

(e) Here is one final question. Assuming that $X_1, \ldots, X_5$ are random variables (and I hope you agree that they are),

   i. Would you expect X_1 ro be related to the other explanatory variables?

   ii. Would you expect the other explanatory variables to be related to each other?