STAT 350

Assignment 1 Solutions

1. An environmental agency sets a standard of 200 ppb for the concentration of cadmium in a lake. The concentration of cadmium in one lake is measured 17 times. The measurements average 211 parts per billion with an SD of 15 parts per billion. Could the real concentration of cadmium be below the standard of 200 ppb?

Solution

Statisticians approach this question as a hypothesis testing problem. It calls for a yes or no answer but we give an answer which is more like "probably not". First we ask how the data were collected so we know what was measured. We hope that the samples were collected in such a way that they can be treated as a simple random sample from a population whose mean value μ is the concentration of cadmium in the lake. We leave it up to the subject area experts to make sure this is true but when consulting we ask questions about it. If so we have a sample of $n = 17$ measurements from a population with mean μ. We observe $\bar{x} = 211$ and $s = 15$.

The question is: is μ above 200 or not? That means the problem is one sided and we have to have either $H_o : \mu \leq 200$ or $H_o : \mu \geq 200$. If we test the former and reject the null hypothesis we would conclude "No the real concentration is (probably) not below 200". If we test the hypothesis $\mu \geq 200$ and accept the null our conclusion is "we have little evidence agains the assertion that $\mu \geq 200$" which is far from providing a definitive answer.

To test the hypotheses we compute

$$t = \frac{211 - 200}{15/\sqrt{17}} = 3.02$$

For the null hypothesis $H_o : \mu \leq 200$ we get a P-value from t tables with 16 degrees of Freedom. I get 0.004 which is very strong evidence against this null and conclude "No, almost certainly not." If you test the other way the P value is 0.996 and the conclusion is far weaker: "I see very little evidence against the assertion that $\mu \geq 200$." Statisticians have a duty to do their best to answer the question asked so the former answer is far better.

2. Consider a population of 200 million people of whom 200 thousand have a certain condition. A test is available with the following properties. Assuming that a person has the condition the probability that the test detects the condition is 0.9. Assuming that a person does not have the condition the test detects (incorrectly) the condition with probability 0.001.

   A person is picked at random from the 200 million people and the test is administered.

   (a) What is the chance that the test detects the condition for this randomly selected person?

   Solution

   I'll use C for the event 'has the condition' and D for 'tests positive' meaning 'test detects the condition'. We are told

$$
\begin{aligned}
P(C) &= \frac{200000}{200000000} = 0.001 \\
P(D|C) &= 0.9 \\
P(D|C^c) &= 0.001
\end{aligned}
$$

   So

$$
\begin{aligned}
P(D) &= P(D \cap C) + P(D \cap C^c) \\
&= P(D|C)P(C) + P(D|C^c)P(C^c) \\
&= 0.9 \times 0.001 + 0.001 \times 0.999 \\
&= 0.001899
\end{aligned}
$$

   (b) Assuming that the condition is detected by the test for this randomly selected person what is the chance that the person has the condition?

   Solution

We use Bayes' Theorem to compute $P(C|D)$ and get

$$\begin{aligned} P(C|D) &= \frac{P(D|C)P(C)}{P(D|C)P(C) + P(D|C^c)P(C^c)} \\ &= \frac{0.9 \times 0.001}{0.9 \times 0.001 + 0.001 \times 0.999} \\ &= \frac{0.0009}{0.001899} \\ &= 0.4739 \end{aligned}$$

(c) A mandatory testing program is contemplated. If all 200 million are tested about how many positive results should be expected? Of these about how many will not have the condition?

Solution

About 200,000,000 times 0.01899 for the first question and about 0.4739 times that for the second. Nearly half of the diagnoses in such a mandatory testing program would be wrong. The numbers are realistic for some tests and conditions.

3. In a study of dietary fat intake 1000 father-son pairs were examined. (Both father and son were adults in all pairs.) For each person the percentage of dietary calories received in the form of fat is measured. The fathers received an average of 35% of their dietary calories in the form of fat with an SD of 6 percentage points. The sons received an average of 30% of their dietary calories in the form of fat with an SD of 8 percentage points. The correlation between father and son was $r = 0.4$.

   (a) About what percentage of the fathers receive more than 40% of their daily caloric intake in the form of fat? Be clear about any assumption you must make to do the problem.

   A: *in order to do this problem it is necessary to assume that the histogram of father's daily caloric intake is roughly normal. Convert 40 to standard units to get*

$$z = \frac{40 - 35}{6} = 0.83$$

Then look up the area to the left of z and get

$$0.7967 \approx 79.7\%$$

The desired area is the area to the right of z which is 1-0.7967 or about 20.3%. I am not concerned with how many digits students round off to.

(b) If a father receives 28% of his calories in the form of fat about what percentage should you predict for the son?

A: *This is a regression problem. Students may proceed in one of two ways:*

i) The estimated slope is

$$b = rs_y/s_x = 0.4\frac{8}{6} = 0.5333$$

The intercept is

$$a = \bar{y} - b\bar{x} = 30 - 0.5333 \times 35 = 11.33$$

Then plug in $x = 28$ to the equation $y = a = bx$ to get the predicted value

$$\hat{y} = 26.26$$

*ii) Alternatively: convert 28 to standard units to get (28-35)/6 = -1.17. Then multiply by r to get -0.467. Convert back to standard units for sons: 30-0.4667*8 =26.27.*

The differences are unimportant round off errors.

Take significant marks away for confusing the two standard deviations or the two averages. This problem is easy except for figuring out which number is which.

(c) Suppose we select 25 families at random from this group of 1000 for a more detailed dietary assessment. You may assume that they are selected with replacement so that the selections are independent. What is the chance that the average percentage of daily dietary calories taken in the form of fat for the 25 selected fathers have comes out between 34 and 36%?

A: *This is a question about the sampling distribution of the mean of a sample of $n = 36$. Students should indicate that they need* *to assume that 36 is not too large compared to 100 or that they assumed we were sampling with replacement.*

The mean of the sampling distribution of $\bar{x}$ is $\mu = 354$. The SD is $\sigma/\sqrt{n} = 6/6 = 1$.

Now convert 34 and 36 to standard deviation units to get -1 and 1. The area between these is 0.6826 or roughly 68.3%

Students who use $n = 25$ and so get 1.2 instead of 1 for standard deviation will have -0.83 and 0.83 so get an area of 0.5935.

(d) In 50 of the families the father received less than 15% of his daily caloric intake in the form of fat. If we eliminate this group of 50 father-son pairs from our study will the correlation coefficient go up or down; that is, is the correlation coefficient for the other 950 pairs more than 0.4, less than 0.4, or still about 0.4? Explain with a graph.

A: *The correlation will go down if you chop off the left part of the scatterplot; the new plot will be less bunched up around the regression line relative to the spread in the x direction.*

Students' answers should include a graph.

(e) Consider the families where the father receives about 28% of his calories in the form of fat. Approximately what would be the standard deviation of the sons' percentage of daily caloric intake in the form of fat in these families?

A: $\sqrt{1-r^2}s_S$ *or* $\sqrt{1-0.4^2} \times 8$ *or 7.33.*

4. An executive of a large supermarket chain discovers that the correlation between the total amount of overtime worked by cashiers at a store and the total number of bad cheques accepted at a store is 0.7. He recommends that a ban be placed on overtime, arguing that cashiers at the end of a long day are less careful. What is wrong with this thinking; your answer should include an alternative explanation of the observed correlation.

Solution

A: *This correlation is probably largely produced by the fact that if the rate at which bad checks are presented is constant then workers doing* *more hours will see more cheques and get more bad ones. You need to compare acceptance rates (cheques per hour worked) with hours worked by individual cashiers to see if any change in overtime hours might be useful.*

5. A large bank has loans outstanding on 100,000 pieces of real estate. At the last audit the average assessed value of the pieces of real estate was \$150,000. The bank suspects that recent economic events mean that the real estate values may have fallen suddenly. A simple random sample of 400 of the outstanding loans shows an average present value of \$139,600 with an SD of \$160,000. Looking more closely at the data collected the bank president goes through the files to find the assessed values of the 400 sampled pieces of real estate at the time of the last audit. The figures average \$145,000 with an SD of \$165,000. He discovers, however, that those properties valued at over \$400,000 at the last audit have decreased in value \$20,000 each on average while those valued at under \$100,000 have not decreased in value at all on average. He develops the following explanation of this observation. Owners of expensive properties have had to sell them. They have then taken the proceeds of the sales and bought less expensive properties thereby keeping up the prices of these properties. Identify a pitfall in the executive's reasoning.

Solution

A: *The effect noted here is the regression effect. The executive picked out properties which are above average in value at last audit. In essentially any scatterplot with a positive relation between x and y the units on the right hand side (above average in x) will be above average in y, but not so much in terms of standard units.*

6. Suppose $Z_1, \ldots, Z_{10}$ are independent random variables each having a $N(0,6)$ distribution. Let $\bar{Z} = \sum Z_i/10$, $U = \sum_{i=1}^{4} Z_i^2/6$, $V = \sum_{i=5}^{10} Z_i^2/6$, $X = Z_1/\sqrt{V}$ and $Y = (3U)/(2V)$. Give the names for the distributions of each of $\bar{Z}$, U, V, X and Y and use tables to find $P(|\bar{Z}| > 1)$, $P(U \leq 9.49)$, $P(-1.2 \leq X \leq 1.2)$, $P(Y > 6.23)$, $P(U \leq V)$, $P(Z_1 - 2Z_2 \geq 10)$.

Solution

$\bar{Z}$ is $N(0, 6/10)$, U is χ^2_4, V is χ^2_6, X is t_6, since Z_1 and V are independent and $Z_1/\sqrt{V} = (Z_1/\sqrt{6})/\sqrt{V/6}$, and $Y = (U/4)/(V/6)$ has an $F_{4,6}$ distribution, using the fact that U and V are independent. Your answer should specify the mean and variance for $\bar{Z}$, the various degrees of freedom and note the required independence for X and Y. The required probabilities are 0.197, 0.95, 0.725, 0.025, 0.688, 0.034. I used SPlus to compute these; if you used tables your answers will be less accurate, particularly for $P(U \leq V) = P(U/V \leq 1) = P(Y < 1.5)$.

You need to remember:

(a) $Z_i/\sqrt{6}$ has a $N(0, 1)$ distribution.

(b) so $Z_i^2/6$ has a χ^2_1 distribution.

(c) $Var(Z_1 - 2Z_2) = 6 + (-2)^2 6 = 30$.

(d) $Var(Z_i) = 6$ (several people thought $Var(Z) = 6^2$).

7. A new process for measuring the concentration of a chemical in water is being investigated. A total of n samples are prepared in which the concentrations are the known numbers x_i for $i = 1, \ldots, n$; the new process is used to measure the concentrations for these samples. It is thought likely that the concentrations measured by the new process, which we denote Y_i, will be related to the true concentrations via

$$Y_i = \beta x_i + \epsilon_i$$

where the ϵ_i are independent, have mean 0 and all have the same variance σ^2 which is unknown.

(a) If this model is fitted by least squares, (that is by minimizing $\sum_i (Y_i - \beta x_i)^2$) show that the least squares estimate of β is

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2}.$$

Solution

Differentiate $\sum_i (Y_i - \hat{\beta} x_i)^2$ with respect to $\hat{\beta}$ and get $-2 \sum x_i (Y_i - \hat{\beta} x_i)$ which is 0 if and only if $\hat{\beta} = \sum x_i Y_i / \sum x_i^2$. The second derivative of the function being minimized is $2\sum x_i^2 > 0$ so this is a minimum.

Note: I sometimes see a dismaying tendency to write things like

$$x_i \sum (Y_i - x_i\beta)$$

moving the x_i inside and out of the sum whenever you felt like it. Since it depends on i and the sum is over different values of i it is **not** a common factor in the sum so you can't take it outside the sum!

(b) Show that the estimator in part (a) is unbiased.

Solution

Let $K = 1/\sum_j x_j^2$. Then $E(\hat\beta) = KE(\sum x_iY_i)$. Use $E(Y_i) = \beta x_i$ to see that

$$E(\hat\beta) = K\sum x_i^2\beta = \beta.$$

(c) Compute (give a formula for) the standard error of $\hat\beta$.

Solution

You have to compute $Var(\hat\beta) = K^2\sum Var(x_iY_i) = \sigma^2 K$ which is simply $\sigma^2/\sum x_i^2$. The standard error is then $\sqrt{\text{Var}(\hat\beta)} = \sigma\sqrt{K}$.

(d) The error sum of squares for this model is $\sum(Y_i - \hat\beta x_i)^2$ which may be shown to have $n-1$ degrees of freedom. If the x_i are the numbers 1, 2, 3 and 4, $\hat\beta = 1$ and the error sum of squares is 0.12 find a 95% confidence interval for β and explain what further assumptions you must make to do so.

Solution

If we assume that the errors are independent $N(0,\sigma^2)$ random variables then $\hat\beta$ is independent of the usual estimate of σ^2, samely $s^2 = 0.12/3 = 0.04$ in this case. The usual t statistic then has a t distribution and the confidence interval is $1 \pm t_{3,0.025}\sqrt{0.04}/\sqrt{1^2+2^2+3^2+4^2}$ which boils down to $1 \pm (3.18)(0.2)/\sqrt{30}$.

Note: sometimes people use formulas for simple linear regression as if they had fitted an intercept as well. They then have 2 degrees of freedom in spite of the fact that the question says there are $n-1$ degrees of freedom for error. Also: many neglect to assume that the errors have a normal distribution.

(e) Show that the estimator

$$\tilde{\beta} = \frac{\sum Y_i}{\sum x_i}$$

is also unbiased.

Solution

Let $K_2 = 1/\sum x_i$; then $E(\tilde{\beta}) = K_2 \sum x_i \beta = \beta$.

(f) Compute (give a formula for) the standard error of $\tilde{\beta}$. Which is bigger, the standard error of $\hat{\beta}$ or that of $\tilde{\beta}$?

Solution

We have $Var(\tilde{\beta}) = K_2^2 \sum Var(Y_i) = n\sigma^2/(\sum x_i)^2$. The difference $Var(\tilde{\beta}) - Var(\hat{\beta})$ is then simply

$$\sigma^2 \frac{n \sum x_i^2 - (\sum x_i)^2}{\sum x_i^2 (\sum x_i)^2}$$

The denominator is positive and the numerator is $n(n-1)$ times the usual sample variance of the x's so this difference of variances is positive. This proves that the standard error of $\hat{\beta}$ is smaller than that of $\tilde{\beta}$. The standard error of $\tilde{\beta}$ itself is

$$\sqrt{\text{Var}(\tilde{\beta})} = \sigma \sqrt{\frac{n}{(\sum x_i)^2}}$$

(g) Show that the mle of β in this model is $\hat{\beta}$, the least squares estimate, if the ϵ_i have normal distributions.

Solution

In this case Y_i is $N(x_i\beta, \sigma^2)$ and the likelihood is

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-(Y_1 - x_1\beta)^2/(2\sigma^2)} \times \cdots \times \frac{1}{\sqrt{2\pi}\sigma} e^{-(Y_n - x_n\beta)^2/(2\sigma^2)}$$

As usual you maximize the logarithm which is

$$-n \log \sqrt{2\pi} - n \log \sigma - \frac{1}{2\sigma^2} \sum (Y_i - x_i\beta)^2$$

Take the β derivative and get the same equation to solve for β as in the first part of this problem.

8. Consider the two-way layout without replicates. We have data Y_{ij} for $i = 1, \ldots I$ and $j = 1, \ldots, J$. We generally fit a so-called additive model

$$Y_{ij} = \mu + \rho_i + \gamma_j + \epsilon_{ij}$$

In the following questions consider the case $I = 2$ and $J = 3$.

(a) If we treat μ, ρ_1, ρ_2, γ_1, γ_2 and γ_3 as the entries in the parameter vector β what is the design matrix $X = X_a$ and what is the rank of X_a?

Writing the data as $Y^T = [Y_{1,1}, Y_{1,2}, Y_{1,3}, Y_{2,1}, Y_{2,2}, Y_{2,3}]$ the design matrix is

$$X_a = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Letting x_i denote column i of X_a we have $x_3 = x_1 - x_2$ and $x_6 = x_1 - x_4 - x_5$ so that the rank of X_a must be no more than 4. But if $a_1 x_1 + a_2 x_2 + a_4 x_4 + a_5 x_5 = 0$ then from row 6 we get $a_1 = 0$. Then from rows 4 and 5 get $a_4 = 0$ and $a_5 = 0$. Finally use row 3 to get $a_2 = 0$. This shows that columns 1, 2 4 and 5 are linearly independent so tha tX_a has rank at least 4 and so exactly 4.

(b) What is the determinant of the matrix $X_a^T X_a$? Is this matrix invertible? How many solutions do the normal equations have?

The matrix $X_a^T X_a$ is 6 by 6 but has rank only 4 so is singular and must have determinant 0. The normal equations are

$$\begin{bmatrix} 6 & 3 & 3 & 2 & 2 & 2 \\ 3 & 3 & 0 & 1 & 1 & 1 \\ 3 & 0 & 3 & 1 & 1 & 1 \\ 2 & 1 & 1 & 2 & 0 & 0 \\ 2 & 1 & 1 & 0 & 2 & 0 \\ 2 & 1 & 1 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} \mu \\ \rho_1 \\ \rho_2 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} = \begin{bmatrix} \sum_{ij} Y_{ij} \\ \sum_j Y_{1j} \\ \sum_j Y_{2j} \\ \sum_i Y_{i1} \\ \sum_i Y_{i2} \\ \sum_i Y_{i3} \end{bmatrix}$$

It may be seen that the second and third rows give equations which add up to the equation in the first row as do the fourth, fifth and sixth rows. Eliminate rows 3 and 6, say and solve. This leaves 4 linearly independent equations in 6 unknowns and so there are infinitely many solutions.

NOTE: You need to check that the equations are consistent. Many students row reduced only the matrix of coefficients forgetting the fact that if A has less than full rank $Ax = b$ might have 0 solutions. My solution above row reduces the **augmented** matrix, checking that there is at least one solution..

(c) Usually we impose the restrictions $\rho_1 + \rho_2 = 0$ and $\gamma_1 + \gamma_2 + \gamma_3 = 0$. Use these restrictions to eliminate ρ_2 and γ_3 from the model equation and, for the parameter vector $\beta^T = (\mu, \rho_1, \gamma_1, \gamma_2)$ find the design matrix X_b.

The restrictions give $\rho_2 = -\rho_1$ and $\gamma_3 = -\gamma_1 - \gamma_2$. In each model equation which mentions either ρ_2 or γ_3 you replace that parameter by the equivalent formula. So, for example,

$$\begin{aligned} Y_{2,3} &= \mu + \rho_2 + \gamma_3 + \epsilon_{2,3} \\ &= \mu - \rho_1 - \gamma_1 - \gamma_2 + \epsilon_{2,3} \end{aligned}$$

The 6th row of the design matrix is obtained by reading off the coefficients of $\mu, \rho_1, \gamma_1, \gamma_2$ which are 1, -1, -1 and -1. This makes

$$X_b = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix}$$

Note: some people eliminated ρ_1 and not ρ_2. This is fine and leads to a similar matrix with the rows in a different order.

(d) An alternate set of restrictions is called corner point coding where we assume $\rho_1 = \gamma_1 = 0$. With this restriction and the parameter vector $\beta^T = (\mu, \rho_2, \gamma_2, \gamma_3)$ what is the design matrix X_c?

This just makes the design matrix X_c just the corresponding columns, 1, 3, 5 and 6 of X_a.

(e) Show that the three design matrices have the same column space by finding a matrix A such that $X_a = X_b A$ and similarly for X_b and X_c and for X_a and X_c.

To write $X_c = X_a A$ just let A be the 6×4 matrix which picks out columns 1,3,5 and 6 of X_a, namely,

$$A_{ca} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

To write $X_a = X_c A$ we just have to put back column 2 and 4 remembering that col 2 is col 1 - col 3 and col 4 is col 1 - col 5 - col 6. Thus

$$A_{ac} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

Similarly

$$A_{bc} = \begin{bmatrix} 1 & 1/2 & 1/3 & 1/3 \\ 0 & -1/2 & 0 & 0 \\ 0 & 0 & -1/3 & -1/3 \\ 0 & 0 & 2/3 & -1/3 \end{bmatrix}$$

A vector in the column space of say X_a is of the form $X_a v$ for a vector of coefficients v. But such a vector is $X_c A_{ac} v$ and so of the form $X_c w$ for the vector $w = A_{ac} v$ and so in the column space of X_c.

(f) Use the previous part to show that the vectors of fitted values $\hat{Y}$ will be the same for any solution of the normal equations for any of the three design matrices.

The easy way to do this is to say: the fitted vector $\hat{\mu}$ is the closest point in the column space of the design matrix to the data vector Y. Since all three have the same column space they all have the same closest point and so the same $\hat{\mu}$.

Algebra is an alternative tactic: The matrix A_{bc} is invertible and we have

$$\hat{\mu}_b = X_b\hat{\beta}_b = X_b(X_b^T X_b)^{-1} X_b^T Y$$

Plug in $X_c A_{bc}$ for X_b and get

$$\hat{\mu}_b = X_c A_{bc}(A_{bc}^T X_c^T X_c A_{bc})^{-1} A_{bc}^T X_c^T Y$$

Use $(BC)^{-1} = C^{-1}B^{-1}$ to see that all occurences of A_{bc} cancel out to give

$$\hat{\mu}_b = X_c(X_c^T X_c)^{-1} X_c^T Y = \hat{\mu}_c \,.$$

The algebraic approach makes it a bit more difficult to deal with the case of X_a because the normal equations have many solutions. Suppose that $\tilde{\beta}$ is any solutions of the normal equations $X_a^T X_a \tilde{\beta} = X_a^T Y$. Then

$$A_{ac}^T X_c^T X_c A_{ac}\tilde{\beta} = A_{ac}^T X_c^T Y$$

or

$$A_{ac}^T \left(X_c^T X_c A_{ac}\tilde{\beta} - X_c^T Y \right) = 0$$

The matrix A_{ac} has rank 4. If any vector v satisfies $A_{ac}^T v = 0$ then

$$A_{ac}^T v = \begin{bmatrix} v_1 \\ v_1 - v_2 \\ v_2 \\ v_1 - v_3 - v_4 \\ v_3 \\ v_4 \end{bmatrix}$$

so that $v_1 = v_2 = v_3 = v_4 = 0$. We apply this with

$$v = X_c^T X_c A_{ac}\tilde{\beta} - X_c^T Y \,.$$

This shows that

$$X_c^T X_c A_{ac} \tilde{\beta} = X_c^T Y$$

so that $A_{ac}\tilde{\beta} = \hat{\beta}_c$.

Thus

$$\hat{\mu}_a = X_a \tilde{\beta} = X_c A_{ac} \tilde{\beta} = X_c \hat{\beta}_c = \hat{\mu}_c$$