

Statistics 403 Problem Set 5

Due in lab on Friday, October 16th

1. Suppose we carry out a survey by selecting 50 residences at random, then interview all adults at each of the residences (for simplicity, assume that there are two adults living in each residence). Since the 50 residences are a random sample, the adults not living in the same residence can be assumed to be independent.
 - (a) If the correlation between the adults living in the same residence is r , and the variance of every response is σ^2 , what is the variance of \bar{X} (the average response for all 100 people in the survey)?

Solution: The upper left 4×4 corner of the matrix looks like this:

$$\begin{pmatrix} \sigma^2 & r\sigma^2 & 0 & 0 \\ r\sigma^2 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & r\sigma^2 \\ 0 & 0 & r\sigma^2 & \sigma^2 \end{pmatrix}.$$

The covariance matrix of the data is a 100×100 matrix with 50 blocks (for the 50 families). Each block contributes two σ^2 's and two $r\sigma^2$'s. The total of all values in this matrix is $100\sigma^2 + 100r\sigma^2$. Thus

$$\text{var}(\bar{X}) = \sigma^2(1 + r)/100.$$

- (b) Now suppose we are considering an alternative survey design, in which we sample individuals at random (rather than sampling households). Suppose it costs \$50 to visit a house and interview one person, and it costs \$60 to visit a house and interview two people. Assuming we have a fixed amount of money to spend, for what values of r would we achieve a better estimate of EX by sampling individuals rather than households?

Solution: The study from part (a) would cost $\$60 \times 50 = \3000 . If we use this \$3,000 to sample individuals, we would be able to survey 60 individuals. The variance of \bar{X} in this case would be

$$\text{var}(\bar{X}) = \sigma^2/60.$$

The two study designs are equivalent if $(1+r)/100 = 1/60$, or $r = 2/3$. If $r > 2/3$, we would have a better result by sampling individuals.

2. Suppose we have four independent random variables A , B , C , and D . All have expected value zero and variance 1. Determine the following correlation coefficients.

(a) $\text{cor}(A + B, A)$

Solution:

$$\text{cov}(A + B, A) = \text{cov}(A, A) + \text{cov}(B, A) = \text{var}(A) + 0 = 1$$

$$\text{SD}(A + B) = \sqrt{\text{var}(A + B)} = \sqrt{\text{var}(A) + \text{var}(B)} = \sqrt{2}$$

$$\text{SD}(A) = 1$$

$$\text{cor}(A + B, A) = 1/\sqrt{2} \approx 0.71$$

(b) $\text{cor}(A + B, C + D)$

Solution:

$$\text{cov}(A + B, C + D) = \text{cov}(A, C) + \text{cov}(A, D) + \text{cov}(B, C) + \text{cov}(B, D) = 0$$

$$\text{cor}(A + B, C + D) = 0$$

(c) $\text{cor}(A + B, A - B)$

Solution:

$$\text{cov}(A + B, A - B) = \text{cov}(A, A) - \text{cov}(A, B) + \text{cov}(A, B) - \text{cov}(B, B) = 1 - 0 + 0 - 1 = 0$$

$$\text{cor}(A + B, A - B) = 0$$

(d) $\text{cor}(A + B, 2B)$

Solution:

$$\text{cov}(A + B, 2B) = \text{cov}(A, 2B) + \text{cov}(B, 2B) = 2$$

$$\text{SD}(A + B) = \sqrt{2}$$

$$\text{SD}(2B) = 2$$

$$\text{cor}(A + B, 2B) = 2/(2\sqrt{2}) = 1/\sqrt{2} \approx 0.71$$

3. Suppose we have an instrument to measure the amount of a certain chemical in a biological sample. We have n independent samples that we can measure to form an estimate \bar{X} of the expected value $EX = \mu$. The instrument measures with a standard deviation of σ , which is known. However, the instrument is miscalibrated, and the true value of the concentrations is $\mu + \delta$.

- (a) If we form a 95% confidence interval based on our data, not knowing about the miscalibration, what is the probability that it contains the true concentration? Express your result as a normal probability in terms of n , σ , etc. Hint: to calculate the coverage probability, you should reverse the calculations on slide 55 of the “background” notes.

Solution: The 95% CI is $\bar{X} \pm 2\sigma/\sqrt{n}$

The coverage probability is

$$\begin{aligned} P(\bar{X} - 2\sigma/\sqrt{n} \leq \mu + \delta \leq \bar{X} + 2\sigma/\sqrt{n}) &= P(-2\sigma/\sqrt{n} - \delta \leq \mu - \bar{X} \leq 2\sigma/\sqrt{n} - \delta) \\ &= P(-2\sigma/\sqrt{n} + \delta \leq \bar{X} - \mu \leq 2\sigma/\sqrt{n} + \delta) \\ &= P(-2 + \delta\sqrt{n}/\sigma \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 2 + \delta\sqrt{n}/\sigma) \\ &= P(-2 + \delta\sqrt{n}/\sigma \leq Z \leq 2 + \delta\sqrt{n}/\sigma) \end{aligned}$$

- (b) Suppose σ and δ are fixed, and n grows large. State what happens to the coverage probability, and briefly explain the reason why it happens.

Solution: The interval $(-2 + \delta\sqrt{n}/\sigma, 2 + \delta\sqrt{n}/\sigma)$ will move off to larger and larger values as n grows, so the probability of Z falling in this interval will go to zero. The coverage probability goes to zero because the interval is centered around the wrong value (μ rather than $\mu + \delta$), and as n grows the width of the CI becomes smaller. Eventually the CI is so narrow that it always misses the true value $\mu + \delta$.

4. Suppose we observe data sequentially, and each data value we observe has correlation coefficient r with the values that come before and after it in the sequence, but is independent of all other values. All values have variance 1. We form an average \bar{X} from the data and use it to construct a 95% CI for EX .

(a) Derive an expression for the width of this CI.

Solution: For example, if $n = 4$, the covariance matrix would look like this:

$$\begin{pmatrix} 1 & r & 0 & 0 \\ r & 1 & r & 0 \\ 0 & r & 1 & r \\ 0 & 0 & r & 1 \end{pmatrix}.$$

In general, the covariance matrix contains n 1's and $2(n - 1)$ r 's. Thus $\text{var}(\bar{X}) = 1/n + 2r(n - 1)/n^2$. The width of the CI is

$$4\sqrt{1/n + 2r(n - 1)/n^2}.$$

(b) Derive an expression for the ratio of the width of this CI relative to the CI that would be obtained if the same number of independent values were observed. What happens to this ratio as n grows large?

Solution: In the independent case, $\text{var}(\bar{X}) = 1/n$, and the width of the CI is $4/\sqrt{n}$. The ratio is

$$\frac{4/\sqrt{n}}{4\sqrt{1/n + 2r(n - 1)/n^2}} = \frac{1}{\sqrt{1 + 2r(n - 1)/n}}$$

For large n , we get $1/\sqrt{1 + 2r}$.