Answers to the 6th problem set

1. Suppose $X_1, X_2 \sim \text{i. i. d. Uniform}(\theta - 1/2,\ \theta + 1/2)$.

   (a) Show that $(\min\{\, X_1, X_2 \,\},\ \max\{\, X_1, X_2 \,\})$ is a 50% confidence interval for θ.

   Answer: We need to show $\mathbf{Pr}(\min < \theta < \max \mid \theta) = 0.5$ regardless of the fixed value of θ. We have

$$\begin{aligned}
\mathbf{Pr}(\min < \theta < \max \mid \theta) &= 1 - \mathbf{Pr}([\min < \theta \ \&\ \max < \theta] \text{ or } [\min > \theta \ \&\ \max > \theta]) \\
&= 1 - \mathbf{Pr}(\min < \theta \ \&\ \max < \theta) - \mathbf{Pr}(\min > \theta \ \&\ \max > \theta) \\
&= 1 - \mathbf{Pr}(X_1 < \theta \ \&\ X_2 < \theta) - \mathbf{Pr}(X_1 > \theta \ \&\ X_2 > \theta) \\
&= 1 - \mathbf{Pr}(X_1 < \theta)\,\mathbf{Pr}(X_2 < \theta) - \mathbf{Pr}(X_1 > \theta)\,\mathbf{Pr}(X_2 > \theta) \\
&= 1 - (0.5)(0.5) - (0.5)(0.5) \\
&= 0.5.
\end{aligned}$$

   (b) Suppose you observe that min= 84.03 and max= 84.04. Would you be confident that θ is within the 50% confidence interval? Suppose you observe that min= 64.52 and max= 65.48? Then would you be confident that θ is within the 50% confidence interval? Carefully explain your reasoning.

   Answer: If min= 84.03 then θ could be anywhere from 83.53 to 84.53, and if max= 84.04 then θ could be anywhere from 83.54 to 84.54. So θ could be anywhere from 83.54 to 84.53. The interval from the observed minimum to the observed maximum is a very tiny subinterval, and the data give no reason to suppose θ is more likely to be in that small subinterval than in any other similarly small subinterval. Thus "50% confidence" may be misleading.

   If min= 64.52 then θ is between 64.02 and 65.02, and if max= 65.48 then θ is between 64.98 and 65.98. Consequently θ must be between 64.98 and 65.02, and so it is certain that θ is in the much larger interval that is in this case the "50% confidence interval."

   (c) Suppose we assign an improper prior distribution to θ, as follows: $f_\theta(u) = c > 0$ for all $u \in \mathbb{R}$. (We fail to notice that $\int_{-\infty}^{\infty} c\,du \not< \infty$.) Let A and B be respectively the first and third quartiles of the posterior distribution. Express A and B as functions of $\min\{\, X_1, X_2 \,\}$ and $\max\{\, X_1, X_2 \,\})$. Find A and B in particular in case min= 84.03 and max= 84.04. Find A and B in case min= 64.52 and max= 65.48.

   Answer: The likelihood function is $L(\theta)$

$$= f_{X_1,X_2|\theta}(x_1, x_1) = f_{X_1|\theta}(x_1) \cdot f_{X_2|\theta}(x_2) = \left\{ \begin{array}{ll} 1 & \text{if } \theta - 1/2 < x_1 < \theta + 1/2 \\ & \quad \text{and } \theta - 1/2 < x_2 < \theta + 1/2 \\ 0 & \text{otherwise} \end{array} \right\}$$

$$= \left\{ \begin{array}{ll} 1 & \text{if } \theta - 1/2 < \min < \max < \theta + 1/2 \\ 0 & \text{otherwise} \end{array} \right\} = \left\{ \begin{array}{ll} 1 & \text{if } \max - 1/2 < \theta < \min + 1/2 \\ 0 & \text{otherwise} \end{array} \right\}.$$

CONTINUED⟶

$$L(\theta) = \left\{ \begin{array}{ll} 1 & \text{if } \max -1/2 < \theta < \min +1/2 \\ 0 & \text{otherwise} \end{array} \right\}.$$

Since the prior density of θ is constant, multiplying it by this likelihood function gives us a function that is constant on the interval from $\{\max -1/2\}$ to $\{\min +1/2\}$. We then need to normalize, i.e., to multiply this by a constant such that

$$\int_{\max -1/2}^{\min +1/2} [\text{constant}]\, d\theta = 1.$$

We conclude that the constant is the reciprocal of the length of the interval, so that

$$f_{\theta|X_1,X_2}(\theta) = \left\{ \begin{array}{ll} \dfrac{1}{(\min +1/2) - (\max -1/2)} & \text{if } \theta \in [\max -1/2,\, \min +1/2], \\ \\ 0 & \text{otherwise.} \end{array} \right.$$

The posterior distribution can be more efficiently characterized by saying that it is uniform on the interval $[\max -1/2,\, \min +1/2]$.

Since it is uniform on the interval, we just need to put $1/4$ of the interval to the left of A and $1/4$ of the interval to the right of B. If we let $C = \max -1/2$ and $D = \min +1/2$ then we get

$$\begin{aligned} A &= (3/4)C + (1/4)D = (3/4)(\max -1/2) + (1/4)(\min +1/2) \\ &= (3/4)\max +(1/4)\min -1/4 \\ \text{and } B &= (1/4)C + (3/4)D = (1/4)(\max -1/2) + (3/4)(\min +1/2) \\ &= (1/4)\max +(3/4)\min +1/4. \end{aligned}$$

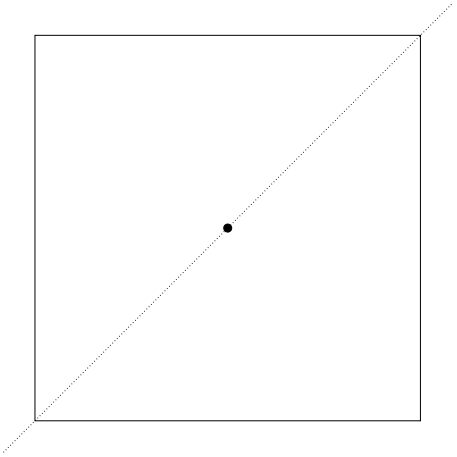
(d) Show that (A, B) is a 50% confidence interval for θ, i.e., when the value of θ is fixed, we have $\mathbf{Pr}(A < \theta < B) = 0.5$.

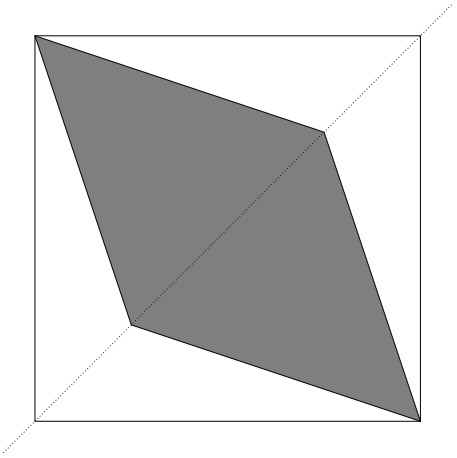
Answer: Notice that it said "... when the value of θ is fixed ...". In other words, we are switching to a frequentist perspective; we should not speak of posterior distributions; no probability distributions are assigned to θ in part (d). Perhaps one of the simplest ways to do this problem is to look at the square

$$\{\, (x_1, x_2) : \theta - 1/2 < x_1 < \theta + 1/2 \quad \& \quad \theta - 1/2 < x_2 < \theta + 1/2 \,\}$$

and consider which points in the square correspond to the event

$$(3/4)\max +(1/4)\min -1/4 < \theta < (1/4)\max +(3/4)\min +1/4.$$





Consider the horizontal axis to be the x_1-axis and the vertical axis to be the x_2-axis. The point right in the center of the square is (θ, θ). The region above the dotted line is the graph of the inequality $x_2 > x_1$ and the region below the dotted line is the graph of the inequality $x_1 > x_2$. In other words, above the dotted line we have $\max = x_2$ and $\min = x_1$ and below the dotted line we have $\max = x_1$ and $\min = x_2$. Therefore, above the dotted line the inequalities

$$(3/4)\max + (1/4)\min - 1/4 < \theta < (1/4)\max + (3/4)\min + 1/4$$

reduce to

$$(3/4)x_2 + (1/4)x_1 - 1/4 < \theta < (1/4)x_2 + (3/4)x_1 + 1/4$$

and below the dotted line they reduce to

$$(3/4)x_1 + (1/4)x_2 - 1/4 < \theta < (1/4)x_1 + (3/4)x_2 + 1/4.$$

The graph of this system of inequalities is shaded in the figure. Since the point (X_1, X_2) is uniformly distributed in the square, one need only observe that the area of the shaded region is exactly half that of the whole square, in order to conclude that the probability of the event that the graph corresponds to, is 1/2.

(e) Briefly discuss ways in which the 50% confidence interval of part (d) is better than the 50% confidence interval of part (a).

Answer: To get full credit for this part, it suffices that you observe that the pathologies that afflict the answer to part (b) do not evidently happen in this case. The problem in part (b) is that, although the 50% confidence interval covers θ 50% of the time, sometimes the data alone make it clear that the particular case is one of those 50% or probably is not.

2. (a) In the regression problem of the 5th problem set, show that there is an $(n-2)$-dimensional space of vectors $c \in \mathbb{R}^n$ such that $\mathsf{E}(c'Y) = \mathsf{E}(c_1Y_1 + \cdots + c_nY_n) = 0$. (The random variable $c'Y$ is called a "linear unbiased estimator of zero.")

Answer: We have

$$0 = \mathsf{E}(c'Y) = c' \mathsf{E}(Y) = c'X\beta$$

regardless of the value of β. In other words

$$0 = \mathsf{E}(c'Y) = c' \mathsf{E}(Y) = c'X\beta \quad \text{for } \underline{\textbf{all}} \text{ values of } \beta \in \mathbb{R}^2.$$

If this works for **all** values of $\beta \in \mathbb{R}^2$, then it works if $\beta = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and it works if $\beta = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. If $0 = c'X \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $0 = c'X \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ then $[0,\, 0] = c'X \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and so we get $c'X = [0,\, 0]$. Depending on how abstract you want to get in doing linear algebra, what you say next could be along these lines:

> The image of the linear transformation $c' \mapsto c'X$ is 2-dimensional.
> The domain of this linear transformation is n-dimensional.

Therefore the kernel of this linear transformation is $(n-2)$-dimensional.

... or it could be along these lines: We have two linear equations in n variables $c_1, \ldots, c_n$:

$$\begin{array}{ccccccc} c_1 & + & \cdots & + & c_n & = & 0 \\ c_1x_1 & + & \cdots & + & c_nx_n & = & 0 \end{array}$$

Then if $x_1 \neq x_2$ we can freely determine $c_3, \ldots, c_n$ and that in turn determines c_1 and c_2. If $x_1 = x_2$ then pick two xs that are not equal. (That not all of them are equal has been tacitly assumed throughout this problem and #5 of the previous problem set.)

(b) Suppose we have $n = 6$ and $x_1 = 2$, $x_2 = 3$, $x_3 = 5$, $x_4 = 5$, $x_5 = 7$, $x_6 = 9$. Find four linear unbiased estimators of zero that are linearly independent of each other, i.e., four vectors that can be put in the role of c in part (a), none of which is a linear combination of the others.

Answer: We need

$$\begin{array}{ccccccccccccc} c_1 & + & c_2 & + & c_3 & + & c_4 & + & c_5 & + & c_6 & = & 0 \\ 2c_1 & + & 3c_2 & + & 5c_3 & + & 5c_4 & + & 7c_5 & + & 9c_6 & = & 0 \end{array}$$

We can set $c_3 = 1, c_4 = c_5 = c_6 = 0$, and we get $c_1 = 2, c_2 = -3$.
We can set $c_4 = 1, c_3 = c_5 = c_6 = 0$, and we get $c_1 = 2, c_2 = -3$.
We can set $c_5 = 1, c_3 = c_4 = c_6 = 0$, and we get $c_1 = 4, c_2 = -5$.
We can set $c_6 = 1, c_3 = c_4 = c_5 = 0$, and we get $c_1 = 6, c_2 = -7$.

The four linearly independent vectors are:

$$(2, -3, 1, 0, 0, 0)$$
$$(2, -3, 0, 1, 0, 0)$$
$$(4, -5, 0, 0, 1, 0)$$
$$(6, -7, 0, 0, 0, 1)$$

The four linearly independent linear unbiased estimators of zero are:

$$2Y_1 - 3Y_2 + Y_3$$
$$2Y_1 - 3Y_2 + Y_4$$
$$4Y_1 - 5Y_2 + Y_5$$
$$6Y_1 - 7Y_2 + Y_6$$

Other systems of four such vectors, or of four linearly independent linear unbiased estimators of zero can also be found, by making some of the choices differently.

(c) Let $\widehat{\beta}_1 = d'Y$ be the least-squares estimator of of the slope in the regression problem of the 5th problem set (in #4 on the 5th problem set, you found the value of d). Show that $\widehat{\beta}_1$ is uncorrelated with every linear unbiased estimator of zero.

Answer: We found that $\widehat{\beta}_1 = [0,\ 1](X'X)^{-1}X'Y$; in other words $d' = [0,\ 1](X'X)^{-1}X'$.
Suppose $c'Y$ is a linear unbiased estimator of zero.
Recall that if $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{\ell\times k}$ are constant

and $U \in \mathbb{R}^{n\times 1}$ and $V \in \mathbb{R}^{k\times 1}$ are random

then $\mathbf{cov}(AU,\ BV) = A(\ \mathbf{cov}(U,\ V)\ )B'$.

So we get

$$\mathbf{cov}(c'Y,\ d'Y) = c'(\ \mathbf{cov}(Y, Y)\ )d = c'(\sigma^2 I_n)d = \sigma^2 c'd = \sigma^2 c'X(X'X)^{-1}\begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

and this $= 0$ because, as we found in part (a), we have $c'X = 0 \in \mathbb{R}^{1\times 2}$.

(d) Let d and c be as above. Show that $(d+c)'Y$ is an unbiased estimator of β_1. Show that $\mathbf{var}((d+c)'Y) = \mathbf{var}(d'Y) + \mathbf{var}(c'Y) \geq \mathbf{var}(d'Y)$. Use the result to show that $d'Y$ is the best linear unbiased estimator of β_1. ("Linear" means linear in Y, i.e., of the form $d'Y$ for some vector d.) (This is not yet the Gauss-Markov theorem mentioned in the summary of March 6th, because that theorem has much weaker assumptions (and so, is a much stronger theorem) than those on which we are relying here.)

Answer:

$$\mathbf{E}(d'Y + c'Y) = \mathbf{E}(d'Y) + \mathbf{E}(c'Y) = E(d'Y) + 0 = \mathbf{E}(d'Y) = \beta_1,$$

so $d'Y + c'Y$ is an unbiased estimator of β_1. Its variance is

$$\mathbf{var}(d'Y + c'Y) = \mathbf{var}(d'Y) + \mathbf{cov}(d'Y, c'Y) + \mathbf{cov}(c'Y, d'Y) + \mathbf{var}(c'Y)$$

and the two terms in the middle are 0, since that was the result of part (c). Since variances of non-constant random variables are positive we can actually get ">" where the problem asked for "$\geq$", unless $c = 0$.

3. Let $X_1, \ldots, X_{15} \sim \text{i.i.d. } N(\mu, \sigma^2)$. Use what you know about the probability distribution of $\sum_{i=1}^{15}(X_i - \overline{X})^2$ to find a 95% confidence interval for σ^2, i.e., find two statistics A and B such that $\mathbf{Pr}(A < \sigma^2 < B) = 0.95$, and $\mathbf{Pr}(\sigma^2 < A) = 0.05/2$ and $\mathbf{Pr}(\sigma^2 > B) = 0.05/2$.

Answer: We know that

$$\frac{\sum_{i=1}^{15}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{14}.$$

From the table of the χ^2 distribution, we conclude that

$$\mathbf{Pr}\left(5.629 < \frac{\sum_{i=1}^{15}(X_i - \overline{X})^2}{\sigma^2} < 26.12\right) = 0.95$$

and the probabilities in the left and right tails are each 0.025. Consequently we have

$$\mathbf{Pr}\left(\frac{\sum_{i=1}^{15}(X_i - \overline{X})^2}{26.12} < \sigma^2 < \frac{\sum_{i=1}^{15}(X_i - \overline{X})^2}{5.629}\right) = 0.95$$

and we can draw a similar conclusion about the two tails.