Statistics 403 Problem Set 6

Due in lab on Friday, October 29th

1. Three processes for collecting data $X_1, \ldots, X_{10}$ are described below. For each, determine the $15^{\text{th}}$, $50^{\text{th}}$, and $90^{\text{th}}$ percentiles of the sampling distribution of $\bar{X}$.

   (a) The data are collected as independent and identically distributed values from a distribution with mean 2 and standard deviation 0.9.

   Solution: The quantile is a solution to an equation of the following form:

   $$ P(\bar{X} \leq Q) = 0.15 $$

   (for the $15^{\text{th}}$ percentile). To solve the equation you need to standardize $\bar{X}$, which will require that you have the expected value $E\bar{X} = 2$ and the variance $\text{var}(\bar{X}) = 0.9^2/10$. For the $15^{\text{th}}$ percentile:

   $$ 0.15 = P(\bar{X} \leq Q) = P(Z \leq \sqrt{10}(Q - 2)/0.9), $$

   then solve

   $$ \sqrt{10}(Q - 2)/0.9 = -1.04 $$

   to get

   $$ Q = 2 - 1.04 \cdot 0.9/\sqrt{10} \approx 1.7. $$

   The other quantiles are $Q = 2$ (for the $50^{\text{th}}$ percentile) and $Q = 2 + 1.28 \cdot 0.9/\sqrt{10} \approx 2.36$ (for the $90^{\text{th}}$ percentile).

   (b) The data are collected as independent values. The $i^{\text{th}}$ data value $X_i$ is collected from a distribution with mean 2 and standard deviation 0.7 (if $i \leq 5$) or 1.1 (if $i > 5$).

   Solution: Again we have $E\bar{X} = 2$, but to get the variance of $\bar{X}$, we need to use the covariance matrix
\[ \Sigma = \begin{pmatrix} 0.7^2 & 0 & 0 & \cdots & \cdots & 0 & 0 \\ 0 & 0.7^2 & 0 & \cdots & \cdots & 0 & 0 \\ 0 & 0 & 0.7^2 & \cdots & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \cdots & 1.1^2 & 0 \\ 0 & 0 & 0 & \cdots & \cdots & 0 & 1.1^2 \end{pmatrix} \]

Thus

\[ \text{var}(\bar{X}) = (5 \cdot 0.7^2 + 5 \cdot 1.1^2)/100 \approx 0.085. \]

So the 15\textsuperscript{th} percentile is

\[ Q = 2 - 1.04 \cdot \sqrt{0.085} \approx 1.70. \]

The median is \( Q = 2 \), and the 90\textsuperscript{th} percentile is \( Q = 2 + 1.28 \cdot \sqrt{0.085} \approx 2.37 \).

(c) Each \( X_i \) has mean 2 and standard deviation 0.9. The \( X_i \) are correlated so that any two values among \( X_1, \ldots, X_5 \) have correlation coefficient 0.5 between them, but the values \( X_6, \ldots, X_{10} \) are independent of each other, and also are independent of \( X_1, \ldots, X_5 \).

**Solution:** In this case, the covariance matrix is

\[ \Sigma = \begin{pmatrix} 0.9^2 & 0.9^2/2 & 0.9^2/2 & \cdots & \cdots & 0 & 0 \\ 0.9^2/2 & 0.9^2 & 0.9^2/2 & \cdots & \cdots & 0 & 0 \\ 0.9^2/2 & 0.9^2/2 & 0.9^2 & \cdots & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \cdots & 0.9^2 & 0 \\ 0 & 0 & 0 & \cdots & \cdots & 0 & 0.9^2 \end{pmatrix}. \]

Note that any pair among the first five observations has covariance

\[ \text{cov}(X_i, X_j) = \text{cor}(X_i, X_j) \text{SD}(X_i) \text{SD}(X_j) = 0.9^2/2. \]

The total of all numbers in this matrix is
10 \cdot 0.9^2 + 20 \cdot 0.9^2/2 = 16.2. \\
Thus, \text{var}(\bar{X}) = 16.2/100 = 0.162, \\
so the 0.15 quantile is \\
\quad \quad 2 - 1.04\sqrt{0.162} \approx 1.58, \\
the median is 2, and the 0.9 quantile is \\
\quad \quad 2 + 1.28\sqrt{0.162} \approx 2.52.

2. The interquartile range (IQR) is the difference between the 75\textsuperscript{th} and 25\textsuperscript{th} percentiles of a distribution. For the normal distribution, the IQR is related to the standard deviation \( \sigma \) via IQR = 1.349\( \sigma \). For the stature (height) data given as a graph in the course notes, estimate the standard deviation of heights for females of age 2, 10, and 20 (treat the heights as being normally distributed, so the given relation between IQR and \( \sigma \) holds).

\textbf{Solution:} The 25\textsuperscript{th} and 75\textsuperscript{th} percentiles at age 2 are 83 and 87.5; at age 10 they are 134 and 142.5; at age 20 they are 159 and 167.5. Thus the IQR’s are 4.5, 8.5, and 8.5 (don’t worry if you got slightly different numbers). Thus the standard deviations are 3.3, 6.3, and 6.3.

3. Suppose we are interested in studying the relationship between the concentration of iron and copper in human blood. We have assays that can be used to measure these quantities, but there is some random measurement error in both of the assays. Let \( I \) and \( C \) denote the true iron and copper concentrations in the blood, and suppose that we observe \( I_{\text{obs}} = I + E_I \) and \( C_{\text{obs}} = C + E_C \), where \( E_I \) and \( E_C \) are the measurement errors. Suppose we also know that the standard deviations of \( E_I \) and \( E_C \) are 0.3, and 0.6, respectively, the standard deviations of \( I \) and \( C \) are 2.5 and 1.5, respectively, and the correlation coefficient between \( I \) and \( C \) is 0.3. What is the correlation coefficient between \( I_{\text{obs}} \) and \( C_{\text{obs}} \)? You can assume that the measurement errors \( E_I \) and \( E_C \) satisfy the conditions given in the notes – they are independent of each other, and also are independent of the true values \( I \) and \( C \).

\textbf{Solution:}
\[
\text{cor}(I_{\text{obs}}, C_{\text{obs}}) = \frac{\text{cov}(I_{\text{obs}}, C_{\text{obs}})}{\text{SD}(I_{\text{obs}}) \cdot \text{SD}(C_{\text{obs}})}
\]
\[
= \frac{\text{cov}(I + E_I, C + E_C)}{\sqrt{\text{var}(I + E_I) \cdot \text{var}(C + E_C)}}^{1/2}
\]
\[
= \frac{\text{cov}(I, C)}{(2.5^2 + 0.3^2) \cdot (1.5^2 + 0.6^2)}^{1/2}
\]
\[
= \frac{0.3 \cdot 2.5 \cdot 1.5}{(2.5^2 + 0.3^2) \cdot (1.5^2 + 0.6^2)}^{1/2}
\]
\[
\approx 0.28.
\]

As a result of the measurement errors \(E_I\) and \(E_C\), there is a slight bias in the correlation coefficient toward zero (an “attenuation”). If the measurement errors were stronger (i.e. if they had greater variance), then the attenuation would become stronger.

4. Suppose we are interested in estimating the difference in mean treatment responses, denoted \(d\), between two medical treatments. We plan to collect data \(X_1, \ldots, X_n\) and \(Y_1, \ldots, Y_n\) on the responses of treated subjects \(X_i\) and untreated subjects \(Y_i\). Assuming that the \(X_i\) and \(Y_i\) are unbiased estimates of the treatment responses, we initially assume \(E\bar{X} - E\bar{Y} = d\), and propose a study based on \(n = 20\) since we are satisfied with the power for the setting when the raw effect size is \(d = EX - EY = 0.7\). However, it turns out that the measurements are biased downward by 10% so that \(E(\bar{X} - \bar{Y}) = 0.9d\). How many additional data points do we need to collect in order to get the same sample size power we would have had with unbiased data?

You can assume that the \(X_i\) and the \(Y_i\) are both independent and identically distributed samples. You can also assume that both variances are known to be 1 (i.e. \(\text{var}(X) = \text{var}(Y) = 1\)).

**Solution:** If the raw effect size were 0.7, then we would have

\[
ET = \frac{0.7}{\sqrt{1/20 + 1/20}} \approx 2.21.
\]

The power for this value of \(ET\) is

\[
P(T > 2) = P(Z > 2 - 2.21)
\]

4
\begin{align*}
  &= P(Z > -0.21) \\
  &= 0.58.
\end{align*}

In reality, the raw effect size is $0.9 \cdot 0.7$, so the power would actually be

\begin{align*}
  ET &= \frac{0.9 \cdot 0.7}{\sqrt{1/20 + 1/20}} \approx 2,
\end{align*}

which gives a power of 0.5. To get the power back up to 0.58, we need to increase the sample size. This will give us a new value

\begin{align*}
  ET &= \frac{0.9 \cdot 0.7}{\sqrt{1/n + 1/n}} \approx 0.45 \sqrt{n}.
\end{align*}

If we set $0.45 \sqrt{n} = 2.21$ (as we thought we had initially), then we get $n = (2.21/0.45)^2 \approx 24$. Thus we need four additional people per group (eight additional people in all), to compensate for the decreased raw effect size.