1. Answer the questions given below for the following probability distribution (your answers may depend on \(a\) and \(p\)):

\[
\begin{array}{c|c|c|c}
 x & -a & 0 & a \\
 P(X = x) & p/2 & 1 - p & p/2 \\
\end{array}
\]

(a) What is the expected value \(EX\) of this distribution?

**Solution:** The expected value is

\[-a \cdot p/2 + 0 \cdot (1 - p) + a \cdot p/2 = 0.\]

(b) What is the variance \(\text{var}(X)\) of this distribution?

**Solution:**

\[
\text{var}(X) = E(X - EX)^2 = a^2 \cdot p/2 + 0^2 \cdot (1 - p) + a^2 \cdot p/2 = a^2 p
\]

(c) What value of \(p\) maximizes the variance of this distribution?

**Solution:**

\[p = 1\]

(d) What is \(P(X \leq 0)\) for this distribution?

**Solution:**

\[P(X \leq 0) = P(X = -a) + P(X = 0) = p/2 + 1 - p = 1 - p/2\]

(e) What value of \(p\) maximizes \(P(X \leq 0)\) for this distribution?

**Solution:** \(p = 0\)
(f) What is $E[1/(1 + X^2)]$ for this distribution?

Solution:

$$\frac{p/2}{1 + a^2} + 1 - p + \frac{p/2}{1 + a^2} = 1 - p + p/(1 + a^2)$$

(g) Suppose the value of $a$ is restricted so $a \geq 1$. What value of $a$ maximizes the value of $E[1/(1 + X^2)]$ for this distribution?

Solution: $a = 1$
2. Suppose we are comparing two samples of data, \( X_1, \ldots, X_n \), and \( Y_1, \ldots, Y_{2n} \) (note that one sample is twice as large as the other). Our goal is to learn about the relationship between their population means \( EX \) and \( EY \). The two populations have the same variance, which we denote \( \sigma^2 \).

(a) If \( \sigma = 1 \), what sample size is required so that the standard deviation of \( \bar{X} - \bar{Y} \) is less than \( 1/10 \)?

Solution:

\[
SD(\bar{X} - \bar{Y}) = \sqrt{\frac{1}{n} + \frac{1}{2n}} = \frac{\sqrt{3}}{\sqrt{2n}}.
\]

Setting \( \frac{\sqrt{3}}{\sqrt{2n}} < 1/10 \) gives us \( n > 150 \).

(b) We observe a difference of \( D = \bar{X} - \bar{Y} = 3 \) between the two sample means. Describe the values of \( \sigma^2 \) under which we would reject the null hypothesis that the population means are equal. Your answer will depend on \( n \).

Solution:

\[
T = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma^2/(2n) + \sigma^2/n}} = \frac{\sqrt{6n}}{\sigma}
\]

Setting \( T > 2 \), we get \( \sigma < \sqrt{6n}/2 \), or \( \sigma^2 < 1.5n \).
3. Suppose we are interested in the amount of pressure $\mu$ that must be applied to a bicycle tire before the tire tears. We have two ways of making the measurement: the first approach, denoted $X$, measures the pressure unbiasedly (so $EX = \mu$), with a standard deviation of 2 units. The second approach, denoted $Y$, is too high on average by one-half of a unit, so $EY = \mu + 1/2$, and has a standard deviation of 1 unit.

(a) Which of the two approaches has a lower probability of producing a value that is two or more units higher than the true value? You can assume that both $X$ and $Y$ are normally distributed.

Solution:

\[
P(X \geq \mu + 2) = P(Z \geq 1)
\]

\[
P(Y \geq \mu + 2) = P(Z \geq 3/2)
\]

So the probability is greater for $X$.

(b) Now suppose that instead of taking one measurement for $X$ and one measurement for $Y$, we take 9 independent measurements for each and average them, producing $\bar{X}$ and $\bar{Y}$. Does your answer to part (a) change? Explain your reasoning.

Solution:

\[
P(\bar{X} > \mu + 2) = P(Z > 3)
\]

\[
P(\bar{Y} > \mu + 2) = P(Z > 9/2)
\]

So $X$ still has the higher probability.
4. Suppose we are interested in estimating the fraction of a population that supports a particular political candidate. We collect \( n \) independent data values in yes/no form (\( X = 1 \) for a “yes, I support the candidate” response, and \( X = 0 \) for a “no, I do not support the candidate” response). We form \( \bar{X} \) and construct an approximate 95% confidence interval for the true level of support \( EX \) in the usual way. Below are descriptions of four graphs, labeled a-d, followed by six actual graphs. Write the letters a-d in four of the graphs, as appropriate. Use each letter only once (two of the graphs will be unused).

(a) The width of the confidence interval as a function of \( EX \)
(b) The width of the confidence interval as a function of \( n \)
(c) The center point of the confidence interval as a function of \( n \)
(d) The lower limit of the confidence interval, as a function of \( n \).
5. Suppose we are comparing two ways of teaching algebra. Two groups of \( n \) high school students are recruited into our study, and each group is taught with one of the two teaching methods. At the end of the year, the students will take a standardized test, and a two-sample mean comparison will be used to assess for a statistically significant difference between the population mean test scores of the two groups. It is believed that one of the teaching methods may raise scores by five points, where the standard deviations of the test scores in either group are 20 points. What sample size \( n \) is needed to have 80% power to detect this effect at the 5% significance level? You don’t need to calculate a numerical answer, but do simplify as much as practical. Some percentiles of the standard normal distribution are given below.

<table>
<thead>
<tr>
<th>( p )</th>
<th>Quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>-1.28</td>
</tr>
<tr>
<td>0.2</td>
<td>-0.84</td>
</tr>
<tr>
<td>0.3</td>
<td>-0.52</td>
</tr>
<tr>
<td>0.4</td>
<td>-0.25</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>0.6</td>
<td>0.25</td>
</tr>
<tr>
<td>0.7</td>
<td>0.52</td>
</tr>
<tr>
<td>0.8</td>
<td>0.84</td>
</tr>
<tr>
<td>0.9</td>
<td>1.28</td>
</tr>
</tbody>
</table>

**Solution:** The power is

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

Since \( P(Z > -0.84) = 0.8 \), to get 80% power, we set \( 2 - ET = -0.84 \). For our situation,

\[
ET = \frac{5}{\sqrt{20^2/n + 20^2/n}} = \frac{5\sqrt{n}}{20\sqrt{2}}
\]

So we set

\[
\frac{5\sqrt{n}}{20\sqrt{2}} = 2.84
\]

and solve for \( n \) to get

\[
n = (2.84 \cdot 4 \cdot \sqrt{2})^2 = 32 \cdot 2.84^2
\]
6. Suppose we are constructing a confidence interval for $EX = \mu$ based on an independent sample $X_1, \ldots, X_n$ in the usual way (i.e. aiming to achieve 0.95 coverage probability). In forming the confidence interval, we use a value $\sigma$ for the standard deviation of $X$, but we are concerned that $\sigma$ might be smaller than the true standard deviation $SD(X) = f\sigma$. How large can $f$ be such that the coverage probability remains above 0.9? You can use the following table of standard normal quantiles.

<table>
<thead>
<tr>
<th>p</th>
<th>Quantile</th>
<th>p</th>
<th>Quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.84</td>
<td>0.9</td>
<td>1.28</td>
</tr>
<tr>
<td>0.825</td>
<td>0.94</td>
<td>0.925</td>
<td>1.44</td>
</tr>
<tr>
<td>0.85</td>
<td>1.04</td>
<td>0.95</td>
<td>1.64</td>
</tr>
<tr>
<td>0.875</td>
<td>1.15</td>
<td>0.975</td>
<td>1.96</td>
</tr>
</tbody>
</table>

**Solution:** The coverage probability is:

$$P(X - 2\sigma/\sqrt{n} \leq \mu \leq X + 2\sigma/\sqrt{n}) = P(-2/f \leq \sqrt{n}\frac{\bar{X} - \mu}{f\sigma} \leq 2/f) = P(-2/f \leq Z \leq 2/f).$$

To get 95% coverage, we need $2/f = 1.64$, so $f = 2/1.64$, or $f \approx 1.22$. Thus as long as the true standard deviation is not more than 22% greater than our “working standard deviation” ($\sigma$), we will have at least 90% coverage.