1. What is the difference between a sample and a population?

- A sample is any subset of measurements selected from the population.
- A population is the set of all measurements of interest to the sample collector.

2. A researcher is interested in the degree of bilateral symmetry in humans. She places a flier around campus to solicit volunteers for the study. For each volunteer she measures the distance between the tip of the middle finger to the palm of the hand, for both hands.

   (a) What is the population? Be very precise.
   - the pair of middle finger lengths for all human beings

   (b) What is the sample?
   - the pair of middle finger lengths for the students who volunteered

   (c) Do you have any criticism of how she collected the data?
   - The sample is not a random selection from the population of interest. People from the US, university students and students who volunteer may be systematically different from the general population of human beings.

   (d) What would the population be if she measured the difference between the lengths of the middle fingers on each hand?
   - The population would now be univariate—the set of all middle finger length differences for all human beings.
3. A recent article in the Gainesville local paper reported that the Center for Disease control was recommending that an HIV test become part of a routine medical exam. Assume that people who go in for routine medical exams can be considered a random sample from the population at large (this is almost certainly not true, but use this assumption as a basis for your calculations). Using rigorous calculations, show why this is a bad idea. Use the following data:

(a) Prevalence of HIV = .00075 = \(P(D)\)
(b) Sensitivity of the test = 0.997 = \(P(P|D)\)
(c) Specificity of the test = 0.997 = \(P(\neg|D^c)\)

Let \(D\) be the event that the tested person has the disease and \(P\) the event that the test is positive. The desired probability \(Pr\{D|P\}\) is obtained by

\[
Pr\{D|P\} = \frac{Pr\{DP\}}{Pr\{P\}}
\]

\[
= \frac{Pr\{P|D\}Pr\{D\}}{Pr\{P|D\}Pr\{D\} + Pr\{P|D^c\}Pr\{D^c\}}
\]

\[
= \frac{(0.997)(0.00075)}{(0.997)(0.00075) + (0.003)(0.99925)} = 0.1996
\]

Thus, only about 20% of those persons whose test results are positive actually have the disease, assuming that these persons are tested at random. This means that about 80% of those testing positive, do not have the disease. Be careful here. This is not the same thing as the probability of a false positive. If we test only persons with symptoms of the disease or those in high risk groups, the results are not so disturbing. If we knew the probability of a person having the disease given that they have symptoms (or are in a high risk group), this probability could be used in the above calculation, thereby increasing the usefulness of the test. This is also the reason most tests are given twice, since the probability of a test being twice positive when the person does not have the disease is usually very low. Hence, testing everyone at routine appointments will not only incur costs from the first test, but also from the second test, not to mention the emotional distress of all those people who will think they have AIDS for some number of days or weeks before they can go back for a second test.
4. A study of the effect of parents’ smoking habits on the smoking habits of students in a high school produced the following table of proportions.

<table>
<thead>
<tr>
<th>Parents</th>
<th>Student Smokes</th>
<th>Student doesn’t smoke</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both parents smoke</td>
<td>0.07</td>
<td>0.26</td>
<td>0.33</td>
</tr>
<tr>
<td>One parent smokes</td>
<td>0.08</td>
<td>0.34</td>
<td>0.42</td>
</tr>
<tr>
<td>Neither parent smokes</td>
<td>0.03</td>
<td>0.22</td>
<td>0.25</td>
</tr>
<tr>
<td>Total</td>
<td>0.18</td>
<td>0.82</td>
<td>1</td>
</tr>
</tbody>
</table>

Suppose a student is randomly selected from this population.

(10) (a) What is the probability the student smokes and at least one parent of the student smokes?

- \[0.07 + 0.08 = 0.15\]

(10) (b) Given the student smokes, what is the probability neither parent smokes?

- \[0.03/0.18 = 0.167\]

(10) (c) What is the probability that either the student smokes or both parents smoke?

- \[0.18 + 0.33 - 0.07 = 0.44\]

(10) (d) Are the event neither parent smokes and the event the student does not smoke independent? Why or why not?

- No, because probability of joint event = 0.22 is not equal product of probabilities of marginal events = 0.82(0.25) = 0.205
5. Systolic blood pressure for the population of males of age 18 to 74 years who live in the United States is well approximated by a normal model with mean 129 mm Hg and standard deviation 19.8 mm Hg. In 2003 guidelines for healthy blood pressure were revised. Under the new guidelines, systolic blood pressure over 120 is considered 'hypertensive'. Systolic blood pressure under 90 mm Hg is considered 'hypotensive'.

(a) According to this model, what proportion of males age 18 - 74 years who live in the U.S.

have

i. hypertension?

\( Z = \frac{120 - 129}{19.8} = -0.455 \) Area to left of \( Z = 0.330 \rightarrow \) area to right of \( Z = 1 - 0.3300 = 0.670 \). Or about 67%.

ii. hypotension

\( Z = \frac{90 - 129}{19.8} = -1.97 \) Area to left of \( Z = 0.025 \rightarrow 0.025 \). Or about 2.5%.

(b) According to this model, what is the 25th percentile of systolic blood pressure?

• 115 mm Hg.

(c) In Batavia, the distribution of systolic blood pressure for the same demographic is also reasonably modeled by a normal distribution but with mean 100 mm Hg and standard deviation 18 mm Hg. Please answer the following question in the 2 ways listed below:

(i) Without doing any calculations: Which man has higher blood pressure relative to the men in his country: a U.S. man with systolic blood pressure of 125 mm Hg or a man from Batavia with systolic blood pressure of 110 mm Hg? Explain in words.

• The man from the Batavia because the US blood pressure is below the US mean and the Batavian’s blood pressure is above the mean in his country.

(ii) Justify your answer in part (i) by performing the appropriate calculations.

• US z-score=(125-129)/19.8=-0.202, Batavian z-score=0.556

The Batavian’s bp is 0.556 std devs above the mean in his country. The US man’s bp is 0.020 std devs below the mean in the US.

(d) Say you do not know the distributions of the US and Batavia, and you observed a sample of size 100 from each country. The US sample average was 128 with standard deviation 18. Batavia had a sample average of 98 with standard deviation 12. Conduct a hypothesis test that the two countries have the same average systolic blood pressure.

\( T = \frac{128-98}{\sqrt{\frac{18^2}{100} + \frac{12^2}{100}}} = 13.86 \) Because this value is so large, we do not need to know the exact degrees of freedom. We reject \( H_0 \) with a p-value << 0.001

6. The mean and standard deviation of the random variables, X, Y, Z are below in the table below.

<table>
<thead>
<tr>
<th>random variable</th>
<th>mean</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>3</td>
<td>0.5</td>
</tr>
<tr>
<td>y</td>
<td>4.6</td>
<td>1.25</td>
</tr>
<tr>
<td>z</td>
<td>10</td>
<td>3</td>
</tr>
</tbody>
</table>

(a) state the assumptions necessary to perform the calculations listed below.
(10) (b) Calculate the mean and standard deviation of: \( T = 0.25X + 0.875Y - 0.5Z \)

- \( E(T) = 0.25(3) + 0.875(4.6) - 0.5(10) = 0.75 + 4.025 - 5 = -0.225 \)
- \( SD(T) = \sqrt{var(0.25X) + var(0.875Y) + var(0.5Z)} = \sqrt{(0.25^2)(0.5^2) + (0.875^2)(1.25^2) + (0.5^2)(3^2)} = 1.87 \)

(10) (c) Calculate the mean and standard deviation of: \( S = 5X - 2Y - 4Z \)

- \( E(S) = 5 \times 3 - 2 \times 4.6 - 4 \times 10 = -34.2 \)
- \( SD(S) = \sqrt{25 \times 0.5^2 + 4 \times 1.25^2 + 16 \times 3^2} = 12.5 \)

(10) (d) Calculate the mean and standard deviation of: \( U = X + Y + Z \)

- \( E(U) = 3 + 4.6 + 10 = 17.6 \)
- \( SD(U) = \sqrt{0.5^2 + 1.25^2 + 3^2} = 3.288 \)

7. Derive the mean and standard deviation of a binomial\((n,p)\) random variable.

- By definition of a binomial random variable, if \( X \sim Bin(n, p) \) then \( X = \) the number of successes in \( n \) bernoulli trials. Hence, let \( Y_1, Y_2, \ldots, Y_n \) be \( n \) bernoulli random variables then \( X = \sum_i Y_i \) is distributed Binomial\((n,p)\). Therefore

\[
E\{X\} = E\left\{\sum Y_i\right\} = E\{Y_1\} + E\{Y_2\} + \cdots + E\{Y_n\} = nE\{Y_1\} = np.
\]

Since all the bernoulli random variables have the same expected value, \( p \).

- Similarly, for the variance.

\[
Var\{X\} = Var\left\{\sum Y_i\right\} = \sum Var\{Y_i\} = np(1 - p).
\]

since the variance of a bernoulli is \( p(1 - p) \).