1. Short answer questions - 9 points each

a. Is there anything wrong with this statement?

> The probability that an overweight person has sleep apnea is 5%. The probability that an overweight person has restless leg syndrome is 1%. Therefore, there is a 6% probability that an overweight person has one of these sleep diseases.

Explain your answer briefly.

Yes, there is potentially something wrong with the statement. The events

$$A = \{\text{Person has sleep apnea}\}$$

and

$$B = \{\text{Person has restless leg syndrome}\}$$

are not mutually exclusive so that $P(A \cup B)$ is not necessarily $P(A) + P(B)$. Formally, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Note that independence between A and B is irrelevant for answering the question.

b. Suppose that $X_1, \ldots, X_{10}$ are iid $N(\mu_1, \sigma_1)$ mutually independent of $Y_1, \ldots, Y_{15}$, which are iid $N(\mu_2, \sigma_2)$. What is the variance of $5(\bar{X} - \bar{Y})$?

$$\text{Var}\{5(\bar{X} - \bar{Y})\} = 25\text{Var}(\bar{X} - \bar{Y}) = 25\{\text{Var}(\bar{X}) + \text{Var}(\bar{Y})\} = 25(\sigma_1^2/10 + \sigma_2^2/15)$$

c. When calculating the sample variance, why is it customary to divide the sum of squared deviations from the mean by $(n-1)$ rather than n? (Be brief.)

It is customary to divide by $n-1$ so that the estimator will be unbiased. That is, when S^2 is calculated using $n-1$ as the divisor, $E[S^2] = \sigma^2$.

d. The number of female children in households with two children follows a probability distribution with population mean 1 and population variance $.6$. What is the population mean and variance variance of the number of male children and why? (Be brief).

Let X and Y be the number of male and female children respectively. Then $X = 2 - Y$ and hence

$$E[X] = E[2 - Y] = 2 - E[Y] = 2 - 1 = 1$$

and

$$\text{Var}(X) = \text{Var}(2 - Y) = \text{Var}(Y) = .6.$$

Note that it is not necessary to make any assumptions about the independence of the children or the probability of a specific gender being $.5$.

Some people, very thoroughly, answered this question correcly as follows. Let $P(Y = i) = p_i$ for $i = 0, 1, 2$. Then we know $E[Y] = 1$ so that

$$1 = 0 \times p_0 + 1 \times p_1 + 2 \times p_2$$

and $E[Y^2] = 1.6$ so that

$$1.6 = 0^2 \times p_0 + 1^2 \times p_1 + 2^2 \times p_2.$$

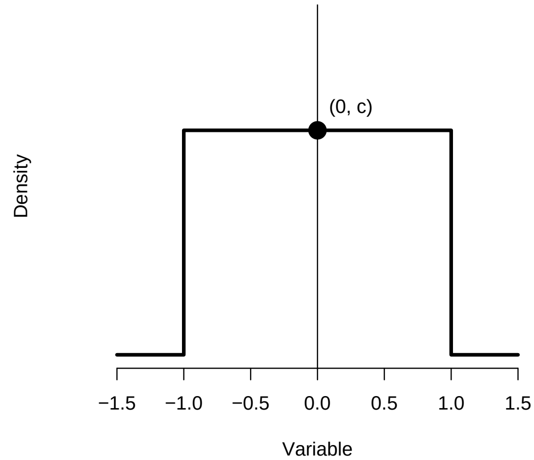
Also, because we are only considering 2 children families,

$$1 = p_0 + p_1 + p_2.$$

This yields three equations with three unknowns, which can be solved to yield $p_0 = .3$, $p_1 = .4$, $p_2 = .3$. Because the number of female children determines the number of male children, you can now calculate the necessary expected values directly.

e. Researchers studying brain volume found that in a random sample of 16 that the average loss in grey matter volume as a person aged four years was $.1mm^3$ with a standard deviation of $.04mm^3$. Calculate the standard error of the mean. Describe (briefly) what $.1mm^3$ and your calculated standard error estimate.

   The standard error is $.04/\sqrt{16} = .01mm^3$. The sample mean, $.1mm^3$ estimates the population mean change in volume. The standard error, $.01mm^3$ estimates the standard deviation of the sampling distribution of the sample mean.



2. The figure above depicts a density that has a constant height c and is positive between -1 and 1.

   a. (9 points) What value of c makes this function a valid density?
   The figure is a box of width 2 and height c and hence area $2c$. To have area one c must then be $1/2$.

   b. (9 points) Argue (you need not calculate) what the mean of this density must be.
   The density is symmetric about 0. Therefore 0 is its mean.

   c. (9 points) What is the 95^{th} percentile of this distribution?
   We want the point x so that $P(X \leq x) = .95$. However, $P(X \leq x) = (x - (-1))/2 = (x+1)/2$. Solving $(x+1)/2 = .95$ for x yields $x = .9$.

   d. (10 points) The variance of this density is $1/3$. Suppose that we sample 10 observations from this density $1,000$ times and take the sample mean of each of the collection of the 10 observations, resulting in $1,000$ sample means. What number would the standard deviation of these $1,000$ number likely be close to?
   It would likely be close to the standard error $\sqrt{1/3/10} = 1/\sqrt{30}$. Note $\sqrt{1/3/1,000} = 1/\sqrt{3000}$ is not correct.

3. Researchers are interested in a new blood test for diagnosing Kryptonite poissoning (a rare disease). For this test a *positive* result is supposed to indicate the *presence* of Kryptonite poissoning. A study found that 85% of patients who are known to have the disease were *positive* on the blood test. In contrast, 21% of individuals who are known not to have the disease, were *positive* on the blood test.

   a. (9 points) A person has a positive blood test result. Interpret this positive test result without knowledge of the disease prevalence.
   The problem gives sensitivity $= .85$ and specificity $= 1 - .21 = .79$. The diagnostic likelihood ratio is $.85/.21 = 4.05 \approx 4$. Therefore the evidence of poisoning is around 4 times higher in light of the positive test result. Or you could say that the positive test result increases the odds of disease fourfold. Note that there were several people who said statements like "This patient is now 4 times as likely to have the disease". This statement is poorly worded. First, the DLR multiplies the odds. Secondly, it could be argued that the person either has the disease or does not. You could more correctly qualify the statement by saying "Our belief (represented by the odds) that the person has the disease is 4 times higher after the positive test result". Anyway, you want to avoid wording that suggests that the test causes the person to have the disease; it does not.

   b. (9 points) Given that the prevalence of the disease is 10% in a subject's population with a positive test result, calculate the probability that the person has the disease.
   The obvious way to do this problem is to use Baye's rule. Note that the statement was corrected during the test so that we know the prevalence of disease is $.1$. Therefore

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^c)P(D^c)} = \frac{.85 \times .1}{.85 \times .1 + .21 \times .9} = .31.$$

   Another interesting approach uses the fact that if p is the probability then $odds = p/(1-p)$ and hence $p = odds/(1 + odds)$. The post test odds of disease is $(.85/.21) \times (.1/.9) = .4497$. Converting these odds to a probability yields $.4497/(1 + .4497) = .31$.