STAT 311: Practice Final

August 18, 2016

1 Take a Hike

I am planning out a schedule for the few weeks I have between the end of summer quarter and the start of school. I have a list of 10 hikes which I would like to try. However, because of limited time, I will only hike 5 of those routes.

1. *How many different schedules of 5 hikes can I make if the order of the hikes matters but I do not want to repeat any hike?*

$$P_1^5 0 = \frac{10!}{5!} = 30240$$

2. *How many different schedules of 5 hikes can I make if the order of the hikes does not matter but I do not want to repeat any hike?*

$$C_1^5 0 = \binom{10}{5}\frac{10!}{5!5!} = 252$$

3. *I find out that the first two days I plan on hiking, only 4 of the trails will be open and the other three days I plan on hiking, only the other 6 trails will be open. What is the number of schedules I can make if I do not want to repeat any hikes and order does not matter?*

$$\binom{4}{2}\binom{6}{3} = 120$$

4. *At Lake Twenty-Two, suppose the number of other people I see on the trail is distributed as a Poisson random variable with a mean of $\lambda = 3$ (if I go early in the morning). What is the probability of seeing exactly 5 other people on the trail? What is the probability of seeing at least one other person on the trail?*

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$P(X = 3) = \frac{3^3 e^{-3}}{5!} = .1008$$

$$P(X > 0) = 1 - P(X = 0) = 1 - \frac{3^0 e^{-3}}{0!} = 1 - .0498 = .9502$$

5. *I am packing a bag of pretzels to eat on the trail. Suppose the number of pretzels I put in the bag is a random variable with a mean of 100 and a variance of 30. Suppose the number of pretzels I eat is a random variable with a mean of 30 and a variance of 10. What is the mean and variance of the number of pretzels I will have remaining?*

$$\mathbb{E}(Remaining) = \mathbb{E}(Pack - Eat) = \mathbb{E}(Pack) - \mathbb{E}(Remain) = 100 - 39 = 70$$

$$Var(Remaining) = Var(Pack - Eat) = Var(Pack) + Var(Remain) = 30 + 10 = 40$$

6. *Suppose I plan on hiking Lake 22 and Mount Pilchuck in one day. The time it takes to hike Lake 22 is a normally distributed random variable with mean of 230 minutes and a standard deviation of 20 minutes. The time it takes to hike Mount Pilchuk is a normally distributed random variable with a mean of 350 minutes and a standard deviation of 25 minutes. The time it takes to travel between the two trails is a normally distributed*

random variable with a mean of 25 minutes and a standard deviation of 5 minutes. What is the distribution of the total time it will take to hike both trails (including travel time between)?

$$\mathbb{E}(Total) = \mathbb{E}(Lake22 + Travel + Pilchuk) = \mathbb{E}(Lake22) + \mathbb{E}(Travel) + \mathbb{E}(Pilchuk) = 230 + 350 + 25 = 605$$

$$Var(Total) = Var(Lake22 + Travel + Pilchuk) = Var(Lake22) + Var(Travel) + Var(Pilchuk) = 20^2 + 25^2 + 5^2 = 1050$$

Furthermore, we know the sum of normal random variable is also normally distributed. So even more than just the mean and variance, we know the shape of the distribution.

$$Total \sim \mathcal{N}(605, 1050)$$

7. *What is the probability that it will take between 600 and 650 minutes to hike both trails (including travel time in between)?*

   We are interested in finding

   $$P(600 < Total < 650)$$

   which we can calculate directly via (it may help to draw out and shade the region under a normal curve. Then note that the shaded region corresponds to the differnce in the two regions stated below)

   $$= P(600 < Total) - P(Total < 650)$$

   Finding the Z-scores of 600 and 650 respectively

   $$\frac{600 - 605}{\sqrt{1050}} = -.15$$

   $$\frac{650 - 605}{\sqrt{1050}} = 1.39$$

   Looking up the areas associated with each z-score yield .918 and .440 respectively. Thus $P(600 < Total < 650) = .918 - .440 = .478$

8. *On each hike I take, there is a .1 chance that I trip and fall. Assuming that the chances of me falling on each hike are independent, what is the distribution of the number of trails I fall on, if I hike 5 trails? Give a name and specify all parameters.*

   Binomial distribution with n = 5 and p = .1

9. *What is the probability that I fall on exactly 2 hikes?* For a binomial we have the following PMF-

   $$P(X = x) = \binom{n}{x} p^x (1-p)^{n-k}$$

   So for our specific example

   $$P(X = 2) = \binom{5}{2} .1^2 .9^3 = .0729$$

2 Crowded Hike

My friend and I are deciding whether to go to Rattlesnake Ledge or Little Si. He claims that Rattlesnake ledge is always more crowded than Little Si, but I think the opposite is true. Being a statistician, I look into the data. Let μ_{RL} be the average number of hikers at Rattlesnake Ledge each day and let μ_{LS} be the average number of hikers at Little Si each day. The parameter of interest is $\mu_{RL} - \mu_{LS}$. I have way too much free time so I spend 10 days at Rattlesnake Ledge and observe an average of 180 hikers and a standard deviation of 12 hikers. My friend counts hikers at Little Si for 15 days and observes an average of 105 hikers and a standard deviation of 10 hikers.

1. *What is the best estimate of the parameter of interest. Write this in symbols and also give a numeric value.*

$$\bar{x}_{rl} - \bar{x}_{ls} = 180 - 105 = 75$$

2. *In order to form a 90% confidence interval for the parameter of interest, what distribution should we use? Give a name and all necessary parameters. What is the value of the multiplier?*

   For a difference of means, we have a T distribution with $\min(n_1 - 1, n_2 - 1)$ degrees of freedom, so in this case we have 9 degrees of freedom. Since we are forming a 90% confidence interval, we find the value which gives .05 to the right. Looking this up in the table yields

$$t^\star = 1.833$$

3. *What is the standard error we should use? Write this in symbols and also give a numeric value.*

   For a difference of means, the standard error is

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{12^2}{10} + \frac{10^2}{15}} = 4.590$$

4. *Form a 90% confidence interval for the parameter of interest.*

$$\bar{x}_{rl} - \bar{x}_{ls} \pm t^\star SE = 75 \pm 1.833 \times 4.590$$

   This gives us an interval of-

$$(66.587, 83.413)$$

5. *Based on the confidence interval you formed, do you have strong evidence about which trail typically has more hikers? Explain.*

   Yes, because the confidence interval we created does not contain 0, we have strong evidence (at the 90% level) that the true mean is positive and thus that Rattlesnake Ledge is busier than Little Si.

6. *Explain in plain English what is implied by a 90% confidence interval.*

   If we repeat this procedure many times, we will get many different confidence intervals since they are based on the samples we take. However, 90% of the time we repeat this procedure, the confidence interval we create will contain the true difference in the average number of hikers on Rattlesnake Ledge and average number of hikers on Little Si.

3 Buying Boots

I am looking to purchase hiking boots for my trips, but I want to ensure that the boots I purchase are durable and well-made. I have decided that I will only purchase a pair of boots, if I am absolutely sure they have over 90% favorable reviews on REI.com (suppose for this problem that the reviews are representative of the overall population). For a particular pair of boots, I see that 250 of the 290 reviews are favorable. Let p denote the proportion of favorable reviews. I am interested in testing

$$H_0 : p = .9$$

$$H_A : p > .9$$

1. *What is the distribution of $\hat{p}$ under the null distribution? Give a name and specify all parameters.*

   $$\hat{p} \sim \mathcal{N}(.9, .9(.1)/290)$$

2. *What test statistic should we use? Write it in symbols and give a numeric value.*

   $$\hat{p} = 250/290 = .86$$

   Note that in class I used the following test statistic-

   $$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.86 - .9}{\sqrt{\frac{.9(.1)}{290}}} = -2.27$$

   This also works if our null distribution is a standard normal distribution, instead of the distribution we gave in part 1. To be consistent with the tables I have given in the formula sheet and the notes, we'll use $\hat{p}$.

3. *What is the p-value for the test statistic?* To convert to a p-value, we first get the z-score

   $$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.86 - .9}{\sqrt{\frac{.9(.1)}{290}}} = -2.27$$

   and look up the area to the right of that value. Note that because the alternative is $p > .9$, our definition of "extreme" is a large value of $\hat{p}$, which is why any value more extreme than what we got is any value with a z-score larger than -2.21.

   $$P(Z > -2.27) = 1 - P(Z < -2.27) = 1 - .0116 = .988$$

4. *For a hypothesis test with level .05, what would you conclude based on the p-value?*

   At a level of .05, we would fail to reject the null hypothesis. Thus we do not have evidence that the true proportion of positive reviews for these boots is different from .9.

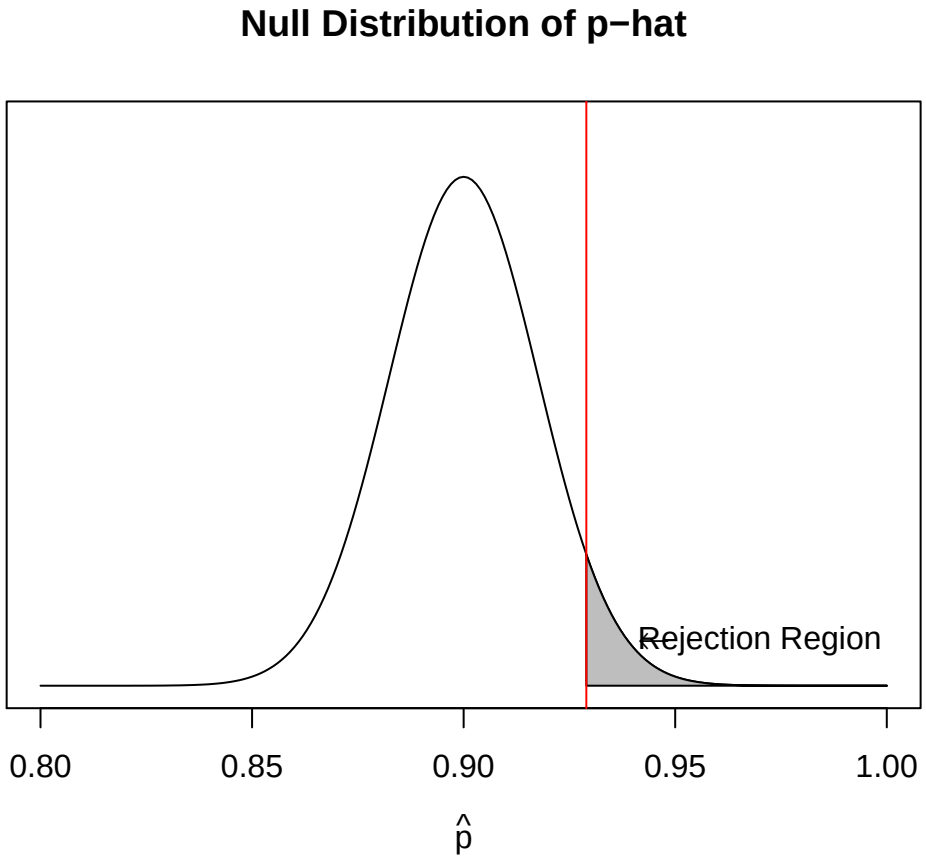
5. *Explain the p-value in plain English.* If the true proportion of all reviews that are favorable is .9, the probability of observing a value at least as large as .86 is .988.

6. *Draw the distribution of $\hat{p}$ under the null distribution. Mark the cutoff and shade the region of possible $\hat{p}$ values for which we would reject the null hypothesis given a level of .05.*

   For a level of .05, we will reject when $\hat{p}$ has a z-score greater than 1.645 (because we need 5% in the upper tail). Thus, we can solve for a cutoff value

   $$1.645 = \frac{p_{cutoff} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{p_{cutoff} - .9}{\sqrt{\frac{.9(1-.9)}{290}}}$$

   Solving for p_{cutoff} yields a value of .929.



7. *For a hypothesis test with level of .05, what is the probability of a type I error when the null hypothesis is true?*

   By definition, the probability of a Type I error is also the level. So in this case it is .05.

8. *If the true proportion of positive reviews is actually .94, what is the power of the hypothesis test?*

   If the true proportion of positive review is actually .94, then

$$\hat{p} \sim \mathcal{N}(.94, \frac{.94(.06)}{290})$$

   We will reject the null hypothesis if $\hat{p} > .929$ (from the question above). The power is the probability of correctly rejecting the null hypothesis when the null hypothesis is false. So in this case, to calculate the power, we just need to find the probability of rejecting the null or equivalently finding the probability that $\hat{p} > .929$ under the true distribution.

   So we find the Z-score of .929 (under the true distribution)

$$\frac{.929 - .94}{\sqrt{\frac{.94(.06)}{290}}} = -.789$$

   and the area to the right of -.789 (because we would reject the null when the $\hat{p}$ is greater than .929) is .785.

   Thus the power when $p = .94$ is .785.

9. *Suppose that if I decide that the true proportion of positive reviews is actually greater than 90%, I will purchase the boots with probability .85. However, if I fail to reject the null hypothesis, the probability that I purchase the boots is .2. Given the power of the test which you calculated above, what is the probability that I will purchase the boots?*

   By the intersection rule

$$P(Reject \bigcap Buy) = P(Reject)P(Buy|Reject) = .785(.85) = .667$$

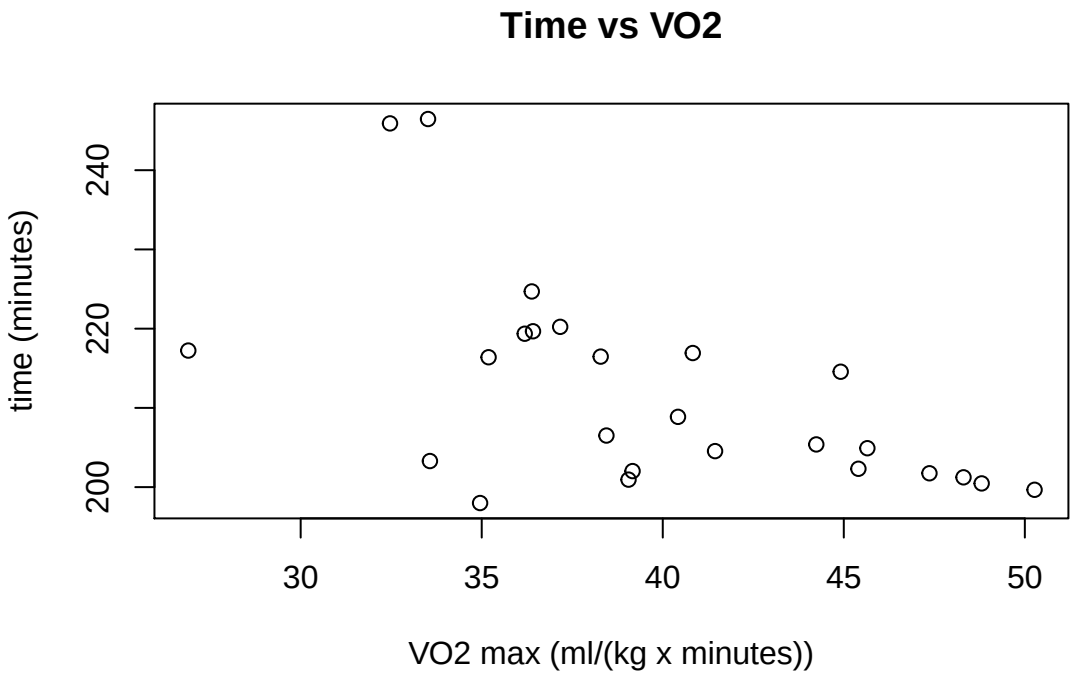
$$P(\text{Fail to Reject} \bigcap Buy) = P(\text{Fail to Reject})P(Buy|\text{Fail to Reject}) = .215(.2) = .043$$

   Then buy the rule of total probability

$$P(Buy) = P(Reject \bigcap Buy) + P(\text{Fail to Reject} \bigcap Buy) = .710$$

4 Hiking Times

VO_2 max is a way to measure the rate at which an individual can consume oxygen and is a way to directly measure aerobic fitness levels. Suppose, I am interested in measuring how an individuals VO_2 max level is related to how quickly they can hike up Rattlesnake Ledge. I take 25 individuals and measure their VO_2 max levels and then time how long it takes them to hike up and down Rattlesnake Ledge. The data is presented below.



Suppose we know the following quantities

- $r_{vo2,time} = -0.576$
- $s_{vo2} = 5.872$
- $s_{time} = 13.006$
- $\bar{y}_{time} = 211.911$
- $\bar{x}_{vo2} = 39.814$

We assume a model-

$$\text{time} = a + b \times vo_2 + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$

1. *Is this study an observational study or experiment? Explain.*
   This is an observational study since we are not manipulating the VO_2 levels.
2. *Give the regression equation for predicting hiking times from vo_2 max levels?*

$$\hat{b} = r\frac{s_y}{s_x} = -.576\frac{13.006}{5.872} = -1.276$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 211.911 + 1.276(39.814) = 262.714$$

So putting everything together yields the prediction equation

$$\hat{y}_i = 262.714 - 1.276x_i$$

or equivalently the equation which specifies the actual underlying process

$$y_i = 262.714 - 1.276x_i + \epsilon_i$$

3. *For an individual with a vo_2 max level of 45, what would you predict their hiking time to be?*

$$\hat{y} = 262.714 - 1.276(45) = 205.294$$

4. *Calculate SS_{total}, $SS_{regression}$ and SS_{error}.*

$$SS_{total} = s_y^2(n-1) = 4059.745$$
$$SS_{reg} = SS_{total}r^2 = 1346.926$$
$$SS_{error} = SS_{total} - SS_{reg} = 2712.819$$

5. *Suppose we want to test whether or not there is a relationship between vo_2 max and hiking time. State the null and alternative hypothesis*

$$H_0 : b = 0$$
$$H_A : b \neq 0$$

6. *Calculate $se(\hat{b})$.*

$$se(\hat{b}) = \frac{s_{reg}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

Note that s_reg is **not** the same as SS_{reg}. In fact, s_{reg} is our estimate of σ when $\epsilon \sim \mathcal{N}(0, \sigma^2)$

$$s_{reg} = \sqrt{\frac{SS_{error}}{n-2}} = \sqrt{\frac{2712.819}{23}} = 10.86$$
$$\sum_i (x_i - \bar{x})^2 = s_x^2(n-1) = 827.529$$

So plugging everything in-

$$se(\hat{b}) = \frac{10.86}{\sqrt{827.529}} = .378$$

7. *What is the appropriate test statistic? Write it in symbols and also give a numeric value.*

$$\frac{\hat{b} - b_0}{se(\hat{b})} = \frac{-1.276}{.378} = 3.376$$

8. *What is the null distribution. Give a name and specify all appropriate parameters. Given this null distribution, what is the p-value?*

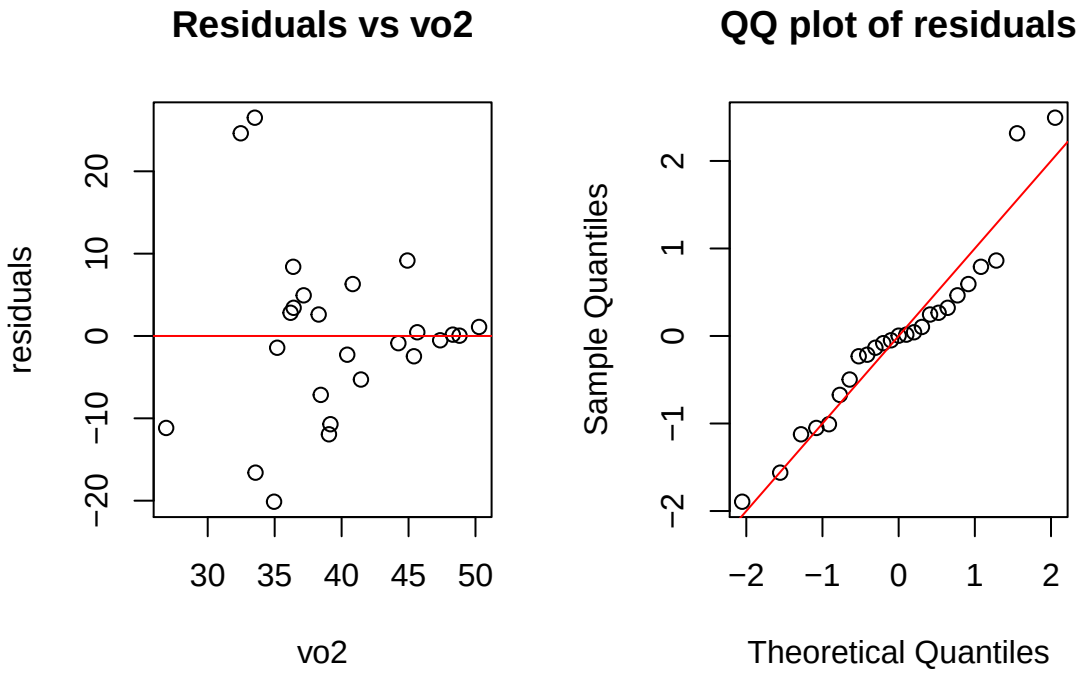
Under the null hypothesis, this should have a T distribution with $n-2$ degrees of freedom. In this case, we have $25 - 2 = 23$ degrees of freedom

In the given T table, we cannot find the p-value directly, however, we know that the cut-off falls between the z-scores for p-values of .0025 and .001. Thus, we know the area to the right of 3.372 is between .0025 and .001. Since we are using a two-sided test, we are interested in

$$P\left(|\frac{\hat{b}}{se(\hat{b})}| > 3.376\right)$$

so we need to consider values which are both extremely large (positive) and extremely small (negative). Thus, we look at the area in both tails and multiply the area we looked up above by 2. Thus, we have a p-value less than $.0025 \times 2 = .005$ but greater than $.001 \times 2 = .002$.

9. *Viewing the residual plots below, state whether or not you think each of the assumptions of linear regression might hold. Explain.*



We can see that the residuals are more variable at the low end of values of vo_2 max when compared to the residuals for higher values. Thus, we see that the homoskedasticity assumption likely does not hold. Other than that, we don't really see any systematic pattern to the residuals, so linearity seems okay. The QQ-plot should yield a roughly straight line if the residuals are normally distributed. Viewing the QQ-plot, they look roughly normal, but could go either way. Just be sure to explain why you think the assumption holds or not.

10. *Give an interval that we would expect to contain the hiking times of 90% of the individuals which have a vo_2 max score of 40.*

In this case, we are looking at a range of values for specific individuals (rather than a conditional mean), so we need a predictive interval.

$$\hat{y} \pm t^{\star}_{n-2}\sqrt{s^2_{reg} + se(fit)^2}$$

$$\hat{y} = 262.714 - 1.276(40) = 211.674$$

We calculated $s_reg = 10.86$ above and for $x = 40$, we have

$$se(fit) = s_{reg}\sqrt{\frac{1}{n} + \frac{x_j - \bar{x})^2)}{\sum_i (x_i - \bar{x})^2}} = 10.86\sqrt{\frac{1}{25} + \frac{(40 - 39.814)}{827.529}} = 2.17$$

For the multiplier, we should get a value from a T distribution with 23 degrees of freedom. Since w e have a predictive interval of 90% we need .05 in each tail, which yields a value of 1.714

Putting everything together yields a predictive interval of

$$211.674 \pm 1.714\sqrt{10.86^2 + 2.17^2}$$

$$(192.692, 230.656)$$

11. *Give a 95% confidence interval for the average hiking times of the individuals which have a vo_2 max level of 40.* This is a confidence interval for the (conditional) mean of all individuals with a vo_2 max level of 40. Thus, we have a confidence interval for conditional means

$$\hat{y} \pm t^{\star}_{n-2} se(fit)$$

Since we have a 95% confidence interval, we need a multiplier with .025 in each tail, which is 2.068. Forming the confidence interval yields-

$$211.674 \pm 2.068(2.17)$$

$$(207.186, 216.162)$$

5 Post Hike Recovery

My friend and I are debating what to drink after the hike, and he suggests that muscle soreness depends on your post-hike beverage. I'm skeptical, but again, we gather data from 300 hikers on their muscle soreness and what they decide to drink after a hike.

	Muscle Condition		
Drink	Sore	Not Sore	Total
Water	20	19	39
Gatorade	33	42	75
Chocolate Milk	10	43	53
Total	63	104	167

1. *From the table, estimate the joint probability of drinking water and not being sore*

$$\frac{19}{167} = .11$$

2. *From the table, estimate the conditional probability of being sore given that an individual drank Gatorade.*

$$\frac{33}{75} = .44$$

3. *Suppose we want to test H_0: No relationship between beverage and soreness versus H_A: There is a relationship between beverage and soreness. Under the null hypothesis, calculate the expected values for each cell*

	Muscle Condition		
Drink	Sore	Not Sore	Total
Water	14.71	24.29	39
Gatorade	28.29	46.7	75
Chocolate Milk	19.99	33.01	53
Total	63	104	167

   We use the formula for expected values under the assumption of independence $\frac{n_{r+}n_{+c}}{n_{++}}$ to calculate expected counts for each cell.

4. *Calculate the appropriate test statistic to test the two hypothesis above. What is the distribution of the test statistic under the null distribution. Give a name and all appropriate parameters.*

   Letting O_{ij} be the observed counts for cell (i,j) and E_{ij} be the expected count for cell (i,j), we calculate the test statistic

$$\chi^2 = \sum_{ij} \left(\frac{O_{ij} - E_{ij}}{E_{ij}} \right)^2 = 12.33$$

   Which should follow a χ^2 distribution with $(R-1)(C-1) = 2 \times 1$ degrees of freedom.

5. *What is the p-value of the test statistic. Based on this p-value what would you conclude?*

   Based on the χ^2 table, we can see that the observed value is much larger than the cut-off for .005, so we can conclude that the p-value is less than .005. So at the .05 level, we reject the null hypothesis that beverage is independent of muscle soreness.