Math 321 – Summer 2019

additional practice for chapter 6, hypothesis testing

1. If we flip a fair coin 100 times, and get 60 heads. Test the hypothesis that the coin is fair at the 5% significance level.

   Solution:

   $H_0 : p = 0.5$
   $H_a : p \neq 0.5$

   A 1-sided test would also be reasonable here.

   The test statistic is $z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.6-0.5}{\sqrt{(0.5)(0.5)/100}} = 2$. For the 2-sided test, we calculate the p-value as $2 \cdot P(Z < -|z|) =$ `2*pnorm(-2)` ≈ 0.04550026. Thus we would reject H_0 at the 5% level, but just barely, so-to-speak. We would also reject it if we instead did a 1-sided test with alternative $H_a : p > 0.5$. So we conclude that 60 heads out of 100 flips for a fair coin is a bit unusual.

   We could also exactly calculate the p-value with the binomial as $P(X \leq 40) + P(X \geq 60) =$ `pbinom(40,100,0.5)+1-pbinom(59,100,0.5)`≈ 0.05688793. Note that this would lead us to not reject the null hypothesis of fairness. However, we would still reject H_0 with a 1-sided "greater" test. Note that if we did not change the 60 to a 59 in this calculation, it would result in a $p < 0.05$!

   This illustrates the fact that when your p-value is close to the desired significance level, you really can't be perfectly sure what the actual truth is. In such cases, it would be better to collect more data or to more carefully think about how the data was gathered. Remember that most statistical tests involve many assumptions and approximations. So any underlying uncertainty or errors might easily result in a different conclusion if our calculated p-value is near α.

2. Consider that a group of 25 people is randomly chosen from of a certain demographic category. The mean height is found to be 5 foot 8 inches and the sample standard deviation is found to be 2 inches. Test the hypothesis that the true mean height of this demographic category is 5 ft 9in at the 5% significance level.

   Solution:

   $H_0 : \mu = 5.75$
   $H_a : \mu \neq 5.75$

   A 1-sided test might also be reasonable here.

   Test statistic: $t = \frac{5.667-5.75}{0.167/\sqrt{25}} = -2.485$.

   Even though our sample size is less than 30, and that we do not know the underlying population standard deviation, using a z-test instead would be reasonable since our sample size is not much below 30 and because height is known to be very close to normally-distributed.

   The p-value is $2 \cdot P(t_{n-1} < t) =$ `2*pt(-2.485,24)` ≈ 0.02 thus we reject the null hypothesis. So our data presents evidence that the true mean height is probably less than the claimed mean, but the evidence is not really that strong. Again, sampling errors, random variation, or other uncertainty could easily cause such a low sample mean.

3. An internet server has data requests arrive at a very high rate. The number of data requests in a minute is collected for 60 randomly chosen 1 minute intervals. The sample mean requests per minute is found to be 982. Test the hypothesis that true mean number of requests per minute is 1000 at the 5% significance level. Consider that the number of requests in a minute can be modeled by a Poisson RV.

   Solution:

   $X \sim Pois(\lambda)$ where λ is the mean number of requests in one minute. We wish to test the hypothesis that $\lambda = 1000$. Either a 1-sided or 2-sided test could be justified here. I will perform a 1-sided test.

$H_0 : \lambda = 1000$
$H_a : \lambda < 1000$. I choose this alternative since our sample mean is below the null mean otherwise we are guaranteed a p-value above 50%.

So under the assumption that the null hypothesis is true, for each one minute interval, $X_i \sim Pois(\lambda = 1000)$ for $i = 1, 2, ..., 60$ each with mean and variance equal to $\lambda = 1000$. Thus $\overline{X} \sim N(\mu = 1000, \sigma^2 = \frac{1000}{60})$. We calculate the probability $P(\overline{X} \leq 982)$. In R this is `pnorm(982,1000,sqrt(1000/60))` $\approx 5.2(10)^{-6}$. So we reject the null hypothesis. This sample is fairly strong evidence for the true mean number of requests in a minute beign less than 1000.

Note that we could also perform this test by just scaling our λ up to 60 minutes. We expect 60,000 total requests in the sample of 60 minutes, but with a sample mean of 982, that gives a total of $982 \times 60 = 58,920$ requests. We can use the Poisson cdf since we know that $Y \sim Pois(\lambda = 60000)$ where Y is the number of requests in 60 minutes. Thus our p-value calculated using this is `ppois(58920,60000)` $\approx 4.9(10)^{-6}$. So we still reject H_0. Often, you will find many optional methods for performing statistical tests. Usually such small p-values means that the type of test used will generally not change the conclusion, but as mentioned already, when p-values are close to *alpha* the type of test used can often change the conclusion.

4. Recent polling data of 5,000 individuals indicate an approval rating for Trump of 43%. At the same time during Obama's first term his approval rating was 45%. Test the hypothesis that Trump's approval rating is at least as good as Obama's to the 5% significance level.

   Solution:

   The claim to test is that $p \geq 0.45$ for the true proportion of voters approving of Trump.

   $H_0 : p \geq 0.45$
   $H_a : p < 0.45$.

   Our test statistic is: $z = \frac{0.43 - \hat{0}.45}{\sqrt{\frac{\hat{0}.45(1-0.45)}{5000}}} \approx -2.842676$. And this gives a p-value of `pnorm(-2.84)` $\approx 0.00224 < 0.05$. Thus we reject the null hypothesis. The data supports the idea that Trump's approval rating is truly below 45%. Note that this doesn't make any specific claim on what the real approval rating is. It could be only slightly below 45% or it could be a lot lower. We can construct a confidence interval, and using the rule of thumb $\hat{p} \pm 1/\sqrt{n}$ gives (0.4160.444). This CI almost captures 45%. In fact, setting α equal to our p-value will give us a confidence interval that captures 0.45.

5. A construction materials manufacturer claims that a particular material of theirs can withstand 5,000 tons before failing. A sample of size 7 is randomly selected and tested until failure. The mean force until failure is found to be 4,900 lb with a standard deviation of 200 lb. Test the manufacturer's claim at the 5% level.

   Solution:

   This is an instance where a t-test is highly advised. The sample size is very small, and we really don't know much about the underlying nature of construction material strength. It is reasonable that it should be approximately normally distributed.

   $H_0 : \mu \geq 5000$
   $H_a : \mu < 5000$.

   Our test statistic is $t = \frac{4900 - 5000}{\frac{200}{\sqrt{7}}} \approx -7.07106$

   This looks to be very far away from the central value of zero, but remember that we are doing a t-test and that the quantiles for a low degrees of freedom t-distribution will be further from the central value.

   The p-value is `pt(-7.07,6)` ≈ 0.0002006034. So we reject the null hypothesis. The data supports the conclusion that the construction materials do not meet the manufacturer's claim.

   Note that a z test would have given a MUCH smaller p-value for such a small sample size. This t-test shows that although we reject H_0, our p-value is really not too small. Random variation, sampling

error, or other uncertainty could possibly change the conclusion. So we are not too sure what the actual true mean value is. It makes sense to reject the claim here since for building material strength it is probably very important to meet specifications.

6. Here is the racial and ethnic makeup of US active duty military in 2016

Am Ind / Ak Nativ	Asian	Blk / Afr Am	Hisp / Lat	Multiple	Nat Haw / Other	White
2100	15861	63380	60466	14142	4556	320801

Here is what the actual general population demographics were at the time:

Am Ind / Ak Nativ	Asian	Blk / Afr Am	Hisp / Lat	Multiple	Nat Haw / Other	White
0.013	0.057	0.13	0.166	0.028	0.002	0.604

Test the hypothesis that the military is chosen randomly from the population, i.e. that the military demographic proportions are actually those of the general population.

Solution:

The hypotheses are:

H_0 : the true military percentages are the same as the general population (i.e. that the military is a random sample from the general population)
H_a : At least 1 demographic group is represented int eh military differently from the general population.

Here are the tabulated observed and expected counts:

Race/Ethn.	AmInd/AkNat	Asian	Blk/Afr	Hisp/Lat	Mult	NatHaw/Oth	White
observed	2100	15861	63380	60466	14142	4556	320801
expected	6256.978	27434.442	62569.78	79896.796	13476.568	962.612	290708.824

The test statistic is $c = \sum \frac{(obs-exp)^2}{exp} = \frac{(2100-6256.978)^2}{6256.978} + \cdots + \frac{(320801-290708.824)^2}{290708.824} \approx 28941.93$.

The p-value is `1-pchisq(28941.93,6)` ≈ 0. Thus we reject the null hypothesis that the military demographics are the same as the general population. The data supports the idea that the military is not randomly selected from the population.

7. The number of casualties in Operation Iraqi Freedom are given below according to gender and military branch.

	Army	Navy	Marine Corps	Air Force
Female	547	6	41	33
Male	21683	547	8573	417

Test for independence of gender and military branch to the 5% level.

Solution:

Here is the tabulated data with row and column totals:

	Army	Navy	Marine Corps	Air Force	TOTALS
Female	547	6	41	33	627
Male	21683	547	8573	417	31220
TOTALS	22230	553	8614	450	31847

Here are the data proportions for each cell:

	Army	Navy	Marine Corps	Air Force
Female	0.0172	0.0002	0.0013	0.0010
Male	0.6808	0.0172	0.2692	0.0131

Here are the hypothetical proportions for each cell under the assumptions of independence. For each cell we multiply the row and column totals and divide by N^2.

	Army	Navy	Marine Corps	Air Force
Female	0.013742633	0.000341866	0.005325193	0.000278191
Male	0.684282299	0.017022407	0.265155543	0.013851868

Notice that these mostly look fairly close to the proportions from our actual data.

Here are the hypothetical counts under the assumption of independence:

	Army	Navy	Marine Corps	Air Force	TOTALS
Female	437.66	10.89	169.59	8.86	31220
Male	21792.34	542.11	8444.41	441.14	31220
TOTALS	22230	553	8614	450	31847

These numbers look mostly similar to our data counts except for Female numbers seem to deviate much from what we expect under the assumption of independence. The counts for females in Army and Air Force in our data are much higher and those in Navy and Marine Corps are much lower compared to the expected counts.

Our test statistic is $c = \sum \frac{(obs-exp)^2}{exp} = \frac{(547-437.66)^2}{437.66} + \cdots + \frac{(417-441.14)^2}{441.14} \approx 196.66$ and our degrees of freedom is $(r-1)(c-1) = 3$.

The p-value is `1-pchisq(196.66,3)` ≈ 0. So we reject the null hypothesis. If we enter our data counts into R stored as `x` then we can perform this test with `chisq.test(x)`.

Even though we reject the null hypothesis, comparing the data proportions and the expected proportions, we might like to say that although gender and military branch are not independent, they are not "too far" from independence. In fact, there is a statistical concept called "effect size". The effect size for this example is actually small. This quantifies the intuition that our factors considered although nto independne, they are not too dependent.