DATA-606
HW-6
Dhananjay Kumar

6.6 2010 Healthcare Law. On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 46% of 1,012 Americans agree with this decision. At a 95% confidence level, this sample has a 3% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.[39]

. (a) We are 95% confident that between 43% and 49% of Americans in this sample support the decision of the U.S. Supreme Court on the 2010 healthcare law.

. (b) We are 95% confident that between 43% and 49% of Americans support the decision of the U.S. Supreme Court on the 2010 healthcare law.

. (c) If we considered many random samples of 1,012 Americans, and we calculated the sample proportions of those who support the decision of the U.S. Supreme Court, 95% of those sample proportions will be between 43% and 49%.

. (d) The margin of error at a 90% confidence level would be higher than 3%.

Solution: a) False: We are 95% confident that between 43% and 46 % American population and NOT just this sample support this decision. The estimates are made for the population and not just for sample.

b) True: This is what the confidence interval means.

c)False: This statistic is about the true population and not of the samples.

d)False: The margin of error for 90% CI would be less. As the Z value for 90% CI is 1.645 and for 95% is 1.96. So the CI will be narrower and the Margin error will also be less.

6.12 Legalization of marijuana, Part I. The 2010 General Social Survey asked 1,259 US residents: “Do you think the use of marijuana should be made legal, or not?” 48% of the respondents said it should be made legal.[44]

. (a) Is 48% a sample statistic or a population parameter? Explain.

. (b) Construct a 95% confidence interval for the proportion of US residents who think marijuana should be made legal, and interpret it in the context of the data.

. (c) A critic points out that this 95% confidence interval is only accurate if the statistic follows a normal distribution, or if the normal model is a good approximation. Is this true for these data? Explain.

. (d) A news piece on this survey's findings states, "Majority of Americans think marijuana should be legalized." Based on your confidence interval, is this news piece's statement justified?

Solution:a) It's a sample statistic.

b) p = 0.48

1-p = 1-0.48 = 0.52
S.E = sqrt(p(1-p)/n)
Sqrt(0.48*0.52/1259) = 0.014

For 95% CI, Z =1.96
0.48 +- 1.96* 0.014
= **(0.45256,0.50744)**

C) There are two conditions for p to be nearly normally distributed:

1) The observations should be independent. The sample taken must be from 10 % of the true population. For this survey we can reasonably rely that this is true.
2) Success-failure condition: The sample size should be large enough such that number of success and failure should be greater than 10. 42% and 52% of 1259 are greater than 10.

d) Based on the confidence interval, we can say majority of Americans think marijuana should be based legalized. This is because the upper limit of confidence interval contains 0.50744.

6.20 Legalize Marijuana, Part II. As discussed in Exercise 6.12, the 2010 General Social Survey reported a sample where about 48% of US residents thought marijuana should be made legal. If we wanted to limit the margin of error of a 95% confidence interval to 2%, about how many Americans would we need to survey?

Solution:

Margin Error(M.E) = 0.02

M.E = Z* standard error(S.E)

0.02 = 1.96 * sqrt(0.48*0.52/n)

Solving the equation we get,

n = 2400

6.28 Sleep deprivation, CA vs. OR, Part I. According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents. Calculate a 95% confidence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived and interpret it in context of the data.[53]

Solution:

Let P1 = proportion of California residents who are sleep deprived
P2 = proportion of Oregon residents who are sleep deprived

P1 = 0.08
1-P1 = 0.92
n1 = 11545
n1p1 = 923 >10 and n1(1-p1) = 10621 both >10

P2 = 0.088
1-P2 = 0.912
n2 = 4691
n2p2 = 412 and n2(1-p2) = 4278 both >10

For both the samples the condition of success failure hold true.
Each group is a sample random sample from less than 10% of the population. The samples are independent both within the group and between the group.
Since both the conditions are met we can use the normal model for estimation.

P1 – P2 = 0.08 – 0.088 = -0.008

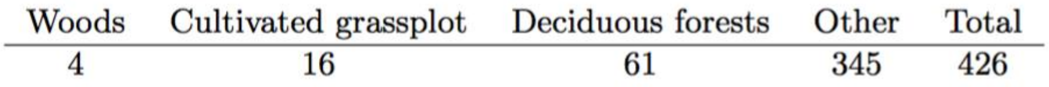
S.E = sqrt(P1(1-P1)/n1 + P2(1-P2)/n2)

= sqrt(0.08* 0.92/11545 + 0.088*0.912/4691)

= 0.0048
95 % CI
-0.008 +- (1.96 * 0.0048)
(-0.017, 0.0015)

The 95% confidence interval between the difference in the proportion of sleep deprivation between Californians and Oregonians is between ~-1.7% to ~0.15%. This means that the proportion of Californians who are sleep deprived can be as much as 1.7% less than Oregonians or the proportions of Californians that are sleep deprived can be as much as 0.15% more than Oregonians for 95% of the samples taken with the given sample sizes. This means that there may also be no difference in the proportion of sleep deprived Californians and Oregonians since the interval includes 0.

6.44 Barking deer. Microhabitat factors associated with forage and bed sites of barking deer in Hainan Island, China were examined from 2001 to 2002. In this region woods make up 4.8% of the land, cultivated grass plot makes up 14.7%, and deciduous forests makes up 39.6%. Of the 426 sites where the deer forage, 4 were categorized as woods, 16 as cultivated grassplot, and 61 as deciduous forests. The table below summarizes these data.[62]



Woods	Cultivated grassplot	Deciduous forests	Other	Total
4	16	61	345	426

. (a) Write the hypotheses for testing if barking deer prefer to forage in certain habitats over others.

. (b) What type of test can we use to answer this research question?

. (c) Check if the assumptions and conditions required for this test are satisfied.

. (d) Do these data provide convincing evidence that barking deer prefer to forage in certain habitats over others? Conduct an appropriate hypothesis test to answer this research question.

<u>Solution: a)</u> H (null): The barking deer doesn't prefer to certain habitat over others.
H (a): The barking deer prefer to certain habitat over others.

. **b)** Since we are given a sample of cases that can be classified into several groups, we can use a chi-square test for this research question.

C) Independence. Each case that contributes a count to the table must be independent of all the other cases in the table.
We assume that all the barking deer habitat variables are independent of each other

All expected counts should be ≥5 The woods habitat has only 4.8 cases.We assume though that this is an acceptable count.

d) H (null): The barking deer doesn't prefer to certain habitat over others.
H (a) : The barking deer prefer to certain habitat over others.

	Woods	**Grassplot**	**Deciduous forests**	**Other**	**Total**
Observed Habitat	**4**	**16**	**61**	**345**	**426**
In %	**0.9**	**3.8**	**14.3**	**81.0**	**100**
Land Make up	**4.8**	**14.7**	**39.6**	**40.9**	**100**

Z (wood) = (0.9-4.8)^2/4.8 = 3.16875

Z (grassplot) = (3.8-14.7)^2/14.7 = 8.0823

Z (forests) = (14.3-39.6)^2/39.6 = 16.1639

Z (other) = (81-40.9)^2/40.9 = 39.3156

Chi-sq = 3.16875 + 8.0823 + 16.1639 + 39.3156 = 66.733055

Df = k-1 = 3

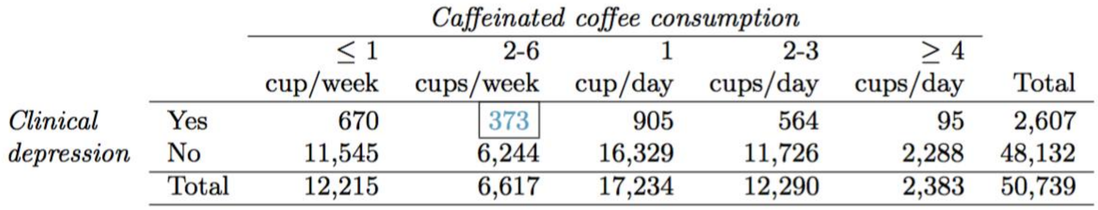
P value using R:

```
pchisq(66.73055,3,lower.tail=FALSE)
```

2.13853e-14

Since this value is small < 0.05 , we can reject the null hypothesis. This means the barking deer prefer to certain habitat over other.

6.48 Coffee and Depression. Researchers conducted a study investigating the relationship between caffeinated coffee consumption and risk of depression in women. They collected data on 50,739 women free of depression symptoms at the start of the study in the year 1996, and these women were followed through 2006. The researchers used questionnaires to collect data on caffeinated coffee consumption, asked each individual about physician-diagnosed depression, and also asked about the use of antidepressants. The table below shows the distribution of incidences of depression by amount of caffeinated coffee consumption.[63]



		Caffeinated coffee consumption					
		≤ 1 cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	≥ 4 cups/day	Total
Clinical depression	Yes	670	373	905	564	95	2,607
	No	11,545	6,244	16,329	11,726	2,288	48,132
	Total	12,215	6,617	17,234	12,290	2,383	50,739

. (a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression?

. (b) Write the hypotheses for the test you identified in part (a).

. (c) Calculate the overall proportion of women who do and do not suffer from depression.

. (d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e. (Observed Expected)^2/Expected.

. (e) The test statistic is ^{2} = 20.93. What is the p-value?

. (f) What is the conclusion of the hypothesis test?

. (g) One of the authors of this study was quoted on the NY Times as saying it was “too early to recommend that women load up on extra coffee” based on just this study.[64] Do you agree with this statement? Explain your reasoning.

Solution: a) Chi-square test for independence is appropriate for evaluating if there is an association between coffee intake and depression.

b) H (null): There is no relationship between coffee intake and depression
H (a): There is relationship between coffee intake and depression

c) Women who are depressed = (2607/50739) * 100 =5.14 %
Women who are not depressed = (48132/50739) *100 = 94.86%

d) Expected count = 2607 * 6617/ 50739 = 339.985
~340
Observed = 373

(373-340)^2/340 = 3.203

e) df = (5-1)*(2-1) = 4
P value using R
pchisq(20.93,4,lower.tail=FALSE)
0.0003269507

f) P value < 0.05 so we reject the null hypothesis. This means we reject that there is no relationship between coffee intake and depression.

g) Yes we agree with the statement. Because even though we rejected the null hypothesis, but it is unclear how strong is the relationship between coffee intake and depression. It is also possible that there are other factors involved.