HW2

Ben Tankus

4/10/2021

R Questions 1. (4 points) It is often argued that victims of violence exhibit more violent behavior toward others. To study this hypothesis, a researcher searched court records to find 908 individuals who had been victims of abuse as children. She then found 667 individuals, with similar demographic characteristics, who had not been abused as children. Based on a search through subsequent years of court records, she was able to determine how many in each of these groups became involved in violent crimes, as shown in the following table. The researcher concluded: ”Early childhood victimization has demonstrable long-term consequences for violent criminal behavior.”

Conduct your own analysis of the data and comment on this conclusion. Is there evidence of a difference between the two groups? Is the strength of the causal implication of this statement justified by the data from this study?

```r
prop.test(c(102,53), c(908,667))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(102, 53) out of c(908, 667)
## X-squared = 4.3205, df = 1, p-value = 0.03765
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.002537413 0.063211651
## sample estimates:
##     prop 1     prop 2
## 0.11233480 0.07946027
```

With a pvalue of 0.03765 and a 95% confidence interval of (0.0025, 0.063) we have evidence to reject the null hypothesis that these two samples are the same.

I do not think the researchers blanket statement is an acceptable conclusion. According to the study methodology the data were collected using a covenience method, NOT a randomized study, therefore causal relationship cannot be made. Also a p-value of 0.03765 while below 0.05 is not a small enough pvalue to make such a strong claim.

I would recommend stating the following:

There is significant evidence that abuse victims in the study are more likely to become involved in violent crime, with a p-value of 0.03765 and confidence interval of (0.0025, 0.063) at a 95% confidence level.

Conceptual Questions

2. (1 point) During an investigation of the U.S. space shuttle Challenger disaster, it was learned that project managers had judged the probability of mission failure to be 0.00001, whereas engineers working on the project had estimated failure probability at 0.005. The difference between these two probabilities, 0.00499, was discounted as being too small to worry about. Is a different picture provided by considering odds? How is that interpreted?

Considering the odds ratio, the engineers estimated the failure rate to be 502.51 times larger than the project managers (0.00001/0.00503). It does seem that this paints a different picture, I certainly wouldn't interpret these things as the same, but I obviously can't speak for NASA.

```
omeg_eng <- 0.00001/(1-0.00001)
omeg_proj <- 0.005/(1-0.005)

odds_ratio <- omeg_proj/omeg_eng
```

3. (2 points) Suppose that 90% of orange tabby cats are male. Determine if the following statements are true or false, and explain your reasoning.

(a) The distribution of sample proportions of random samples of size 30 is left skewed.

TRUE, proportion is so high that it is still left skewed regardless of sample size 30. It also does not satisfy the $np > 5$ and $n(1-p) > 5$ equations

(b) Using a sample size that is 4 times as large will reduce the standard error of the sample proportion by one-half.

TRUE, this is correct because of the standard error equation denominator of sqrt(n).

(c) The distribution of sample proportions of random samples of size 140 is approximately normal.

TRUE, random samples of a sample proportion should be normally distributed with samples of this size (Central Limit Theorem)

(d) The distribution of sample proportions of random samples of size 280 is approximately normal.

TRUE, random samples of a sample proportion should be normally distributed with samples of this size (Central Limit Theorem)

4. (2 points) A 2010 survey asked 827 randomly sampled registered voters in California, ”Do you support or do you oppose drilling for oil and natural gas off the coast of California? Or do you not know enough to say?” Below is the distribution of responses, separated based on whether or not the respondent graduated from college.

(a) What percents of college graduates and non-college graduates in this sample do not know enough to have an opinion on drilling for oil and natural gas off the coast of California (i.e., report two percent values)?

104/438 (23.74%) of college grads and 131/389 (33.68%) of non-grads “do not know”.

(b) Conduct a hypothesis test to determine whether there is evidence that the proportion of college graduates who do not have an opinion on this issue is different from that of non-college graduates.

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

```r
# Determine if sample size is large enough
npNo = 438
np2No = 104*(1-0.2374) #79.3104

npYes = 389
np2Yes = 131*(1-0.3368) #86.88

prop.test(c(104,131), c(438,389))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(104, 131) out of c(438, 389)
## X-squared = 9.5084, df = 1, p-value = 0.002045
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.16333768 -0.03529832
## sample estimates:
##    prop 1    prop 2
## 0.2374429 0.3367609
```

There is significant evidence (pvalue 0.002) to reject the null hypothesis that these two proportions are the same with a 95% confidence interval between (-0.163 and -0.035).

5. (1 point) A study of British male physicians noted that the proportion who died from lung cancer was 0.00140 per year for cigarette smokers and 0.00010 per year for nonsmokers. Additionally, the proportion who died from heart disease was 0.00669 for smokers and 0.00413 for nonsmokers. Which response (lung cancer or heart disease) is more strongly related to cigarette smoking, in terms of the reduction in deaths that could occur with the absence of smoking?

```
cig_death_lung <- 0.00140
noCig_death_lung <- 0.00010

cig_death_heart <- 0.00669
noCig_death_heart <- 0.00413

total_death_smoking <- cig_death_lung + cig_death_heart
total_death_noSmoking <- noCig_death_lung + noCig_death_heart

death_lung <- cig_death_lung + noCig_death_lung
death_heart <- cig_death_heart + noCig_death_heart


cig_death_heart/death_heart
```

```
## [1] 0.6182994
```

```
cig_death_lung/death_lung
```

```
## [1] 0.9333333
```

From these data we can see that 93.33% of the deaths from lung cancer were from the smoking group, while only 62% of the deaths from heart disease were from the smoking group. This means that smoking is a much higher risk factor for lung cancer death than it is for heart disease.