HW1

Nikhil Gopal

1/25/2021

Question 1

- 1
  - nominal
  - ordinal
  - ordinal
  - nominal
  - nominal
  - ordinal

Question 2

- 2
  - response: attitude toward GC, explanatory: gender/education
  - response: heart disease, explanatory: blood pressure/cholesterol
  - response: vote for president, explanatory: race/religion

Question 3

Expected value of a binomial distribution = E[xi] = $n * \pi = 100(0.25) = 25$ Variance of binomial distribution = $n * \pi(1 - \pi) = 100(0.25)(1\text{-}0.25) = 18.75$ Standard deviation = (variance)^0.5 = 4.3301270189

Observing 50 correct responses is more than 3 standard deviations above the mean, meaning there is less than a 1% chance of this happening due to random chance alone. It would be extremely surprising to see these results. Note that the pnorm function which gives us the distribution function of the normal distribution gives an extremely low (near zero value) probability for finding a value greater than 50.

```
pnorm((50-25) / 4.330127, lower.tail  = FALSE)
```

```
## [1] 3.882018e-09
```

Question 4

*4a

```
#P(Y=0)
dbinom(0, 4, 0.6)
```

```
## [1] 0.0256
```

```r
#P(Y=1)
dbinom(1, 4, 0.6)
```

```
## [1] 0.1536
```

```r
#P(Y=2)
dbinom(2, 4, 0.6)
```

```
## [1] 0.3456
```

```r
#P(Y=3)
dbinom(3, 4, 0.6)
```

```
## [1] 0.3456
```

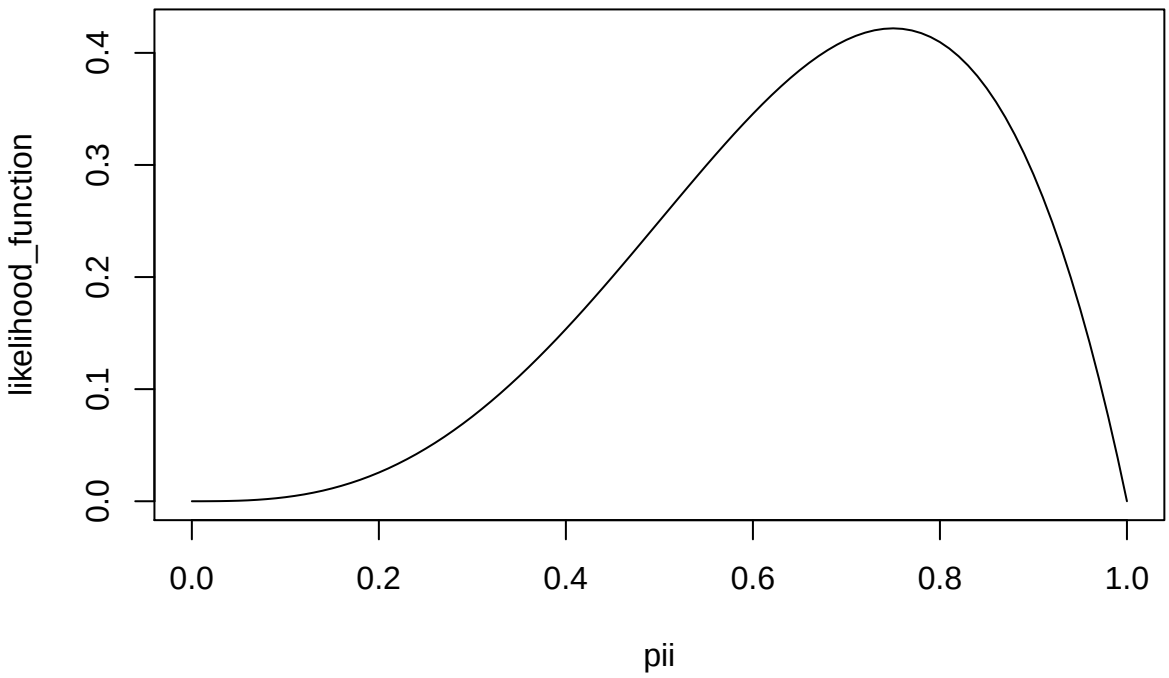
```r
#P(Y=4)
dbinom(4, 4, 0.6)
```

```
## [1] 0.1296
```

- Mean of binomial random variable = np = *4* 0.6 = 2.4 people support an increase in minimum wage
- Standard deviation of binomial random variable = sqrt[nP(1 - P)] = 0.97979589711
- 4b
  - $L(\pi \mid n = 4, y = 3) = ln(\frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y}) = ln(\frac{4!}{3!(4-3)!}\pi^3(1-\pi)^{4-3})$

To find the maximum of this likelihood function we could derive and optimize but I will use R to make it simple.

```r
pii = seq(0,1, 0.01)
likelihood_function = (factorial(4)/factorial(3))*(pii^3)*(1-pii)
plot(pii, likelihood_function, type = "l")
```




```
#MLE
max(likelihood_function)
```

```
## [1] 0.421875
```

```
#parameter
pii[which.max((likelihood_function))]
```

```
## [1] 0.75
```

The MLE of the parameter pi when Y = 3 is 0.75, which means that our likelihood function's maximum is at pi = 0.75. I obtained this value by looking at the graph, and also by using a built in R function.

Question 5

The proportion estimate is 486/1374 = 0.35371179

Confidence interval = (0,0.3879369):

```
prop.test(486, 1374, conf.level = 0.99, alternative = "less")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  486 out of 1374, null probability 0.5
```

```
## X-squared = 117.03, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is less than 0.5
## 99 percent confidence interval:
##  0.0000000 0.3846085
## sample estimates:
##         p
## 0.3537118
```

The hypothesis test for Ho: p != 0.5 and Ha: p < 0.5 returned a chi square statistic of 117.03 on 1 degrees of freedom. The p-value was 2.2e^-16, which is extremely close to zero. Thus, it is highly likely that a minority of the population would say yes to accept cuts in their standard of living to protect the environment.

Question 6

```
binom.test(8, 10, p=0.5, alternative = "greater", conf.level = 0.99)
```

```
##
##  Exact binomial test
##
## data:  8 and 10
## number of successes = 8, number of trials = 10, p-value = 0.05469
## alternative hypothesis: true probability of success is greater than 0.5
## 99 percent confidence interval:
##  0.3882571 1.0000000
## sample estimates:
## probability of success
##                    0.8
```

The p value for this test was 0.05469. This is a very low p value, and depending on the alpha value used for the test, we may be able to reject the null hypothesis. However, given that this binomial test is being done in the context of a clinical trial, which often have high standards for data verification, I will use alpha = 0.05 and thus fail to reject the null hypothesis that the sample proportion is different from the expected proportion. This does not mean that the sample proportion is not significantly different, just that we are unable to conclude that using the data given. The 99% confidence interval is 0.39 to 1.0, and the sample proportion is within the confidence interval, further strengthening my argument that we are unable to conclude with certainty that the sample proportion is statistically different from the expected proportion.

Question 7

True:

7a:

Sensitivity = how often test generates a positive result for people with the condition = 0.75 Specificity = how often test generates a negative result for people that don't have the condition = 0.9

Part c:

```
cancer <- matrix(c(0.75,0.25,0.1,0.9), ncol = 2, byrow = TRUE)
colnames(cancer) <- c("Test Positive", "Test Negative")
rownames(cancer) <- c("Has Cancer", "No Cancer")
cancer
```

```
##            Test Positive Test Negative
## Has Cancer          0.75          0.25
## No Cancer           0.10          0.90
```

Part D:

P(positive test has cancer) = P(+|C)P(C) = (3/4)(0.04) = 0.03

P(Positive test but no cancer) = P(+|C^-)P(C^-) = (1/10)(1-0.04) = 0.096

P(test neg and has cancer) = P(negative|C) * P(C) = (1/4)(0.04) = 0.01

P(test negative and no cancer) = P(negative|C^-)P(C^-) = (9/10)(1-0.04) = 0.864

```
test_pos_has_cancer <- 0.03

test_pos_no_cancer <- 0.096


test_neg_has_cancer <- 0.01


test_neg_no_cancer <- 0.864

probability_table <- matrix(c(test_pos_has_cancer, test_pos_no_cancer, test_neg_has_cancer, test_pos_no
colnames(probability_table) <- c("Has Cancer", "No Cancer")
rownames(probability_table) <- c("Test Positive", "Test Negative")

probability_table
```

```
##               Has Cancer No Cancer
## Test Positive       0.03     0.096
## Test Negative       0.01     0.096
```

The probability of having cancer given a positive test is 0.238:

$P(C|+) = \frac{P(C \cap +)}{P(+)} = \frac{0.03}{0.126} = 0.238$

Question 8

Let X = true status (1=disease, 2 = no disease)

Let Y = diagnosis (1=positive, 2 = negative)

Let $\pi_1 = P(Y = 1|X = 1)$

Let $\pi_2 = P(Y = 1|X = 2)$

Let γ = P(someone has the disease)

Show:

$P(X = 1|Y = 1) = \pi_1\gamma/[\pi_1 + \pi_2(1 - \gamma)]$

Thus:

$\pi_1 = P(Y = 1|X = 1) = P(Y \cap X)/P(X)$

$P(Y = 1) = P(Y = 1|X = 1) + P(Y = 1|X = 2)$

Question 9

Difference in proportions:

```
diff_in_prop = 62.4/1000000 - 1.3/1000000
diff_in_prop
```

```
## [1] 6.11e-05
```

So the UK has 6.11e^-5 deaths from gun homicide than the USA.

Relative Risk:

```
rr = (62.4/1000000)/(1.3/1000000)
rr
```

```
## [1] 48
```

The relative risk ratio is 48. This means that dying by gun homicide is 48 times more likely in the US than the UK

Part b:

This depends on weather you want to say that there is a significant or insignificant difference between the UK/US homicide rates. The difference in proportions isn't useful when talking about small differences like this example, where the difference is almost negligigble. However, the relative risk demonstrates that there is a much higher chance of dying by gun homicide relatively in the US.

Question 10

The difference in proportions for smokers - non smokers was 0.0013 for lung cancer and 0.00256 for heart disease, meaning there is likely a stronger association with cigarette smoking and heart disease.

Odds ratios:

```
odds_lung = (0.00140/(1-0.00140))/(0.00010/(1-0.00010))

odds_lung
```

```
## [1] 14.01823
```

```
odds_heart = (0.00669/(1-0.00669))/(0.00413/(1-0.00413))

odds_heart
```

```
## [1] 1.624029
```

The odds of dying from lung cancer are nearly 14 times higher for smokers than non smokers, and the odds of dying from heart disease are nearly 1.62 times higher for smokers than non smokers. This means that using the odds ratio, lung cancer seems to have a higher association with smoking than the heart disease.

The difference in proportions metric only gives the difference of proportions, and does not actually quantify how many people will not die as a result of abstaining from smoking that have lung cancer or heart disease. Obviously there are other confounding variables. However this is also the case with the odds ratio. The difference in proportion is a better metric, because it focuses on comparing differences between groups, and thus provides a better estimator of the true number of deaths avoided by abstaining from smoking. Since heart disease has a higher difference (0.00256 vs 0.0013), it likely has a higher association with reduction in deaths that occur from abstaining from smoking.