Stat 477 - HW 3

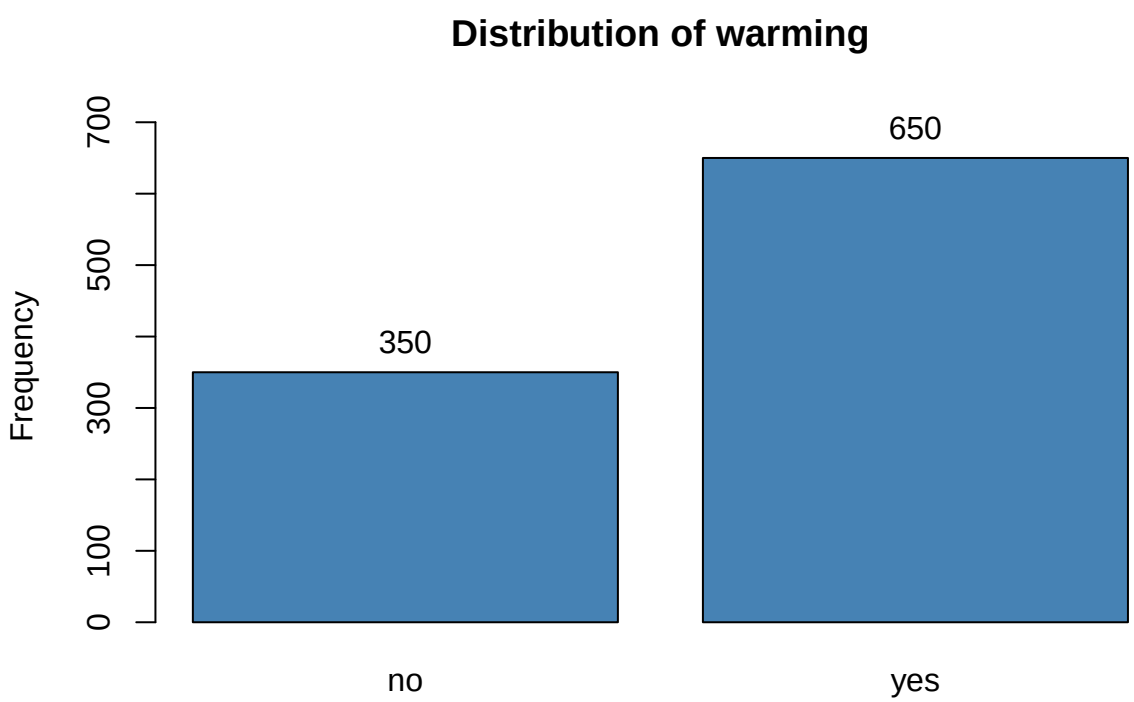
Jay Maxwell

2/8/2022

1. According to a Gallup poll, of 1,000 randomly selected adults aged 18 or older in the United States, 65% believe that global warming is more a result of human actions than natural causes.

   (a) Use R to obtain a summary table and bar graph of the sample data.

```
warming <- as.factor(c(rep("yes", 650), rep("no", 350)))
tab1(warming, graph = TRUE, cum.percent = FALSE, col = "steelblue")
```




```
## warming :
##         Frequency Percent
## no            350      35
## yes           650      65
##   Total      1000     100
```

(b) Describe the population proportion of interest p in words.

In this example, the population proportion (p) is the faction of the entire US population that believes global warming is caused more by people than natural causes.

(c) Use R to calculate a 95% confidence interval for the population proportion p using the normal approximation method.

```
prop.ci(650, 1000, type = "normal", 0.95)
```

```
## 0.6204377 0.6795623
```

(d) Give the interpretation of the 95% confidence interval you calculated in part (c) in the context of the problem.

We are 95% confident that the true population proportion of people in the USE who believe global warming is caused more by people than nature is between 62.04% to approximately 67.95%.

(e) Use R to calculate a 95% confidence interval for the population proportion p using the Wilson's score method. Compare the center and width of this interval to the one you calculated in part (c).

```
prop.ci(650, 1000, type = "score", 0.95)
```

```
## 0.6199147 0.6789373
```

```
(0.6795623 + 0.6204377)/2
```

```
## [1] 0.65
```

```
(0.6789373 + 0.6199147)/2
```

```
## [1] 0.649426
```

The results of the score method have a slightly smaller range than the results of the normal approxmation method. I would call the centers for each calculation approximately equal at .65.

(f) Gallup is planning to conduct another poll on global warming. They would like to have a 95% confidence interval with a margin of error of no more than 2.5%. What sample size do they need to obtain this margin of error? Make sure to specify how you are calculating the sample size

```
cat("Sample size needed:", nprop.ci(0.65, 0.025, 0.95))
```

```
## Sample size needed: 1399
```

To calculate the sample size, we solve for $n \geq (\frac{z_{1-\frac{\alpha}{2}}}{M})^2 \hat{p} * (1 - \hat{p})$**. Because this is a follow up survey, the reasearches might use their prior findings to estimate** $\hat{p} = .65$**.**

2. Unlike confidence intervals for other parameters... What do you notice about the coverage rates of the two methods? How do these results depend on the values of n and p?

	Normal			Wilson's		
p	n=25	n=250	n=1000	n=25	n=250	n=1000
0.5	95.651	95.145	94.656	95.766	95.008	94.729
0.75	89.468	93.999	94.66	93.901	94.365	94.758
0.9	91.851	93.104	95.17	96.33	95.54	94.893

When we compare the coverages calcuated above it appears that for most of our simulations the Wilson's score method results in coverage closer to the 95% goal more sore than the Normal approximation method. I have a hard time determining a pattern of relation between n and p for both methods. There is a lot of variation in coverage (betwen 89% and 95%) for the normal method for all the combinations of n and p. While the Wilson's score method seems to stay consistently close to our ideal coverage of 95% no matter what the combinations of n and p are changed to.

```
# coverage.ci(25,.5,'normal',0.95) coverage.ci(250,.5,'normal',0.95)
# coverage.ci(1000,.5,'normal',0.95) coverage.ci(25,.5,'score',0.95)
# coverage.ci(250,.5,'score',0.95) coverage.ci(1000,.5,'score',0.95)
# coverage.ci(25,.75,'normal',0.95) coverage.ci(250,.75,'normal',0.95)
# coverage.ci(1000,.75,'normal',0.95) coverage.ci(25,.75,'score',0.95)
# coverage.ci(250,.75,'score',0.95) coverage.ci(1000,.75,'score',0.95)
# coverage.ci(25,.9,'normal',0.95) coverage.ci(250,.9,'normal',0.95)
# coverage.ci(1000,.9,'normal',0.95) coverage.ci(25,.9,'score',0.95)
# coverage.ci(250,.9,'score',0.95) coverage.ci(1000,.9,'score',0.95)
```

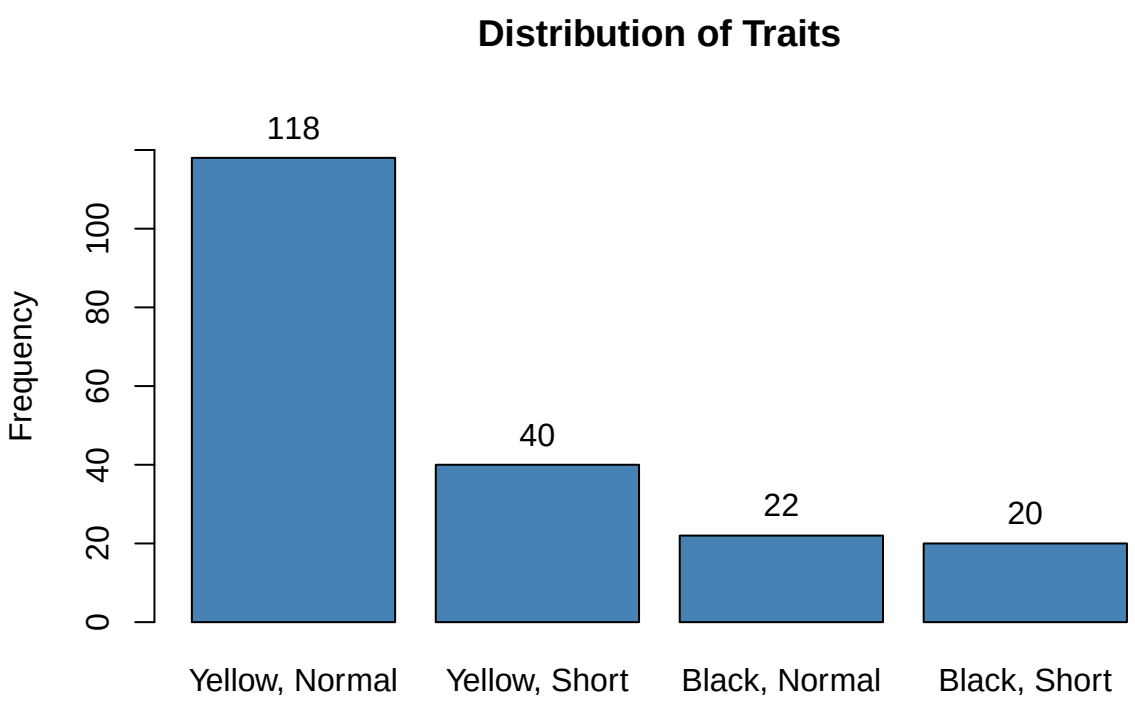
3. Offspring of certain fruit flies may have either yellow or black bodies and either normal or short wings. Genetic theory predicts that these traits will appear with the following probabilities: Yellow, Normal Yellow, Short Black, Normal Black, Short 9/16 3/16 3/16 1/16 A researcher examines 200 flies and identifies the traits for each fly. These data can be found in the file flies.csv.

Yellow, Normal	Yellow, Short	Black, Normal	Black, Short
9/16	3/16	3/16	1/16

(a) Use R to give the summary table and a bar graph of the sample data.

```
flies <- read_csv("flies.csv", col_names = TRUE, show_col_types = FALSE)
flies$Traits <- factor(flies$Traits, levels = c("Yellow, Normal", "Yellow, Short",
    "Black, Normal", "Black, Short"))


tab1(flies$Traits, cum.percent = FALSE, sort.group = "decreasing", main = "Distribution of Traits",
    col = "steelblue")
```




```
## flies$Traits :
##                Frequency Percent
## Yellow, Normal       118      59
## Yellow, Short         40      20
## Black, Normal         22      11
## Black, Short          20      10
##   Total              200     100
```

(b) Give the null and alternative hypotheses for the goodness of fit test for the correctness of the genetic theory.

$H_0 : p_{yn} = p_{ys} = p_{bn} = p_{bs}$
H_1 : At least one of the probabilities in the null hyptothesis is different

(c) Calculate the expected number of fruit flies from the "Yellow, Normal" category under the assumption the genetic theory is true. Only calculate this expected value here, and show your work. You may use R to verify your answer.

$E(Y_{Yellow,Normal}) = n * p_{Yellow,Normal} = 200 * \frac{9}{16} = 112.5$

```
200 * 9/16
```

```
## [1] 112.5
```

(d) Calculate the contribution of the "Yellow, Normal" category to the test statistic X2 . Only calculate this contribution here, and show your work. You may use R to verify your answer.

$X^2 = \sum_{j=1}^{J} \frac{(Y_j - E(Y-j))^2}{E(Y_j)}$

Thus the portion of the sum contributed by "Yellow, Normal" would be:

$\frac{(118-112.5)^2}{112.5} = 0.2689$

```
(((118 - 112.5)^2)/112.5) %>%
    round(4)
```

```
## [1] 0.2689
```

(e) Use R to find the test statistic and p-value for this hypothesis test.

```
yn <- 9/16
ys <- 3/16
bn <- 3/16
bs <- 1/16
modelp <- c(yn, ys, bn, bs)
flies.summary <- plyr::count(flies, var = "Traits")
flies.goodtest <- chisq.test(flies.summary[2], p = modelp)
cat("Test statistic", flies.goodtest$statistic %>%
    round(4))
```

```
## Test statistic 11.3422
```

```
cat("p-value", flies.goodtest$p.value %>%
    round(4))
```

```
## p-value 0.01
```

(f) Write a conclusion about the correctness of the genetic theory.

Based on the sampled data, there is strong evidence to suggest that at least one of our types is inconsistent with the probabilities determined by the genetic theory.