Module 06 Homework 01

HDS

2024-08-30

DSE5001 Module 06, Homework 01

Material from Diez et al

checked 01/03/2025

first several problems using confidence interval methods

Problem 1

> 5.13 Website registration. A website is trying to increase registration for first-time visitors, exposing 1% of these visitors to a new site design. Of 752 randomly sampled visitors over a month who saw the new design, 64 registered.

(a) Check any conditions required for constructing a confidence interval.

The two conditions that must be satisfied are Independence and the Success-Failure condition. We know this is independent because the users were randomly sampled. Now, let's compute the Success-Failure condition to verify that the sample is sufficiently large (np and n(1-p) are both greater than 10).

```
n=752
p=64/752
c1=n*p
c2=n*(1-p)
c1
```

```
## [1] 64
```

```
c2
```

```
## [1] 688
```

Both conditions are met.

(b) Compute the standard error.

SE=(p(1-p)/n)^0.5

```
SE=((p*(1-p))/n)^0.5
SE
```

```
## [1] 0.01017554
```

The standard error in p is ~0.0102

(c) Construct and interpret a 90% confidence interval for the fraction of first-time visitors of the site who would register under the new design (assuming stable behaviors by new visitors over time).

use the mean, std error and qnorm to find the interval boundaries, assuming two tails

Source for z critical values in r: https://www.geeksforgeeks.org/how-to-find-z-critical-values-in-r/# (https://www.geeksforgeeks.org/how-to-find-z-critical-values-in-r/#)

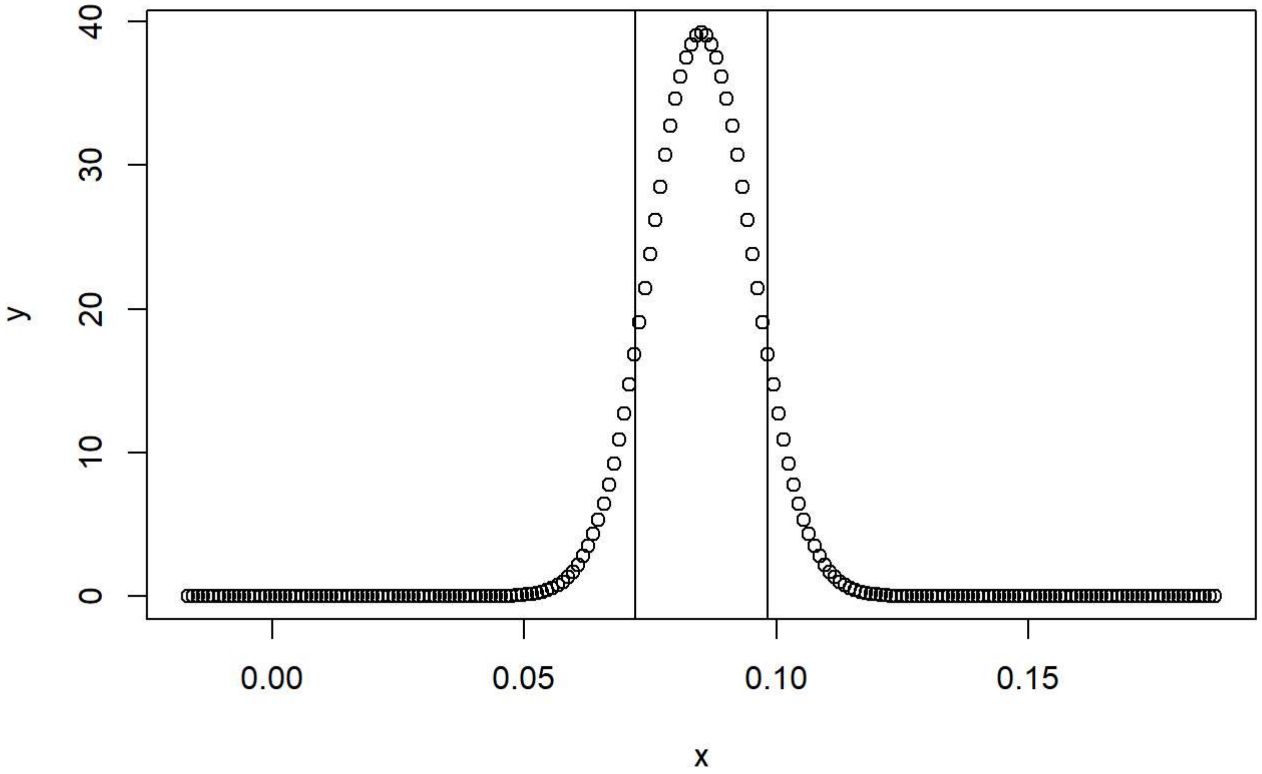
```
#First, plot the normal distribution with the mean (which is equal to the proportion) and standard error from this distribution. This will help me understand the shape of the distribution.
pad = 10
x=seq(p-(SE*pad), p+(SE*pad), SE/pad)
y=dnorm(x, p, SE)

#Get the z critical values using the qnorm function
z_star_lower=qnorm(0.1)
z_star_higher=qnorm(0.9)

#Find the values of the upper and lower bounds
LCL=p+(z_star_lower*SE)
UCL=p+(z_star_higher*SE)

LCL
```

```
## [1] 0.0720659
```

```
UCL
```

```
## [1] 0.09814686
```

```
#Plot the distribution with the bounds drawn
plot(x, y)
abline(v=LCL)
abline(v=UCL)
```



Problem 2

> 5.22 Getting enough sleep. 400 students were randomly sampled from a large university, and 289 said they did not get enough sleep.

(a) Conduct a hypothesis test to check whether this represents a statistically significant difference from 50%, and use a significance level of 0.01.

From the Statistical Thinking textbook with a null hypothesis difference of 0.5:

$$Z = \frac{\hat{P} - 0.5}{\sqrt{0.5(1-0.5)/n}}$$

```
#Define the proportion of interest for the null hypothesis
p_hat=289/400
p_hat
```

```
## [1] 0.7225
```

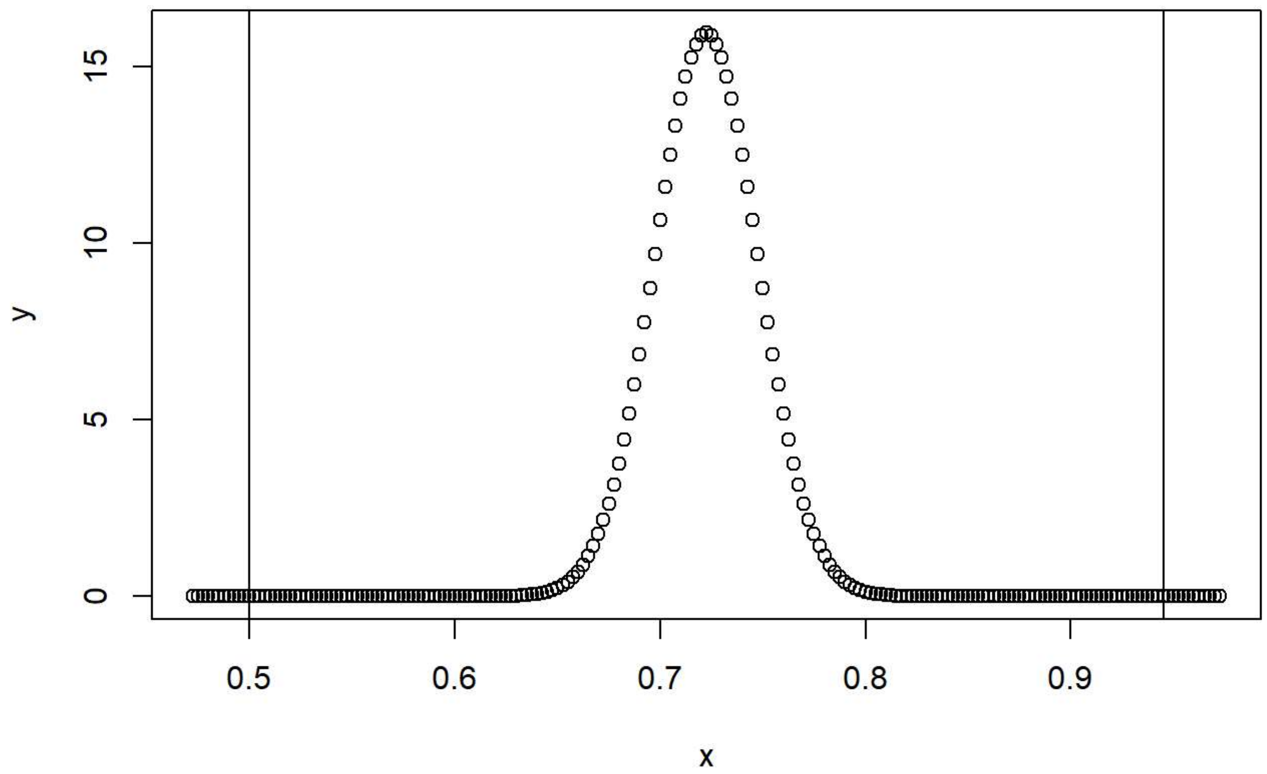
```
#Define the null hypothesis
p_H_0=0.5

#Define the sample size
n=400

#Find the standard error based on the null hypothesis.
SE=((p_H_0*(1-p_H_0))/n)^0.5

#Find Z
Z=(p_hat-p_H_0)/sqrt((p_H_0*(1-p_H_0))/n)

#Plot all of this so it makes a bit more sense.
pad = 10
x=seq(p_hat-(SE*pad), p_hat+(SE*pad), SE/pad)
y=dnorm(x, p_hat, SE)

#Find the values of the upper and lower bounds
LCL=p_hat+(-Z*SE)
UCL=p_hat+(Z*SE)

#Print the LCL and UCL
LCL
```

```
## [1] 0.5
```

```
UCL
```

```
## [1] 0.945
```

```
#Plot the distribution with the bounds drawn
plot(x, y)
abline(v=LCL)
abline(v=UCL)
```



Based on this plot, it appears that the null hypothesis p-values are toward the end of the distribution. Let’s see what they actually are. SOURCE: https://www.geeksforgeeks.org/how-to-find-p-value-from-test-statistic/ (https://www.geeksforgeeks.org/how-to-find-p-value-from-test-statistic/)

```
p_value <- 2 * (1 - pnorm(Z))
p_value
```

```
## [1] 0
```

A p-value of 0 is lower than the significance, therefore we reject the null hypothesis, suggesting that the proportion of students not getting enough sleep is different than 50%. Based on the p_hat value, it appears to be about 72.25%.

I also found this in the Statistical Thinking textbook a bit further down, which does the same thing as the long hand calculation, above.

```
prop.test(289, 400)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  289 out of 400, null probability 0.5
## X-squared = 78.323, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.6753634 0.7652837
## sample estimates:
##      p
## 0.7225
```

Problem 3 - discussion only, no code needed

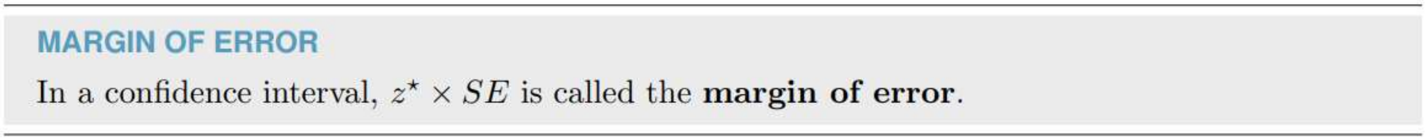
> 5.26 Which is higher? In each part below, there is a value of interest and two scenarios (I and II). For each part, report if the value of interest is larger under scenario I, scenario II, or whether the value is equal under the scenarios.

(a) The standard error of p when (I) n = 125 or (II) n = 500.

The standard error is greater in scenario I because n is significantly smaller, so the value of interest will represent a larger proportion of the total population.

(b) The margin of error of a confidence interval when the confidence level is (I) 90% or (II) 80%.

The margin of error is greater in I because it scales with the critical value, which, in turn, scales with the confidence interval. Margin of error can be calculated with the equation ME = z_star x SE



MARGIN OF ERROR

In a confidence interval, $z^\star \times SE$ is called the **margin of error**.

Margin of Error

(c) The p-value for a Z-statistic of 2.5 calculated based on a (I) sample with n = 500 or based on a (II) sample

with n = 1000.

This will be equal because the Z-statistic is normalized to the size of the distribution, therefore the p-value is unaffected by sample size.

(d) The probability of making a Type 2 Error when the alternative hypothesis is true and the significance

level is (I) 0.05 or (II) 0.10.

The probability will be lower in case (II) because the criteria for rejecting the null hypothesis is greater, reducing the likelihood that the null hypothesis is not rejected when H_A is true.

Problem 4

> 5.32 Nearsighted. It is believed that nearsightedness affects about 8% of all children. In a random sample of 194 children, 21 are nearsighted.

(a) Conduct a hypothesis test for the following question: do these data provide evidence that the 8% value is inaccurate?

```
prop.test(21, 194)
```

```
## 
##  1-sample proportions test with continuity correction
## 
## data:  21 out of 194, null probability 0.5
## X-squared = 117.53, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.06981131 0.16277150
## sample estimates:
##         p 
## 0.1082474
```

Yes, based on the proportion test, above, the null hypothesis is rejected. The true proportion is likely nearer to 10.8%.

Problem 5

ANOVA

Load the built-in InsectSprays data set

```
data("InsectSprays")
```

It has the insect counts seen under different types of sprays

1.) Load the data, how many different sprays are present?

```
head(InsectSprays)
```

```
##   count spray
## 1    10     A
## 2     7     A
## 3    20     A
## 4    14     A
## 5    14     A
## 6    12     A
```

```
length(unique(InsectSprays$spray))
```

```
## [1] 6
```

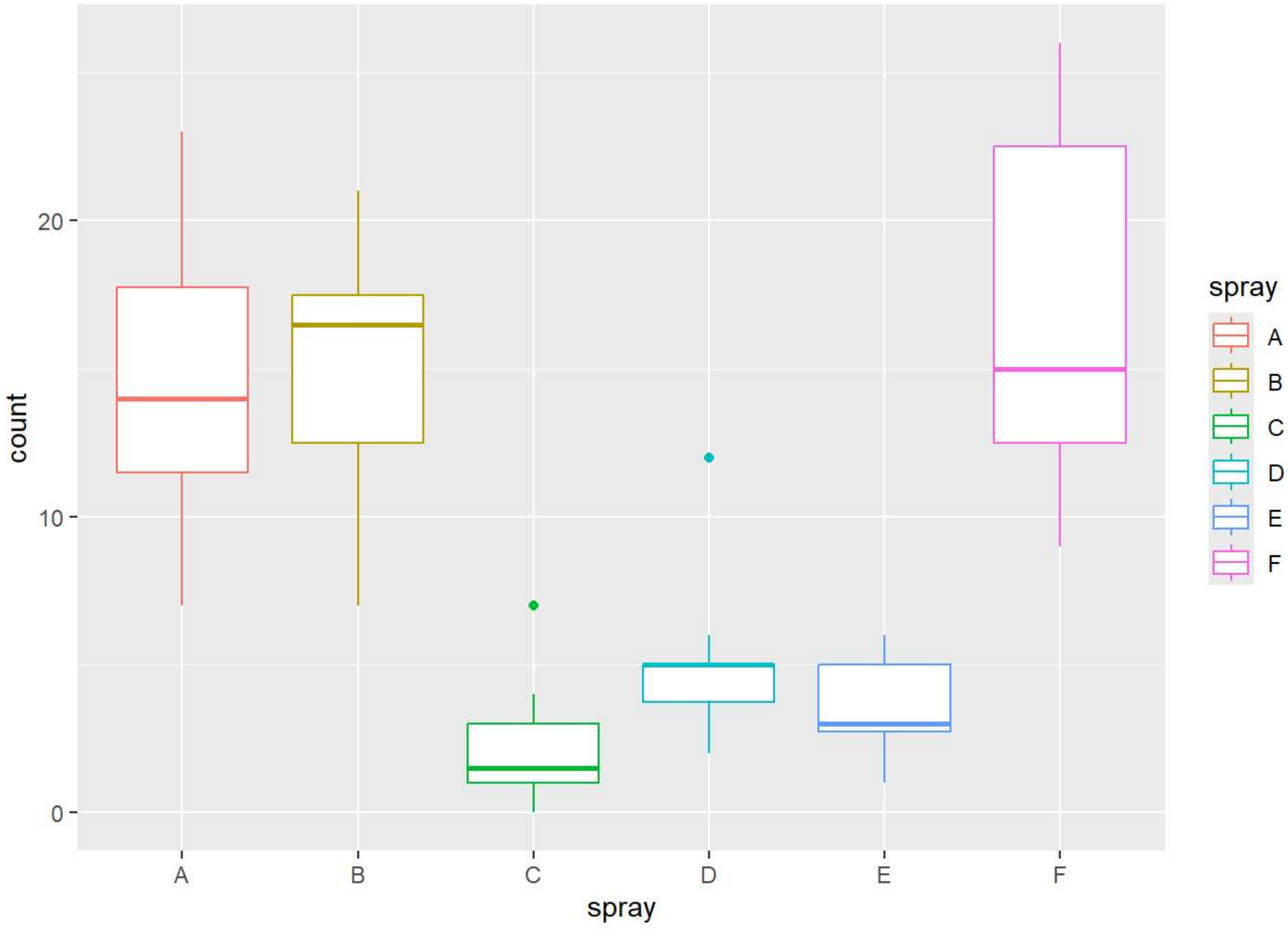
2.) Set spray to be a factor,

sprays is already a factor

3.) Create a boxplot of count vs spray, do the sprays look like they are equally effective or not?

```
library(ggplot2)
```

```
ggplot(
  InsectSprays,
  aes(x=spray, y=count, color = spray)
) + geom_boxplot()
```



4.) Run an ANOVA to determine if the sprays differ in insect counts or not

```
aov_result<-aov(count~spray, data=InsectSprays)
summary(aov_result)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)    
## spray        5   2669   533.8    34.7 <2e-16 ***
## Residuals   66   1015    15.4                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is essentially 0, indicating that the spray is directly related to the variance in insect count.

Problem 6

Contingency Table

Load the HairEyeColor set

Here it is split into male and female

```
data("HairEyeColor")

male_HEC=HairEyeColor[,,1]
female_HEC=HairEyeColor[,,1]

summary(male_HEC)
```

```
## Number of cases in table: 279
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 41.28, df = 9, p-value = 4.447e-06
##  Chi-squared approximation may be incorrect
```

```
summary(female_HEC)
```

```
## Number of cases in table: 279
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 41.28, df = 9, p-value = 4.447e-06
##  Chi-squared approximation may be incorrect
```

```
print(male_HEC)
```

```
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    32   11    10     3
##   Brown    53   50    25    15
##   Red      10   10     7     7
##   Blond     3   30     5     8
```

```
print(female_HEC)
```

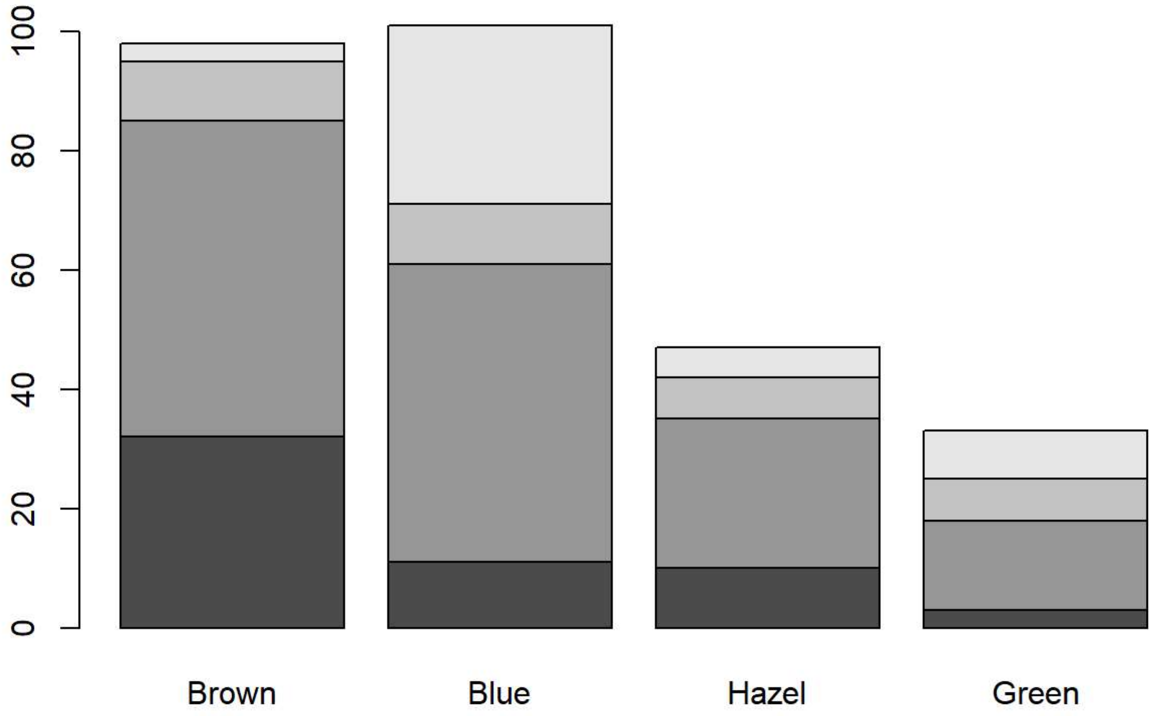
```
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    32   11    10     3
##   Brown    53   50    25    15
##   Red      10   10     7     7
##   Blond     3   30     5     8
```

For both males and females, test the Hypothesis that the counts are randomly distributed, with no interaction of hair or eye color.

Show bar plots of both tables

Use chi-square and fisher’s exact tests

```
library("MASS")

barplot(male_HEC)
```



```
barplot(female_HEC)

male_HEC_summed<-addmargins(male_HEC)
female_HEC_summed<-addmargins(female_HEC)

chisq.test(male_HEC_summed)
```

```
## Warning in chisq.test(male_HEC_summed): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  male_HEC_summed
## X-squared = 41.28, df = 16, p-value = 0.0005048
```

```
chisq.test(female_HEC_summed)
```

```
## Warning in chisq.test(female_HEC_summed): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  female_HEC_summed
## X-squared = 41.28, df = 16, p-value = 0.0005048
```

```
fisher.test(male_HEC_summed, simulate.p.value = TRUE)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  male_HEC_summed
## p-value = 0.0009995
## alternative hypothesis: two.sided
```

```
fisher.test(female_HEC_summed,  simulate.p.value = TRUE)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  female_HEC_summed
## p-value = 0.0009995
## alternative hypothesis: two.sided
```

Based on the large X-squared values and small p-values, there is strong evidence against the null hypothesis, indicating that there is a relationship between hair and eye color, regardless of sex.