Lab Assignment 2

Paul J Anderson

2025-04-13

```
library(vcdExtra)
```

```
## Loading required package: vcd
```

```
## Loading required package: grid
```

```
## Loading required package: gnm
```

```
library(Sleuth3)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x dplyr::summarise() masks vcdExtra::summarise()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(gnm)
```

What to submit (gradescope): 1. .Rmd 2. pdf Note: the document mentions the .Rmd must generate a reasonable HTML output, but we are checking and submitting a pdf. Is this a mistake?

Questions:

NOTE: The first two questions are for credit. The next two are optional/for fun.

Q1.

The following data come from one of Gregor Mendel's famous experiments with pea plants. In this particular experiment, Mendel examined two categorical traits in pea seeds, each with two possible values: seed color – yellow or green – and seed shape – round or angular. According to Mendel's hypothesis, the inheritance pattern of these traits was characterized by "independent assortment" – that is, there was no "genetic linkage" between these traits.

```
mendel <- as.table(
  matrix(c(315, 108, 101, 32),
         nrow = 2,
         dimnames =
           list(Color= c("Yellow", "Green"),
```

```
                Shape = c("Round", "Angular"))))
mendel
```

```
##         Shape
## Color    Round Angular
##   Yellow   315     101
##   Green    108      32
```

How consistent are these data with Mendel's hypothesis of independence? Notice that in this particular case, we're not interested in a *departure* from independence, but rather a *confirmation* of independence. Is a Chi-squared test appropriate here? Why or why not?

ANSWER: Highly consistent. Mendel hypothesized that color and shape were genetically independent. There are two tests we learned in this lab to test independence of categorical data. Chi-squared test of independence and Fisher's Exact Test. Here is a breakdown of the assumptions: Test Use Case When to Use Chi-squared test of independence Tests whether two categorical variables are independent (e.g. gender vs. outcome) Large sample sizes (expected counts 5) Fisher's Exact Test Tests independence in a 2×2 table without relying on large-sample approximations Small sample sizes (especially when expected cell counts < 5)

In this case, total cell counts are >=5, T/F we can use Chi-squared test of independence to "confirm" independence.

Let's run the test in R, since this is a 2x2 and we're using the default of Yates' per instruction, we'll use Pearson's Chi-squared test with Yates' continuity correction. The ouput of this test (p-value = 0.8208) indicates there is insufficient evidence to reject the null hypothesis of independence of distributions. These findings are highly consistent with Mendel's hypothesis of independence.

```
chisq.test(mendel)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  mendel
## X-squared = 0.051332, df = 1, p-value = 0.8208
```

Q2.

We mentioned previously, and it's mentioned in the narrated lecture materials that there's an equivalence between the Chi-squared test and the difference in proportions test in the case of 2x2 tables. Take a look for the Mendel data:

```
chisq.test(mendel, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  mendel
## X-squared = 0.11634, df = 1, p-value = 0.733
```

```
prop.test(mendel, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  mendel
## X-squared = 0.11634, df = 1, p-value = 0.733
## alternative hypothesis: two.sided
```

```
## 95 percent confidence interval:
##  -0.09506177  0.06662771
## sample estimates:
##    prop 1    prop 2
## 0.7572115 0.7714286
```

The output from the second chunk refers to a `two.sided` test, with an interpretation in terms of a signed difference between two proportions, whereas the `chisq.test()` is a one-sided test, with an interpretation involving deviation from the hypothesis of independent rows and columns. It's not *entirely* obvious why these results should be the same. The following exercise can help you see what's going on.

Use `rnorm()` and `rchisq()` to produce and store the following vectors:

```
set.seed(201)
v1 <- rnorm(1000)
v2 <- rchisq(1000, df = 1)
```

1000 draws from the standard normal distribution, 1000 draws from the Chi-squared distribution with 1 degree of freedom.

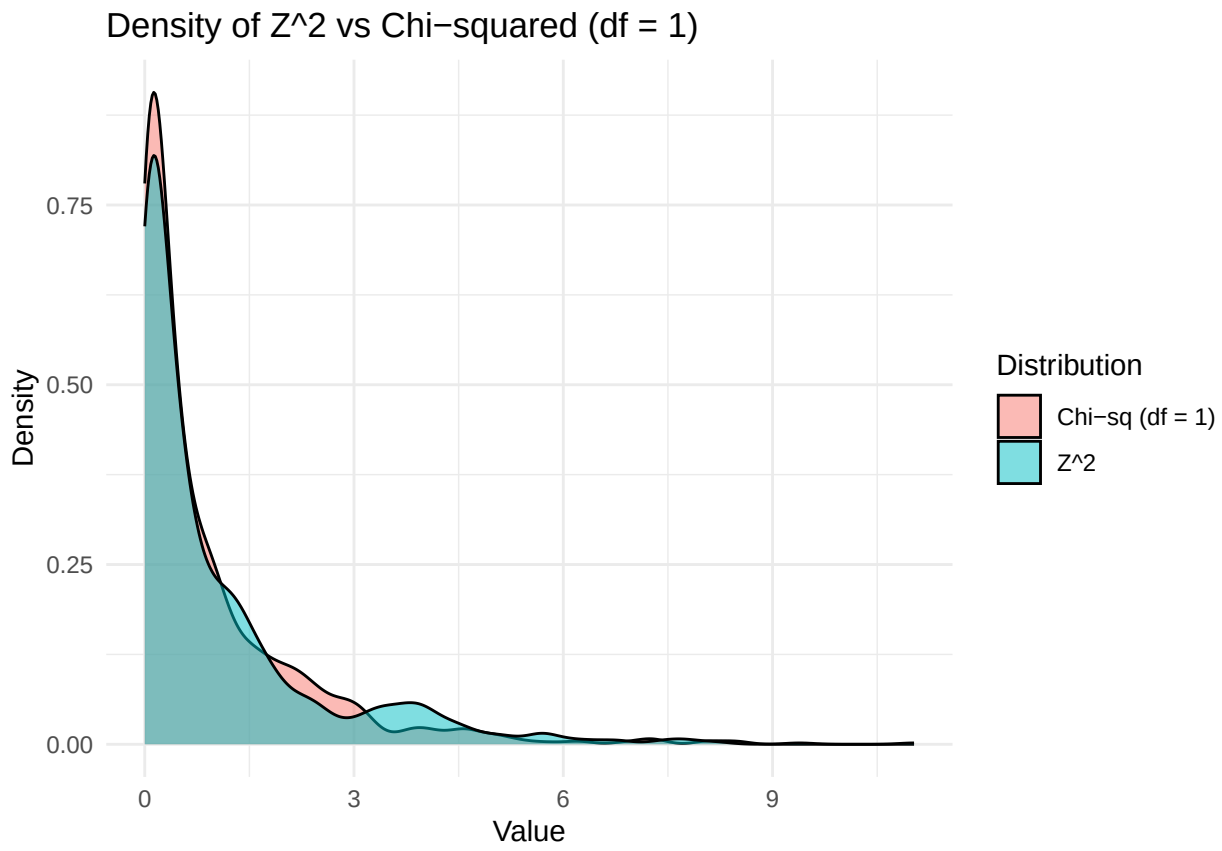
(a) Construct separate histograms/density plots of the raw chi-squared values and the *squares* of the s

ANSWER: While there are some slight variation in data due to random sampling, the squared standard normal distribution (Z^2) and the chi-shared distribution with 1 degree of freedom are closely matched. We expect this, since the square of the standard normal variable follows a chi-squared distribution with 1 DF.

```
# square the standard normal values
v1_squared <- v1^2

# creating the df
df <- data.frame(
  value = c(v1_squared, v2),
  type = rep(c("Z^2", "Chi-sq (df = 1)"), each = 1000)
)

# plot
ggplot (df, aes(x = value, fill = type)) +
  geom_density(alpha = 0.5) +
  labs(title = "Density of Z^2 vs Chi-squared (df = 1)",
       x = "Value", y = "Density", fill = "Distribution") +
  theme_minimal()
```



(b) Recall (or obtain from `qnorm(0.975)` that the *2-sided 95%* critical values for the standard normal

ANSWER: We see that the critical values from a 2-sided 95% test are +/- 1.96. We square this value and get 3.8416, our Z^2. We run a one-sided 95% critical value chi^2 test with 1 DF and get 3.8416, the same value. This shows that the square of the standard normal variable (Z^2) follows a chi-squared distribution with 1 degree of freedom.

```
# critical values for 2-sided 95% test
(crit_95 <- qnorm(0.975))
```

```
## [1] 1.959964
```

```
# critical values from the standard normal or Z^2 distribution
(sq_crit_95 <- crit_95^2)
```

```
## [1] 3.841459
```

```
# one-sided 95% crit value for chi-squared with 1 DF
(qchisq(0.95, df = 1))
```

```
## [1] 3.841459
```