lab3-Chen

Weixuan Chen

2/15/2023

```
library(ggplot2)
```

Task1

1

We can have a sample size of 100 seeds for both GMO and wild type, and plant them in the same field. Here, we control the environment they will grow and the only difference is the type of themselves. And we can count the number of germinated seeds in each group. And, for the purpose of repliaction, we can repeat the experiment for 100 times, or we have 100 trials going on simutaneously.

2

```
# Download the data
d1 <- read.csv("http://faraway.neu.edu/biostats/lab3_dataset1.csv")
head(d1)
```

```
##   trial gmo.germinated wild.germinated gmo.notgerminated wild.notgerminated
## 1     1             74              21                26                 79
## 2     2             91              23                 9                 77
## 3     3             91              22                 9                 78
## 4     4             83              24                17                 76
## 5     5             66              29                34                 71
## 6     6             84              29                16                 71
```

3

Plot for gmo.germinated

```
mean_gmo_germ = mean(d1$gmo.germinated)
upper_gmo_germ = mean_gmo_germ + 1.96*sd(d1$gmo.germinated)/sqrt(length(d1$gmo.germinated))
lower_gmo_germ = mean_gmo_germ - 1.96*sd(d1$gmo.germinated)/sqrt(length(d1$gmo.germinated))

mean_gmo_notgerm = mean(d1$gmo.notgerminated)
upper_gmo_notgerm = mean_gmo_notgerm + 1.96*sd(d1$gmo.notgerminated)/sqrt(length(d1$gmo.notgerminated))
lower_gmo_notgerm = mean_gmo_notgerm - 1.96*sd(d1$gmo.notgerminated)/sqrt(length(d1$gmo.notgerminated))
```

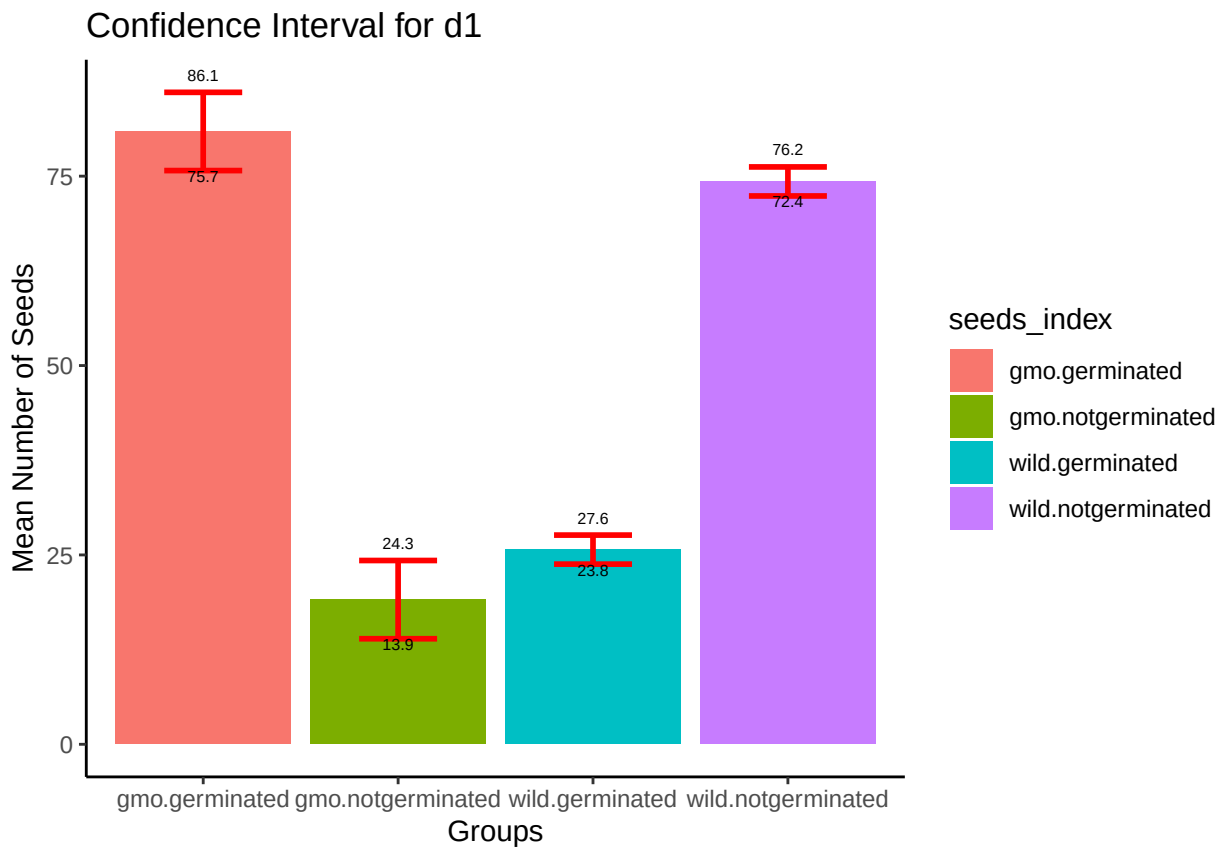
```r
mean_wild_germ = mean(d1$wild.germinated)
upper_wild_germ = mean_wild_germ + 1.96*sd(d1$wild.germinated)/sqrt(length(d1$wild.germinated))
lower_wild_germ = mean_wild_germ - 1.96*sd(d1$wild.germinated)/sqrt(length(d1$wild.germinated))

mean_wild_notgerm = mean(d1$wild.notgerminated)
upper_wild_notgerm = mean_wild_notgerm + 1.96*sd(d1$wild.notgerminated)/sqrt(length(d1$wild.notgerminat
lower_wild_notgerm = mean_wild_notgerm - 1.96*sd(d1$wild.notgerminated)/sqrt(length(d1$wild.notgerminat


seeds_mean <- c(mean_gmo_germ, mean_wild_germ, mean_gmo_notgerm, mean_wild_notgerm)
seeds_lower_ci <- c(lower_gmo_germ, lower_wild_germ, lower_gmo_notgerm, lower_wild_notgerm)
seeds_upper_ci <- c(upper_gmo_germ, upper_wild_germ, upper_gmo_notgerm, upper_wild_notgerm)
seeds_index <- c('gmo.germinated', 'wild.germinated', 'gmo.notgerminated', 'wild.notgerminated')

df <- data.frame(seeds_index, seeds_mean, seeds_lower_ci, seeds_upper_ci)

df
```

```
##          seeds_index seeds_mean seeds_lower_ci seeds_upper_ci
## 1     gmo.germinated       80.9       75.73536       86.06464
## 2    wild.germinated       25.7       23.78294       27.61706
## 3  gmo.notgerminated       19.1       13.93536       24.26464
## 4 wild.notgerminated       74.3       72.38294       76.21706
```

4

```r
ggplot(data = df) +
  geom_bar(aes(x=seeds_index, y=seeds_mean, fill = seeds_index), stat = "identity") +
  geom_errorbar(aes(x=seeds_index, ymin = seeds_lower_ci, ymax = seeds_upper_ci), width = 0.4, color =
  geom_text(aes(x=seeds_index, y = seeds_lower_ci, label = round(seeds_lower_ci, 1)), size = 2, vjust =
  geom_text(aes(x=seeds_index, y = seeds_upper_ci, label = round(seeds_upper_ci, 1)), size = 2, vjust =
  theme_classic() +
  labs(title = "Confidence Interval for d1") +
  labs(x="Groups", y = "Mean Number of Seeds")
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
```



5

By the plot, we can see the mean of germinated GMO outperformed the wild type. It is definitely worth toiling in the lab.

6

H0: GMO seeds and wild seeds are equally likely to germinated H1: Not H0

7

```
germinated <- c(sum(d1$gmo.germinated), sum(d1$wild.germinated))

notgerminated <- c(sum(d1$gmo.notgerminated), sum(d1$wild.notgerminated))

fisher_index <- c('GMO', 'Wild')

fisher_df <- data.frame(germinated, notgerminated)
rownames(fisher_df) <- fisher_index

fisher_df
```

```
##      germinated notgerminated
## GMO         809           191
## Wild        257           743
```

```
fisher_tst <- fisher.test(fisher_df)
fisher_tst
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  fisher_df
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   9.850881 15.226085
## sample estimates:
## odds ratio
##   12.22468
```

8

We can see the p-value is 2.2e-16, which is so small that we should reject the null hypothesis whichever the significance level we choose. So there is evidence that seed germination is different in GMO and Wild groups.

Task 2

```
 # Download the data
d2 <- read.csv("http://faraway.neu.edu/biostats/lab3_dataset2.csv")
head(d2)
```

```
##      countries gmo.disease gmo.nodisease nogmo.disease nogmo.nodisease
## 1        India          45            40            15              31
## 2      Vietnam          59            42            27              23
## 3       Brazil          58            44            30              31
## 4 South Africa          52            44            21              29
## 5     Cambodia          39            51            22              25
## 6  Ivory Coast          53            50            23              24
```

1

H0: two variables are independent (GMO and disease have no association) H1: Not H0

2

```
have_disease <- c(sum(d2$gmo.disease), sum(d2$nogmo.disease))

have_no_disease <- c(sum(d2$gmo.nodisease), sum(d2$nogmo.nodisease))
```

```r
disease_index <- c('GMO', 'Not GMO')

disease_df <- data.frame(have_disease, have_no_disease)
rownames(disease_df) <- disease_index

disease_df
```

```
##         have_disease have_no_disease
## GMO              506             465
## Not GMO          228             260
```

```r
disease_tst <- fisher.test(disease_df)
disease_tst
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  disease_df
## p-value = 0.05242
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.9920017 1.5524887
## sample estimates:
## odds ratio
##   1.240706
```

3

We can see the p-value is 0.05242, which is greater than 0.05 level. So we say that we don't have enough evidence to show that these two groups have no association at 0.05 level.

4

```r
pvals <- numeric(NROW(d2))
for (i in 1:NROW(d2)) {
  have_disease <- c(d2$gmo.disease[i], d2$nogmo.disease[i])
  have_no_disease <- c(d2$gmo.nodisease[i], d2$nogmo.nodisease[i])
  disease_index <- c('GMO', 'Not GMO')
  disease_df <- data.frame(have_disease, have_no_disease)
  rownames(disease_df) <- disease_index
  disease_tst <- fisher.test(disease_df)
  pvals[i] = disease_tst$p.value
}

df_pvals_countries <- data.frame(pvals)
rownames(df_pvals_countries) <- d2$countries
df_pvals_countries
```

```
##                  pvals
## India        0.0288584
## Vietnam      0.7271239
## Brazil       0.4170731
## South Africa 0.2219837
## Cambodia     0.7204465
## Ivory Coast  0.8606783
## Pakistan     0.3042649
## Laos         0.7393854
## Peru         0.7224751
## Venezuela    1.0000000
```

5

Except for India, there is evidence of an association between GMO and disease at 0.05 significance level.