HW1

Solutions

Problem 1: Adapted from Sleuth3 1.17

Seven students volunteered for a comparison of study guides for an advanced course in mathematics. They were randomly assigned to use one of two study guides, four to study guide A and three to study guide B. All were instructed to study independently. Following a two-day study period, all students were given an examination about the material covered by the guides, with the following results:

Study Guide A scores: 68 77, 82, 85

Study Guide B scores: 53, 64, 71

(a) What is the difference between sample averages for the two groups?

The mean for group A is 78

The mean for group B is 62.67

The difference in group means is 15.33.

(b) State relevant null and alternative hypotheses to compare the performance of the students assigned to each group.

You might use μ_1 and μ_2 or δ as your parameters. Whichever parameter(s) you choose to use in your statement of the hypotheses, please write a brief sentence defining what the parameter(s) are in the context of this example.

Option 1: in terms of μ_1 and μ_2.

Define μ_1 to be the mean score among people who might be assigned to group A in the population of people similar to those enrolled in this study, and μ_2 to be the mean score among people who might be assigned to group B in the population of people similar to those enrolled in this study.

Our hypotheses are:

$H_0 : \mu_1 = \mu_2$. The mean scores are the same for both groups.

$H_A : \mu_1 \neq \mu_2$. The mean scores are different for group A and group B.

Option 2: in terms of δ.

Define δ to be the difference in mean scores between people assigned to group A and group B in the population of people similar to those enrolled in this study; alternatively/equivalently, define δ to be the average difference in scores for an individual between the settings where they are assigned to group A and to group B, in the population of people similar to those enrolled in this study.

$H_0 : \delta = 0$. The average difference in scores when an individual is assigned to group A and when they are assigned to group B is 0.

$H_A : \delta \neq 0$. The average difference in scores when an individual is assigned to group A and when they are assigned to group B is different from 0.

(c) Perform a randomization test of the hypotheses you stated in part (b). There are 35 possible ways that these students could have been randomized to two groups, listed below. For each such randomization, the difference between sample averages for the two groups is shown. Use this information to calculate the two-sided p value for the test.

```
Group A Scores: 68, 77, 82, 85       Group A mean: 78
Group B Scores: 53, 64, 71           Group B mean: 62.667
Group A Mean - Group B Mean: 15.333
Group A Scores: 68, 77, 82, 53       Group A mean: 70
Group B Scores: 85, 64, 71           Group B mean: 73.333
Group A Mean - Group B Mean: -3.333
Group A Scores: 68, 77, 82, 64       Group A mean: 72.75
Group B Scores: 85, 53, 71           Group B mean: 69.667
Group A Mean - Group B Mean: 3.083
Group A Scores: 68, 77, 82, 71       Group A mean: 74.5
Group B Scores: 85, 53, 64           Group B mean: 67.333
Group A Mean - Group B Mean: 7.167
Group A Scores: 68, 77, 85, 53       Group A mean: 70.75
Group B Scores: 82, 64, 71           Group B mean: 72.333
Group A Mean - Group B Mean: -1.583
Group A Scores: 68, 77, 85, 64       Group A mean: 73.5
Group B Scores: 82, 53, 71           Group B mean: 68.667
Group A Mean - Group B Mean: 4.833
Group A Scores: 68, 77, 85, 71       Group A mean: 75.25
Group B Scores: 82, 53, 64           Group B mean: 66.333
Group A Mean - Group B Mean: 8.917
Group A Scores: 68, 77, 53, 64       Group A mean: 65.5
Group B Scores: 82, 85, 71           Group B mean: 79.333
Group A Mean - Group B Mean: -13.833
Group A Scores: 68, 77, 53, 71       Group A mean: 67.25
Group B Scores: 82, 85, 64           Group B mean: 77
Group A Mean - Group B Mean: -9.75
Group A Scores: 68, 77, 64, 71       Group A mean: 70
Group B Scores: 82, 85, 53           Group B mean: 73.333
Group A Mean - Group B Mean: -3.333
Group A Scores: 68, 82, 85, 53       Group A mean: 72
Group B Scores: 77, 64, 71           Group B mean: 70.667
Group A Mean - Group B Mean: 1.333
Group A Scores: 68, 82, 85, 64       Group A mean: 74.75
Group B Scores: 77, 53, 71           Group B mean: 67
Group A Mean - Group B Mean: 7.75
Group A Scores: 68, 82, 85, 71       Group A mean: 76.5
Group B Scores: 77, 53, 64           Group B mean: 64.667
Group A Mean - Group B Mean: 11.833
Group A Scores: 68, 82, 53, 64       Group A mean: 66.75
Group B Scores: 77, 85, 71           Group B mean: 77.667
Group A Mean - Group B Mean: -10.917
Group A Scores: 68, 82, 53, 71       Group A mean: 68.5
Group B Scores: 77, 85, 64           Group B mean: 75.333
Group A Mean - Group B Mean: -6.833
Group A Scores: 68, 82, 64, 71       Group A mean: 71.25
Group B Scores: 77, 85, 53           Group B mean: 71.667
Group A Mean - Group B Mean: -0.417
Group A Scores: 68, 85, 53, 64       Group A mean: 67.5
```

```
Group B Scores: 77, 82, 71           Group B mean: 76.667
Group A Mean - Group B Mean: -9.167
Group A Scores: 68, 85, 53, 71       Group A mean: 69.25
Group B Scores: 77, 82, 64           Group B mean: 74.333
Group A Mean - Group B Mean: -5.083
Group A Scores: 68, 85, 64, 71       Group A mean: 72
Group B Scores: 77, 82, 53           Group B mean: 70.667
Group A Mean - Group B Mean: 1.333
Group A Scores: 68, 53, 64, 71       Group A mean: 64
Group B Scores: 77, 82, 85           Group B mean: 81.333
Group A Mean - Group B Mean: -17.333
Group A Scores: 77, 82, 85, 53       Group A mean: 74.25
Group B Scores: 68, 64, 71           Group B mean: 67.667
Group A Mean - Group B Mean: 6.583
Group A Scores: 77, 82, 85, 64       Group A mean: 77
Group B Scores: 68, 53, 71           Group B mean: 64
Group A Mean - Group B Mean: 13
Group A Scores: 77, 82, 85, 71       Group A mean: 78.75
Group B Scores: 68, 53, 64           Group B mean: 61.667
Group A Mean - Group B Mean: 17.083
Group A Scores: 77, 82, 53, 64       Group A mean: 69
Group B Scores: 68, 85, 71           Group B mean: 74.667
Group A Mean - Group B Mean: -5.667
Group A Scores: 77, 82, 53, 71       Group A mean: 70.75
Group B Scores: 68, 85, 64           Group B mean: 72.333
Group A Mean - Group B Mean: -1.583
Group A Scores: 77, 82, 64, 71       Group A mean: 73.5
Group B Scores: 68, 85, 53           Group B mean: 68.667
Group A Mean - Group B Mean: 4.833
Group A Scores: 77, 85, 53, 64       Group A mean: 69.75
Group B Scores: 68, 82, 71           Group B mean: 73.667
Group A Mean - Group B Mean: -3.917
Group A Scores: 77, 85, 53, 71       Group A mean: 71.5
Group B Scores: 68, 82, 64           Group B mean: 71.333
Group A Mean - Group B Mean: 0.167
Group A Scores: 77, 85, 64, 71       Group A mean: 74.25
Group B Scores: 68, 82, 53           Group B mean: 67.667
Group A Mean - Group B Mean: 6.583
Group A Scores: 77, 53, 64, 71       Group A mean: 66.25
Group B Scores: 68, 82, 85           Group B mean: 78.333
Group A Mean - Group B Mean: -12.083
Group A Scores: 82, 85, 53, 64       Group A mean: 71
Group B Scores: 68, 77, 71           Group B mean: 72
Group A Mean - Group B Mean: -1
Group A Scores: 82, 85, 53, 71       Group A mean: 72.75
Group B Scores: 68, 77, 64           Group B mean: 69.667
Group A Mean - Group B Mean: 3.083
Group A Scores: 82, 85, 64, 71       Group A mean: 75.5
Group B Scores: 68, 77, 53           Group B mean: 66
Group A Mean - Group B Mean: 9.5
Group A Scores: 82, 53, 64, 71       Group A mean: 67.5
Group B Scores: 68, 77, 85           Group B mean: 76.667
Group A Mean - Group B Mean: -9.167
Group A Scores: 85, 53, 64, 71       Group A mean: 68.25
```

```
Group B Scores: 68, 77, 82           Group B mean: 75.667
Group A Mean - Group B Mean: -7.417
```

In our actual experiment, the difference in group means was about 15.33. Since we specified a two-sided alternative hypothesis, the p-value will be the proportion of the 35 randomizations where the difference in group means is at least as large in magnitude as 15.33, in either direction. There were three such randomizations: the ones with differences in group means of 15.33, -17.33, and 17.08. Our p-value is therefore $3/35 \approx 0.086$.

(d) Write a sentence or two interpreting the p-value in context. (I am not looking for a conclusion for the test here, but a restatement of the definition of the p-value in the context of this example. We will talk about using p-values to draw conclusions in the next Chapter.)

If there was actually no difference in population mean scores for the two groups, the probablity of obtaining a difference in group means at least as large as the difference we observed in this study due to randomization alone is approximately 0.086.

Problem 2: Adapted from Sleuth3 1.26

Each year, the League of Conservation Voters (LCV) identifies legislative votes taken in each house of the U.S. Congress – votes that are highly influential in establishing policy and action on evironmental problems. The LCV then publishes whether each member of Congress cast a pro-environment or an anti-environment vote. The following R code reads in a data set with the LCV's ratings for each member of House of Representatives from 2005 to 2007, and filters it to include results for only Democrats and Republicans (not Independents).

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

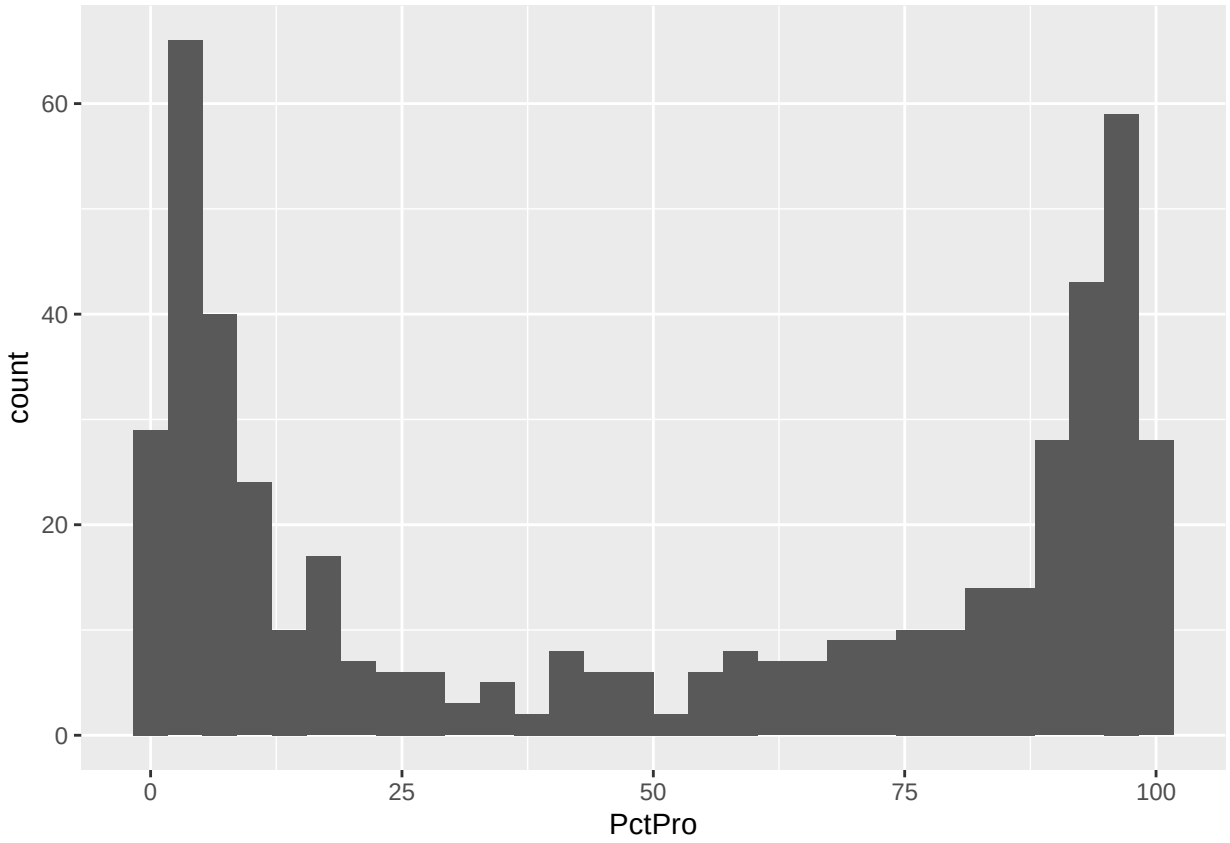
```
library(ggplot2)
lcv_ratings <- read.csv("http://www.evanlray.com/data/sleuth3/ex0126_lcv_house.csv") %>%
  filter(
    Party %in% c("D", "R")
  )
```

(a) Variations on a histogram

The purpose of this part of the exercise is to explore a bunch of options for how to build histograms using `ggplot2` in R. As a starting point, here's a histogram summarizing the distribution of the overall percentage of "pro-environment" votes for all representatives in the data set:

```
ggplot(data = lcv_ratings, mapping = aes(x = PctPro)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



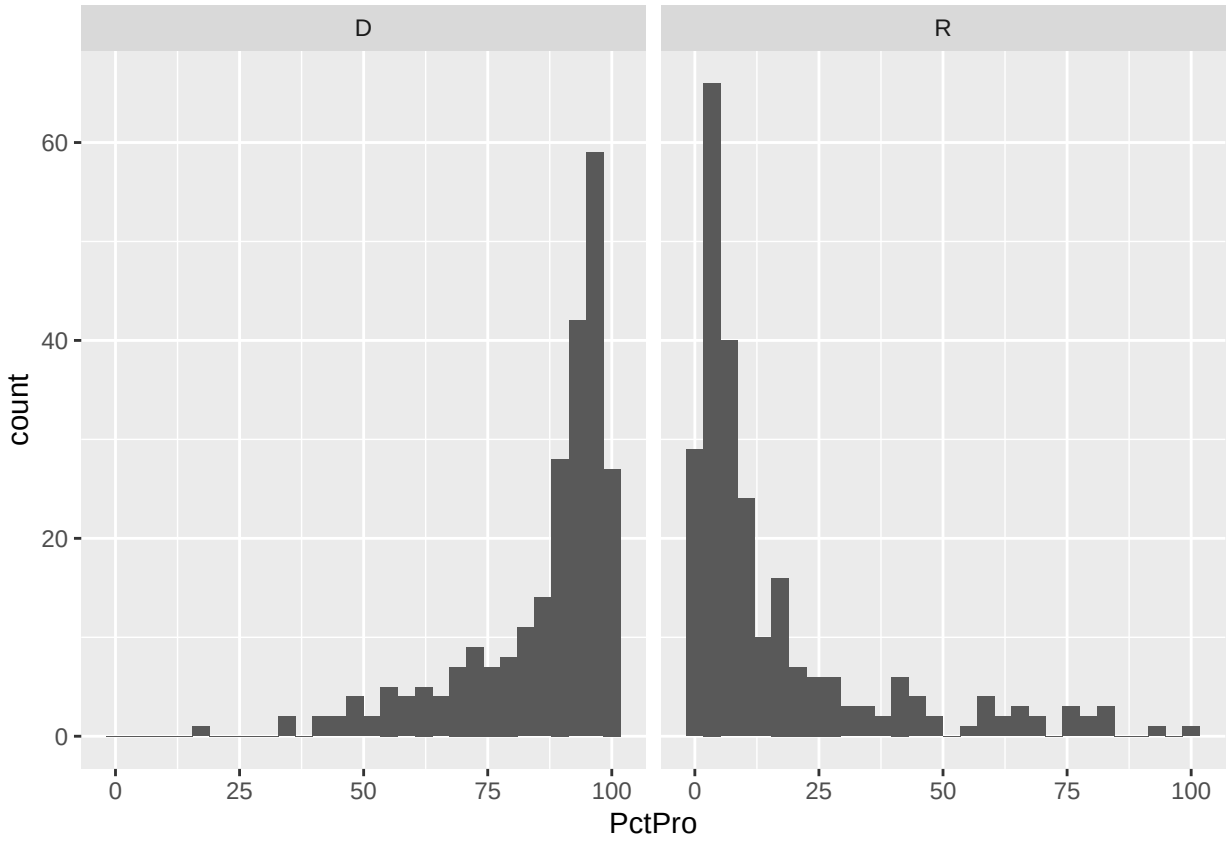
Recall from your previous work or the DataCamp assignment that the first line of this code sets up a new plot based on the `lcv_ratings` data frame. It also sets up an aesthetic mapping, putting the `PctPro` variable on the `x` (horizontal) axis. The geometry type of the plot is a `histogram`.

i. Add facetting by the representatives' `Party`

Add facetting to the plot code below using `facet_grid` or `facet_wrap` (either way, your choice). If you don't remember how to do this, check out the module about this on the DataCamp assignment.

```
ggplot(data = lcv_ratings, mapping = aes(x = PctPro)) +
  geom_histogram() +
  facet_grid( ~ Party)
```

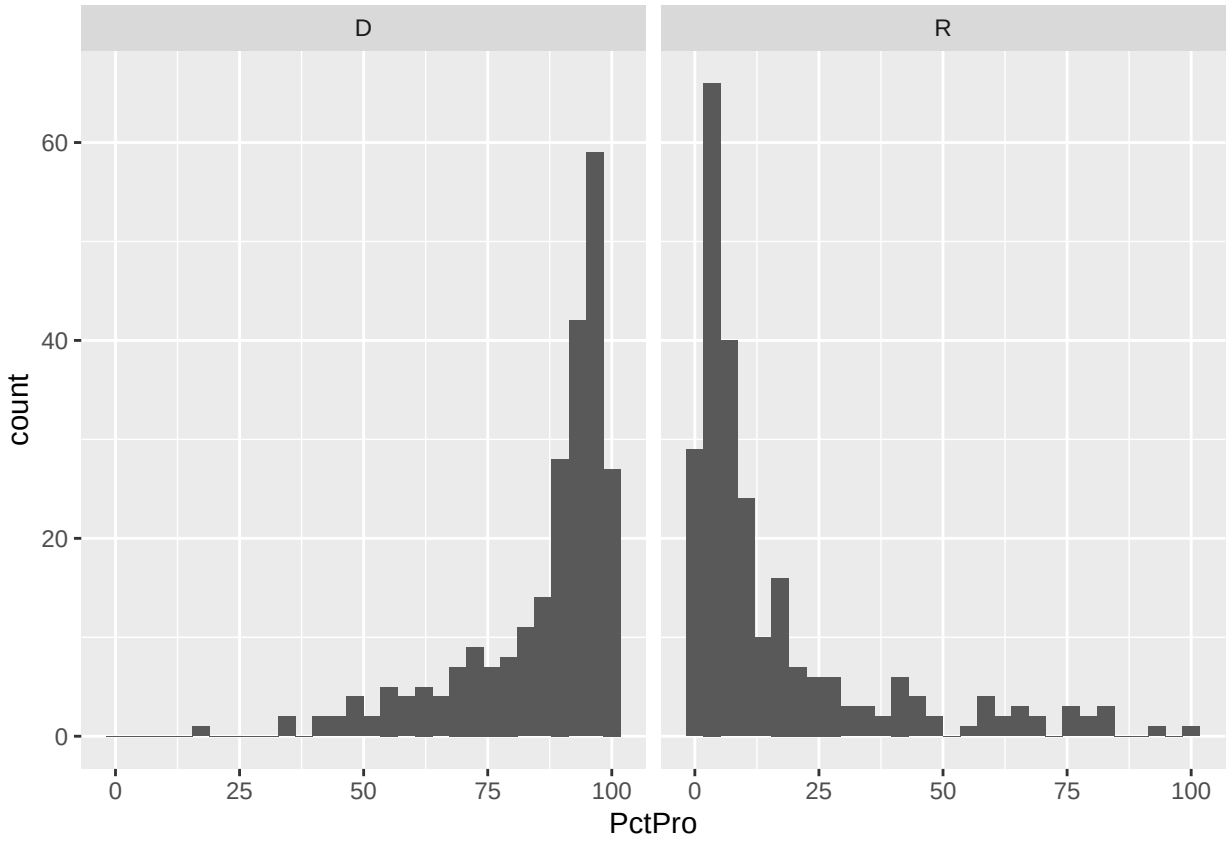
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



. . . or . . .

```
ggplot(data = lcv_ratings, mapping = aes(x = PctPro)) +
  geom_histogram() +
  facet_wrap( ~ Party)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

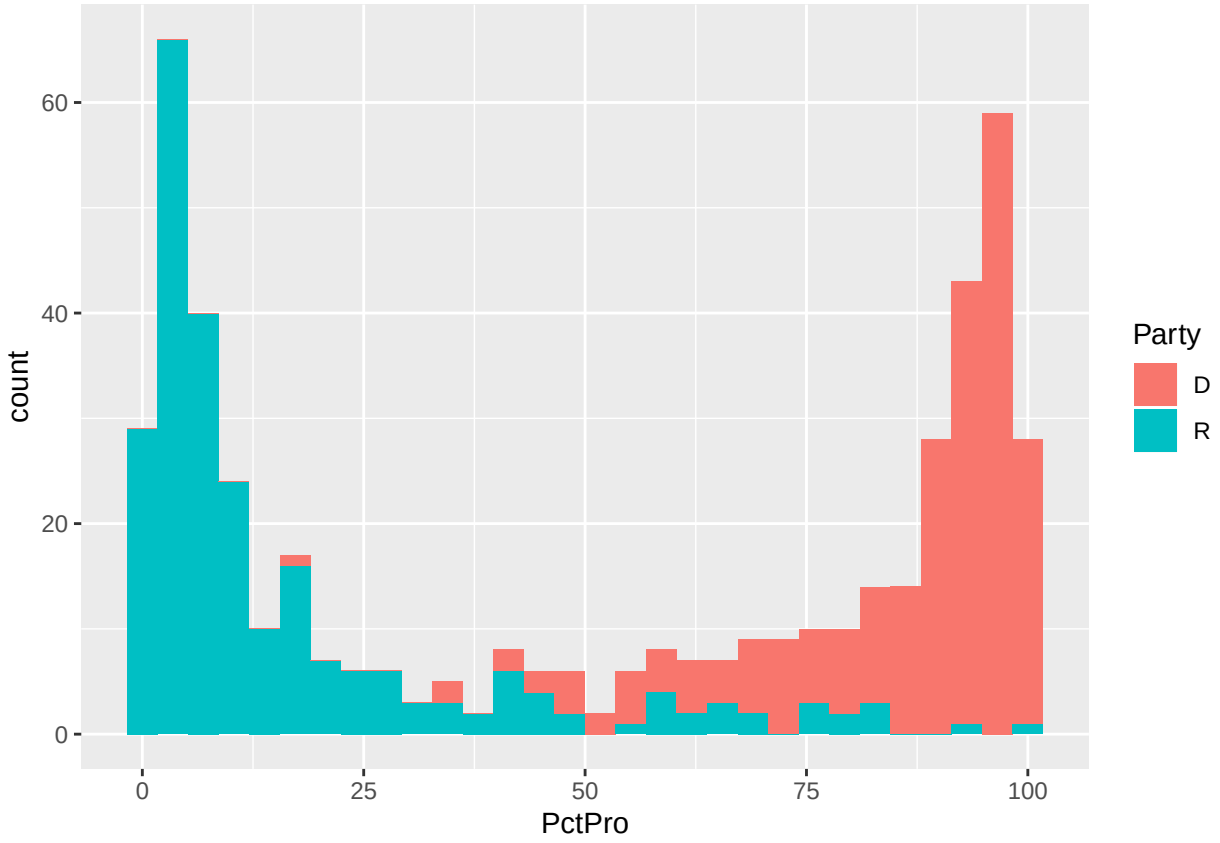


ii. Add a fill by the representatives' `Party`

Add a second aesthetic mapping by adding `, fill = Party` after `x = PctPro`.

```
ggplot(data = lcv_ratings, mapping = aes(x = PctPro, fill = Party)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

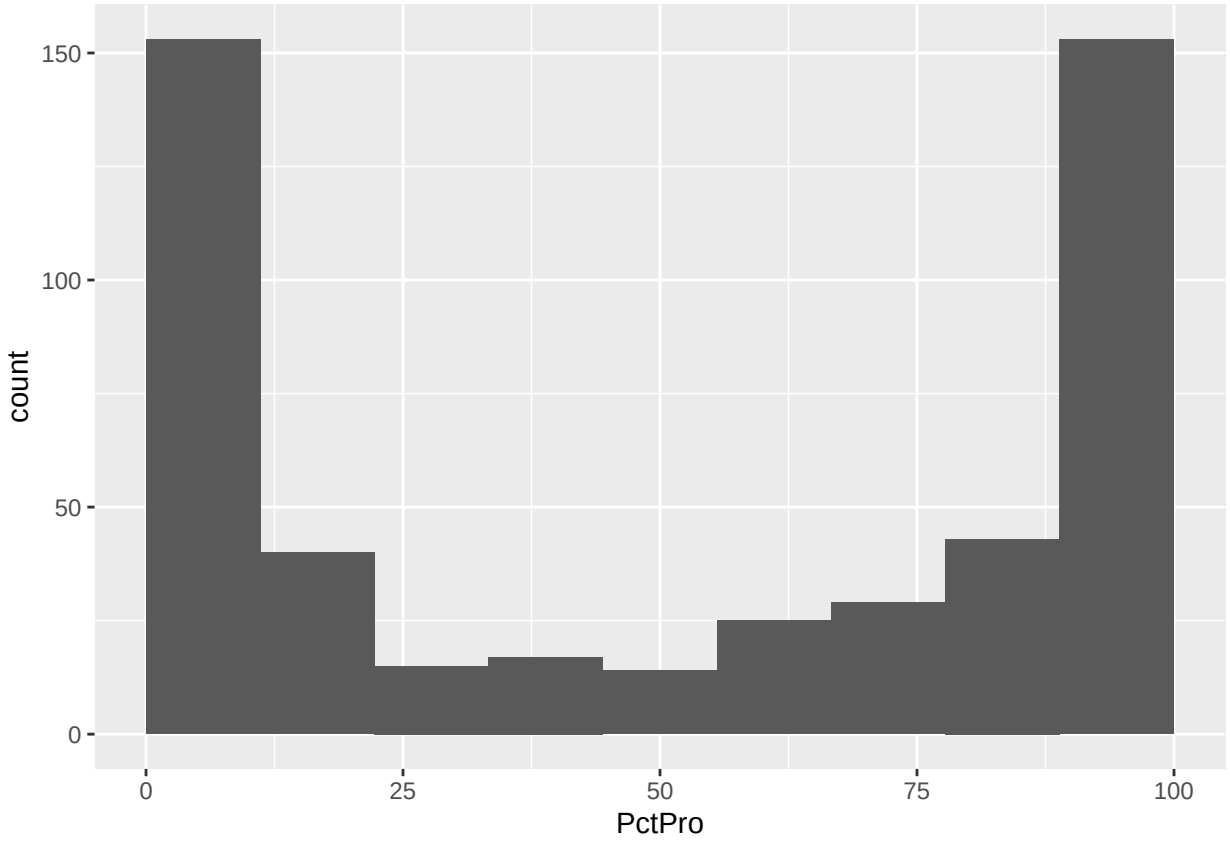


iii. Set the number of bins used and the boundary of the bins

Add a `bins = 10` argument to the `geom_histogram()` function call (between the parentheses after `geom_histogram`). Experiment with the number of bins used until you find a number of bins that looks like it provides a good summary of the data.

Also add a `boundary = 0` argument to the `geom_histogram()` function call, separated by a comma from the `bins` argument. This argument specifies that one of the bins will have its left endpoint at 0 on the horizontal axis. The locations of the other bins are determined by the width of the bins.

```
ggplot(data = lcv_ratings, mapping = aes(x = PctPro)) +
  geom_histogram(bins = 10, boundary = 0)
```



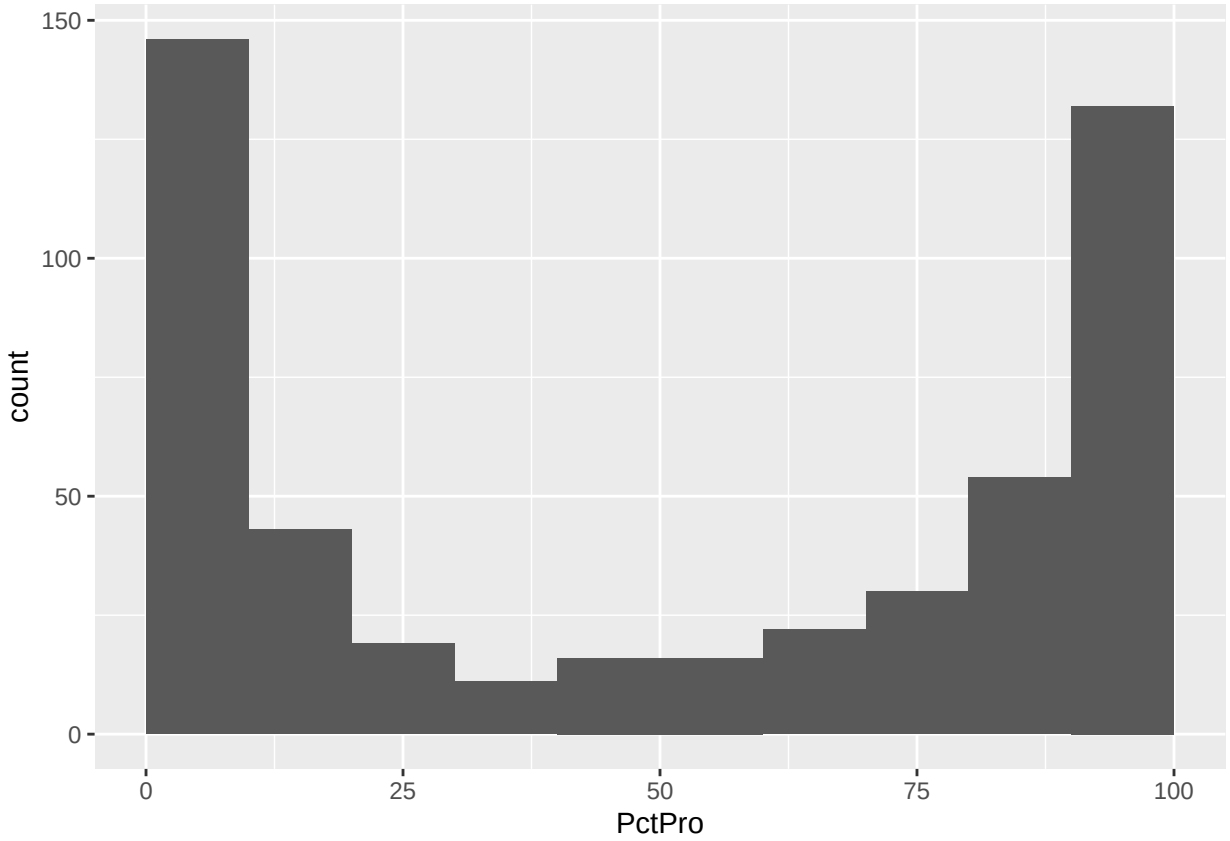
iv. Set the width of the bins used and the boundary of the bins

This is another alternative for how to set the size of the bins.

Add a `binwidth = 10` argument to the `geom_histogram()` function call (between the parentheses after `geom_histogram`). Experiment with the width of the bins used until you find a bin width that looks like it provides a good summary of the data.

Also add a `boundary = 0` argument to the `geom_histogram()` function call, separated by a comma from the `bins` argument. This argument specifies that one of the bins will have its left endpoint at 0 on the horizontal axis. The locations of the other bins are determined by the bin width.

```
ggplot(data = lcv_ratings, mapping = aes(x = PctPro)) +
  geom_histogram(binwidth = 10, boundary = 0)
```

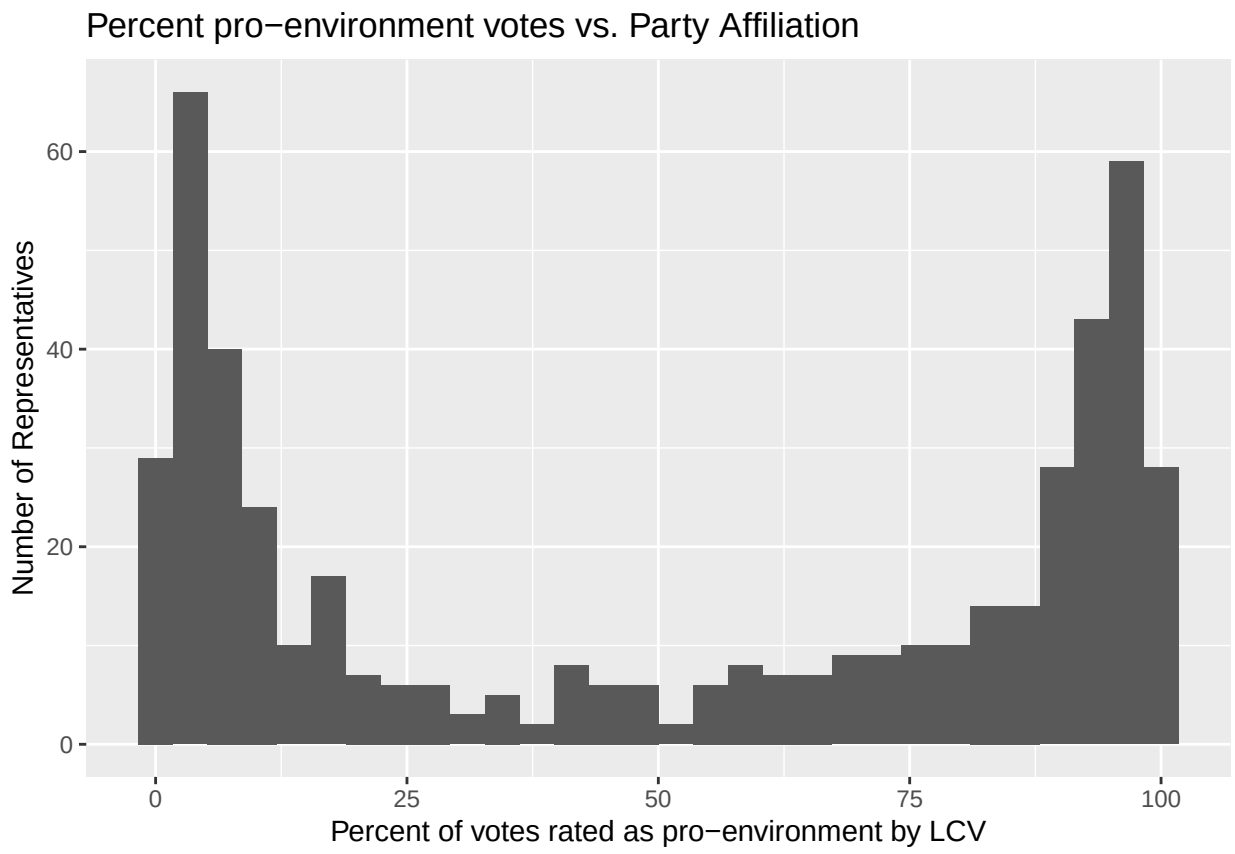


v. Set a more informative plot title and axis labels

I've provided the set up for this below, but you should fill in appropriate values between the quotes.

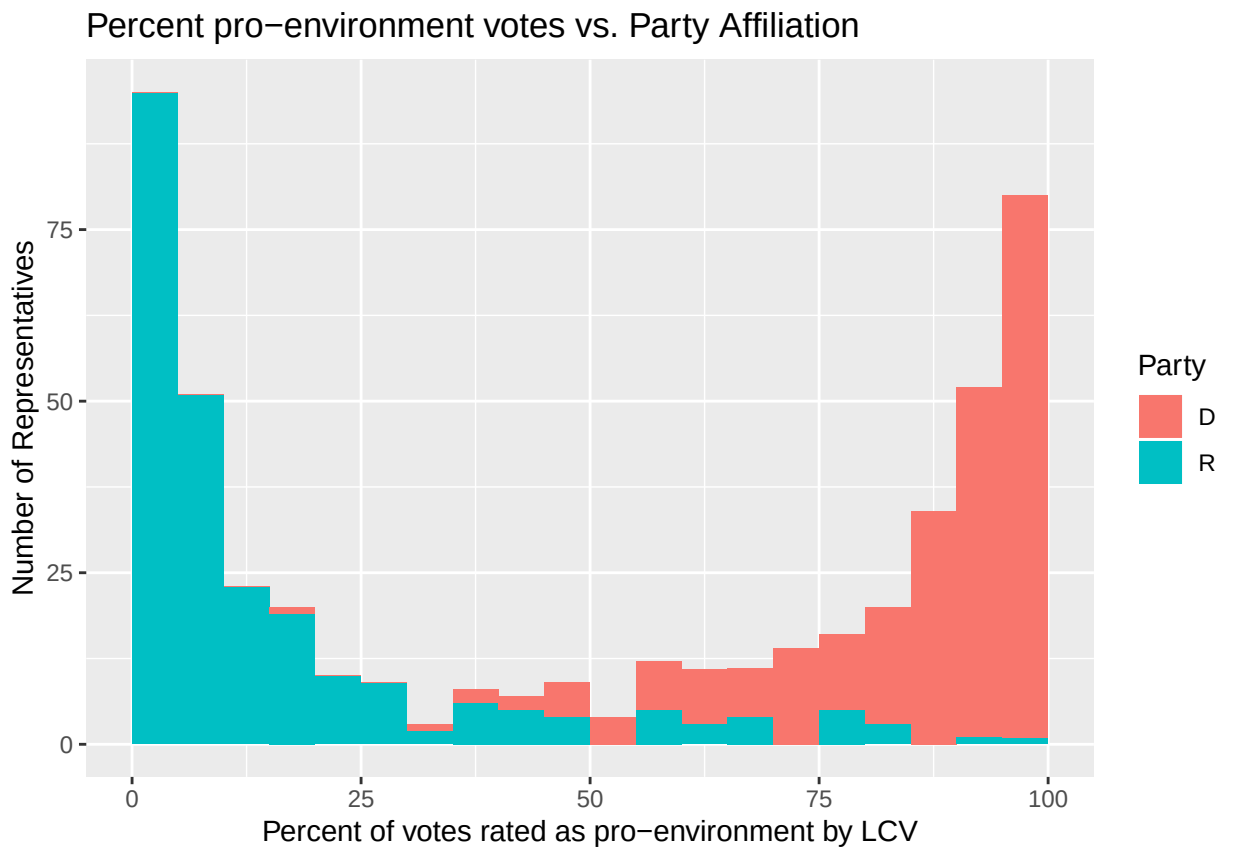
```
ggplot(data = lcv_ratings, mapping = aes(x = PctPro)) +
  geom_histogram() +
  ggtitle("Percent pro-environment votes vs. Party Affiliation") +
  xlab("Percent of votes rated as pro-environment by LCV") +
  ylab("Number of Representatives")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



vi. By combining your choice of some of the different ideas above, create a histogram that you think does the best job of conveying what's going on in the data set.

```
ggplot(data = lcv_ratings, mapping = aes(x = PctPro, fill = Party)) +
  geom_histogram(binwidth = 5, boundary = 0) +
  ggtitle("Percent pro-environment votes vs. Party Affiliation") +
  xlab("Percent of votes rated as pro-environment by LCV") +
  ylab("Number of Representatives")
```



(b) Conduct a hypothesis test to compare pro-environment votes for Democratic and Republican members of the house of representatives.

i. Define the parameter or parameters you are making inference about.

μ_D = average proportion of votes rated as pro-environment by LCV, among Democratic members of the house of representatives (or people who might be elected to be Democratic members of the house of representatives).

μ_R = average proportion of votes rated as pro-environment by LCV, among Republican members of the house of representatives (or people who might be elected to be Democratic members of the house of representatives).

ii. State the null and alternative hypotheses for a two-sided test.

$H_0 : \mu_D = \mu_R$

$H_A : \mu_D \neq \mu_R$

If you went into this with specific ideas about the choices likely to be made by members of the two parties, you might also have specified a one-sided alternative:

$H_A : \mu_D > \mu_R$

iii. The code below is adapted from the code in our first lab, and calculates an approximate p-value from a permutation test with 1,000 permutations. Run this code to get the approximate p-value (you don't need to modify the code at all). (Just so you're not confused by the output - the approximate p-value is 0!)

```
library(mosaic)
```

```
## Warning: package 'mosaic' was built under R version 3.6.2
```

```
## Loading required package: lattice
```

```
## Loading required package: ggformula
```

```
## Warning: package 'ggformula' was built under R version 3.6.2
```

```
## Loading required package: ggstance
```

```
## Warning: package 'ggstance' was built under R version 3.6.2
```

```
##
## Attaching package: 'ggstance'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     geom_errorbarh, GeomErrorbarh
```

```
##
## New to ggformula?  Try the tutorials:
##  learnr::run_tutorial("introduction", package = "ggformula")
##  learnr::run_tutorial("refining", package = "ggformula")
```

```
## Loading required package: mosaicData
```

```
## Warning: package 'mosaicData' was built under R version 3.6.2
```

```
## Loading required package: Matrix
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.
##
## Note: If you use the Matrix package, be sure to load it BEFORE loading mosaic.
```

```
##
## Attaching package: 'mosaic'
```

```
## The following object is masked from 'package:Matrix':
##
##     mean
```

```
## The following object is masked from 'package:ggplot2':
##
##     stat

## The following objects are masked from 'package:dplyr':
##
##     count, do, tally

## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median,
##     prop.test, quantile, sd, t.test, var

## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```r
# set a seed to get reproducible results
set.seed(513945)
# find observed difference in means for the actual sample data
group_means <- lcv_ratings %>%
  group_by(Party) %>%
  summarize(mean_PctPro = mean(PctPro))
observed_group_means_difference <- group_means$mean_PctPro[1] - group_means$mean_PctPro[2]
# simulate 10000 random assignments of legislators to the different parties,
# and for each simulation calculate the mean
simulation_results <- data.frame(
  group_means_difference = rep(NA, 1000)
)
for(i in seq_len(1000)) {
  shuffled_group_means <- lcv_ratings %>%
    mutate(
      Party = shuffle(Party)
    ) %>%
    group_by(Party) %>%
    summarize(mean_PctPro = mean(PctPro))

  simulation_results$group_means_difference[i] <- shuffled_group_means$mean_PctPro[1] - shuffled_group_m
}
count_greater <- simulation_results %>%
  summarize(
    count = sum(group_means_difference >= observed_group_means_difference)
  )
count_less <- simulation_results %>%
  summarize(
    count = sum(group_means_difference <= -observed_group_means_difference)
  )
approximate_pval <- (count_greater$count + count_less$count) / 1000
approximate_pval
```

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

iv. Write a sentence or two interpreting the p-value in context. (I am not looking for a conclusion for the test here, but a restatement of the definition of the p-value in the context of this example. We will talk about using p-values to draw conclusions in the next Chapter.)

If in fact the average proportion of votes rated as pro-environment by LCV, among people who might be elected to be Democratic members of the house of representatives was equal to the average proportion of votes rated as pro-environment by LCV among people who might be elected to be Republican members of the house of representatives, the probability of obtaining a difference in group averages at least as large as the difference observed in this sample due to randomization alone would be approximately 0.

v. Explain why the approximate p-value from part iii is just an approximate p-value. What would we have to do to find the exact p-value? (This is related to the definition of a sampling distribution.)

A calculation of the exact p-value for the test would be based on the exact sampling distribution of the statistic: in this case, the values of the statistic that could be obtained through all possible assignments of representatives (with their corresponding vote percentages) to the different parties. However our calculation in part iii was based on only 1000 possible assignments of the representatives to the different parties. This gives us a pretty good sense of what the sampling distribution looks like, but it does not get us the full sampling distribution. Our calculation based on this approximate sampling distribution is just an approximation of the p-value.