Problem Set 1: Data and Inference

Linh Vu   Collab: Brice Laurent

Due: September 22, 2023

This assignment is **due Friday, September 22 at 11:59pm**, handed in through Gradescope. There will be two separate submissions, one for your pdf, and the other for you rmd file. Show your work and provide clear, convincing, and succinct explanations when asked. **Incorporate the relevant R output in this R markdown file**; choose the included R wisely. Only the key output should be displayed for each problem and the relevant parts should be **highlighted** in some way. Make sure that you write-up any interpretation of R-code in your own words (don't just provide the output). Think of this as a document intended for a collaborator.

When performing a formal hypothesis test throughout this class, be sure to explicitly state (1) hypotheses, (2) the calculated test statistic (and degrees of freedom if appropriate), (3) the calculated p-value or critical value, and (4) the conclusion in context of the problem along with the scope of inference. Use Type I error rates of $\alpha = 0.05$ and confidence levels of 95% unless explicitly stated otherwise. You can assume all tests are two-sided unless otherwise specified.

Collaboration policy (for this and all future homework): You are encouraged to discuss the problems with other students, but you must write up your solutions yourself and in your own words. Copying someone else's solution, or just making trivial changes is not acceptable. Please indicate the students with whom you worked.

Problem 1. A nutritionist is interested in determining whether the consumption of foods coming from plants in the Solanaceae family (commonly called *nightshades*) is truly associated with increases in inflammation. This association is often reported in online media sources, with maybe the most famous example being a self-proclaimed fitness and nutrition guru closely linked to the NFL quarterback Tom Brady.

The nutritionist enrolls 80 volunteers from their practice into the randomized study: 40 are randomly assigned to each of two treatments groups. The active treatment group is put on the strict TB12 diet (which restricts nightshade consumption) while the control group is assigned to a similar diet calorically and nutritionally but with no nightshade restriction (similar to a Paleo diet). The nutritionist measures two outcomes at two time points: (1) before entering the study (aka, at baseline) and (2) after three months on the diet in which the volunteers prepared their own meals. The two outcomes for inflammation that measured were:

i. Self-perceived joint inflammation/pain on a 0 to 100 scale.
ii. Level of C-reactive protein in the blood.

Both groups saw a statistically significant improvement in both outcomes, but this improvement was similar in the two groups (not statistically significant).

(a) Does this study suggest that the restriction of nightshades in someone's diet causes lowering levels of inflammation? How do you know? Explain in 2 or 3 sentences.

This study does not suggest a causal relationship between restriction of nightshades in diet and decrease in inflammation level. It is a paired study (comparing someone's pre-study inflammation level with their own post-study inflammation level), and although the improvement was significant, it was similar across the two groups. Because of this, we know that improvement in inflammation can't be purely attributed to less consumption of nightshades.

(b) How could proponents of restricting nightshade consumption see the results of this study and interpret it as confirming their beliefs? Explain in 2 or 3 sentences.

They could ignore the results of the control group, focus on the experimental group's result that we see improvement in inflammation levels among those on the TB12 diet, and conclude that low nightshades consumption is linked to less joint inflammation. They could also go further and establish a causal link (because we do have random assignment of treatment among subjects).

(c) Discuss the generalizability of this study to a larger population in 2 or 3 sentences.

I'm hesitant to generalize this study to a larger population because the subjects are from the nutritionist's practice. Because this nutrition has a clear biased perspective going into the study, their clients are likely to hold similar views, and (particularly if they know the purposes of the study going in) might be inclined to report (overly) favorable results about the TB12 diet.

The fact that the subjects are volunteers at a clinic also makes the cohort not representative of the population because patients at this nutritionist's practice might be more health-conscious and wealthier and thus exercise more/are in better shape than the general population, for example.

(d) The nutritionist did not restrict or measure any other known contributors to inflammation by the participants (like intense exercise, alcohol and tobacco consumption, etc.). How would this influence the study results? Briefly discuss the pros and cons of this decision.

In terms of pros, the study has random assignment of treatment, so it controls for confounding variables already. In terms of cons, since the study only has 80 participants (40 for each treatment group), this small sample size means that it's possible we didn't control for the counfounding factors. For example, if we divide the groups in a way that we have many more smokers in the treatment group than in the control group, we might not see a pronounced effect.

(e) Propose two ways to take into account that participants may have had varying levels of other contributors to inflammation: 1) by modifying the design the study and 2) by leaving the original design of the study as is and modifying the planned analysis of the results.

(1) Modify study design: We can categorize study participants into three groups (low, medium, and high levels of contributors to inflammation). We can conduct a multiple comparisons test (eg. ANOVA) to see if different groups see different levels of improvement.

(2) Modify planned analysis: To control for the varying levels of inflammation contributors, we can run multiple linear regressions and include the inflammatory contributor levels as predictors in the model. That way the results will show how much the inflammation level improves depending on the different levels of inflammation contributors.

Problem 2. A pharmaceutical company is surveying through 100 different targeted compounds to try to determine whether any of them may be useful in treating migraine headaches. From previous experiments like this, they believe that each compound independently has a 2/100 chance of truly being effective, and 98/100 chance of having zero effect. For each potential compound, they perform a hypothesis test to determine whether it is truly effective at the $\alpha = 0.01$ level. A truly effective drug will be statistically significant based on this hypothesis test 80% of the time (which is called statistical *power*).

(a) What is the expected number of compounds that will be shown to be statistically significant based on these 100 separate hypothesis tests?

Let S denote the event where the compound is found to have a statistically significant effect, and let E denote the event where the compound is actually effective. Let c denote complementary event.

We interpret the α level to be the probability of the event where we find statistically significance when the compound is actually not effective. Mathematically, this is written as $p(S|E_c) = 0.01$

Using LOTE, considering that we are conducting 100 hypothesis tests, we can expect to get to find

$$E(S) = E(S|E)p(E) + E(S|E_c)p(E_c)$$

$$= 100 \cdot [(0.08)(\frac{2}{100}) + (0.01)(\frac{98}{100})] = 2.58$$

compounds to be statistically significant.

(b) Given a compound is flagged as statistically significant, what is the probability that it is actually effective in treating migraine headaches?

Using Bayes rule and the same notation as in part (a), we get:

$$p(E|S) = \frac{p(S|E) \cdot p(E)}{p(S)}$$

$$= \frac{p(S|E) \cdot p(E)}{p(S|E) \cdot p(E) + p(S|E_c) \cdot p(E_c)}$$

$$= \frac{0.8 \cdot 0.02}{0.8 \cdot 0.02 + 0.01 \cdot (1 - 0.02)} = 0.62$$

(c) After testing all 100 potential compounds, the company has exactly 1 compound that was deemed to be statistically significant based on the tests. Let π be the probability that this one compound is actually effective in treating migraine headaches. How does π compare to your result in part (b)? Explain briefly. Note: you do not need to calculate π, just compare it to your answer in the previous part.

In this scenario, $p(E)$ is closer to 0.01 instead of 0.02 like we initially assumed. Using the same formula as in (b), we find that the probability decreases as $p(E)$ decreases from 0.02 to 0.01, so π should be smaller than 0.62. For example, when we plug in $p(E) = 0.01$, we get $p(E|S) = 0.44$, which is smaller than 0.62.

Problem 3. Let identically distributed observations $X_i \sim N(\mu, \sigma^2)$ for $i = 1, 2, ..., n$. Let $\bar{X}$ be the usual sample mean.

(a) Assume that all X_i are independent. Determine $E(\bar{X})$ and $\text{Var}(\bar{X})$. Feel free to just reference results from class.

According to lecture notes and CLT, $E(\bar{X}) = \mu$, and $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$

(b) Now assume that $\text{Cov}(X_i, X_j) = \rho\sigma^2$ (where ρ is the correlation) for all $i \neq j$. Determine $E(\bar{X})$ and $\text{Var}(\bar{X})$.

$E(\bar{X}) = \mu$. This is the same as in (a) because linearity of expectation holds regardless of whether X_i are independent or not.

$$\text{Var}(\bar{X}) = \text{Var}(\frac{X_1 + X_2 + ... + X_n}{n})$$

$$= \frac{1}{n^2}[\sum_{i=1}^{n} \text{Var}(X_i) + \sum_{i,j diff} \text{Cov}(X_i, X_j)]$$

$$= \frac{1}{n^2}[n\text{Var}(X_1) + \frac{n(n-1)}{2} 2\rho\sigma^2]$$

$$= \frac{\sigma^2}{n}(1 + \rho(n-1))$$

(c) For what values of ρ will $\text{Var}(\bar{X})$ be large? When will it be small? Make an intuitive argument why this makes sense (2 or 3 sentences).

$\text{Var}(\bar{X})$ is large when ρ is large (near 1), $\text{Var}(\bar{X})$ is small when ρ is small (near -1, although $\text{Var}(\bar{X})$ can't be negative). This makes sense intuitively because when ρ is large, X_i and X_j are very positively correlated with each other; in the extreme case where ρ is 1, X_i and X_j are perfectly correlated and equal to each other, so $\bar{X}$ can be very off from the true mean, making $\text{Var}(\bar{X})$ very large. When ρ is small (negative), X_i and X_j are negatively correlated and there is more randomness/variation in values, so the sample means can be quite close to each other, making $\text{Var}(\bar{X})$ small.

(d) What implications does this have for a real life data set? When will ρ be non-zero? What bearing does this have on the usual one-sample t-test to determine if $H_0 : \mu = 0$ is reasonable?

For a real life dataset, we prefer when we get independent observations ($\rho = 0$) because then we get a smaller variance when we try to obtain an estimate for the population mean from the sample mean. ρ is non-zero when observations are not independent and can influence each other's value.

When the data are dependent (instead of independent), ρ is non-zero. If the data are actually negatively correlated, ρ is negative, so we get a smaller variance than in actuality. The T-statistic $= \frac{\bar{X}-0}{\text{Var}(\bar{X})/\sqrt{n}}$ becomes larger than in reality, making it less likely for us to get a statistically significant result. This means we're less likely to reject the null hypothesis, leading to a lower type I error rate (which happens when we erroneously reject the null hypothesis when it's actually true). If the data are positively correlated, the trend is reverse and we get a higher type I error rate.

Problem 4. Let $X_1 \sim N(\mu = 1, \sigma^2 = 3^2)$ and independently $X_2 \sim N(\mu = -1, \sigma^2 = 4^2)$. Let $Y = (X_1 + X_2)^2$.

(a) Find $P(X_1 < X_2)$.

$$X_1 - X_2 \sim N(\mu = 1 - (-1) = 2, \sigma^2 = 3^2 + 4^2 = 5^2)$$

To simplify:

$$X_1 - X_2 \sim N(2, 25)$$

Thus, $P(X_1 < X_2) = P(X_1 - X_2 < 0) = 0.34$

```
round(pnorm(0, mean = 2, sd = 5), 2)
```

```
## [1] 0.34
```

(b) What distribution does $X_1 + X_2$ have? What distribution is Y based on? Be explicit about parameters and scaling factors.

$$X_1 + X_2 \sim N(\mu = 1 + (-1) = 0, \sigma^2 = 3^2 + 4^2 = 5^2)$$

To simplify:

$$X_1 + X_2 \sim N(0, 25)$$

We also know that

$$\frac{X_1 + X_2}{5} \sim N(0, 1)$$

hence

$$\frac{(X_1 + X_2)^2}{25} \sim \chi^2_1$$

equivalently,

$$\frac{(X_1 + X_2)^2}{25} \sim \gamma(\frac{1}{2}, \frac{1}{2})$$

Thus,

$$Y = (X_1 + X_2)^2 \sim \gamma(\frac{1}{2}, \frac{1}{50})$$

(c) Determine the mean and variance of Y.

Using the Table of distributions:

$E(Y) = \frac{\alpha}{\gamma} = \frac{\frac{1}{2}}{\frac{1}{50}} = 25$

$\text{Var}(Y) = \frac{\alpha}{\gamma^2} = \frac{\frac{1}{2}}{(\frac{1}{50})^2} = 1250$

(d) Without using R, determine $P(Y > E(Y))$. How does this compare to 1/2? Why does this result compared to 1/2 make sense?

Hint: after some analytical work, this can be calculated numerically using the empirical rule for a standard normal distribution.

$$P(Y > E(Y)) = P(Y > 25) = P((X_1 + X_2)^2 > 25)$$

$$= P(X_1 + X_2 > 5 | X_1 + X_2 < -5) = 2P(\frac{X_1 + X_2}{5} > 1) = 0.32$$

since we know that $\frac{X_1+X_2}{5} \sim N(0,1)$ from part (b). We should get a probability smaller than 0.5, and this makes sense because the Gamma distribution is right skewed, so it is uncommon to randomly draw a value larger than the mean.

(e) Approximate the mean and variance of Y along with $P(Y > E(Y))$ using R. Hint: start by generating and storing $n = 1,000,000$ replicates from the Normal distributions above and transform appropriately. Then calculate the mean and variance of y using the commands `mean(y)`, `var(y)`, and `mean(y>`μ_Y`)`, where μ_Y is the true mean of the random variable Y.

```
# set seed and number of simulations
set.seed(139)
nsims <- 10^6

# generate x_1 and x_2
x_1 <- rnorm(nsims, 1, 3)
x_2 <- rnorm(nsims, -1, 4)

# generate y
y <- (x_1 + x_2)^2

# calculate mean, var, mean(y>mean) of y
mean(y)
```

```
## [1] 25.003
```

```
var(y)
```

```
## [1] 1249.383
```

```
mean(y > mean(y))
```

```
## [1] 0.317328
```

Problem 5. What are the the differences in demographics and work habits between people who believe marijuana should be legal versus those who do not. The 2018 General Social Survey (GSS), which was introduced in Lecture 1, has the following relevant variables measured:

Variable	Description
grass	indicator for whether the respondent believes marijuana should be legal. 1 = Yes, 2 = No.
partyid	self-identified political party affiliation. See https://gssdataexplorer.norc.org/variables/141/vshow for full details.
usetech	% of time at work respondent typically spends using different types of electronic technologies. See https://gssdataexplorer.norc.org/variables/2840/vshow for full details.

(a) Define political groups. Create a factor variable named `partycat` with 3 categories: `dem`, `rep`, and `ind` to indicate whether a respondent is in groups 0 & 1, groups 5 & 6, or groups 2, 3, & 4 respectively, in `partyid`. Describe the breakdown of party affiliation after performing these steps.

The biggest group (more than 1000 people, 45.6%) is independent. There are fewer Democrats (731 people, 31%). People who identify as independent make up about 23%.

```
# load data
gss18 = read.csv("data/gss18.csv")

# create x
x <- gss18$partyid

# recode x
x[x %in% c(0,1)] = "dem"
x[x %in% c(5,6)] = "rep"
x[x %in% c(2,3,4,7)] = "ind"

# create partycat
gss18$partycat <- factor(x, levels = c("dem", "rep", "ind"), labels=c("dem", "rep", "ind"))

# count
table(gss18$partycat)
```

```
##
##  dem  rep  ind
##  731  527 1057
```

```
# proportion
prop.table(table(gss18$partycat))
```

```
##
##       dem       rep       ind
## 0.3157667 0.2276458 0.4565875
```

(b) Formally test whether there is a statistically significant difference in `grass` between the two user-defined political parties of `dem` and `rep`.

We're conducting a 2-sample proportions test. The hypotheses are:

Null hypothesis: The proportion of people who believe in legalization of marijuana is the same among the Democrat group and the Republican group.

Alternative hypothesis: The proportion of people who believe in legalization of marijuana is different among the Democrat group and the Republican group.

```r
# subset data
grass.df <- gss18[gss18$partycat %in% c("dem", "rep"), ]

# conduct test
grass.test <- prop.test(table(grass.df$partycat=="dem", grass.df$grass))

# print result
grass.test
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  table(grass.df$partycat == "dem", grass.df$grass)
## X-squared = 58.571, df = 1, p-value = 1.961e-14
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.3434409 -0.2030444
## sample estimates:
##    prop 1    prop 2
## 0.4523810 0.7256236
```

Conducting the `prop.test` above gives us the test statistic of 58.57 (df = 1). The p-value is 0. Because the p-value is much smaller than our α level, we reject the null hypothesis and conclude that a higher proportion of the Democrat group supports legalization of marijuana than among the Republican group.

(c) Provide a relevant 95% confidence interval for the test in the previous part. Interpret this interval and compare to the test result.

The 95% confidence interval for the difference in proportions of marijuana legalization supporters is (-0.34, -0.2). Because this confidence interval is entirely smaller than 0, the difference in proportions can't be 0. This means there is a statistically significant difference in `grass` between the two parties (i.e. more Democrats support legalization of marijuana than Republicans). This conclusion aligns with what we found in (b).

(d) In 2-3 sentences, use the data to address the assumptions to the inferences in pars (b) and (c).

We have the following assumptions:

(1) Independence of observations within each group (political party): This assumption is violated when family members of the same political affiliation influence each other's belief about marijuana legalization.

(2) Independence of observations between each group (political party): This assumption holds because realistically, a Republican's belief on marijuana should not influence a Democrat's belief on marijuana.

(3) The underlying distribution of the observations is normal: we don't need to worry about this assumption since we have a large sample size (per the table below, each category has more than 10 observations).

```r
# assumption 3: check sample size
table(grass.df$partycat=="dem", grass.df$grass)
```

```
##
##           1   2
##   FALSE 152 184
##   TRUE  320 121
```

(e) Formally test whether there is a statistically significant difference in `usetech` between the three political party groups defined in `partycat`. Provide a visual to support your results.

We're conducting an ANOVA F test. The hypotheses are:

Null hypothesis: The mean of `usetech` is the same among across the three political groups.

Alternative hypothesis: The mean of `usetech` is different for at least one political group.

```r
# subset data
usetech.df <- gss18[!is.na(gss18$usetech),]

# run anova test
usetech.test <- aov(usetech~partycat, data=usetech.df)
summary(usetech.test)
```

```
##               Df  Sum Sq Mean Sq F value Pr(>F)
## partycat       2    6175    3088   2.159  0.116
## Residuals   1390 1987547    1430
## 19 observations deleted due to missingness
```

Conducting the `aov` test above gives us the F test statistic of `2.064` (df = `2`). The p-value is `0.127`. Because the p-value is greater than our α level, we fail to reject the null hypothesis and conclude that the average % of time at work spent using technology is comparable across the three political groups.

(f) In 2-3 sentences, use the data to address the assumptions to the inference performed in part (e).

We have the following assumptions:

(1) Independence of observations within each group (political party): This assumption holds because whether someone uses technology at work or not is not influenced by whether another person in the same political party uses technology at work.

(2) Independence of observations between each group (political party): This assumption holds because realistically, a Republican's use of tech at work should not influence a Democrat's use of tech (or lack thereof).

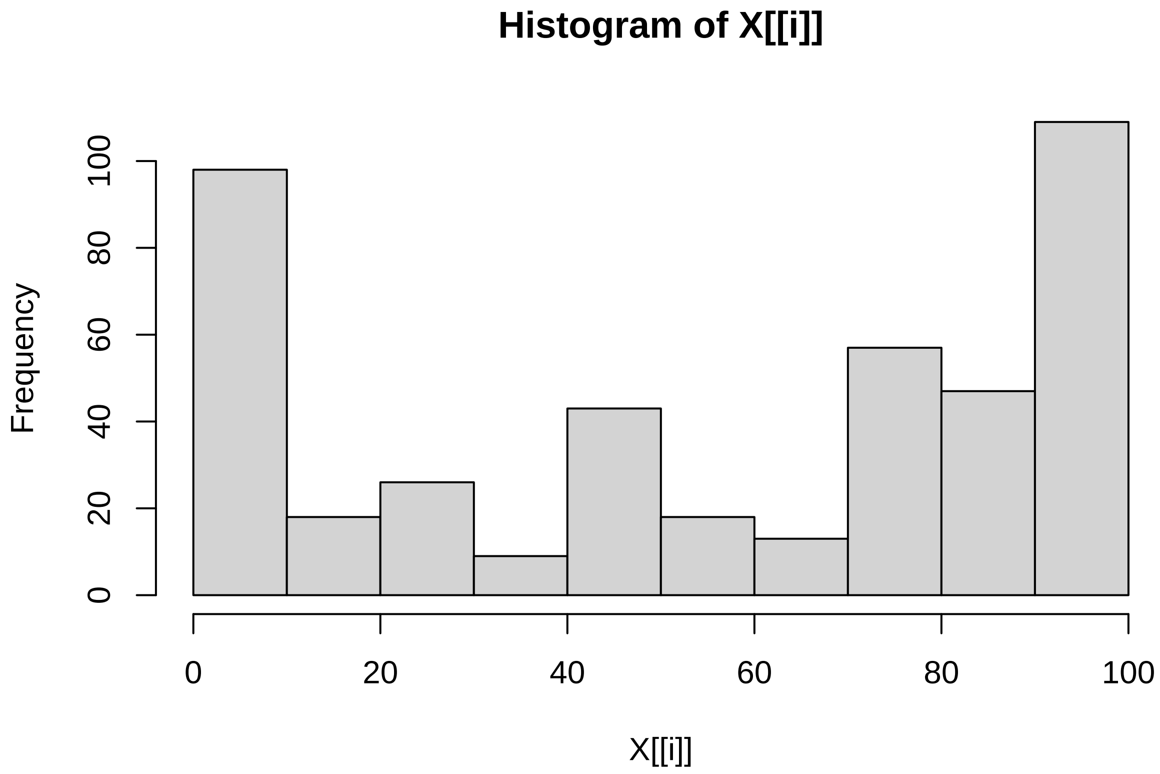
(3) Equal variances between groups: the variances are relatively similar; the independent group has higher variance but not drastically different from the other two groups.

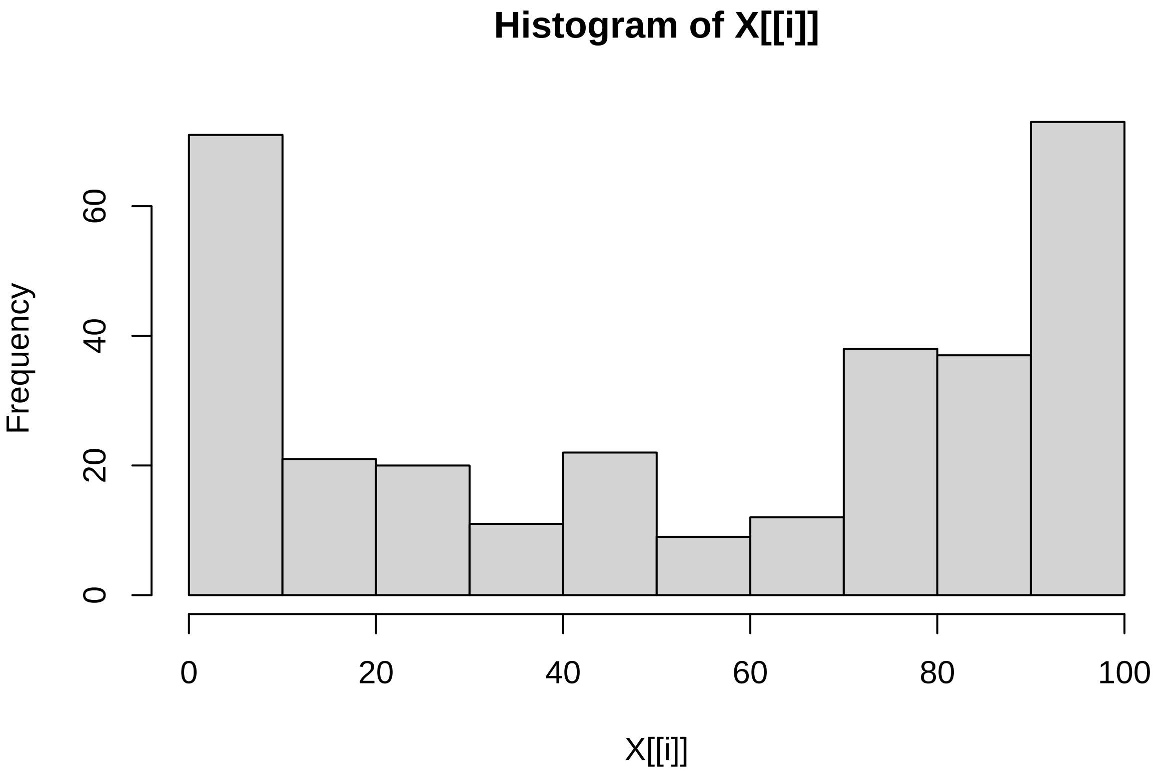
(4) Observations are normally distributed around group mean: This assumption is violated (per histograms below, the order is dem, rep, ind) since the distributions are bimodal. Most people either use a lot of technology at work or not at all. However, our sample sizes are big enough (per part (a)) that we can still proceed with the test.

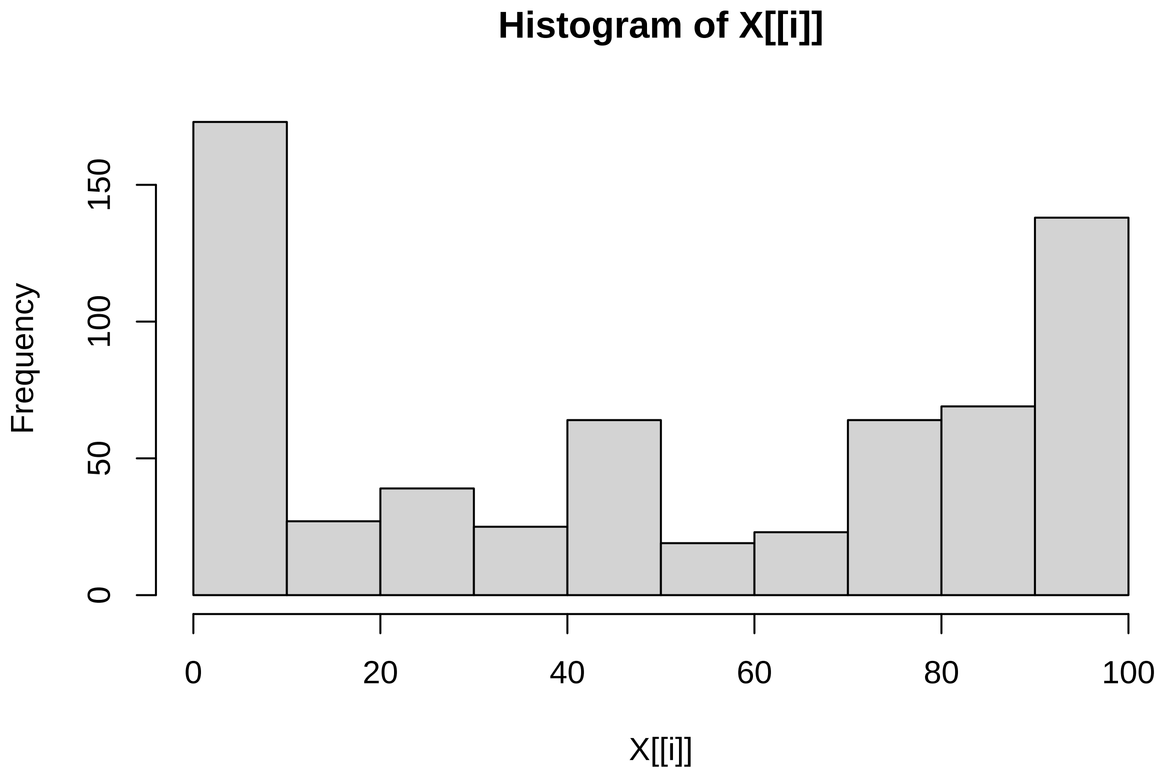
```
# assumption 3: calculate variance
tapply(usetech.df$usetech, usetech.df$partycat, var)
```

```
##      dem      rep      ind
## 1411.835 1408.963 1452.454
```

```
# assumption 4: normal dist
tapply(usetech.df$usetech, usetech.df$partycat, hist)
```




Histogram of X[[i]]
Frequency
0
20
40
60
0
20
40
60
80
100
X[[i]]



(g) Formally test whether there is a statistically significant difference in `usetech` between the two groups in `grass`. Provide a visual to support your results.

We're conducting a Two Sample t-test. The hypotheses are:

Null hypothesis: The mean of `usetech` is the same for those who support and oppose legalization of marijuana.

Alternative hypothesis: The mean of `usetech` is different for those who support and oppose legalization of marijuana.

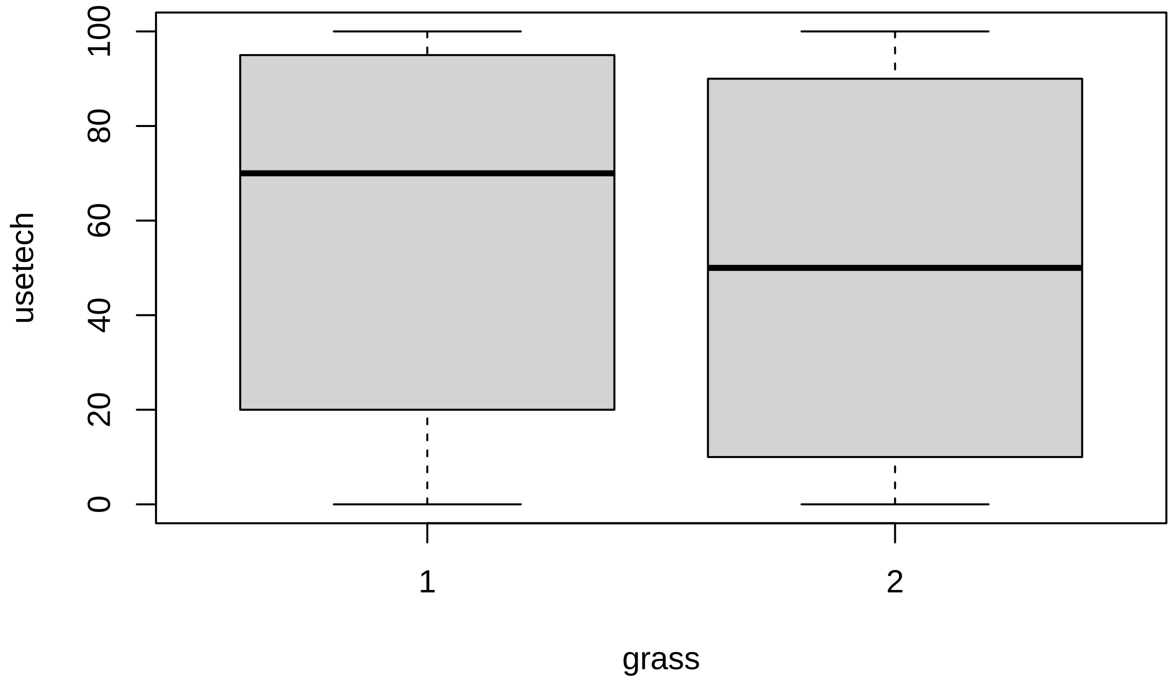
```
# subset data
techgrass.df <- gss18[!is.na(gss18$grass) & !is.na(gss18$usetech),]

# conduct t test
techgrass.test <- t.test(usetech~grass, data = techgrass.df)

# result
techgrass.test
```

```
##
##  Welch Two Sample t-test
##
## data:  usetech by grass
## t = 2.5268, df = 460.41, p-value = 0.01184
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##   1.606398 12.846878
## sample estimates:
## mean in group 1 mean in group 2
##        58.10403        50.87739
```

```
# visual
boxplot(usetech~grass, data = techgrass.df)
```



Conducting the `t.test` above gives us the test statistic of 2.53 (df = 460.41). The p-value is 0.01. Because the p-value is much smaller than our α level, we reject the null hypothesis and conclude that those who support legalization of marijuana tend to use technology more often at work.