AMCS 210 - Applied Statistics and Data Analysis
Hernando Catequista Ombao
Applied Mathematics and Computational Sciences Program
Computer, Electrical and Mathematical Sciences & Engineering (CEMSE) Division
King Abdullah University of Science and Technology (KAUST)

HOMEWORK
2

Henrique Aparecido Laureano

Fall Semester 2017

Contents

Exercise 1

A politician running for office is interested in estimating the proportion among legal residents of the state of California who support the DREAMER's act (which we denote as π). An organization was instructed to conduct a poll with the constraint that the estimator should be within a tolerance limit of 0.05.

Suppose that the true population proportion of supporters is $\pi = 0.65$. The organization is now at a planning stage to determine the desired sample size so that there is only a slim chance of 10% that an estimator (sample proportion) falls outside of the designated tolerance limit (i.e., it falls outside of 0.60 and 0.70).

(a)

Conduct a simulation study to determine if a sample size of n = 200 respondents will be sufficient to meet the target? As a guide to approaching this problem: (i.) simulate at least B = 10.000 samples each of size n = 200 from a binomial distribution; (ii.) for each sample record the sample proportion; (iii.) determine the proportion (out B = 10.000 samples) of sample proportion that fall within the designated tolerance limit.

Solution:

```
# <code r> ============================================================================ #
## defining some variables
n <- 200 # sample size
B = 1e4 # number of replications
p <- .65 # true proportion
tol <- .05 # tolerance level
in_inter <- numeric(B) # creating empth object of size B

## simulation study code
for (i in 1:B){
  samp <- rbinom(n, 1, p) # simulating samples from a binomal dist.
  prop <- mean(samp) # sample proportion
  abs_samp <- abs(prop - p) # difference to the tolerance level
  in_inter[i] <- ifelse(abs_samp <= tol, "in", "out") # if is in or out
}
# proportion of sample proportion that fall within (or out) the tol. level
prop.table(table(in_inter))
# </code r> ============================================================================ #
```

```
in_inter
    in    out
0.8633 0.1367
```

The proportion is of 15%, bigger than the desired proportion of 10%.

(b)

If the sample size of n = 200 does not satisfy the requirement above continue with procedure of finding the smallest sample size that would satisfy the tolerance limit.

Solution:

```
# <code r> ============================================================================ #
final_prop <- prop.table(table(in_inter)) ; prop_out <- final_prop[[2]]
while (final_prop[[2]] > .1){
  for (i in 1:B){
    samp <- rbinom(n + 1, 1, p)
    prop <- mean(samp) ; abs_samp <- abs(prop - p)
    in_inter[i] <- ifelse(abs_samp <= tol, "in", "out")}
  final_prop <- prop.table(table(in_inter))
  prop_out[length(prop_out)+1] <- final_prop[[2]] ; n <- n + 1}
library(latticeExtra)
xyplot(prop_out ~ 200:n, type = c("h", "p"), pch = 16
       , xlab = "Sample size", ylab = "Tolerance limit"
       , main = paste0(
         "Smallest sample size: ", n, " (tolerance limit: ", min(prop_out), ")")
       , scales = list(x = list(at = c(200, n)))
       , panel = function(...){
         panel.abline(h = .1, col = 2, lty = 2)
         panel.xyplot(...)
         panel.segments(n, 0, n, min(prop_out), col = 2, lwd = 2)
         panel.points(n, min(prop_out), col = 2, pch = 16)})
# </code r> ============================================================================ #
```

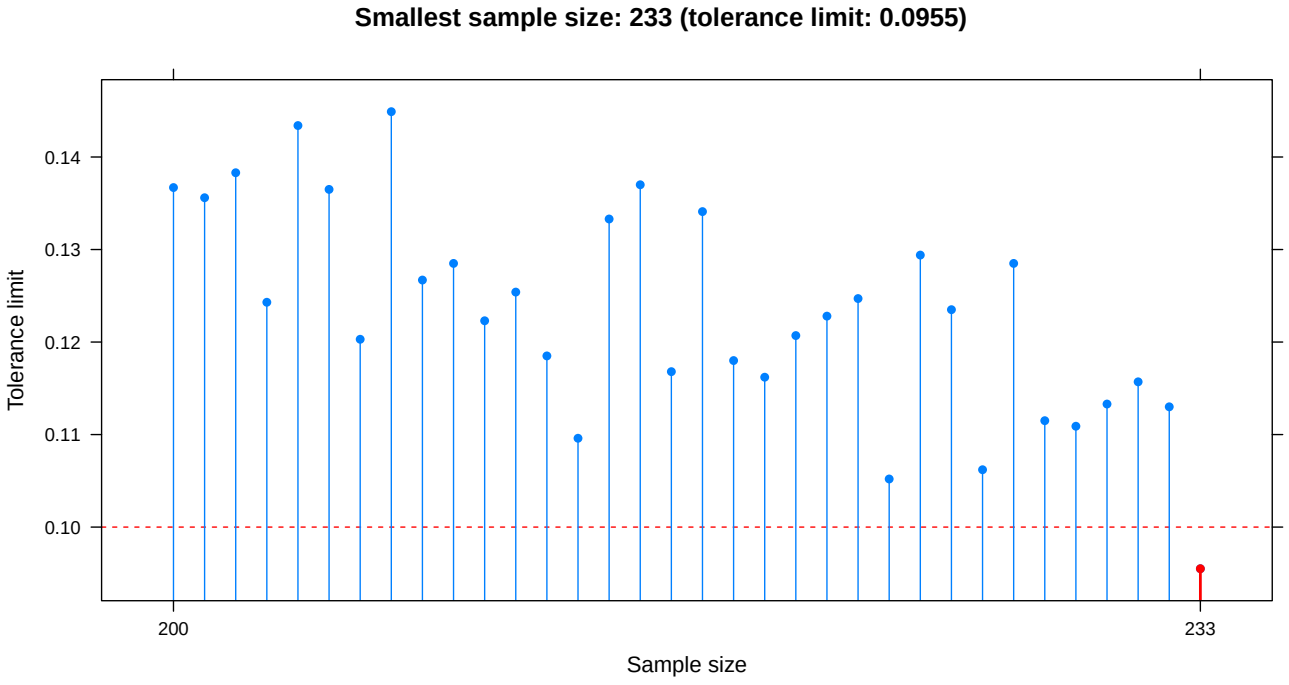

Figure 1: Sample sizes vs. tolerance limits.

Exercise 2

Exercises on calculating binomial probabilities. Let X be a binomial random variable with $n = 5$ and probability parameter $\pi = 0.3$. Think of this as counting the number of heads obtained by tossing $n = 5$ identical coins each of which has the probability of $\pi = 0.3$ that it lands in heads.

(a)

Enumerate the elements of the sample space of X.

Solution:

$$\Omega = \{0, 1, 2, 3, 4, 5\}.$$

(b)

Obtain the table of probability mass function, i.e., compute the probabilities $\mathbb{P}(X = x)$ for each x in the sample space.

Solution:

$$\mathbb{P}(X = x) = \binom{n}{x}\pi^x(1-\pi)^{n-x} = \binom{5}{x}0.3^x(0.7)^{5-x}.$$

x	$\mathbb{P}(X = x)$
0	0.16807
1	0.36015
2	0.3087
3	0.1323
4	0.02835
5	0.00243

(c)

Compute $\mathbb{P}(X \geq 4)$.

Solution:

$$\mathbb{P}(X \geq 4) = \mathbb{P}(X = 4) + \mathbb{P}(X = 5) = 0.02835 + 0.00243$$

$$\boxed{\mathbb{P}(X \geq 4) = 0.03078.}$$

(d)

Compute $\mathbb{P}(X < 3)$.

Solution:

$$\mathbb{P}(X < 3) = 1 - \mathbb{P}(X \geq 3) = 1 - [\mathbb{P}(X = 3) + \mathbb{P}(X \geq 4)] = 1 - [0.1323 + 0.03078] = 1 - 0.16308$$

$$\boxed{\mathbb{P}(X < 3) = 0.83692.}$$

Exercise 3

Exercises on calculating probabilities for the normal distribution. Some of these require the use of the standard normal probability table which has been uploaded in the subfolder "Admin" in the class dropbox.

(a)

Suppose that the systolic blood pressure (SBP) among white teens follows a normal distribution with mean $\mu = 120$ units (mm Hg) and standard deviation $\sigma = 4$ units.

(i.) What are the first and third quartile SBP values in this population?

Solution:

First quartile: Q_1, third quartile: Q_3

$$Z = \frac{Q_1 - \mu}{\sigma} = \frac{Q_1 - 120}{4}, \quad Z = \frac{Q_3 - \mu}{\sigma} = \frac{Q_3 - 120}{4}$$

The Z value that correspond to 0.25 (first quartile) is -0.67 (using standard normal probability table), and to 0.75 (third quartile) is 0.67. So,

$$Q_1 = Z \cdot \sigma + \mu = -0.67 \cdot 4 + 120, \quad Q_3 = 0.67 \cdot 4 + 120. \quad \boxed{Q_1 = 117.32, \quad Q_3 = 122.68.}$$

(ii.) What is the 90-th percentile SBP value?

Solution:

The Z value that correspond to 0.9 (90-th percentile, perc_{90}) is 1.28 (using standard normal probability table). So,

$$1.28 = \frac{\text{perc}_{90} - 120}{4}, \quad \text{perc}_{90} = 1.28 \cdot 4 + 120, \qquad \boxed{\text{perc}_{90} = 125.12.}$$

(iii.) Use the standard normal table to calculate the proportion of teens with SBP between 112 and 126 units.

Solution:

X = proportion of teens with some quantite of SBP (in units).

$$\mathbb{P}(112 \leq X \leq 126) = \mathbb{P}\bigg(\frac{112-120}{4} \leq \frac{X-120}{4} \leq \frac{126-120}{4}\bigg) = \mathbb{P}\bigg(-2 \leq Z \leq \frac{3}{2}\bigg)$$
$$= \mathbb{P}(Z \leq 1.5) - \mathbb{P}(Z \leq -2) = 0.9332 - 0.0228$$

$$\boxed{\mathbb{P}(112 \leq X \leq 126) = 0.9104.}$$

(iv.) A teen was randomly selected from the population. What is the probability that this teen has SBP that is greater than 130 units?

Solution:

$$\mathbb{P}(X > 130) = 1 - \mathbb{P}(X \leq 130) = 1 - \mathbb{P}\bigg(\frac{X-120}{4} \leq \frac{130-120}{4}\bigg) = 1 - \mathbb{P}(Z \leq 2.5) = 1 - 0.9938$$

$$\boxed{\mathbb{P}(X > 130) = 0.0062.}$$

(v.) An investigator randomly selected 5 teens. What is the probability that none of them has SBP tha exceeds 130 units?

Solution:

Y = number of teens with SBP that exceeds 130 units.

$$\mathbb{P}(Y = 0) = \binom{5}{0} 0.0062^0 (1 - 0.0062)^5, \qquad \boxed{\mathbb{P}(Y = 0) = 0.969382.}$$

(vi.) Again on the 5 teens: what is the probability that exactly 1 has SBP that exceeds 130 units.

Solution:

$$\mathbb{P}(Y = 1) = \binom{5}{1} 0.0062^1 (1 - 0.0062)^4, \qquad \boxed{\mathbb{P}(Y = 1) = 0.0302383.}$$

(b)

A graduate program will admit only students who score on the top 90-th percentile in an international examination. There is some complication in the admissions process because the grading scheme for the examination changed in 2016. Prior to 2016, the scores ranged from 0 to 800. From 2016 onwards the scores were recalibrated so

that they range from 0 to 150. Suppose that the pre-2016 scores followed a normal distribution with mean 600 and variance of 50. From 2016 onwards the distribution remained normal but the mean is 120 and the variance is 15. One applicant who took the test in 2015 scored 700 (out of 800) and another who took the test in 2017 scored 130 (out of 200). Check if either of these applicants should be automatically denied admission based on the examination scores.

Solution:

We have two Normal's, we can call $X \sim \mathrm{N}(600, \sqrt{50})$ and $Y \sim \mathrm{N}(120, \sqrt{15})$. In a Standard Normal, $\mathrm{N}(0, 1)$, the Z value that correspond to the 90-th percentile, perc_{90}, is 1.28.

For each Normal distributions the 90-th percentiles are:

$$\mathrm{perc}_{90} = 1.28 \cdot \sqrt{50} + 600 = 609.0509668, \quad \mathrm{perc}_{90} = 1.28 \cdot \sqrt{15} + 120 = 124.9574187.$$

Based on the examination scores both applications shouldn't be automatically denied, because both scores are on the top 90-th percentile in the international examination.

Exercise 4

Exercises on calculating binomial probabilities when the sample size n is large. A center is tasked with understanding the attitude of young college-age adults towards socialist ideas. Suppose that the true proportion in this population that has a positive attitude towards socialism is $\pi = 0.65$. A random sample of $n = 1.000$ students were selected.

(a)

Compute the probability that at least 800 of the students in a sample of size $n = 1000$ will have a positive view of socialism.

Solution:

$X =$ number of students with positive view of socialism.

$$\begin{aligned}
\mathbb{P}(X \geq 800) &= \mathbb{P}\left(\frac{X - n\pi}{\sqrt{n\pi(1-\pi)}} \geq \frac{800 - n\pi}{\sqrt{n\pi(1-\pi)}}\right) = \mathbb{P}\left(\frac{X - 1000 \cdot 0.65}{\sqrt{1000 \cdot 0.65 \cdot 0.35}} \geq \frac{800 - 1000 \cdot 0.65}{\sqrt{1000 \cdot 0.65 \cdot 0.35}}\right) \\
&= \mathbb{P}\left(\frac{X - 650}{15.0831031} \geq 9.9449032\right) = \mathbb{P}(Z \geq 9.9449032) = 1 - \mathbb{P}(Z < 9.9449032) \\
&= 1- \approx 1
\end{aligned}$$

$$\boxed{\mathbb{P}(X \geq 800) \approx 0.}$$

(b)

What is the probability that between 500 and 800 students (in a sample of $n = 1000$) will have a positive view of socialism.

Solution:

$$\begin{aligned}\mathbb{P}(500 \leq X \leq 800) &= \mathbb{P}\left(\frac{500 - 1000 \cdot 0.65}{\sqrt{1000 \cdot 0.65 \cdot 0.35}} \leq \frac{X - 1000 \cdot 0.65}{\sqrt{1000 \cdot 0.65 \cdot 0.35}} \leq \frac{800 - 1000 \cdot 0.65}{\sqrt{1000 \cdot 0.65 \cdot 0.35}}\right) \\ &= \mathbb{P}\left(-9.9449032 \leq \frac{X - 650}{15.0831031} \leq 9.9449032\right) \\ &= \mathbb{P}(Z \leq 9.9449032) - \mathbb{P}(Z \leq -9.9449032) = \approx 1 - \approx 0\end{aligned}$$

$$\boxed{\mathbb{P}(500 \leq X \leq 800) \approx 1.}$$

Exercise 5

Two populations of patients are treated with two different drugs for diabetes for a period of 6 months. The first population is given drug A while the second is given drug B. At the end of 6 months, a researcher is interested in comparing the proportions of patients in each of the two populations who experience a significant drop in their blood pressure. Denote these population proportion parameters to be π_A and π_B respectively.

A random sample of size $n_A = 100$ was taken from population A and of these $Y_A = 70$ had significant drops in blood pressure. Moreover a random sample of size $n_B = 500$ was taken from population B and of these $Y_B = 300$ had significant drops in blood pressure.

(a)

Compute the proportion of patients from each sample who had significant drops in their blood pressure. Denote these to be $\hat{\pi}_A$ and $\hat{\pi}_B$.

Solution:

$$\hat{\pi}_A = \frac{Y_A}{n_A} = \frac{70}{100} = 0.7, \quad \hat{\pi}_B = \frac{Y_B}{n_B} = \frac{300}{500} = 0.6. \qquad \boxed{\hat{\pi}_A = 0.7, \quad \hat{\pi}_B = 0.6.}$$

(b)

Conduct a test of hypothesis $H_0 : \pi_A = \pi_B$ using the test statistic

$$T_1 = \frac{\hat{\pi}_A - \hat{\pi}_B}{\sqrt{\frac{\hat{\pi}_A(1-\hat{\pi}_A)}{n_A} + \frac{\hat{\pi}_B(1-\hat{\pi}_B)}{n_B}}}.$$

To conduct this test you need to

(i.) compute the observed test statistic $T_{1,\text{obs}}$.

Solution:

$$T_{1,\text{obs}} = \frac{0.7 - 0.6}{\sqrt{\frac{0.7 \cdot 0.3}{100} + \frac{0.6 \cdot 0.4}{500}}} = \frac{0.1}{\sqrt{0.0021 + 4.8 \times 10^{-4}}} = \frac{0.1}{0.0507937}, \quad \boxed{T_{1,\text{obs}} = 1.9687481}$$

(ii.) state the reference distribution of T_1.

Solution:

The construction of the reference distribution is over H_0, with means that we considerer that don't exist difference between the samples (the sample proportions). For this reason in the code below we put the samples together and randomly selected members and assume that this members are from the population A or B, because over the null hyphotesis they don't differ.

```
# <code r> ============================================================================ #
samp_a <- rbinom(100, 1, .7) # sampling from population A
samp_b <- rbinom(500, 1, .6) # sampling from population B
samp <- c(samp_a, samp_b) # putting the samples together
B <- 1e3 # number of replication
# computing the test statistic
t1 <- function(prop_a, n_a, prop_b, n_b){
  (prop_a - prop_b)/sqrt((prop_a*(1-prop_a))/n_a + (prop_b*(1-prop_b))/n_b)}
store <- numeric(B) # empth vector to store the test statistics
for (i in 1:B){ # generating the reference distribution
  # choosing randomly the members of population A
  a_i <- sample(length(samp), size = length(samp_a), replace = FALSE)
  samp_ai <- samp[a_i] ; n_ai <- length(samp_ai)
  # the rest go to population B
  samp_bi <- samp[-a_i] ; n_bi <- length(samp_bi)
  # computing proportions
  prop_ai <- mean(samp_ai) ; prop_bi <- mean(samp_bi)
  # computing test statistic
  t1_i <- t1(prop_a = prop_ai, n_a = n_ai, prop_b = prop_bi, n_b = n_bi)
  store[i] <- t1_i # storing test statistic
}
# histogram
hist(store
     , xlab = "Proportion difference", main = "Reference distribution"
     , las = 1, xlim = range(store)
     , col = "#0080ff", border = "orange")
# </code r> ============================================================================ #
```

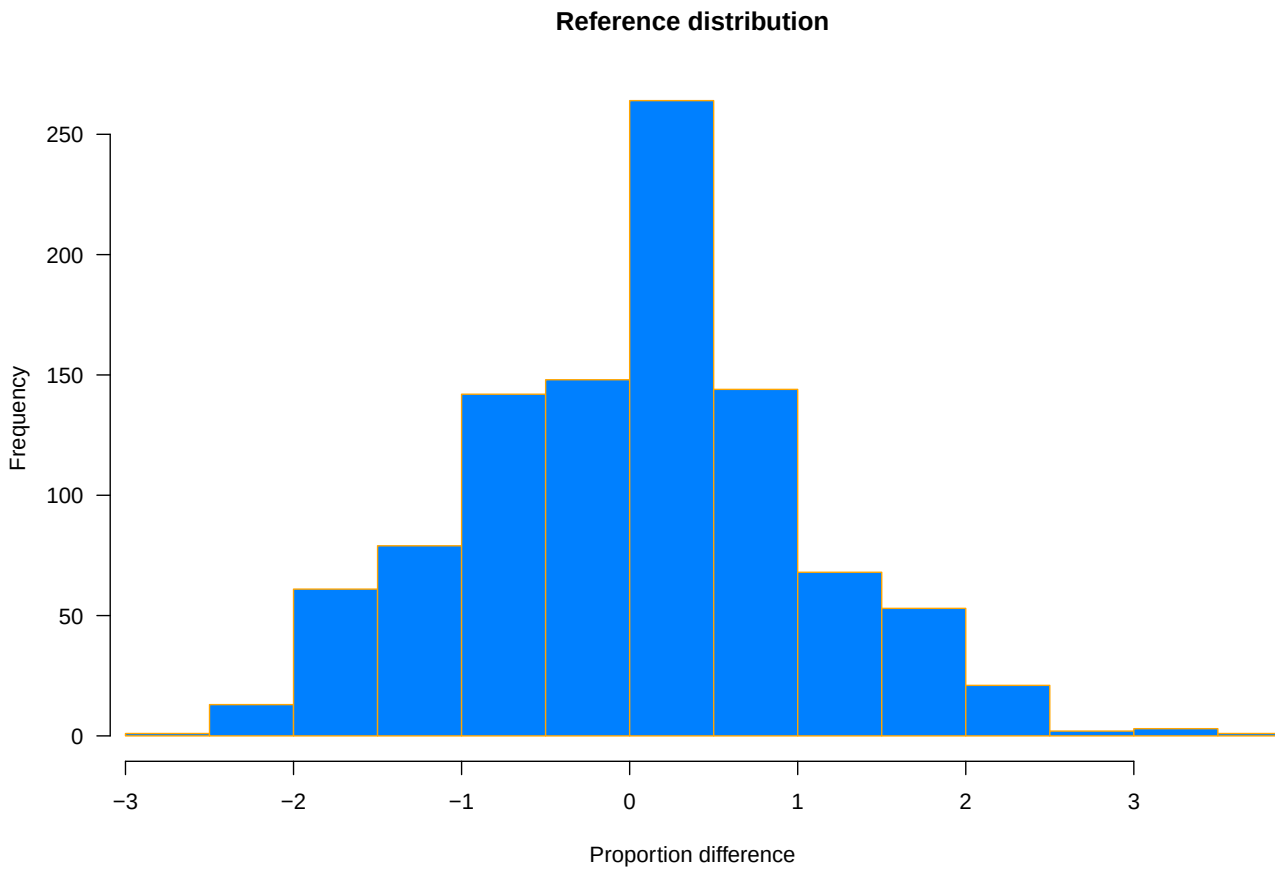

Figure 2: Histogram of the reference distribution (1000 relications).

(iii.) identify the acceptance and rejection regions for a given probability of type I error.

Solution:

The type I error rate or significance level is the probability of rejecting H_0 given that it's true. We will use an type I error rate of 5%. It means, the 95-th percentile of the reference distribution. We can see this region at red in the histogram below. To the left of the 95-th percentile we have the acceptance region, in the right we have the rejection region.

```
# <code r> ============================================================================ #
hist(store
     , xlab = "Proportion difference"
     , main = "Reference distribution"
     , las = 1
     , xlim = range(store)
     , col = "#0080ff"
     , border = "orange")
abline(v = quantile(store, .95), col = 2, lwd = 2)
# </code r> ============================================================================ #
```

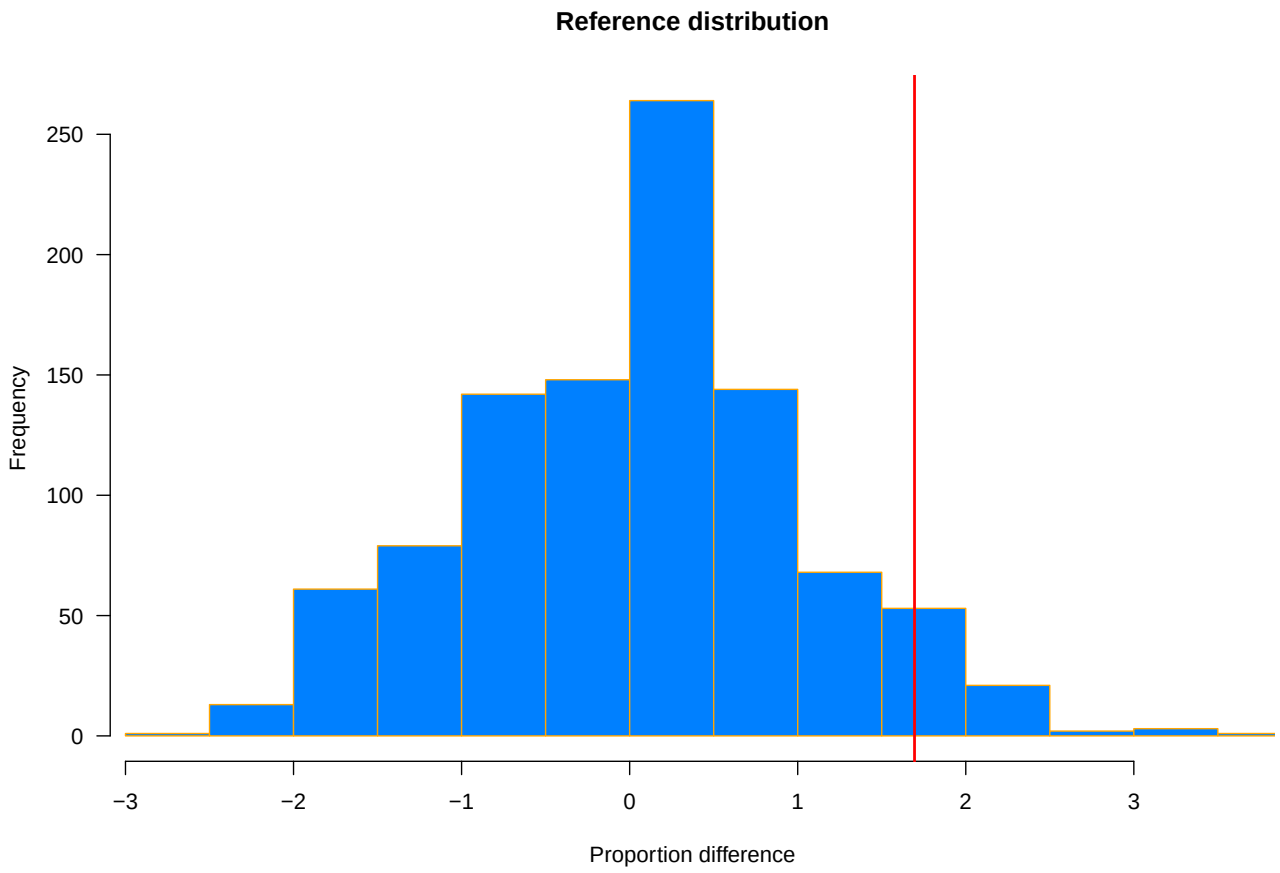

Figure 3: Histogram of the reference distribution. To the left of the red line we have the acceptance region, to the right of the red line we have the rejection region.

(iv.) decision and conclusion.

Solution:

In the histogram below we can see that the observed test statistic is in the rejection region, to the right of the 95-th percentile. This means that we reject the H_0, which means that we don't have statistical evidence that the sample proportions are equal (statistically equal).

```
# <code r> ============================================================================ #
hist(store
     , xlab = "Proportion difference"
     , main = "Reference distribution"
     , las = 1
     , xlim = range(store)
     , col = "#0080ff"
     , border = "orange")
abline(v = quantile(store, .95), col = 2, lwd = 2)
abline(v = t1(prop_a = .7, n_a = 100, prop_b = .6, n_b = 500), col = 3, lwd = 3)
# </code r> ============================================================================ #
```

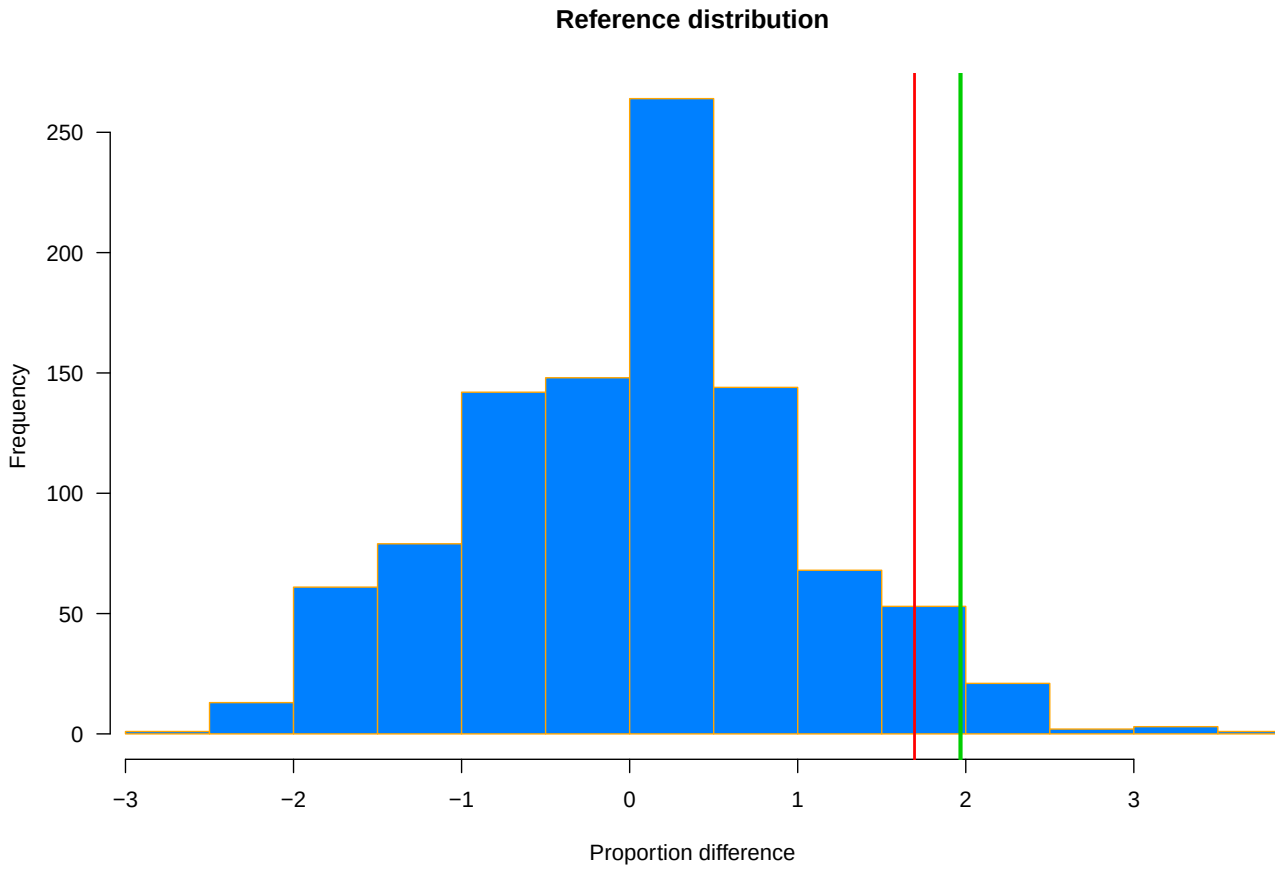

Figure 4: Histogram of the reference distribution. In red we have the 95-th percentile and in green the observed test statistic value.

(c)

Construct the table of (estimated) <u>expected</u> counts of patients with significant drop under the null hypothesis H_0.
Table of expected counts.

	Expected count with significant drop	Expected count with no significant drop
Popn A	$\hat{Y}_A$	$n_A - \hat{Y}_A$
Popn B	$\hat{Y}_B$	$n_B - \hat{Y}_B$

<u>Solution:</u>

Construct the table of <u>observed</u> counts.
Table of observed counts.

	Observed count with significant drop	Observed count with no significant drop
Popn A	Y_A	$n_A - Y_A$
Popn B	Y_B	$n_B - Y_B$

Solution:

Conduct a test of the null hypothesis $H_0 : \pi_A = \pi_B$ using the test statistic

$$T_2 = \frac{(Y_A - \hat{Y}_A)^2}{\hat{Y}_A} + \frac{(n_A - Y_A - (n_A - \hat{Y}_A))^2}{(n_A - \hat{Y}_A)} + \frac{(Y_B - \hat{Y}_B)^2}{\hat{Y}_B} + \frac{(n_B - Y_B - (n_B - \hat{Y}_B))^2}{(n_B - \hat{Y}_B)}.$$

To conduct this test you need to

Solution:

(i.) compute the observed test statistic $T_{2,obs}$.

Solution:

(ii.) state the reference distribution of T_2.

Solution:

(iii.) identify the acceptance and rejection regions for a given probability of type I error.

Solution:

(iv.) decision and conclusion.

Solution:

(d)

Conceptual question: suppose you "do not reject" the null hypothesis H_0. Does this mean that H_0 is "correct"?

Solution:

(e)

Another conceptual question: suppose you "reject" the null hypothesis H_0. Does this mean that H_0 is "wrong"?

Solution:

■