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(1, n = n, p = 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
0.8622 0.1378
```

The proportion is of 14%, 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(1, n = n + 1, p = 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> ============================================================================ #
```

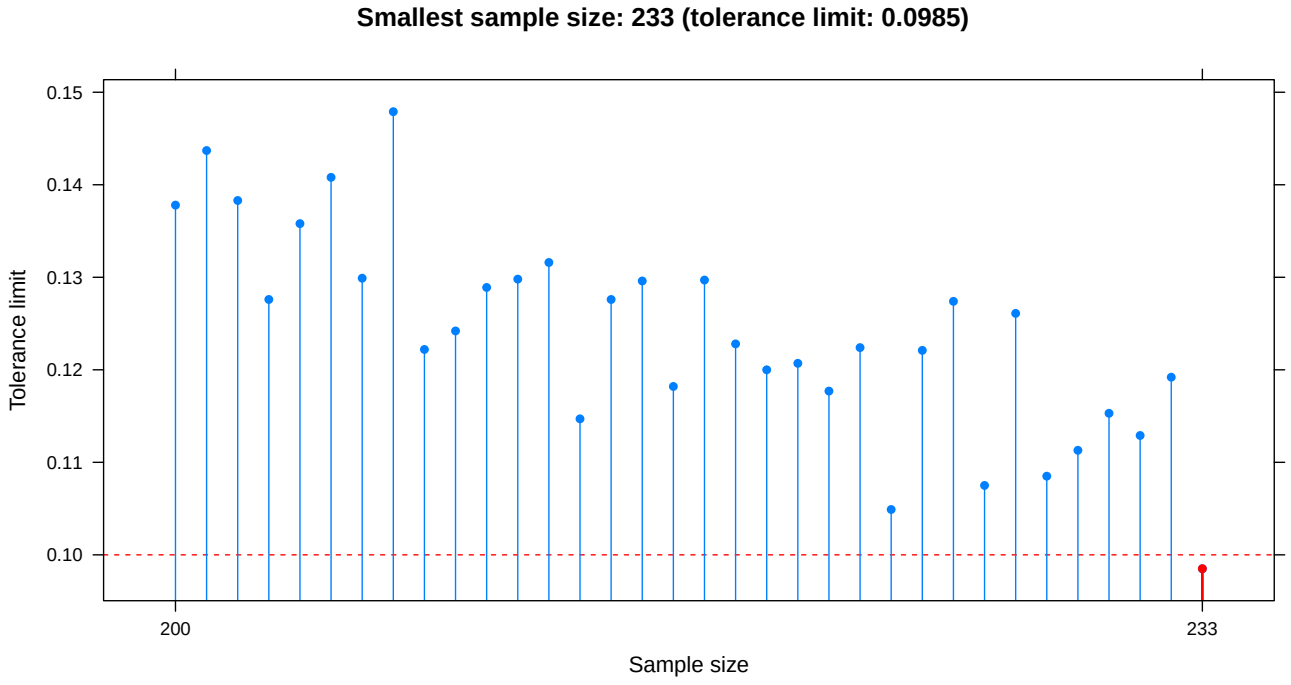

Figure 1: Sample sizes vs. tolerance limits.

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

Consider this dataset derived from the Framingham Heart Study where one of the goals was to study possible links between high systolic blood pressure (SBP) and coronary heart disease (CHD). Participants with SBP $\geq$ 165 mm Hg were put in the "high" SBP category.

	CHD	No CHD
High SBP	144	62
Normal SBP	120	419

(a)

Marginal proportions. **From the table, what is the proportion of participants who had CHD? What is the proportion of participants who had high SBP?**

Solution:

CHD = 144 + 120 = 264, NoCHD = 62 + 419 = 481, n = CHD + NoCHD = 745

Proportion of participants who had CHD: CHD/n = 264/745 = 0.3543624 (35%).

HiSBP = 144 + 62 = 206, NorSBP = 120 + 419 = 539,
n = CHD + NoCHD = HiSBP + NorSBP = 745

Proportion of participants who had high SBP: HiSBP/n = 206/745 = 0.2765101 (28%).

(b)

Conditional proportions. **Among the participants with high SBP, what is the proportion of those who also have CHD? Among the participants with normal SBP, what is the proportion of those who also have CHD? Does this provide some evidence of a link between elevated systolic blood pressure and coronary heart disease? Can one now claim that elevated systolic blood pressure causes coronary heart disease?**

Solution:

Among the participants with high SBP, the proportion of those who also have CHD:
CHD/HiSBP = 144/206 = 0.6990291 (70%).

Among the participants with normal SBP, the proportion of those who also have CHD:
CHD/NorSBP = 120/539 = 0.2226345 (22%).

Does this provide some evidence of a link between elevated systolic blood pressure and coronary heart disease? Yes. Because in the patients with normal SBP the proportion with CHD is 22 percent, almost 1/5, while in the patients with high SBP the proportion with CHD is 70 percent, much higher.

Can one now claim that elevated systolic blood pressure causes coronary heart disease? No. With this descriptive analysis we can see a possible correlation between high SBP and CHD, but this not implies in causality. we don't have enough information to claim this type of result. Correlation is different form causality.

(c)

Odds and odds ratios. **Compute the odds of having CHD among the high SBP group. Compute the odds of having CHD among the normal SBP group. Compute the odds ratio of having CHD for the high SBP vs normal SBP groups.**

Solution:

Odds of having CHD among the high SBP group:

$$\frac{144/206}{1-144/206} = 2.322581.$$

Odds of having CHD among the normal SBP group:

$$\frac{120/539}{1-120/539} = 0.2863962.$$

Odds ratio of having CHD for the high SBP vs normal SBP groups:

$$\frac{2.322581}{0.286396} = 8.109678.$$

The odds of having CHD for the high SBP groups is 8.11 times of the odds of CHD for the normal SBP group.

Exercise 6

The beetle mortality data. Groups of beetles were exposed to varying doses of toxins and the number of deaths (`ytotal`) out of the total exposure (`ntotal`) were recorded. Enter the variables in R:

$$\begin{aligned} \texttt{dose} &= \text{c}(1.69, 1.72, 1.75, 1.78, 1.81, 1.84, 1.86, 1.88) \\ \texttt{ntotal} &= \text{c}(59, 60, 62, 56, 63, 59, 62, 60) \\ \texttt{ytotal} &= \text{c}(6, 13, 18, 28, 52, 53, 61, 60) \end{aligned}$$

```
# <code r> ============================================================================ #
dose <- c(1.69, 1.72, 1.75, 1.78, 1.81, 1.84, 1.86, 1.88)
ntotal <- c(59, 60, 62, 56, 63, 59, 62, 60)
ytotal <- c(6, 13, 18, 28, 52, 53, 61, 60)
# </code r> ============================================================================ #
```

(a)

Calculate the proportion of deaths for each dose-group.

Solution:

```
# <code r> ============================================================================ #
rbind(dose, "death proportion" = round(ytotal/ntotal, 2))
# </code r> ============================================================================ #
```

```
                 [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
dose             1.69 1.72 1.75 1.78 1.81 1.84 1.86 1.88
death proportion 0.10 0.22 0.29 0.50 0.83 0.90 0.98 1.00
```

(b)

Plot the proportion of deaths against dose. Describe the trend: is it increasing/ decreasing, is there an asymptote feature?

Solution:

```
# <code r> ============================================================================ #
xyplot(ytotal/ntotal ~ dose, type = c("p", "l")
       , pch = 19, lwd = 1.5, xlab = "Dose", ylab = "Proportion of deaths"
       , scales = list(x = list(at = dose))
       , panel = function(...){
         panel.abline(v = dose, h = seq(.2, 1, .2), col = "gray70")
         panel.xyplot(...)})
# </code r> ============================================================================ #
```

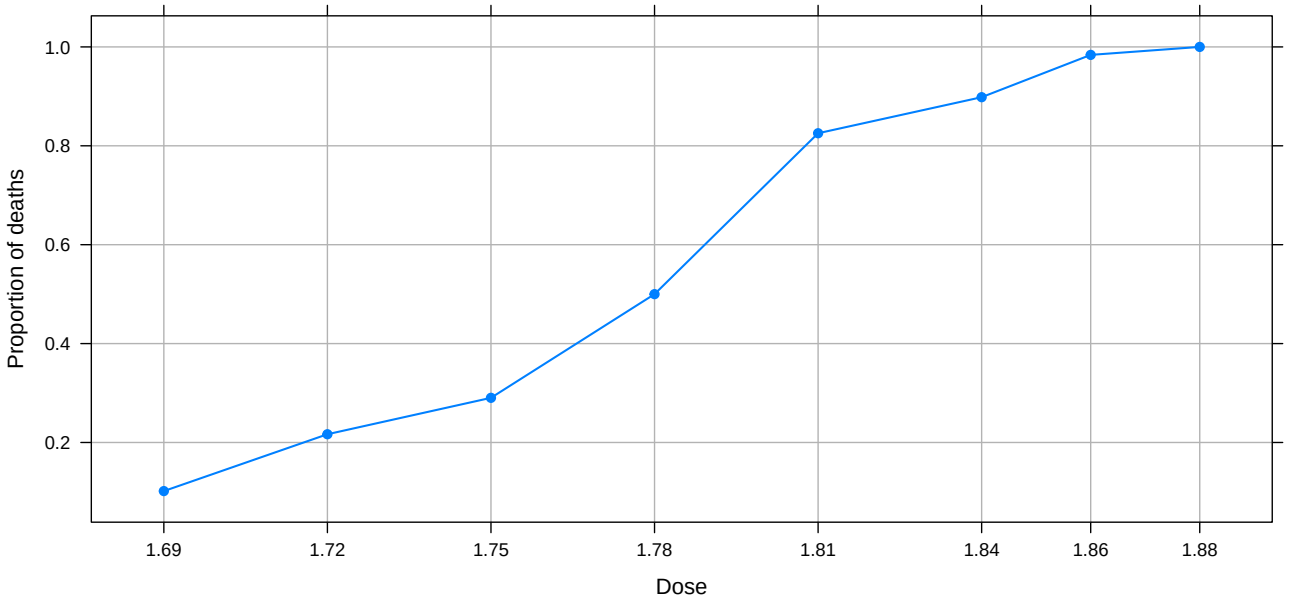

Figure 2: Proportion of deaths against dose.

The trend is increasing. For small doses the proportions are very close to zero, with the biggest doses the proportions are very close to one. With the last dose we have a proportion equal to one, the maximum possible. Whats means that with the biggest dose all the beetles are dead.

■