FinalAssessment

Louise Branzuela

2022-06-15

Setting seed for random sampling in simulations

```
set.seed(20121865)
```

Question 1

Probability the temperature is at most 21C

```
pnorm(21, mean=25, sd=4)
```

```
## [1] 0.1586553
```

The probability that in a random day the temperature is at most 21 degrees Celsius is 0.1586553.

Verifying at most 21C probability through simulation

```
a <- rnorm(1e6, mean=25, sd=4)
a <- table(a)
sum(a[names(a)<=21])/1e6
```

```
## [1] 0.158795
```

Probability the temperature is at least 28C

```
1-pnorm(28, mean=25, sd=4)
```

```
## [1] 0.2266274
```

The probability that in a random day the temperature is at least 28 degrees Celsius is 0.2266274.

Verifying that at least 28C probability through simulation

```
sum(a[names(a)>=28])/1e6
```

```
## [1] 0.226972
```

Probability that temperature is between 28 and 32 degrees

```
pnorm(32, mean=25, sd=4)-pnorm(28, mean=25, sd=4)
```

```
## [1] 0.1865682
```

The probability that in a random day the temperature is between 28 degrees and 32 degrees Celsius is 0.1865682

Verifying that between 28C and 32C probability through simulation

```
(sum(a[names(a)<=32]) - sum(a[names(a)<=28]))/1e6
```

```
## [1] 0.186964
```

Excellence Questions

We can say that the distribution for days that are greater than 30 degrees celsius and days that are below 30 degrees celsius follows a Bernoulli distribution.

Probability that the temperature is at least 30 degrees Celsius

```
1-pnorm(30, mean=25, sd=4)
```

```
## [1] 0.1056498
```

First we get the probability that in a random day the temperature is at least 30 degrees Celsius which is 0.1056498.

Probability that 5 days have temperatures greater than 30 degrees

```
dbinom(x=5, size=14, 0.1056498)
```

```
## [1] 0.009646637
```

The calculation for the probability that 5 days at random have temperatures greater than 30 degrees Celsius given that we chose 14 random days is 0.009646637.

Verifying the probability of 5 days having temperatures greater than 30 degrees

```
b <- table(rbinom(1e6, 14, 0.1056498))
sum(b[names(b)==5])/1e6
```

```
## [1] 0.009584
```

Expected days with temperature greater than 30C

```
dbinom(1, 14, 0.1056498)
```

```
## [1] 0.3464142
```

I would expect at most 10 days that have temperatures that exceed 30 degrees Celsius. Only having 1 day that has a temperature that is greater than 30 degrees has the highest probability of happening given that there are 14 days chosen at random.

Verification via simulation

```
sum(b[names(b)==1])/1e6
```

```
## [1] 0.346914
```

Question 2

Null Hypothesis

A sensible null hypothesis would be that there would be no difference in the probability of either the English cows or Scottish cows catching the disease. Therefore H0: P(English|Healthy) = P(Scottish|Healthy).

P-value and more extreme definition

P-value is the probability, if the null is true, of obtaining the observation or an observation more extreme. "More extreme" in this case would means that the p-value is larger
than the observed sample mean.

One-sided or Two-sided

In this case, we need a one-sided(right) test to see if the population mean is larger than the null mean, because we need to see if English cows are healthier than Scottish cows.

Fisher's test

```
cows <- matrix(c(8, 2, 1, 4),2,2)
dimnames(cows) <- list(
  cow=c("English","Scottish"),state=c("Healthy","Sick"))
cows
```

```
##           state
## cow        Healthy Sick
##   English        8    1
##   Scottish       2    4
```

```
fisher.test(cows, alternative="greater")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  cows
## p-value = 0.04695
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  1.031491      Inf
## sample estimates:
## odds ratio
##   12.49706
```

Since the p-value is less than 0.05, we reject the null hypothesis and prove that English cows are healthier than the Scottish cows.

Probability that an English cow is sick

```
1/9
```

```
## [1] 0.1111111
```

Probability that an Scottish cow is sick

```
4/6
```

```
## [1] 0.6666667
```

Using dhyper() to verify fisher's test

```
dhyper(8, 9, 6, 10)
```

```
## [1] 0.04495504
```

Question 3

```
temp <- c(17.1, 19.2, 15.2, 18.1, 15, 17.8, 17.8, 15.2, 17.4, 15.7, 17.8,
          16.9, 17.1, 17, 18, 15.9, 16.5, 17.3, 16.7, 15.9, 16.1, 19, 16)
fail <- c(28, 30, 27, 33, 27, 34, 31, 29, 29, 26, 33, 31, 27, 31, 32,
          28, 30, 33, 32, 26, 30, 34, 27)
```

Null Hypothesis

A sensible null hypothesis would be that there is no correlation between the temperature and the occurrence of failures.

P-value and More extreme definition

P-value is the probability, if the null is true, of obtaining the observation or an observation more extreme. "More extreme" in this case would means that the p-value is larger than the observed sample mean.

One-sided or Two-sided test

In this case, we need a two-sided test because we want to determine if there is any difference between the temp and fail.

Linear Regression

```
df <- data.frame(temp, fail)
df
```

```
##    temp fail
## 1  17.1   28
## 2  19.2   30
## 3  15.2   27
## 4  18.1   33
## 5  15.0   27
## 6  17.8   34
## 7  17.8   31
## 8  15.2   29
## 9  17.4   29
## 10 15.7   26
## 11 17.8   33
## 12 16.9   31
## 13 17.1   27
## 14 17.0   31
## 15 18.0   32
## 16 15.9   28
## 17 16.5   30
## 18 17.3   33
## 19 16.7   32
## 20 15.9   26
## 21 16.1   30
## 22 19.0   34
## 23 16.0   27
```

```
df.linearModel <- lm(fail ~ temp, df)
summary(df.linearModel)
```

```
##
## Call:
## lm(formula = fail ~ temp, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5300 -1.5985  0.3571  1.2724  2.6716
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.3364     5.8375   0.572 0.573696
## temp          1.5726     0.3446   4.563 0.000169 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.88 on 21 degrees of freedom
## Multiple R-squared:  0.4979, Adjusted R-squared:  0.4739
## F-statistic: 20.82 on 1 and 21 DF,  p-value: 0.0001692
```

Since the p-value is 0.0001692, we reject the null hypothesis. This means that, there is evidence that computer failures occur more often during hotter weather.

Question 4

```
y <- scan(text = "
37.3 37.5 37.2 38.0 37.6 37.2 37.6 37.7 37.7 37.4 37.9 37.6 37.3
36.8 37.8 37.5 37.5 37.8 37.7 37.7 37.8 37.7 37.5 36.9 37.7 37.5
37.4 37.0 37.3 37.6 37.9 37.5 37.6 37.5 37.1 37.4 37.4 37.5 37.8
37.7 37.4 37.4 37.7 37.7 37.3 37.3 37.6 37.7 37.5 37.8 37.6 37.3
37.6 37.2 37.9 38.1 37.4 37.2 37.7 37.4 38.2 37.5 37.7 37.5 37.3
37.5 36.9 37.9 37.5 38.1 37.6 37.3 37.7 37.2 37.1 37.6 37.4 37.5
37.5 37.3")
y.table <- table(y)
```

Successes

```
successes = sum(y.table[names(y.table)<=38])
successes
```

```
## [1] 77
```

Failures

```
failures = sum(y.table[names(y.table)>38])
failures
```

```
## [1] 3
```

Random Variable and Distribution

The random variable would be the probability that the medicine is effective. Since the data is basically a sequence of Bernoulli trials, the distribution of the random variable should be a beta distribution.

Prior

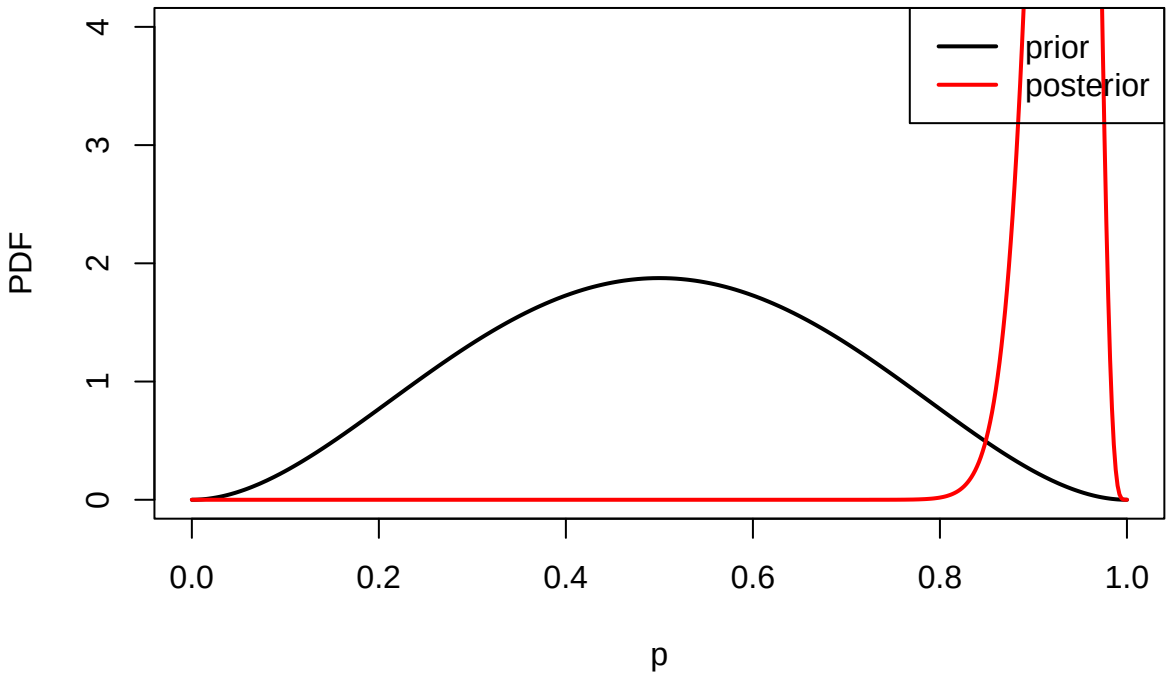
My prior would be B(3, 3), because assuming that we're in a skeptical position, we would believe that the medicine would not be effective half the time. I also chose it to show and represent uncertainties.

Posterior

The posterior would be B(3+77, 3+3)

Graph of both posterior and prior

```
p <- seq(from=0,to=1,len=500)
plot(p,p*0,ylim=c(0,4),ylab="PDF",type="n")
points(p,dbeta(p,3,3),lwd=2,type="l",col='black')
points(p,dbeta(p,3+77, 3+3),lwd=2,type="l",col='red')
legend("topright",lwd=2,col=c("black","red"),legend=c("prior","posterior"))
```



We can see that on the plot, my prior is wide which reflects our skeptical position in our beliefs that the medicine would only work half the time, but after our posterior, we see that the distribution is narrower, and more skewed to the left which shows us that the medicine is actually quite effective.