Tutorial 1 Solutions

Tutorial Problem 1

1.

The probability is **0.5**, because either true or false answers are equally likely to be correct.

2.

This is the probability of answering three questions correctly in a row: $0.5^3 = 1/8 =$ **0.125**.

3.

A reasonable simple test statistic could be the **number of contestant wins**: fewer wins is more extreme, so this is a logical way to quantify the extreme-ness of the unexpected result we achieved. We'll abbreviate this statistic as W. Its value for the result we observed is $\mathbf{W = 1}$.

(Note: we could also have used the number of contestant losses, or divide the number of wins/losses by the number of games to express the statistic as a proportion, or one could compare the number of wins to the number of expected wins, etc. These solutions will just use the number of wins.)

4.

The null hypothesis is that the probability of a contestant winning a *game* is P_{win}=1/8, as we calculated in part 2. (Alternatively, one could take a step further back and state that the probability of answering a question correctly is P_q=1/2, but of course one implies the other.)

5.

The most conservative alternative hypothesis would be that the probability is not as calculated above, i.e. P_{win} $\neq$ 1/8. This alternative hypothesis turns out to be somewhat difficult to test via the methods we have learned so far, and furthermore a result in which contestants do better than the "guessing score" isn't particularly surprising (it just implies contestants might actually know some answers). So we will make a more specific alternative hypothesis which is that the probability of winning a game is P_{win}<1/8.

6.

The probability of the most extreme possible result under the null hypothesis is the probability that every contestant loses the game, i.e. that for all 45 games, the result is a loss (W = 0). The probability of a loss is P_{loss} = 1-P_{win} = 1-1/8 = 7/8, so the probability of 45 consecutive losses is:

```r
P0win = (7/8)^45
P0win
```

```
## [1] 0.002456758
```

The next most extreme result, which is also the result actually obtained, is that 1 contestant wins and the remaining 44 contestants lose (W = 1). Any of the contestants (first, second, third… etc.) can be the winner. First calculate the probability that the very first contestant wins and then all the other contestants lose, which is $P_{win} \times P_{loss}^{(45-1)}$, or:

```
P_firstwin_otherslose = (1/8)*(7/8)^44
P_firstwin_otherslose
```

```
## [1] 0.0003509654
```

This is the same as the probability that the second contestant wins, and the same as the probability that the third contestant wins, etc. (all the way on up to the 45th contestant). These are all assumed to be independent outcomes, so simply add these all up (sum over all 45 contestants, i.e. multiply the probability above by 45) to get the probability of a single win during the season:

```
P1win = 45*(1/8)*(7/8)^44
P1win
```

```
## [1] 0.01579344
```

So the probability of an "as or more extreme" result (W=0 or W=1) is:

```
p = P0win + P1win
p
```

```
## [1] 0.0182502
```

Note that we are not treating anomalously *large* number of wins (45 wins, 44 wins, etc.) as "extreme" because we decided these were not "interesting" under our alternative hypothesis, although this decision could certainly be called into question as we might have been suspicious if all or even most candidates had won. (In statistical terms, our alternative hypothesis has led us to consider only the left "tail" of the probability distribution as interesting.)

Had we wanted to do a two-tailed test that treats unexpectedly large numbers of wins as "interesting", we wouldn't be able to test it yet as we have not yet covered two-tailed hypothesis tests involving asymmetric distributions. However, we we will learn how to treat these situations later in the module (Lecture 5 for continuous data, and Lecture 9 for discrete/binomial data of the type encountered here).

7.

Because the p-value is less than 0.05, we would rule out the null hypothesis and conclude in favour of the alternative hypothesis were this a blind experiment. However, a significant concern is that we have proposed our test *after* noticing/suspecting that something was amiss in the data ("a-posteriori statistics"). Our result may not be as significant as we think, because our calculation does not take into account the other potential "anomalous" results that could have been obtained in this data, or other data that we encountered in our daily lives (perhaps this is not the only game show we watch, etc.)

8.

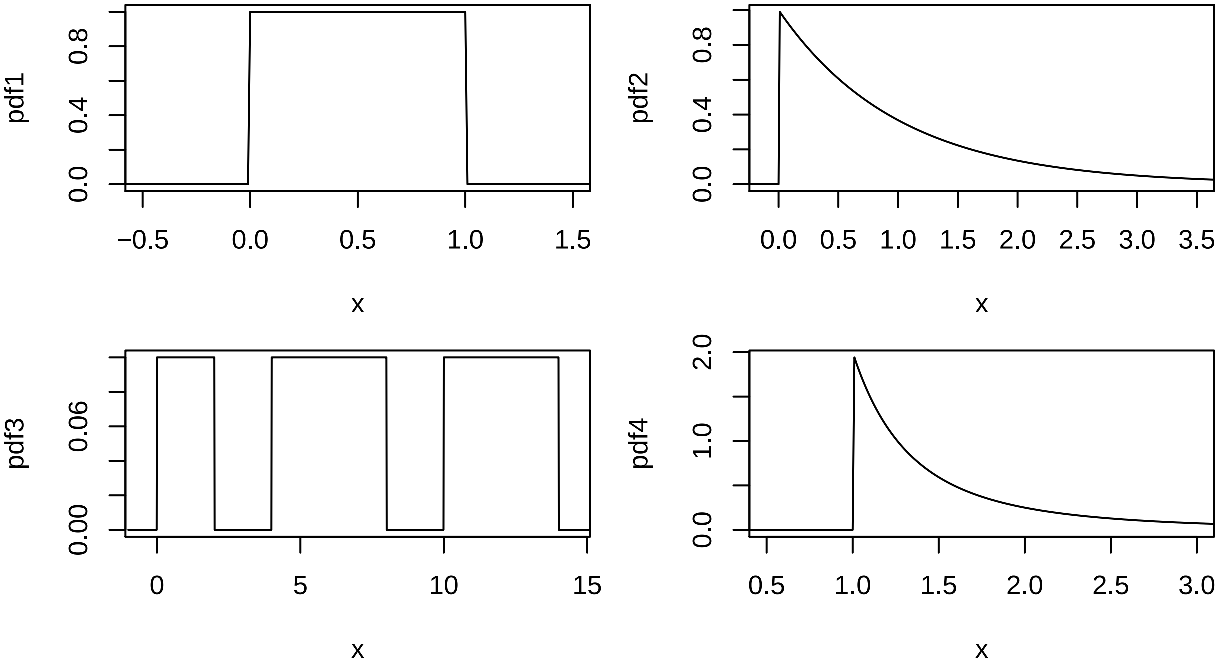
Since the second season has not happened yet, we can now declare our test and prediction in advance, eliminating the dependence on a-posteriori statistics. If a significant result is obtained again by this specific test, this would be much stronger evidence in favour of the alternative hypothesis. Of course, as with all statistical tests, some risk of Type I error would remain (a 5% risk, if we test at a significance level of α=0.05).

Tutorial Problem 2

1.

There are many ways to produce these plots; a fully general one would be to define each PDF as its own function and then fill it out using a loop, or vectorize it using R's special vectorization tools (which we will not cover in this module). The solution below takes a simpler approach, which is to use logical subscripts to calculate the values of each PDF segment by segment.

```
x = seq(-1,20,0.01)   # create a plotting X variable covering a domain appropriate
                      # for all the PDFs (-1 to 20) and with good resolution (0.01)

pdf1 = rep(0,length(x))      # create a plotting Y variable to fill in
pdf1[x >= 0 & x <= 1] = 1

pdf2 = rep(0,length(x))
pdf2[x > 0] = exp(-x[x > 0])

pdf3 = rep(0,length(x))
pdf3[x >= 0 & x <= 2] = 0.1
pdf3[x >= 4 & x <= 8] = 0.1
pdf3[x >= 10 & x <= 14] = 0.1

pdf4 = rep(0,length(x))
pdf4[x > 1] = 2*x[x>1]^(-3)

par(mfrow=c(2,2))
par(mar=c(4,4,1,1))
plot(x,pdf1, typ='l', xlim=c(-0.5,1.5))
plot(x,pdf2, typ='l', xlim=c(-0.1,3.5))
plot(x,pdf3, typ='l', xlim=c(-0.5,14.5))
plot(x,pdf4, typ='l', xlim=c(0.5,3))
```



```r
par(mfrow=c(1,1))
```

2.

The integral from negative infinity to positive infinity of a probability density function is always 1.

3.

Each of these is solved by integration over the PDF. It is important to remember that the integral is a *definite* integral over the PDF from negative infinity to x, which means that the equation for each segment includes a constant term that is not necessarily zero. The constant term can be determined by integrating carefully over all segments, or by ensuring that the "boundary conditions" are satisfied (each segment must connect to the previous segment with no gaps, and the CDF must have a range of zero to one.)

$$CDF1(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } 0 < x < 1, \\ 1 & \text{if } x > 1, \end{cases}$$

$$CDF2(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 - e^{-x} & \text{if } x < 1, \end{cases}$$

$$CDF3(x) = \begin{cases} 0 & \text{if } x < 0, \\ 0.1x & \text{if } 0 < x < 2, \\ 0.2 & \text{if } 2 < x < 4, \\ 0.1x - 0.2 & \text{if } 4 < x < 8, \\ 0.6 & \text{if } 8 < x < 10, \\ 0.1x - 0.4 & \text{if } 10 < x < 14, \\ 1 & \text{if } x > 14, \end{cases}$$

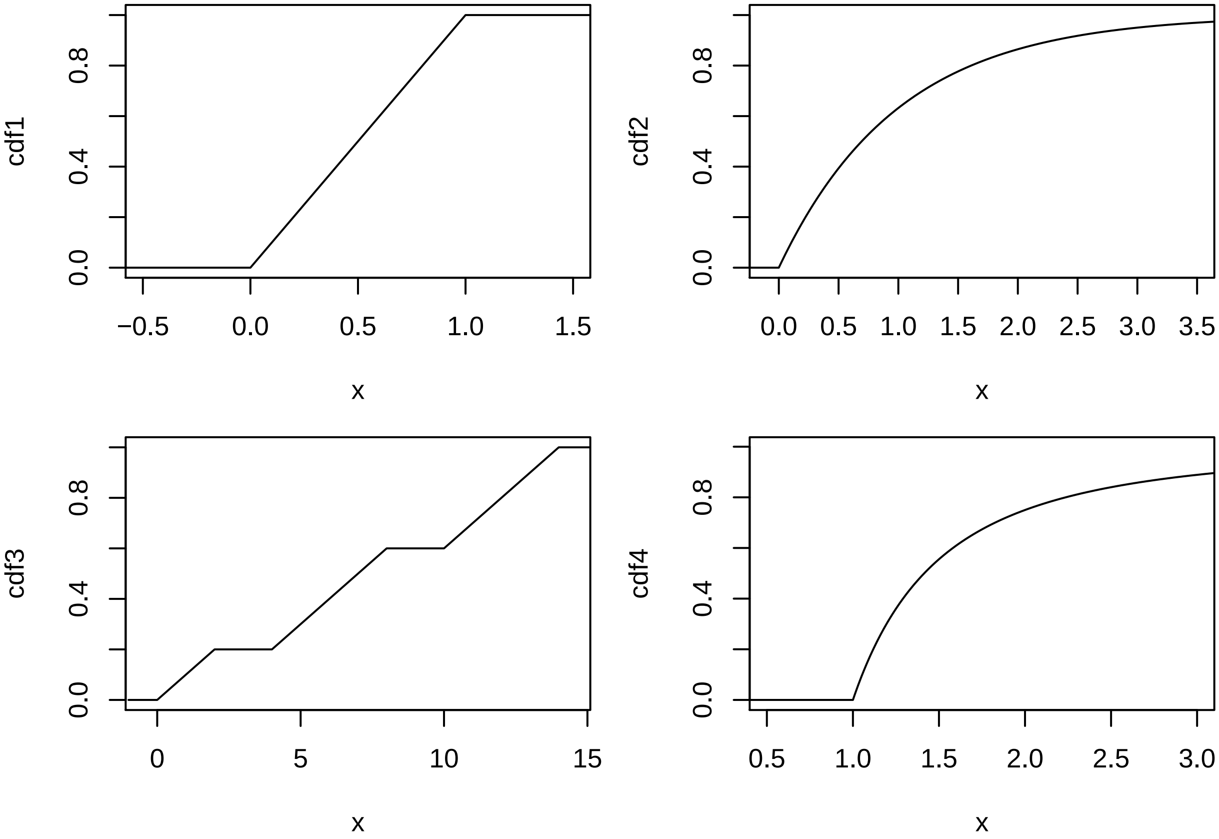
$$CDF4(x) = \begin{cases} 0 & \text{if } x < 1, \\ 1 - x^{-2} & \text{if } x >= 1, \end{cases}$$

4.

The CDFs can be plotted using the same techniques used to plot the PDFs:

```r
cdf1 = rep(0,length(x))
cdf1[x >= 0 & x <= 1] = x[x >= 0 & x <= 1]
cdf1[x > 1] = 1

cdf2 = rep(0,length(x))
cdf2[x > 0] = 1 - exp(-x[x > 0])

cdf3 = rep(0,length(x))
cdf3[x >= 0 & x <= 2] = 0.1*x[x >= 0 & x <= 2]
cdf3[x >= 2 & x <= 4] = 0.2
cdf3[x >= 4 & x <= 8] = 0.2+0.1*(x[x >= 4 & x <= 8]-4)
cdf3[x >= 8 & x <= 10] = 0.6
cdf3[x >= 10 & x <= 14] = 0.6+0.1*(x[x >= 10 & x <= 14]-10)
cdf3[x > 14] = 1

cdf4 = rep(0,length(x))
cdf4[x > 1] = 1-x[x>1]^(-2)
```

```
par(mfrow=c(2,2))
par(mar=c(4,4,1,1))
plot(x,cdf1, typ='l', xlim=c(-0.5,1.5))
plot(x,cdf2, typ='l', xlim=c(-0.1,3.5))
plot(x,cdf3, typ='l', xlim=c(-0.5,14.5))
plot(x,cdf4, typ='l', xlim=c(0.5,3))
```



```
par(mfrow=c(1,1))
```

5.

PDF1: CDF1(1.5) - CDF1(0.5) = 1-0.5 = 0.5

PDF2: CDF2(1.5) - CDF2(0.5) = 1-exp(-1.5) - (1-exp(-0.5)) = 0.3834

PDF3: CDF3(1.5) - CDF3(0.5) = 0.15 - 0.05 = 0.1

PDF4: CDF4(1.5) - CDF4(0.5) = (1-1.5^(-2)) - 0 = 0.5556

6.

(a) True (from symmetry)

(b) True: CDF2(ln(2)) = 1-exp(-ln(2)) = 1-(1/2) = 0.5, so this must be the median.

(c) True: CDF3(4.5) = 0.2 + 0.1×(4.5-4) = 0.2 + 0.05 = 0.25, so this must be the first quartile.

(d) False: PDF4 has no single mode (nor do any of these functions!)

Part (a) could also have been solved with an integral, and parts (b) and (c) by calculating and then evaluating the quantile function.

Tutorial Problem 3

Start by defining some constants:

```
mu = 50
sigma = 10
```

1.

```
1-pnorm(75, mean=mu,sd=sigma)
```

```
## [1] 0.006209665
```

2.

```
pnorm(58, mean=mu,sd=sigma)-pnorm(52, mean=mu,sd=sigma)
```

```
## [1] 0.2088849
```

3.

```
(1-pnorm(80, mean=mu,sd=sigma))/(1-pnorm(75, mean=mu,sd=sigma))
```

```
## [1] 0.2173866
```

4.

```
qnorm(0.99, mean=mu, sd=sigma)
```

```
## [1] 73.26348
```

5.

```
z = (56-mu)/(sigma/sqrt(12))
2*(1-pnorm(z))
```

```
## [1] 0.03766692
```

The change is significant.

6.

```
ci = 46 + (7/sqrt(25)) * qt(c(0.025,0.975), 25-1)
ci
```

```
## [1] 43.11054 48.88946
```

7.

Yes, because 50 is not within the confidence interval. To obtain a p-value:

```
t = (46-mu)/(7/sqrt(25))
2*pt(t,25-1)
```

```
## [1] 0.008691035
```

8.

The sample size equation (for the SE of a mean) is: $n = (s/SE)^2$.

First estimate the SE we need. We want the difference between upper and lower limits of the confidence interval to be 2. To a reasonable approximation when n is large, the upper limit of a 95% confidence interval is $\bar{x} + 2 \times SE$ and the lower limit is $\bar{x} + 2 \times SE$, so the difference between upper and lower limits is $4 \times SE$. We want this to be about 2, so we need to aim for an SE of about 0.5.

Plugging in the numbers, we when have $n = (7/0.5)^2 = 14^2 = 196$.

Rounding up (this equation is an approximation, after all) we probably want about 200 students to obtain the needed precision.

(This could also be solved quickly using ratio logic: for large n, sample size goes as the square of the precision, however the precision is defined. Our current precision is $P = 48.8 - 43.1 = 5.7$ for a sample of 25. To achieve a precision of $P = 2.0$, we need to increase our sample by a factor of $(5.7/2)^2$ or about 8. So instead of 25 students we need about 200.)