STA 570 Exam 1 (take-home part)

Anirban Chetia

October 14, 2021

My solution to problem 0

I ran `data()`, browsed through the subsequently shown data sets, and then picked the `economics` dataset.

Loading my picked data set:

```
data(economics, package = "ggplot2")
```

Creating a data frame with a continuous variable of my interest (personal consumption expenditures, in billions of dollars) from my picked data set:

```
economics.data <- data.frame(personalConsumptionExpenditures = economics$pce)
```

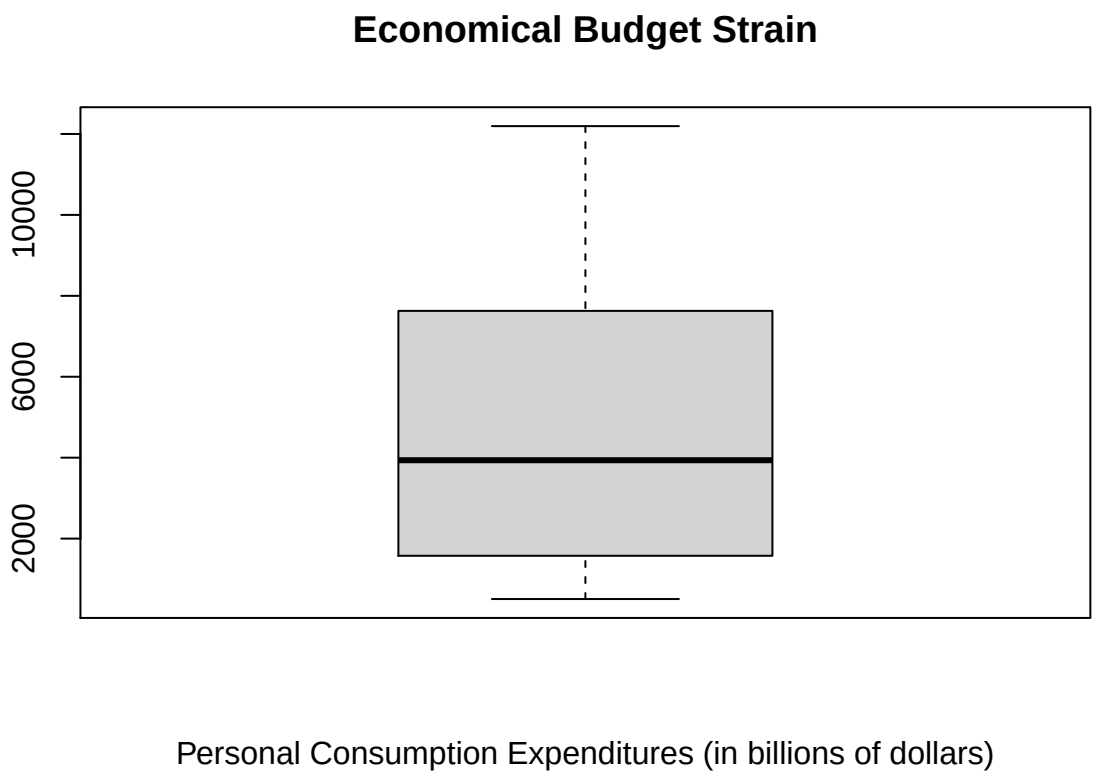
Computing and appropriately **naming** the sample mean, median and standard deviation of my chosen variable:

```
economics.data %>%
  summarise(Mean = mean(personalConsumptionExpenditures),
            Median = median(personalConsumptionExpenditures),
            Standard.Deviation = sd(personalConsumptionExpenditures)) %>% kable()
```

Mean	Median	Standard.Deviation
4820.093	3936.85	3556.804

Presenting a box plot of my chosen variable:

```
boxplot(economics.data, main = "Economical Budget Strain",
        xlab = "Personal Consumption Expenditures (in billions of dollars)")
```



There are outliers present (mostly) towards the end of the `personalConsumptionExpenditures` column which deviate the mean value away from the median (or the 50th percentile) with a comparatively greater value.

My solution to problem 1

Removing the outliers:

```
lower.limit <- 1000
upper.limit <- 5000
economics.data <- economics.data %>% filter(personalConsumptionExpenditures < upper.limit &
                                            personalConsumptionExpenditures > lower.limit)
```

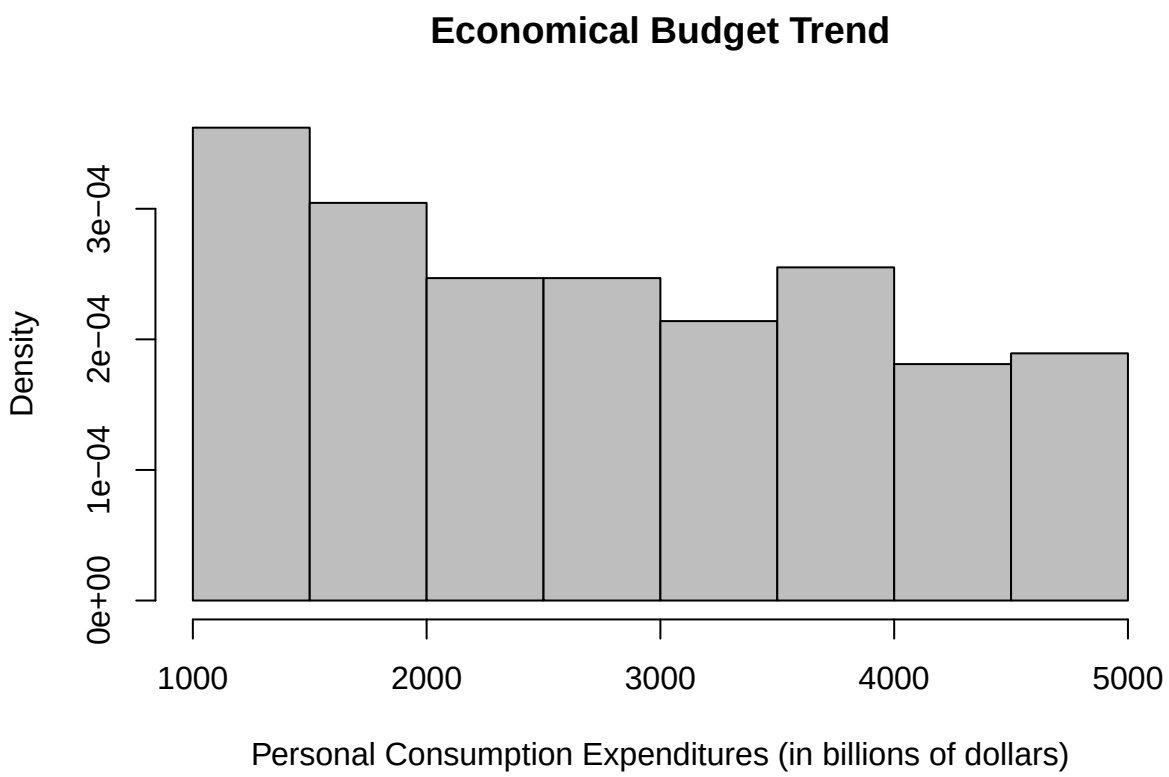
Checking the mean and median values for similarity, now that the outliers have been removed:

```
economics.data %>% summarise(Mean = mean(personalConsumptionExpenditures),
                             Median = median(personalConsumptionExpenditures)) %>% kable()
```

Mean	Median
2776.729	2696.4

Presenting a histogram of the outlier-free data:

```
hist(economics.data$personalConsumptionExpenditures,
     main = "Economical Budget Trend", col = "Grey",
     prob = TRUE, xlab = "Personal Consumption Expenditures (in billions of dollars)")
```



The data definitely **does not** appear to be normal.

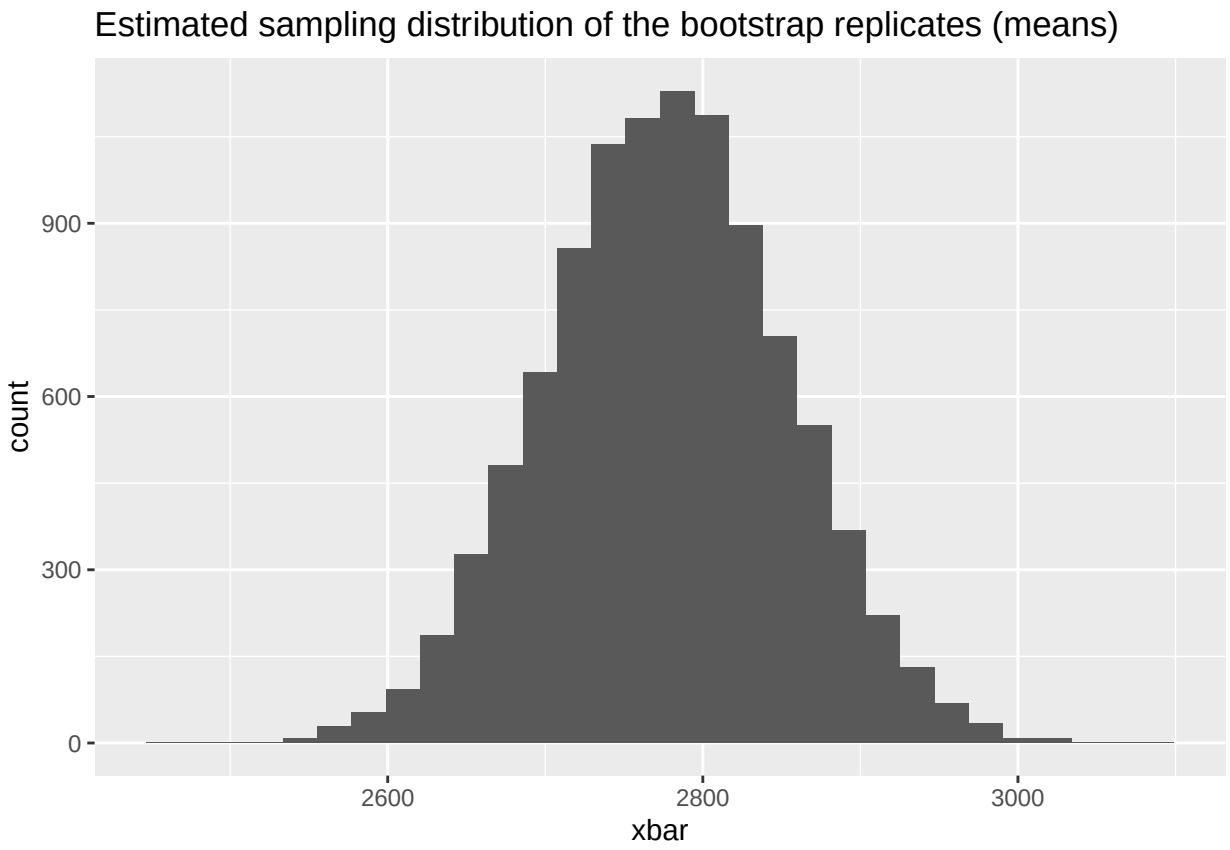
My solution to problem 2

Creating 10000 bootstrap replicates of the mean statistic (computed from the resample, i.e. $\bar{x}^*$) from the sample mean ($\bar{x}$):

```
mean.function <- function(x, index) {
    d <- x[index]
    return(mean(d))
}
Boot.economics.data <- boot(data = economics.data$personalConsumptionExpenditures,
                            statistic = mean.function, R = 10000)
head(Boot.economics.data$t)
```

```
##          [,1]
## [1,] 2778.013
## [2,] 2775.232
## [3,] 2737.514
## [4,] 2759.532
## [5,] 2706.556
## [6,] 2826.574
```

```
Boot.economics.data.plot <- data.frame(xbar = Boot.economics.data$t)
ggplot(Boot.economics.data.plot, aes(x = xbar)) + geom_histogram() +
       ggtitle('Estimated sampling distribution of the bootstrap replicates (means)')
```



This distribution indeed **does** appear to be normal.

My solution to problem 3

An asymptotic confidence interval is given by:

$$\bar{x} \pm t_{df}^{(1-\alpha)} * \sigma/\sqrt{n}$$

or,

$$(\bar{x} - t_{df}^{(1-\alpha)} * \sigma/\sqrt{n},\ \bar{x} + t_{df}^{(1-\alpha)} * \sigma/\sqrt{n})$$

Creating a 92% confidence interval for the population mean (μ) of my variable of interest using the asymptotic approach: (using the formula that I wrote above)

```
n <- length(economics.data$personalConsumptionExpenditures)
df <- n - 1
confidence.level = 0.92
alpha = 1 - confidence.level
lower.limit <- mean(economics.data$personalConsumptionExpenditures) -
               qt(1 - (alpha/2), df) * (sd(economics.data$personalConsumptionExpenditures)/sqrt(n))
upper.limit <- mean(economics.data$personalConsumptionExpenditures) +
               qt(1 - (alpha/2), df) * (sd(economics.data$personalConsumptionExpenditures)/sqrt(n))
interval.one <- paste0("(", lower.limit, ", ", upper.limit, ")")
interval.one
```

```
## [1] "(2644.93177754882, 2908.52665866517)"
```

This interval when **interpreted** implies that we are 92% confident that the sample mean would have a value that exists within the range given by the interval, or a value that lies in between the bounds (lower and upper limits) of the interval.

Constructing a similar confidence interval using the percentiles of the bootstrap replicates that I created above:

```
interval.two <- quantile(Boot.economics.data$t, probs = c(.04, .96))
interval.two
```

```
##       4%      96%
## 2644.294 2908.447
```

My logic on choosing the lower and upper limits for the middle 92% interval above:

From the practice problems that I did, I observed a pattern for getting the lower and upper limits for the 90%, 95% and 99% confidence intervals: (in order to get the middle 90/95/99%)

90% -> 10% from 100% -> $(\frac{10}{2})(\frac{1}{100}) = 0.05$ -> This is the lower bound, and I can obtain the upper bound by subtracting it from 1, as their sum should be equal to 1 (i.e., $lower\ bound + upper\ bound = 1$). Hence, the upper bound will be $1 - 0.05 = 0.95$, and the interval will be $(0.05, 0.95)$.

95% -> 5% from 100% -> $(\frac{5}{2})(\frac{1}{100}) = 0.025$ -> Again, this is the lower bound, and I can obtain the upper bound by subtracting it from 1 $(1 - 0.025 = 0.975)$. Thus, the interval here would be $(0.025, 0.975)$.

99% -> 1% from 100% -> $(\frac{1}{2})(\frac{1}{100}) = 0.005$ -> Likewise, this is the lower bound, and I can obtain the upper bound by subtracting it from 1 $(1 - 0.005 = 0.995)$. The interval here would thus be $(0.005, 0.995)$.

X% -> (100-X)% from 100% -> $(\frac{X}{2})(\frac{1}{100}) = Y(say)$ -> This is the lower bound, and I can obtain the upper bound by subtracting it from 1. Conversely, I can add Y to X% or $\frac{X}{100}$ to get the upper bound, and then I can obtain the lower bound by subtracting it from 1.

Applying the same pattern, I got the bounds for constructing the middle 92% interval:

92% -> 8% from 100% -> $(\frac{8}{2})(\frac{1}{100}) = 0.04$ -> This is the lower bound, and I can obtain the upper bound by subtracting it from 1. $(1 - 0.04 = 0.96)$ Thus, the required interval here is $(0.04, 0.96)$.

Now that I've computed two confidence intervals with different approaches, I can **compare** them:

```
# CI using the asymptotic approach:
interval.one
```

```
## [1] "(2644.93177754882, 2908.52665866517)"
```

```
# CI using the bootstrap method:
interval.two
```

```
##       4%      96%
## 2644.294 2908.447
```

I observe that I get near about the same results, which is as I expected, given that our sample mean from the bootstrap is normally distributed (as can be seen from the graph in my solution to problem 3 above), and hence the percentiles of the bootstrap replicates fall under the same page as what the result of an asymptotic confidence interval would.

There are no significant discrepancies for the same reason.

My solution to problem 4

Running the required commands with the required values: (as asked in the question/problem statement)

```
m <- mean(economics.data$personalConsumptionExpenditures)
s <- sd(economics.data$personalConsumptionExpenditures)
n <- length(economics.data$personalConsumptionExpenditures)
sidn <- 6168656
set.seed(sidn)
mu0 <- rnorm(1, m, 2*s/sqrt(n))
mu0
```

```
## [1] 2907.862
```

I am considering the null hypothesis to be that the mean of the personal consumption expenditures is equal to 2907.862 billions of dollars, i.e., symbolically:

$$H_o : \mu = (2907.862 * 10^9)\$$$

or

$$H_o : \mu = 2907.862 \ Billion \ Dollars$$

I then **pick** a **right-tailed** alternative hypothesis, stating that the mean of the personal consumption expenditures is greater than 2907.862 billions of dollars, i.e., symbolically:

$$H_a : \mu > (2907.862 * 10^9)\$$$

or

$$H_a : \mu > 2907.862 \ Billion \ Dollars$$

The reason why I am considering a right-sided test is because a greater mean for the personal consumption expenditures (than the considered value for the null hypothesis) would indicate that people spend more on the goods and services as a consumer than we would expect, which could possibly lead to economical budget strain.

Making a type I error (based on the chosen value for α) here would indicate that the mean of the personal consumption expenditures is not equal to 2907.862 billions of dollars (and thus we reject the null hypothesis, albeit incorrectly) as per our test, when in fact, it is actually equal.

This implies that our value for α should be decreased, or that we should consider a lower significance level than the one we are considering, that led to this wrong conclusion.

On the other hand, making a type II error (β) here would indicate that the mean of the personal consumption expenditures is equal to 2907.862 billions of dollars (supporting the null hypothesis, albeit incorrectly) as per our test, when in fact, it is actually greater than that value (and we should have instead gone with the alternative hypothesis).

This error implies we failed to reject our null hypothesis, which is actually false (another form of a wrong conclusion). Contrary to type I, we can avoid this error by increasing our significance level or value of α, and also by increasing our sample size.

I am **choosing** a significance level of $\alpha = 0.05$, based on the relative importance of avoiding both of these errors above: (i.e., a value not too small, and neither too big)

```
alpha = 0.05
conf.level = 1 - alpha
```

Conducting a one-sample t-test for μ_0 (the hypothesized value of the true population mean μ) using the `t.test()` function with the confidence level I chose above (95%) and extracting the p-value from it:

```
t.test(x = economics.data$personalConsumptionExpenditures, mu = mu0,
       alternative = "greater", conf.level = conf.level)$p.value
```

```
## [1] 0.9592385
```

I got a large p-value which is greater than 0.05 (my chosen value for α), which tells that the probability to reject the null hypothesis is not statistically significant and indicates strong evidence for it to be true, i.e. this suggests that the mean of the personal consumption expenditures is not equal to 2907.862 billions of dollars, or at least the fact that we cannot reject our null hypothesis based on the taken test statistic. (i.e., the conclusion **drawn** here is that it supports the null hypothesis)

My solution to problem 5

Determining the power of the test with the supplied values: (sample size n, a standard deviation of s, a difference among the null and alternative means of $\delta = s/2$ and finally, a significance levevl of 0.001)

```
result <- power.t.test(n = n, delta = s/2, sd = s, sig.level = 0.001,
                       type = "one.sample", alternative = "one.sided")
result$power
```

```
## [1] 0.9999981
```

I know that power is the probability of rejecting the null hypothesis correctly (in which case the specified alternative hypothesis becomes true), and in this case it turns out to be ~0.99 - this gives the implication that we are extremely likely to successfully reject the null when it is in fact, false and that our Type II error (β) is almost equal to zero.

Interpreting this with context to my chosen variable of interest, this implies that the mean of the personal consumption expenditures would most certainly (given that power is almost equal to 1, with a value of ~0.99) be greater than 2907.862 billions of dollars, while indicating that the null is false.

Two ways via which I can yield greater power:

1) I can increase α or the significance level - for e.g. I can run the above code at 90%, 95% and 99% significance levels (i.e. with `sig.level` values of 0.1, 0.05 and 0.01 respectively), and for all of them the resultant power will be 1: (which is greater than what I observed above)

```
# 99% significance level:
result <- power.t.test(n = n, delta = s/2, sd = s, sig.level = 0.01,
                       type = "one.sample", alternative = "one.sided")
result$power
```

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

```
# 95% significance level:
result <- power.t.test(n = n, delta = s/2, sd = s, sig.level = 0.05,
                       type = "one.sample", alternative = "one.sided")
result$power
```

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

```
# 90% significance level:
result <- power.t.test(n = n, delta = s/2, sd = s, sig.level = 0.1,
                       type = "one.sample", alternative = "one.sided")
result$power
```

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

2) I can increase my sample size `n`, say by 100 more observations to increase the power to 1: (which again, is greater than what I observed above)

```
n <- n + 100
result <- power.t.test(n = n, delta = s/2, sd = s, sig.level = 0.001,
                       type = "one.sample", alternative = "one.sided")
result$power
```

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

The power being 1 (in both the cases above) also implies that we are certain to successfully reject the null hypothesis when it is actually false, and that our type II error (β) is equal to zero.

My solution to problem 6

One topic that I struggled with the most during the test was the bootstrap sampling part, wherein I had to create a bootstrap sample of a variable given some measurements of it. The reason I felt this as bemusing and difficult is because of the way I went through chapter 3 and bootstrapping in general - I previously understood how to create a bootstrap sample in R, based on a statistic (mean, standard deviation etc.), and how to create a CI through the replicates or the percentiles using the quantile function. What I was lacking was the know-how of the principles behind it, and how I could create such a sample without R, going by the theory.

Re-iterating briefly on what I learnt:

Bootstrapping is a type of resampling where large numbers of smaller samples of the same size are repeatedly drawn, with replacement, from a single original sample. A bootstrap sample is a smaller sample that is "bootstrapped" from a larger sample.

For example, let's consider a sample made up of five numbers: 66, 12, 31, 45, 50. I randomly draw two numbers, say 50 and 45. I then replace those numbers into the sample and draw two numbers again. I repeat this process of drawing n numbers (where n is 2 for what I just said) m (here m is also two until this point, given that I replaced and resampled twice) times. (as far as I've observed, the original samples are much larger than this simple example of mine, and m is usually taken in hundreds to tens of thousands or more!). After a large number of iterations, the bootstrap statistics are compiled into a bootstrap distribution (which is what I created in R).

I bridged this gap myself over the weekend following the in-person/written test, and my comprehension was solidified when you mentioned this in class today. (I also cleared my misconception today that the standard deviation is equal to the standard error for most cases, and that the terms can be used interchangeably. I understood that the later is the statistic's standard deviation, and that it is calculated by taking the standard deviation and dividing it by the square root of the sample size for the problems we are concerned with)

My solution to problem 7

Yes! Probability was one topic I digged into deep, practicing a handful of problems from resources available online. Here's what I know about the subject:

The very basic notion of probability is a value for the occurrence of an event - one that is either one of the binary forms (i.e. 0 and 1, or boolean false and true if I were to interpret it in my domain of computer science!), or it is a value in between them. The former accounts for the case of Bernoulli random variables (i.e., they can only take the values of 0 or 1).

These two values are extreme values and thus do not fall under the category of 'an event that is *likely/probable* to occur', as because if the event doesn't, its a 0 and if it does, its a 1. These cases are denoted as an 'impossible' and a 'certain' event respectively.

The sum of the probabilities of two events occurring together (intersection) and either of the events occurring (union) is equal to the sum of the individual probabilities of those events to occur, i.e. $P(A \cap B) + P(A \cup B) = P(A) + P(B)$.

The sum of the probabilities of an event and its complement must equal 1, i.e. say for an event A, $P(A) + P(A') = 1$ or $P(A) + P(A^c) = 1$, depending upon how one writes/denotes the complement.

Conditional probability gives the measure of an event occurring, given that another event has already occurred. Considering two events 'A' and 'B', the probability of A given B is given by $P(A/B) = \frac{P(A \cap B)}{P(B)}$, and conversely, the probability of B given A is given by $P(B/A) = \frac{P(B \cap A)}{P(A)}$.

If A and B independent (unrelated events), the probability of them occurring together (intersection) becomes the product of their respective probabilities, and hence their conditional probabilities will simply be the probability of the occurring event (given that another event has already occurred, but it doesn't change anything) i.e. to say, $P(A \cap B) = P(A).P(B) => P(A/B) = P(A)$ and $P(B/A) = P(B)$

If A and B are disjoint (mutually exclusive), the probability of them occurring together becomes 0, and so does their conditional probabilities; i.e. to say, $P(A \cap B) = \phi => P(A/B) = P(B/A) = 0$

For discrete random variables:

- The expected value (weighted average of the possible values, where the weights are the proportions with which those values occur) and the variance (weighted average of the squared-deviations that could occur) are respectively given by:

$$\mu = E[S] = \sum_{\text{possible } s} s \cdot P\left(S = s\right)$$

$$\sigma^2 = V[S] = \sum_{\text{possible } s} (s - \mu)^2 \cdot P\left(S = s\right)$$

- Their probabilities are additive, for e.g.: $P(1 \cup 2) = P(1) + P(2)$
- There exists a probability mass function, which describes how the probability is spread across the possible outcomes.

For instance, the probability mass function for a:

- Binomial random variable X taken from n trials each with probability of success π is:

$$P(X = x) = \frac{n!}{x!(n-x)!}\, \pi^x\, (1 - \pi)^{n-x}$$

- Poisson random variable for some count variable Y with expected number of events over an unit of time/space λ: (e is a constant with a value $2.718\ldots$)

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

On the other hand for continuous random variables, there exists a probability density function instead, which is a function whose value at any given sample in the sample space can be interpreted as providing a relative likelihood that the value of the random variable would be close to.

The probability density function for the:

- Uniform (variation of binomial) distribution is:

$$f(x) = \begin{cases} \frac{1}{B-A} & A \leq x \leq B \end{cases}$$

- Standard uniform distribution (follows standardization) is:

$$f(x) = \begin{cases} 1 & 0 \leq x \leq 1 \end{cases}$$

- Exponential (continuous analog of poisson) distribution is:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \text{ and } \lambda > 0 \\ 0 & \text{otherwise} \end{cases}$$

- Normal distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

In R, there exists cumulative probability distribution functions of the form 'px', where 'x' is the probability distribution considered (for instance, x = 'norm' for normal, x = 'binom' for binomial and x = 'pois' for poisson).

That's nearly everything I know about probability in general, and everything that came to my mind while writing this. I would like to mention that I spent time on set theory problems as well, since it's something you need to know for probability. Some problems that I did revolved around 3 variables as well, the formula for which is: (I figured this might not come, given that you told us we do not need to focus on memorizing formulas - but I practiced until this point anyway)

$$n(A \cup B \cup C) = n(A) + n(B) + n(C) \breve{} n(A \cap B) \breve{} n(B \cap C) \breve{} n(C \cap A) + n(A \cap B \cap C)$$

That's it for my solutions for the problems on this take-home test. Thanks for reading!