BMI 713: Computational Statistics for Biomedical Sciences

Assignment 2

1 Random variables and distributions

1. Assume that a die is fair, i.e. if the die is rolled once, the probability of getting each of the six numbers is 1/6. Calculate the probability of the following events.

   - Rolling the die once, what is the probability of getting a number less than 3?

     Sol'n. *Let X be a random variable denoting the value of a die roll. The probability that one particular die roll satisfies some criteria is:*

     $$\frac{\textit{(\# of die rolls that satisfy the criteria)}}{\textit{(\# of possible die rolls)}}.$$

     We can easily count both of these values. $X < 3$ means $X = 1$ or $X = 2$. So the number of die rolls satisfying $X < 3$ is 2. There are 6 possible die rolls. Hence

     $$P(X < 3) = 2/6 = 1/3.$$

   - (Optional) Rolling the die twice, what is the probability that the sum of two rolling numbers is less than 3?

     Sol'n. *Let X and Y be random variables such that X is the value of the first die roll and Y is the value of the second. We wish to find the number of rolls where $X + Y < 3$. It is clear that $X + Y < 3$ occurs* only *when $X = 1$ and $Y = 1$. That is, only a single roll satisfies the criteria. There are 6 possible values for both X and Y, hence there are $6 \times 6 = 36$ possible rolls. It follows that the probability of rolling a sum less than 3 is*

     $$\frac{\textit{(\# of rolls with } X + Y < 3)}{\textit{(total \# of possible rolls)}} = \frac{1}{36}.$$

2. Let p be the probability of obtaining a head when flipping a coin. Suppose that Bob flipped the coin n ($n \geq 1$) times. Let X be the total number of head he obtained.

   - What distribution does the random variable X follow? Is X a discrete or continuous random variable?

     Sol'n. *X follows the binomial distribution with n trials and probability of success p. This is a discrete distribution.*

   - What is the probability of $X = k$, i.e. what is $Pr(X = k)$ ($0 \leq k \leq n$)? (Write down the mathematical formula for calculating this probability.)

     Sol'n. *In general, the probability that a binomially distributed random variable $X \sim Bin(n, p)$ is equal to some value k is*

     $$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

   - What is the probability of $X \geq k$, i.e. what is $Pr(X \geq k)$?

     Sol'n.

     $$P(X \geq k) = P(X = k) + P(X = k+1) + \ldots + P(X = n).$$

     Equivalently, $P(X \geq k)$ can be written as $1 - P(X < k)$:

     $$P(X \geq k) = 1 - P(X < k) = 1 - [P(X = 0) + P(X = 1) + \ldots + P(X = k-1)].$$

- Suppose $p = 0.4$ and $n = 10$. Calculate the probabilities $Pr(X = 3)$ and $Pr(X \geq 3)$. (You may need the functions `dbinom` and `pbinom` in R to calculate these two probabilities. Use `?dbinom` and `?pbinom` to get help information of these two functions).

  Sol'n. *Using the general formula above with $n = 10$ and $p = 0.4$:*

$$P(X = 3) = \binom{10}{3} \cdot 0.4^3 \cdot 0.6^7 \approx 0.215.$$

  Using R's `dbinom`:

```
> dbinom(x=3, size=10, prob=0.4)
[1] 0.2149908
```

  As shown above, we may compute $P(X \geq 3)$ either by summing the probabilities for $X = 3$, $X = 4$, ..., $X = 10$ or we could subtract the probabilities for $X = 0$, $X = 1$ and $X = 2$ from 1. The latter method is more convenient since it requires fewer terms:

$$\begin{aligned} P(X \geq 3) &= 1 - P(X < 3) = 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &= 1 - \left[\binom{10}{0} \cdot 0.4^0 \cdot 0.6^{10}\right] - \left[\binom{10}{1} \cdot 0.4^1 \cdot 0.6^9\right] - \left[\binom{10}{2} \cdot 0.4^2 \cdot 0.6^8\right] \\ &\approx 1 - 0.006 - 0.040 - 0.121 \\ &\approx 0.833. \end{aligned}$$

  Again, using R:

```
> 1 - pbinom(q=2, size=10, prob=0.4)
[1] 0.8327102
```

3. In a population of certain type of fish, the lengths of the individual fish follow a normal distribution. The mean length of the fish is 54.00 mm and the standard deviation is 4.50 mm. Answer the following questions.

   - What percentage of the fish are less than 63 mm?

     Sol'n. *We know that the lengths of the fish follow a normal distribution with mean $\mu = 54$ and standard deviation $\sigma = 4.50$. More succinctly: if X is a random variable denoting the length of a randomly selected fish, then:*

$$X \sim N(54, 4.50^2).$$

     The solution we are searching for is then $P(X < 63)$. We can compute this directly using R:

```
> pnorm(q=63, mean=54, sd=4.5)
[1] 0.9772499
```

     However, the typical method for computing such a probability is to standardize *the normal variable and then determine the probability from a table. That is, let $Z = \frac{X-54}{4.50}$. Then Z is a standard normal random variable: $Z \sim N(0, 1)$. We also know that:*

$$P(X < 63) = P\left(\frac{X - 54}{4.50} < \frac{63 - 54}{4.50}\right) = P(Z < 2).$$

     Since Z follows the standard normal distribution, we can look up the value for $P(Z < 2)$. The table at:

     `http://en.wikipedia.org/wiki/Standard_normal_table#Partial_Table`

     gives $P(Z < 2) = 0.9772$.

   - What percentage of the fish are more than 50 mm?

Sol'n. *In this case, we wish to find* $P(X > 50)$.

$$P(X > 50) = P\left(\frac{X - 54}{4.50} > \frac{50 - 54}{4.50}\right) = P(Z > -8/9).$$

Since the standard normal distribution is symmetric around 0, we know that $P(Z > -8/9) = P(Z < 8/9)$. *Using a table we find:*

$$P(X > 50) = P(Z < 8/9) \approx 0.811.$$

In `R`*, we may use the following commands to calculate the probability*

```
> 1-pnorm(50,mean=54,sd=4.5)
[1] 0.8129686
```

or

```
> pnorm(50, mean=54, sd=4.5, lower.tail=F)
[1] 0.8129686
```

With the standardized value, we may use

```
> pnorm(-8/9, lower.tail=F)
[1] 0.8129686
```

- Suppose that you randomly selected 10 fish from the population. What is the probability that the average length of the 10 fish is between 52 mm to 56 mm?

  Sol'n. *Let* $X_1, X_2, \ldots, X_{10}$ *denote the lengths of 10 randomly selected fish. Then the average is given by:*

$$Y = \frac{1}{10}\sum_{i=1}^{10} X_i,$$

which is itself a random variable. As we saw in lecture, Y *approximately follows* $N(54, \frac{4.50^2}{10})$. *Now we can compute the probability* $P(52 < Y < 56) = P(Y < 56) - P(Y < 52)$ *by standardizing. Let* $Z = \frac{Y-54}{4.50/\sqrt{10}}$*:*

$$P(Y < 56) = P\left(Z < \frac{56 - 54}{4.50/\sqrt{10}}\right) \approx P(Z < 1.41) \approx 0.9207.$$

and

$$P(Y < 52) = P\left(Z < \frac{52 - 54}{4.50/\sqrt{10}}\right) = P(Z < -1.41).$$

Again, we use symmetry to handle the negative value since it is often not available in tables:

$$P(Z < -1.41) = P(Z > 1.41) = 1 - P(Z < 1.41) \approx 1 - 0.9207 = 0.0793.$$

Finally,

$$P(52 < Y < 56) = P(Y < 56) - P(Y < 52) \approx 0.9207 - 0.0793 = 0.8414.$$

We may also calculate the probability using the following command in `R`*:*

```
> pnorm(56, mean=54, sd=4.5/sqrt(10)) - pnorm(52, mean=54, sd=4.5/sqrt(10))
[1] 0.8401145
```

2 Hypothesis testing

1. In each of the following situations, state an appropriate null hypothesis H_0 and alternative hypothesis H_1. Be sure to identify the parameters that you use to state the hypotheses.

   (a) An experiment on learning in animals measures how long it takes a mouse to find its way through a maze. The mean time is 18 seconds for one particular maze. A researcher thinks that a loud noise will cause the mice to complete the maze slower. She measures how long each of 10 mice takes with a noise as stimulus.

   Sol'n. *H_0: The loud noise does not affect the speed of the mice, i.e. the mean completion time with noise is still 18s.*
   H_1: The loud noise causes the mice to complete the maze slower, i.e. the mean completion time with noise is greater than 18s.

   (b) A pharmaceutical company developed a new drug for certain type of cancer and the company believed that the drug can significantly increase patient's survival time after surgery. The mean survival time after surgery is 16 months. The company selected 20 volunteers who had the surgery, treated them with the drug and records their survival time.

   Sol'n. *H_0: The mean survival time of the treated patients is 16 months.*
   H_1: The drug increases patients' survival time, i.e. the mean survival time of the treated patients is greater than 16 months.

2. Determine if the following statements are true.

   (a) P-value is the probability that the null hypothesis is true. **False.**

   (b) P-value is the probability, computed assuming that H_0 is true, that the test statistics will take a value at least as extreme as that actually observed. **True.**

   (c) Standard normal distribution has fatter tail than student t-distribution. **False. The t-distribution has a fatter tail.**

   (d) When the sample size is small, t-test is more accurate than z-test. **True.**

3. The mean level of calcium in the blood in healthy young adults is about 9.5 milligrams per deciliter. A clinic in Boston measures the blood calcium level of 10 healthy pregnant women as follows

   $$9.09,\ 9.82,\ 9.58,\ 9.03,\ 10.48,\ 9.35,\ 9.85,\ 9.36,\ 9.64,\ 9.43.$$

   Is this an indication that the mean calcium level in the population from which these women come differs from 9.5?

   - State the null hypothesis H_0 and and the alternative hypothesis H_1.

     Sol'n. *H_0: The women are sampled from a population with mean blood calcium level 9.5.*
     H_1: The women are sampled from a population with mean blood calcium level greater or less than 9.5.

   - Calculate the mean and standard deviation of the blood calcium level of the 10 women.

     Sol'n. *The mean is:*

     $$\bar{X} = \frac{9.09 + 9.82 + 9.58 + 9.03 + 10.48 + 9.35 + 9.85 + 9.36 + 9.64 + 9.43}{10} = 9.563.$$

     Let $X_1, X_2, \ldots, X_{10}$ denote the 10 blood calcium levels. Then the sample standard deviation is:

     $$s = \sqrt{\frac{1}{10-1}\sum_{i=1}^{10}(X_i - 9.563)^2} = 0.422.$$

- Calculate the z-score and perform the z-test based on the z-score. What is the P-value?

  Sol'n. *As we have seen in lecture, the z-score is*

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

where $\mu_0 = 9.5$ is the population mean, $n = 10$ is the sample size and σ is the population standard deviation. However, we do not know σ. We use the sample standard deviation s as an estimate for σ. This gives us the z-score

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{9.563 - 9.5}{0.422/\sqrt{10}} = 0.472.$$

The z-score approximately follows standard normal distribution (where the approximation improves as the number of samples grows). Therefore, the p-value given by the z-test is

$$P(|Z| > |z|) = P(|Z| > 0.472) = P(Z > 0.472) + P(Z < -0.472).$$

Using the standard normal distribution, we get

$$P(Z > 0.472) = 1 - P(Z < 0.472) = 0.3184$$

and

$$P(Z < -0.472) = 0.3184.$$

Therefore, the p-value is $P(|Z| > z) = 0.3184 + 0.3184 = 0.6369$.
Note that since the standard normal distribution is symmetric about 0, we have $P(Z > |z|) = P(Z < -|z|)$.
In `R`, we may use the following command to compute the p-value

```
> 2*pnorm(-0.472)
[1] 0.6369268
```

If we choose a significance level of 0.05, we will fail to reject the null hypothesis H_0.

- Calculate the t-statistic and perform the t-test. What is the P-value? (Optional) What is the degree of freedom of this t-test?

  Sol'n. *If $\bar{X}$ denotes the sample mean, μ_0 denotes the mean of the null hypothesis, s denotes the sample standard deviation and n is the number of samples, the t-statistic is given by*

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{9.563 - 9.5}{0.422/\sqrt{10}} = 0.472.$$

The probability of obtaining a more extreme t-statistic is

$$P(T < -0.472) + P(T > 0.472),$$

where T follows the t-distribution with $10 - 1 = 9$ degrees of freedom. Using `R`, we can find these probabilities:

```
# P(T < -0.472)
> pt(-0.472, df=9)
[1] 0.3240804
# P(T > 0.472)
> 1 - pt(0.472, df=9)
[1] 0.3240804
```

$P(T < -0.472) + P(T > 0.472) \approx 0.648$. *Thus, under the null hypothesis, the probability of selecting 10 random individuals with mean blood calcium level more extreme than the observed mean 9.563 is roughly 65%. As with the z-test, we would not be able to reject the null hypothesis if we select a significance level of 0.05.*

To perform the t-test in R*:*

```
> v = c(9.09, 9.82, 9.58, 9.03, 10.48, 9.35, 9.85, 9.36, 9.64,9.43)
> t.test(v,mu=9.5,alternative=c("two.sided"))

        One Sample t-test

data:  v t = 0.4714, df = 9, p-value = 0.6486

alternative hypothesis: true mean is not equal to 9.5

95 percent confidence interval:
 9.260663 9.865337
sample estimates: mean of x 9.563
```