HW2

Pei Tian, pt2632

9/30/2023

Problem 1

Suppose the probability of having at least one dental checkup during a two-year period is 73%. 56 random individuals are being studied for their health care utilization practices over a two-year period. Compute "part a" by hand and show key steps. For all other parts you may hand calculate or use R.

a) What is the probability that exactly 40 of these individuals will have at least one dental checkup?

   According to the problem description, let X be the count of people who have at least one dental checkup during a two-year period among the studied population, then $X \sim Bin(0.73, 56)$.

   $P(X = 40) = \binom{56}{40} \cdot 0.73^{40} \cdot 0.27^{16} = 0.1133$

```r
# calculate the probability that X equals 40
dbinom(40, 56, 0.73)
```

```
## [1] 0.1133284
```

b) What is the probability that at least 40 of these individuals will have at least one dental checkup?

   $P(X \geq 40) = \sum_{i=40}^{56} \binom{56}{i} \cdot 0.73^{i} \cdot 0.27^{56-i} = 0.6679$

```r
# calculate the probability that X is at least 40
p <- 0.73
n <- 56
sum(dbinom(seq(40, 56), n, p))
```

```
## [1] 0.6678734
```

c) Could you use an approximation method to calculate the probabilities above? If yes, calculate the probabilities using approximations and compare to exact values; otherwise, explain why approximation methods are not appropriate.

   The poisson distribution could not be used to approximate this binomial distribution because this condition doesn't meet the following requirements:

   - Sample size n is not large enough (n should be larger than 100)
   - Probablity of success p is not samll enough (p should be less 0.01)

d) How many individuals do you expect to have at least one dental checkup?

   $E(X) = n * p = 0.73 * 56 = 41$

```r
# calculate expectation
n * p
```

```
## [1] 40.88
```

e) What is the standard deviation of the number of individuals who will have at least one dental checkup?

   $std(X) = \sqrt{n * p * (1-p)} = 3$

```r
# calculate standard deviation
sqrt(n * p * (1-p))
```

```
## [1] 3.322288
```

Problem 2

Suppose the number of tornadoes in the United States follows a Poisson distribution with parameter $\lambda = 6$ tornadoes per year. Compute using tables or R. Show the formula for "part a" (it can be handwritten and embedded in the pdf file).

a) What is the probability of having fewer than 3 tornadoes in the United States next year?

   According to the problem description, let X be the number of tornadoes in the United States, then $X \sim Poisson(6)$.

   $P(X < 3) = e^{-6} \sum_{n=0}^{2} \frac{6^n}{n!} = e^{-6} \cdot (\frac{6^0}{0!} + \frac{6^1}{1!} + \frac{6^2}{2!}) = 0.0620$

```r
# calculate probability that X is less than 3
sum(dpois(seq(0, 2), 6))
```

```
## [1] 0.0619688
```

b) What is the probability of having exactly 3 tornadoes in the United States next year?

   $P(X = 3) = e^{-6} \frac{6^3}{3!} = 0.0892$

```r
# calculate probability that X equals to 3
dpois(3, 6)
```

```
## [1] 0.08923508
```

c) What is the probability of having more than 3 tornadoes in the United States next year?

   $P(X > 3) = 1 - P(X \leq 3) = 1 - e^{-6} \sum_{n=0}^{3} \frac{6^n}{n!} = 0.8488$

```r
# calculate probability that X is more than 3
1 - sum(dpois(seq(0, 3), 6))
```

```
## [1] 0.8487961
```

Problem 3

Assume the systolic blood pressure of 20-29 year old American males is normally distributed with population mean 128.0 and population standard deviation 10.2.

a) What is the probability that a randomly selected American male between 20 and 29 years old has a systolic blood pressure above 137.0?

   According to problem description, let X be the systolic blood preasure of 20-29 years old American males, then $X \sim N(128.0, 10.2^2) \Rightarrow Z = \frac{X-128.0}{10.2} \sim N(0,1)$.

   $P(X > 137) = P(Z > \frac{137-128}{10.2}) = 1 - P(Z \leq 0.8824) = 1 - z_{0.8824} = 0.1888$

```
# calculate the probability that the blood pressure of a random sample is above 137
m <- 128
s <- 10.2
1 - pnorm((137-m)/s)
```

```
## [1] 0.188793
```

b) What is the probability that the sample mean for blood pressure of 50 males between 20 and 29 years old will be less than 125.0?

   According to the desciption and CLT, then $\overline{X} \sim N(128, \frac{10.2^2}{50}) => Z = \frac{\overline{X}-128}{\frac{10.2}{\sqrt{50}}} \sim N(0,1)$.

   $P(\overline{X} < 125) = P(Z < \frac{125-128}{\frac{10.2}{\sqrt{50}}}) = z_{-2.0797} = 0.0188$

```
# calculate the probability the sample mean is less than 125
n <- 50
z_score <- (125-m) / (s / sqrt(n))
pnorm(z_score)
```

```
## [1] 0.01877534
```

c) What is the 90th percentile of the sampling distribution of the sample mean X for a sample size of 40?

   In this scenario, $X \sim N(128, \frac{10.2^2}{40}) \Rightarrow Z = \frac{X-128}{\frac{10.2}{\sqrt{40}}} \sim N(0,1)$

   $P(X < t) = P(Z < \frac{t-128}{\frac{10.2}{\sqrt{40}}}) = 0.9 \Rightarrow z_{0.9} = 1.282 = \frac{t-128}{\frac{10.2}{\sqrt{40}}} \quad \therefore t = 130.0668$

   The 90th percentile of the sampling distribution of the sample mean is 129.8486 when sample size is 40.

```
# calculate the 90th percentile
n <- 40
qnorm(0.9) * (s / sqrt(n)) + m
```

```
## [1] 130.0668
```

Problem 4

Some researchers are interested in the mean pulse of young women suffering from fibromyalgia. They selected a random sample of 40 young females suffering from fibromyalgia. The sample mean of their pulses was 80 and the sample standard deviation was 10.

a) Compute the 95% confidence interval for the population mean pulse rate of young females suffering from fibromyalgia.

   Denote X as the pulse rate of young females suffering from fibromyalgia, the sample mean of pulse rate of young females suffering from fibromyalgia equals to $\overline{X} = \hat{\mu} = 80$, and the sample variance of young females suffering from fibromyalgia equals to $\hat{\sigma}^2 = s^2 = 10^2$

   According to the problem description, $\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$, so the 95% confidence interval is $(\overline{X} + t_{39,0.025}\frac{s}{\sqrt{n}}, \overline{X} + t_{39,0.975}\frac{s}{\sqrt{n}})$, which equals to (76.80184, 83.19816).

```r
# calculate 95% CI
m <- 80
s <- 10
n <- 40
c(qt(0.025, n-1), qt(0.975, n-1)) * (s / sqrt(n)) + m
```

```
## [1] 76.80184 83.19816
```

b) Interpret the calculated confidence interval.

   We are 95% confident that the **true population mean** lies between 76.80184 and 83.19816.

c) Suppose the researchers now want to test the null hypothesis that the mean pulse of young women suffering from fibromyalgia is equal to 70, against the alternative that the mean pulse is not equal to 70, at the $\alpha = 0.01$ significance level. Conduct this hypothesis test, and interpret the results.

   $H_0 : \mu = 70$ vs $H_1 : \mu \neq 70$

   Test statistic: $t = \frac{\hat{\mu}-70}{s/\sqrt{n}}$

   Reject H_0 when $|t| > t_{n-1,1-\alpha/2} = t_{39,0.995} \Rightarrow t < -2.707913$ or $t > 2.707913$

   Because $t = 6.324555 > 2.707913$, so we could reject null hypothesis.

   Interpretation: The mean pulse of young women suffering from fibromyalgia is NOT equal to 70 at significance level 0.01.

```r
# reject region calculation
a <- 0.01
m0 <- 70
# t-score
t <- (m - m0) / (s / sqrt(n))
c(qt(a/2, n-1), qt(1-a/2, n-1))
```

```
## [1] -2.707913  2.707913
```