**Question 1**

**New medication and side effects**

A systematic review of drug safety shows that 5% of adults who take a new medication experience negative side effects.

For the surveillance of drug safety, your hospital is interested in the chance of seeing more than 10 side effects among 100 randomly selected patients who need to take this medication now.

1. Find the probability that more than 10 patients in the random sample of 100 will experience negative side effects. You can use the binomial formula or normal approximation.

**State the numbers: n = 100, p = 0.05, and we are interested in P(X > 10)**

**Equivalently, 1 - P(X <= 10)**

**So, you can use pbinom to get P(X <= 10) OR,**

**You can use dbinom 10 times for P(X = 0) + P(X=1) + … + P(X=10)**

**Or, you can use dbinom() like this**

**1-sum(dbinom(x = 0:10, prob = 0.05, size = 100)) = 0.01147241**

**If you like the normal approximation approach, do:**

**Mean = 5, SD = 2.179449**

**And just compute area under the Z-curve as**

**(10 - 5)/2.179449 = 2.294158**

**And get the area under the Z-curve as**

**1-pnorm(mean = 0, sd = 1, q = 2.294158) = 0.01089071**

**A bit far from the exact value …why??**

1. Find the expected number of side effects in this random sample.

***µ = n\*p =* 100\*0.05**

1. The cost for the side effects is high from the patient's QoL (Quality of Life) and monetary perspective, but the hospital will have 1000 patients who need this medication next year and need to be certain about the potential loss. The hospital cannot use this medication if over 5% of people develop the side effects. What is the probability that less than 5% of patients will have side effects next year?

**But…I did not state this as a random sample. It is possible that these people come from the same neighborhood or share some biological or environmental features related to the probability of side effect. In such case, I did not capture people randomly from the hospital’s catchment area. So, the answer here is somewhat questionable, from the independent assumption perspective in binomial.**

**We want to know….**

**P( < 0.05) for the next year. And 5% of 1000 is 50,**

**So…**

**P(X < 50), when n = 1000 and p = 0.05**

**pbinom(size = 1000, p = 0.05, q = 49)**

**sum(dbinom(size = 1000, p = 0.05, x = 0:49))**

**Question 2 – Confidence Intervals**

Your client is interested in estimating the self-reported health score (ranging from 0-100; see the pet and pandemic study in Part 1) among low-income adults in Montreal.

From a large-scale population-representative survey conducted recently, the mean health score of adults in Montreal was 70, and the standard deviation was 20.

You randomly sampled 50 participants from the low-income population in Montreal, and found the sample mean among the low-income groups as 60.

a) State your population of interest and write down -values (lower and upper side) corresponding to 95% of the normal curve.

**=1.96**

**The last notation indicates the Z-score for the middle 95%, excluding 2.5% in both sides.**

**Your target population of interest is low-income adults in Montreal (how you define low-income, there are many ways).**

b) Calculate the margin of error at a 95% confidence level. Use the population standard deviation of 20 (in the next class, you will see how to calculate this from your sample)

**m = \***

**so,**

**1.96\*20/sqrt(50) = 5.543717**

c) Make an interpretation for the 95% confidence interval for the health score.

**Mean is 60, and the 95% CI is (54.45, 65.54).**

**If we repeat the sampling and the calculation of CI many times, 95% of the interval will contain the true value in a long run. We have one such value here.**

d) What are the approximate bounds on the values of health scores that would include 68% of the sampling distribution among the low-income population in Montreal?

**One standard error away from the center of the sampling distribution indicates middle 68% of sample means. The standard error is the standard deviation of sampling distribution. Thus, we are talking about away from the center of the sampling distribution. The center of the sampling distribution, is approximated by , if the Law of Large Number and Central limit Theorem holds (CLT requires the sample size of 50 to be safe). Thus, 60 plus minus would include approximately 68% of sample means (you have one such value here, as 60).**

e) Calculate the margin of error at a 90% confidence level.

**Just substitute with 1.645**

f) Calculate the margin of error when the sample size is n = 200.

**Narrow the margin by the factor of two.**

g) Do you think that the low-income population in Montreal has the same average health score as all adults in Montreal? Explain at 95% confidence level.

**The 95% CI (54.45, 65.54) from our sample does not include the Montreal mean of 70, so there it is likely to be different from the Montreal mean.**

h) State two assumptions to construct CI (hint, one has something to do with sampling and the other is related to the Central Limit Theorem)

**Random sampling, and sample size of at least 50 (in reality, sample size of <10 would work unless the sample mean is very skewed, but to be safe, we tend to state >30 or >50. In this class, we will stick to n > 50).**

1. Finally, you want to set the confidence level to 99%. To keep the same width of the margin of error as question b), do you need a larger or smaller sample size? Explain in one line.

**Making the confidence level 99% make the margin of error wider as increases. To make it narrower, we need a larger sample size.**

FYI, the z-scores for the commonly used confidence level are 90%, 95%, and 99%. This implies that the green areas take 10%, 5%, and 1% of the extreme side of the curve for CI bands that are expected to hit the true value 90%, 95%, and 99% of the time in the long run.

A diagram of a normal curve

Description automatically generated

**Questions about hypothesis testing**

Refer to the problem setting above i.e., health score. Our research question is: whether the health score among the low-income population in Montreal is the same as the mean score from all adults in Montreal or not. We will use a hypothesis test. Again, the mean score among all adults is 70, and the population standard deviation is 20. The sample mean obtained from low-income participants from random sampling is with *n* = 50.

1) State the null Hypothesis and the alternative hypothesis.

***H0*: thus,**

***Ha*:**

2) Do you think that the two-sided test is more appropriate than the one-sided test? If so, why?

**Most of the time, two-sided is more appropriate, as we typically are not 100% certain about the direction. There are a few exception (e.g., superiority trial, if you studied RCT already, but in this class, we assume the two-sided version as the way to go).**

3) Compute the test statistic (Z-score corresponding to your sample mean, standardized by the sampling distribution assuming that the null hypothesis is true). Show your work to derive the Z-score.

Z = = = -3.535534

4) Calculate the p-value from the resulting Z-score. Note that the value depends on whether your hypothesis is based on a 2-sided test or not.

**pnorm(q = -10/(20/sqrt(50)), mean = 0, sd = 1)= 0.000203476**

**This is the extreme lower (left) side of the table, and seems very far away from the null (center). This is one-sided, so we need to multiple as**

**p-value = 2\*P(Z < -3.535534) = 2\*0.000203476**

5) Provide an interpretation for the computed p-value.

**Assuming that the null hypothesis is true, p-value is the probability of a test statistic observing as extreme or more extreme than the obtained test statistic (Z=-3.54) in this sample.**

6) Draw the location of your Z-score on the plot of Z-distribution and shade the p-value (the area as extreme or more extreme than the obtained Z-value from the center).

**Very tiny area on the left tail of Z curve**

7) On a separate image, point to the location of the Z-score (both sides of the curve) that indicates the center 68%, 95%, and 99% of the Z-distribution.

**I have pasted this image already**

8) Based on the calculated p-value, do you think you will reject the null hypothesis or retain it at the threshold probability of rejection for 0.32, 0.05, and 0.01?

**The Z-score from our sample is much further away from the mean than any of Z-scores (called critical values) corresponding to 0.32, 0.05, and 0.01 right?**

9) ~~Provide an interpretation for the p-value~~ and state your conclusion based on the 0.05 threshold for the rejection of the null hypothesis.

**We will reject the null hypothesis based on the 0.05 threshold.**

10) Does this conclusion agree with that of 95% CI?

**Yes (see the answer for CI above)**