Indian Institute of Technology Madras

Department of Data Science and Artificial Intelligence

DA5000: Mathematical Foundations of Data Science

Tutorial VIII - Solutions

# Problem 1:

1. Given $\alpha = 0.05$. Since it is 1 tailed, $Z = -1.645 . Z = \frac{X-\mu}{\sigma'}$ $\sigma' = \frac{\sigma}{\sqrt{n}} = \frac{100}{10} = 10$

$$-1.645 = \frac{\bar{X} - 28}{10}$$

$$-16.45 = \bar{X} - 28$$

$$X_t = 11.55$$

   Now, given

$$\mu = 25$$

   . Probability of $N(25, 10^2$ providing a value greater than 11.55 is type 2 error. We can find the CDF and then find its complement to get the final value.

$$Z' = \frac{11.55 - 25}{10} = -1.845$$

   . CDF gives 0.0325. Therefore probability of type 2 error is $1 - 0.0325 = 0.91069$. Therefore there is a 91.069% chance of getting a type 2 error

2. Now when $\mu = 30$, it falls in the fail to reject region because it is above the hypothesized threshold of $H_o = 28$. Therefore the alternate hypothesis is already false. So the probability of type 2 error is 0.

3. Type 1 error is atmost given by $\alpha = 0.05$. Therefore p=0.05

# Problem 2:

1. The probability of committing a Type I error is equal to the significance level ($\alpha$). Therefore, we have:

$$P(\text{Type I Error}) = \alpha = 0.05$$

2. The power of the test is defined as $1 - \beta$, where $\beta$ is the probability of committing a Type II error. Given the power of the test is 0.80, we can calculate $\beta$ as follows:

$$\beta = 1 - \text{Power} = 1 - 0.80 = 0.20$$

3. Increasing the sample size generally decreases the probability of committing a Type II error ($\beta$) because larger samples provide more accurate estimates of the population parameters, leading to better discrimination between $H_0$ and $H_1$. However, the probability of committing a Type I error ($\alpha$) remains constant at the significance level set for the test.

4. If the significance level ($\alpha$) is reduced, the confidence interval (CI) will increase. This is because a lower $\alpha$ corresponds to a more stringent criterion for rejecting the null hypothesis, resulting in a wider interval that reflects greater uncertainty about the parameter estimate.

5.
$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Precision} = \frac{1-\beta}{1-\beta+\alpha}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$
$$\text{Recall} = 1 - \beta$$

   (a) If $\alpha$ increases, Precision decreases, and Recall is unaffected.

   (b) If $\beta$ increases, both Recall and precision decreases.

# Problem 3:

1. Let $X \sim N(\mu, \sigma^2)$ with $\sigma^2 = 0.062$. Then, the sample mean $\bar{X} \sim N(\mu, \sigma^2/36)$.

   We have hypotheses:
$$H_0 : \mu = 32 \quad \text{and} \quad H_A : \mu \neq 32$$
   with a significance level $\alpha = 0.05$.

   The rejection region is:
$$\alpha = P(\bar{X} \geq 32 + a) + P(\bar{X} \leq 32 - a) = 2P\left(\frac{\bar{X} - 32}{\sigma/6} \geq \frac{a}{\sigma/6}\right) = 2P\left(Z \geq \frac{6a}{\sigma}\right)$$
   where $Z \sim N(0, 1)$.

   Setting $\frac{6a}{\sigma} = z_{0.025} = 1.96$, we find:
$$a = \frac{1.96 \times \sigma}{6} = \frac{1.96 \times 0.062}{6} = 0.0196$$
   Thus, the rejection regions are:
$$32 \pm a = 32.0196, \quad 31.9804$$

2.
$$\text{Power}_1 = P(\bar{X} > 32.0196 \mid \mu = 31.97) + P(\bar{X} < 31.9804 \mid \mu = 31.97) = 0.8508$$
   The following are the calculations for power:
$$\text{Power}_2 = P(Z > 2.96) + P(Z < -0.96) = 0.1700$$
$$\text{Power}_3 = P(Z > 1.96) + P(Z < -1.96) = 0.05$$
$$\text{Power}_4 = P(Z > 0.96) + P(Z < -2.96) = 0.1700$$
$$\text{Power}_5 = P(Z > -1.04) + P(Z < -4.96) = 0.8508$$

3.
$$\beta = 1 - \text{Power}_5 = 0.1492$$

# Problem 4:

$$H_0 : \text{The patient does not have the disease.}$$

$$H_1 : \text{The patient has the disease.}$$

1. **Type-I and Type-II Errors:**

   In this context:

   - A **Type-I error** (false positive) occurs if the test incorrectly indicates that the patient has the disease when they actually do not. This error could lead to unnecessary worry for the patient and possibly unnecessary follow-up tests or treatments, which could have side effects or involve financial costs.

   - A **Type-II error** (false negative) occurs if the test incorrectly indicates that the patient does not have the disease when they actually do. This could have serious consequences for the patient, as they may miss the chance for early treatment, potentially leading to worse health outcomes as the disease progresses untreated.

2. **Reason for Prioritizing Minimizing Type-I Error:**

   Given the rarity of the disease, the designers of the test might prioritize minimizing the Type-I error (false positive) because the disease is rare, and a high rate of false positives would mean that a large number of people without the disease are misidentified as having it. This could lead to a substantial number of unnecessary follow-ups, causing emotional and financial burdens for many individuals.

   Prioritizing a low Type-I error rate typically increases the probability of a Type-II error (false negative), as making the test stricter to avoid false positives may result in some true cases of the disease being missed.

3. **Calculating the Test's Power:**

   The power of a test is the probability of correctly rejecting the null hypothesis when it is false, which can be expressed as:

   $$\text{Power} = 1 - \text{P(Type-II error)}.$$

   $$\text{Power} = 1 - 0.2 = 0.8.$$

   Thus, the test has a power of 0.8, meaning it has an 80% probability of correctly identifying a patient with the disease.

4. **Impact of Low Disease Prevalence on Test Reliability:**

   When the prevalence of the disease is very low (e.g., 0.1%), even a small probability of a Type-I error could lead to many false positives in absolute terms. For instance, with a low disease prevalence, the majority of positive test results could be false positives. This phenomenon is known as the *base rate fallacy*, where the low prevalence causes a positive result to be less reliable in indicating the presence of the disease.

   Therefore, in the case of a positive result, it is important to interpret it with caution. Further confirmatory testing may be necessary to verify the result, as the probability that a positive result is actually correct (the *positive predictive value*) may be low due to the rarity of the disease.

## Problem 5:

(a) $X \sim \text{Binomial}(500, P)$

$$H_0 : \mu = 0.7$$
$$H_A : \mu > 0.7$$

(b)

$$\alpha = P(X \geq 375 \mid H_0) \approx P\left(\frac{X - 500 \times 0.7}{\sqrt{500 \times 0.7 \times 0.3}} \geq \frac{375 - 500 \times 0.7}{\sqrt{500 \times 0.7 \times 0.3}} \mid H_0\right)$$
$$\approx P(Z > 2.4398)$$
$$= 0.073$$

(c)

$$\text{Power} = P(X > 375 \mid C = 0.8) \approx P\left(Z \geq \frac{375 - 500 \times 0.8}{\sqrt{500 \times 0.8 \times 0.2}}\right)$$
$$= P(Z \geq -2.80)$$
$$= 0.9974$$

(d)

$$P = P(X \geq 395 \mid H_0) \approx P\left(Z \geq \frac{395 - 500 \times 0.7}{\sqrt{500 \times 0.7 \times 0.3}}\right)$$
$$= P(Z > 4.39) < 0.0001$$