Hypothesis Testing

May 24, 2020

Required concepts from sampling theory

We start by reviewing some of the terminology involved in sampling theory.

- A random variable, X, its distribution (e.g. $N(\mu, \sigma^2)$) and one specific occurrence x
- Population vs sample
- The random variable of sample means, $\overline{X}$, its distribution (e.g. $N(\mu, \frac{\sigma^2}{n})$) and a particular sample mean $\overline{x}$.

The **population** refers to the entire group under consideration. A **sample** refers to a subset of the population. Different terms and symbols are used when we discuss characteristics of a population and a sample so it is key we are able to distinguish them.

POPULATION	SAMPLE
	n: **Sample size**
μ: **Population mean**	$\overline{x}$: **Sample mean** Also used as the **unbiased estimator of population mean**
σ^2: **Population variance**	s^2: **Unbiased estimator of population variance**
	Sample variance

$\overline{x}$ is the unbiased estimator of μ as $\mathrm{E}(\overline{X}) = \mu$ while s^2 is the unbiased estimator of σ^2 as $\mathrm{E}(S^2) = \sigma^2$.

Estimations of μ and σ^2

We have the following formulas:

$$\begin{aligned}
\overline{x} &= \frac{1}{n}\sum x \\
s^2 &= \frac{1}{n-1}\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) \\
&= \frac{n}{n-1}\left(\frac{\sum (x-\overline{x})^2}{n}\right) \\
&= \frac{n}{n-1}\,\text{(Sample Variance)}
\end{aligned}$$

Distributions of sample means and the central limit theorem

If X is normally distributed,

$$X \sim \mathrm{N}(\mu, \sigma^2)$$

and X_i refers to independent observations of X, we have the following:

$$X_1 + X_2 + \cdots + X_n \sim \mathrm{N}(n\mu, n\sigma^2)$$
$$\overline{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$$

If, however, X is not normally distributed (or we may simply not know the distribution of X), the distribution of the sample mean may not be normal.

However, if our sample size (n) is large, we can invoke the **central limit theorem**:

For any distribution X, if n is large,

$$X_1 + X_2 + \cdots + X_n \sim \mathrm{N}(n\mu, n\sigma^2) \quad \text{approximately}$$
$$\overline{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n}) \quad \text{approximately}$$

Definitions and symbols in hypothesis testing

- H_0: Null hypothesis.
- H_1: Alternative hypothesis.
- $\mu, \overline{x}, n$: Population mean, sample mean, sample size.
- σ^2, s^2: Population variance, unbiased estimate of population variance.
- $\alpha\%$: Level of significance.
  The level of significance refers to the **probability that the hypothesis test results in the rejection of the null hypothesis when in fact the null hypothesis is true**.
- p-value: Probability of obtaining a test statistic more/less/more extreme than the observed results given that H_0 is true.
- If n is large, CLT ensures that $\overline{X}$ is normally distributed approximately so X need not be (assumed to be) normally distributed.

Typical structure of a hypothesis test

The steps for carrying out a hypothesis test is as follows:

1. Define random variable (if question has not defined it).
2. Set up H_0 vs H_1.
3. State the level of significance, $\alpha\%$ and the type of test.
4. Under H_0, write down the distribution of the test statistic.
5. Calculate/write down the appropriate values ($\overline{x}, n, \sigma$ or s).
6. Get p-value using GC.
7. Compare with level of significance.
   - p-value $< \frac{\alpha}{100} \Rightarrow$ Reject H_0. Sufficient evidence at α level of significance to reject H_0 and conclude H_1.
   - p-value $\geq \frac{\alpha}{100} \Rightarrow$ Do not reject H_0. Insufficient evidence at α level of significance to reject H_0.
8. Conclude, phrasing in the context of the question.

To do a hypothesis test correct, we must

- Identify H_0, H_1 and μ_0 correctly.
- Calculate or identify $\overline{x}, n, \alpha$ and σ. If σ is not given, then we will have to calculate s.
- Determine if we need to invoke CLT.
- Type appropriate data correctly into GC.
- Present the information correctly.
- Make the appropriate conclusion in context of the question.

Types of questions

Most hypothesis test can be broken into 3 types:

1. Carry out a hypothesis test. (Method: Press GC to get p-value)
2. Given the conclusion, find α. (Method: Press GC to get p-value)
3. Given the conclusion, find $\overline{x}, n, \mu_0, \sigma$ or s. (Method: Use standardize, invNorm and the bell curve)

Example type 1. The coca-cola company claims that, on average, the volume inside a can of soft drink is 300ml. It is known that the volume inside a can of soft drink is normally distributed with a variance of 30.
A sample of 50 cans is obtained and it was noted that the total volume of liquid in the 50 cans is 14,926ml.
Test, at the 5% level of significance, whether the coca-cola company is overstating the volume.

Let X be the random variable denoting the volume in a can of soft drink.

$\mathrm{H}_0: \quad \mu = 300$
$\mathrm{H}_1: \quad \mu < 300$

We perform a left-tail Z test at the 10% level of significance.

Under H_0, test statistic, $Z = \dfrac{\overline{X} - 300}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$

$\overline{x} = \frac{14926}{50} = 298.52, \sigma = \sqrt{30}, n = 50.$

From GC, using a one-tail Z-test, p-value $= 0.0280 < 0.05 \Rightarrow \mathrm{H}_0$ rejected

Hence there is sufficient evidence at the 5% level of significance to conclude that the coca-cola company is overstating the volume.

Example type 2. The coca-cola company claims that, on average, the volume inside a can of soft drink is 300ml. It is known that the volume inside a can of soft drink is normally distributed with a variance of 30.
A test is carried out to determine if the average volume inside a can is 300ml.
A sample of 50 cans is obtained and it was noted that the total volume of liquid in the 50 cans is 14,926ml.
Given that the test determined that the average volume inside a can is not 300ml, find the range of possible values of α, the level of significance.

$\mathrm{H}_0: \quad \mu = 300$
$\mathrm{H}_1: \quad \mu \neq 300$ *(depending on question, may be $\mu > 300$ or $\mu \neq 300$.)*

We perform a two-tail Z test at the α% level of significance.

Under H_0, test statistic, $Z = \dfrac{\overline{X} - 300}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$

$\overline{x} = \frac{14926}{50} = 298.52, \sigma = \sqrt{30}, n = 50.$

From GC, using a one-tail Z-test, p-value $= 0.0560$
Since H_0 is rejected, p-value $< \frac{\alpha}{100} \Rightarrow \alpha > 5.60.$

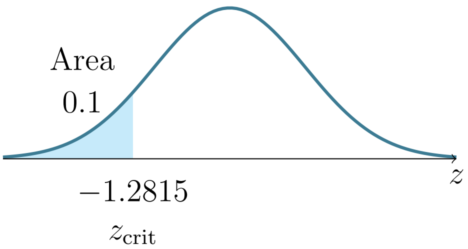
Example type 3. The coca-cola company claims that, on average, the volume inside a can of soft drink is 300ml. It is known that the volume inside a can of soft drink is normally distributed with a variance of 30.
A sample of 50 cans is obtained and it was noted that the average volume in the 50 cans is $\bar{x}$.
A test was carried out to determine whether the coca-cola company is overstating the volume. Given that the null hypothesis was rejected, what is the range of possible values for $\bar{x}$

$\mathrm{H}_0: \quad \mu = 300$
$\mathrm{H}_1: \quad \mu < 300$

We perform a left-tail Z test at the 10% level of significance.

Under H_0, test statistic, $Z = \dfrac{\overline{X} - 300}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$

$\sigma = \sqrt{30}, n = 50.$
From GC, critical value $= -1.2815$.

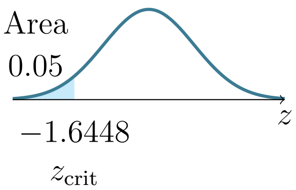
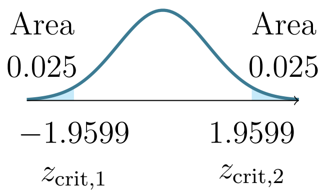
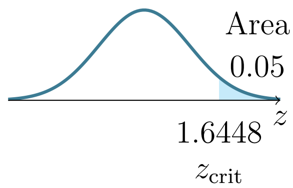


Since H_0 is rejected, $\dfrac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < -1.28$.

$\dfrac{\overline{x} - 300}{\sqrt{\frac{30}{50}}} < -1.2815.$
$\overline{x} < 299.$

Summary of question type 3

Assume level of significance of 0.05.

Tail(s)	Left	Two	Right
Picture			
Reject H_0	$\dfrac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{\text{crit}}$	$\dfrac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < -1.9599$ or $\dfrac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} > 1.9599$	$\dfrac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} > z_{\text{crit}}$
Do not reject H_0	$\dfrac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} > z_{\text{crit}}$	$-1.9599 < \dfrac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < 1.9599$	$\dfrac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{\text{crit}}$

* Warning: find the z_{crit} value(s) by invNorm using the actual question's level of significance.
** If σ^2 is unknown, we use the unbiased estimate s^2 instead.

Common key words to decide H_1

$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$
less than decreased smaller	is (equal to) is not (equal to) changed	more than increased larger
overstated		understated
at least not less than		at most not less than