DATA ANALYTICS

Class #9

Statistical Inference

Dr. Sreeja S R

Assistant Professor

Indian Institute of Information Technology

IIIT Sri City

QUOTE OF THE DAY..

Live as if you were to die tomorrow. Learn as if you were to live forever.

- MAHATMA GANDHI, father of nation of India

Errors in Hypothesis Testing

- In hypothesis testing, there are two types of errors.

Type I error: A type I error occurs when we incorrectly reject H_0 (i.e., we reject the null hypothesis, when H_0 is true).

Type II error: A type II error occurs when we incorrectly fail to reject H_0 (i.e., we accept H_0 when it is not true).

Decision	Observation	
	H_0 is true	H_0 is false
H_0 is accepted	Decision is correct	Type II error
H_0 is rejected	Type I error	Decision is correct

PROBABILITIES OF MAKING ERRORS

Type I error calculation

$\boldsymbol{\alpha}$: denotes the probability of making a Type I error

$$\boldsymbol{\alpha} = \mathbf{P}(\text{Rejecting } H_0 | H_0 \text{ is true})$$

Type II error calculation

$\boldsymbol{\beta}$: denotes the probability of making a Type II error

$$\boldsymbol{\beta} = \mathbf{P}(\text{Accepting } H_0 | H_0 \text{ is false})$$

Note:

- $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are not independent of each other as one increases, the other decreases
- When the sample size increases, both to decrease since sampling error is reduced.
- In general, we focus on Type I error, but Type II error is also important, particularly when sample size is small.

Calculating α

- Assuming that we have the results of random sample. Hence, we use the characteristics of sampling distribution to calculate the probabilities of making either Type I or Type II error.

Example 6.6:

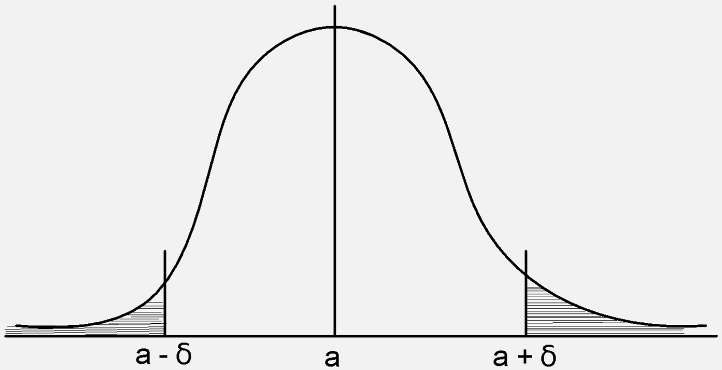
Suppose, two hypotheses in a statistical testing are:

$$H_0: \mu = a$$

$$H_1: \mu \neq a$$

Also, assume that for a given sample, population obeys normal distribution. A threshold limit say $a \pm \delta$ is used to say that they are significantly different from a.

Calculating α



Here, shaded region implies the probability that,
$\overline{X} > a + \delta \;\; or \;\; \overline{X} < a - \delta$

Thus the null hypothesis is to be rejected if the mean value is less than $a - \delta$ or greater than $a + \delta$.

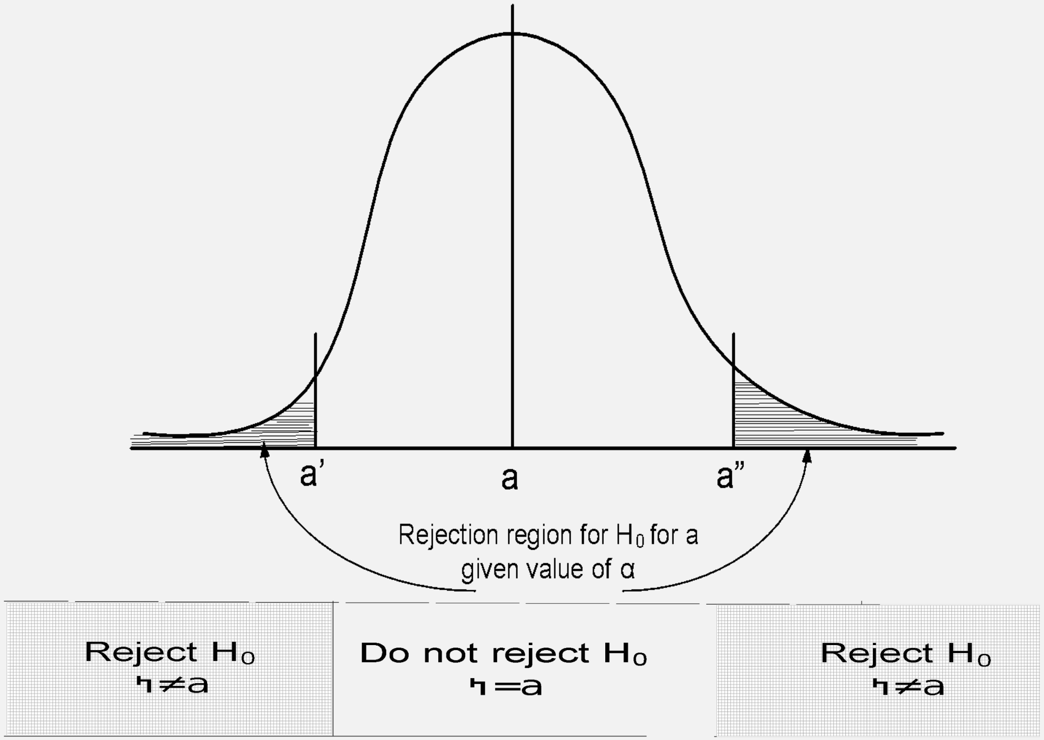
If $\overline{X}$ denotes the sample mean, then the Type I error is

$$\alpha = P(\overline{X} < a - \delta \; or \; \overline{X} > a + \delta, \qquad when \; \mu = a \,, \qquad \text{i.e.,} \; H_0 \text{ is true})$$

THE REJECTION REGION

The rejection region comprises of value of the test statistics for which

1. The probability when the null hypothesis is true is less than or equal to the specified α.
2. Probability when H_1 is true are greater than they are under H_0.



Two-Tailed Test

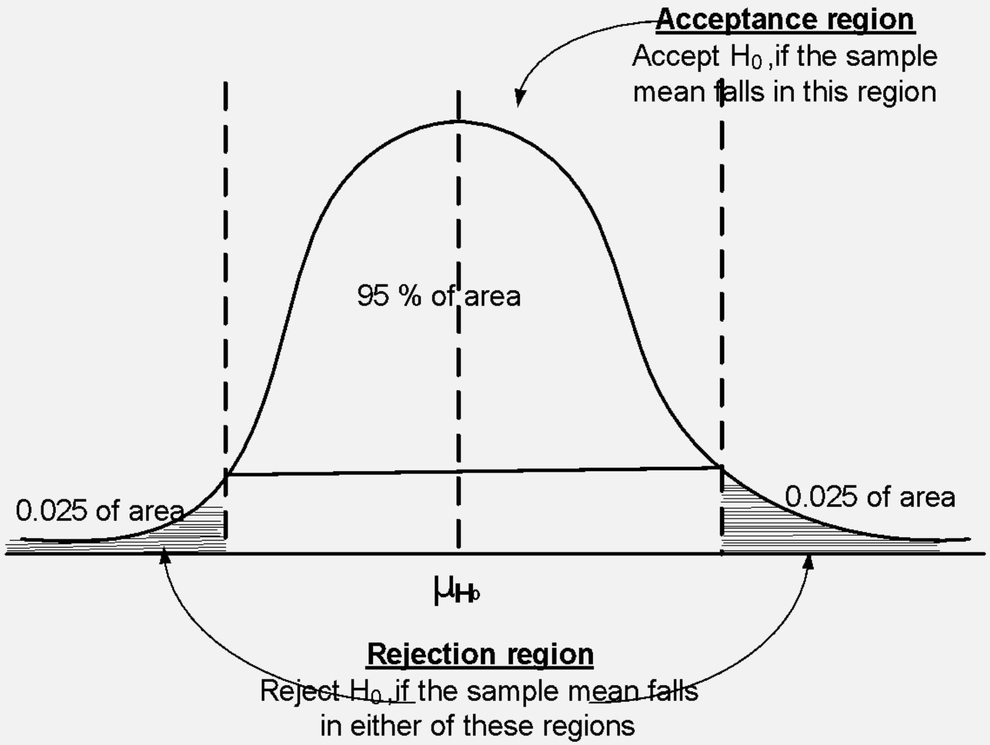
- For two-tailed hypothesis test, hypotheses take the form

$$H_0 : \mu = \mu_{H_0}$$
$$H_1 : \mu \neq \mu_{H_0}$$

In other words, to reject a null hypothesis, sample mean $\mu > \mu_{H_0}$ or $\mu < \mu_{H_0}$ under a given α.

Thus, in a two-tailed test, there are two rejection regions (also known as critical region), one on each tail of the sampling distribution curve.

Two-Tailed Test



Acceptance and rejection regions in case of a two-tailed test with 5% significance level.

One-Tailed Test

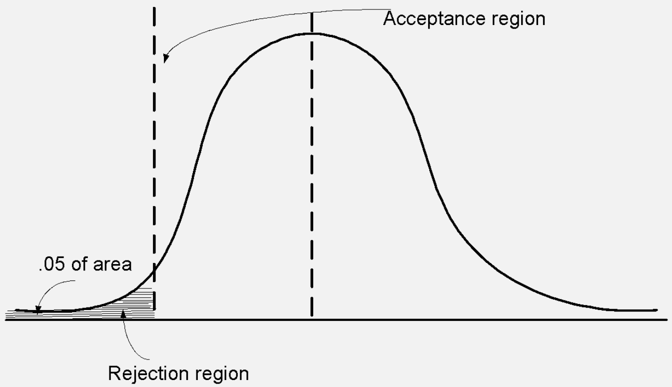
A one-tailed test would be used when we are to test, say, whether the population mean is either lower or higher than the hypothesis test value.

Symbolically,

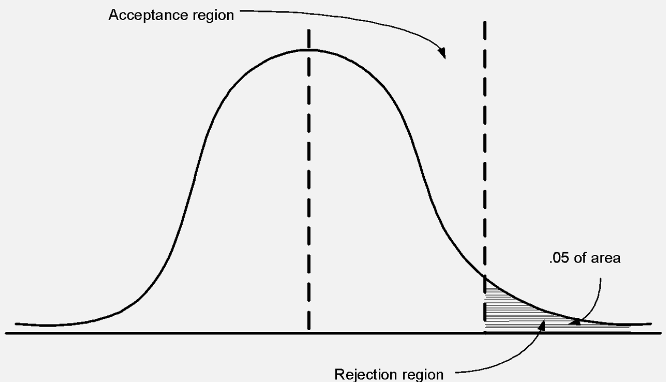
$$H_0 : \mu = \mu_{H_0}$$

$$H_1 : \mu < \mu_{H_0} \qquad\qquad [or\ \mu > \mu_{H_0}]$$

Wherein there is one rejection region only on the left-tail (or right-tail).



Left – tailed test



Right – tailed test

EXAMPLE 6.7: CALCULATING α

- Consider the two hypotheses are

The null hypothesis is

$$H_0: \mu = 8$$

The alternative hypothesis is

$$H_1: \mu \neq 8$$

Assume that given a sample of size 16 and standard deviation is 0.2 and sample follows normal distribution.

EXAMPLE 6.7: CALCULATING α

- We can decide the rejection region as follows.

Suppose, the null hypothesis is to be rejected if the mean value is less than 7.9 or greater than 8.1. If $\overline{\mathrm{X}}$ is the sample mean, then the probability of Type I error is

$$\alpha = P(\overline{\mathrm{X}} < 7.9 \; or \; \overline{\mathrm{X}} > 8.1, \text{when } \mu = 8)$$

Given σ, the standard deviation of the sample is 0.2 and that the distribution follows normal distribution.

Thus,

$$P(\overline{\mathrm{X}} < 7.9) = P\left[Z = \frac{7.9 - 8}{0.2/\sqrt{16}}\right] = P[Z < -2.0] = 0.0228$$

and

$$P(\overline{\mathrm{X}} > 8.1) = P\left[Z = \frac{8.1 - 8}{0.2/\sqrt{16}}\right] = P[Z > 2.0] = 0.0228$$

Hence, $\alpha = 0.0228 + 0.0228 = 0.0456$

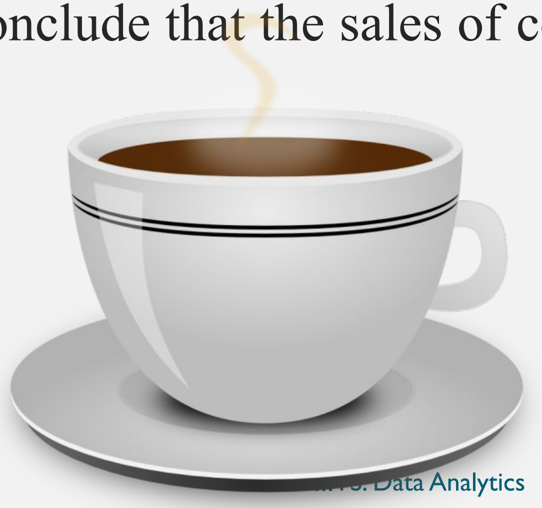
CASE STUDY 1: **COFFEE SALE**

A coffee vendor nearby Chennai central railway station has been having average sales of 500 cups per day. Because of the development of a bus stand nearby, it expects to increase its sales. During the first 12 days, after the inauguration of the bus stand, the daily sales were as under:

550 570 490 615 505 580 570 460 600 580 530 526

On the basis of this sample information, can we conclude that the sales of coffee have increased?

Consider 5% level of confidence.



HYPOTHESIS TESTING : 5 STEPS

- The following five steps are followed when testing hypothesis

1. Specify H_0 and H_1, the null and alternate hypothesis, and an **acceptable level of α.**
2. Determine an appropriate sample-based test statistics and the **rejection region** for the specified H_0.
3. Collect the sample data and calculate the test statistics.
4. Make a decision to either reject or fail to reject H_0.
5. Interpret the result in common language suitable for practitioner.

CASE STUDY 1: **STEP 1**

Step 1: Specification of hypothesis and acceptable level of α

Let us consider the hypotheses for the given problem as follows.

$H_0 : \mu = 500$ cups per day

The null hypothesis that sales average 500 cups per day and they have not increased.

$H_1 : \mu > 500$

The alternative hypothesis is that the sales have increased.

Given the acceptance level of $\alpha = 0.05$ $(i.e., 5\%\ level\ of\ significance)$

CASE STUDY 1: **STEP 2**

Step 2: Sample-based test statistics and the rejection region for specified H_0

Given the sample as

550 570 490 615 505 580 570 460 580 530 526

Since the sample size is small and the population standard deviation is not known, we shall use $t - test$ assuming normal population. The test statistics t is

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

To find $\bar{X}$ and S, we make the following computations.

$$\bar{X} = \frac{\sum x_i}{n} = \frac{6576}{12} = 548$$

CASE STUDY 1: **STEP 2**

Case Study 1: **Step 2**

$$S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{23978}{12-1}} = 46.68$$

Hence, $t = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{48}{46.68/\sqrt{12}} = \frac{48}{13.49} = 3.558$

Note:

Statistical table for t-distributions gives a *t*-value given *n*, the degrees of freedom and α, the level of significance and vice-versa.

Case Study 1: Step 3

Step 3: Collect the sample data and calculate the test statistics

$$Degree\ of\ freedom = n - 1 = 12 - 1 = 11$$

As H_1 is one-tailed, we shall determine the rejection region applying one-tailed in the right tail because H_1 is more than type) at 5% level of significance.

Case Study 1: Step 3

Step 3: Collect the sample data and calculate the test statistics

$$Degree\, of\ freedom = n - 1 = 12 - 1 = 11$$

As H_1 is one-tailed, we shall determine the rejection region applying one-tailed in the right tail because H_1 is more than type) at 5% level of significance.

Using table of $t - distribution$ for 11 degrees of freedom and with 5% level of significance,

$$R: t > 1.796$$

Case Study 1: **Step 4**

Step 4: Make a decision to either reject or fail to reject *H_0*

The observed value of $t = 3.558$ which is in the rejection region and thus H_0 is rejected at 5% level of significance.

Case Study 1: **Step 5**

Step 5: Final comment and interpret the result

We can conclude that the sample data indicate that coffee sales have increased.

CASE STUDY 2: MACHINE TESTING

- A medicine production company packages medicine in a tube of 8 ml with a standard deviation of 0.2. In maintaining the control of the amount of medicine in tubes, they use a machine. To monitor this control a sample of 16 tubes is taken from the production line at random time interval and their contents are measured precisely. The mean value of 16 tube is 7.89. The mean amount of medicine in these 16 tubes will be used to test the hypothesis that the machine is indeed working properly.

CASE STUDY 2: **STEP 1**

Step 1: Specification of hypothesis and acceptable level of α

The hypotheses are given in terms of the population mean of medicine per tube.

The null hypothesis is

$$H_0: \mu = 8$$

The alternative hypothesis is

$$H_1: \mu \neq 8$$

We assume α, the significance level in our hypothesis testing ≈ 0.05.

(This signifies the probability that the machine needs to be adjusted less than 5%).

CASE STUDY 2: **STEP 2**

Step 2: Sample-based test statistics and the rejection region for specified H_0

Rejection region: Given $\alpha = 0.05$, which gives $|Z| > 1.96$ (obtained from standard normal calculation for $n(Z: 0,1) = 0.025$ for a rejection region with two-tailed test).

CASE STUDY 2: **STEP 3**

Step 3: Collect the sample data and calculate the test statistics

Sample results: $n = 16$, $\bar{x} = 7.89$, $\sigma = 0.2$

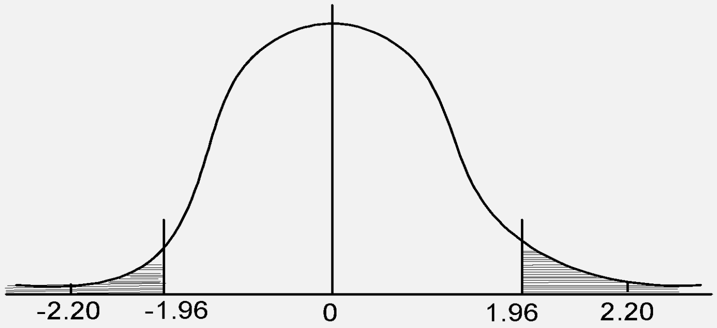
With the sample, the test statistics is

$$Z = \frac{7.89 - 8}{0.2/\sqrt{16}} = -2.20$$

Hence, $|Z| = 2.20$

CASE STUDY 2: **STEP 4**

Step 4: Make a decision to either reject or fail to reject H_0



Since $Z > 1.96$, we reject H_0

CASE STUDY 2: **STEP 5**

Step 5: Final comment and interpret the result

We conclude $\mu \neq 8$ and recommend that the machine be adjusted.

CASE STUDY 2: ALTERNATIVE TEST

Suppose that in our initial setup of hypothesis test, if we choose $\alpha = 0.01$ instead of 0.05, then the test can be summarized as:

1. $H_0 : \mu = 8$, $H_1 : \mu \neq 8$ $\alpha = 0.01$

2. Reject H_0 if $Z > 2.576$

3. Sample result n =16, $\sigma = 0.2$, $\overline{X}$=7.89, Z $= \frac{7.89-8}{0.2/\sqrt{16}} = -2.20$, $|Z| = 2.20$

4. $|Z| < 2.20$, we fail to reject H_0= 8

5. We do not recommend that the machine be readjusted.

Hypothesis Testing Strategies

- The hypothesis testing determines the validity of an assumption (technically described as null hypothesis), with a view to choose between two conflicting hypothesis about the value of a population parameter.

- There are two types of tests of hypotheses

✔ Non-parametric tests (also called distribution-free test of hypotheses)

✔Parametric tests (also called standard test of hypotheses).

Parametric Tests : Applications

- Usually assume certain properties of the population from which we draw samples.
  - Observation come from a normal population
  - Sample size is small
  - Population parameters like mean, variance, etc. are hold good.
  - Requires measurement equivalent to interval scaled data.

Parametric Tests

Important Parametric Tests

The widely used sampling distribution for parametric tests are

- $Z - test$
- $t - test$
- $\chi^2 - test$
- $F - test$

Note:

All these tests are based on the assumption of normality (i.e., the source of data is considered to be normally distributed).

Parametric Tests : Z-test

$Z - test$: This is most frequently test in statistical analysis.

- It is based on the normal probability distribution.
- Used for judging the significance of several statistical measures particularly the mean.
- It is used even when *binomial distribution* or $t - distribution$ is applicable with a condition that such a distribution tends to normal distribution when n becomes large.
- Typically it is used for comparing the mean of a sample to some hypothesized mean for the population in case of large sample, or when **population variance** is known.

Parametric Tests : t-test

-

$t - test$: It is based on the t-distribution.

- It is considered an appropriate test for judging the significance of a sample mean or for judging the significance of difference between the means of two samples in case of
  - small sample(s)
  - **population variance is not known** (in this case, we use the variance of the sample as an estimate of the population variance)

Parametric Tests : χ^2 -test

-

$\chi^2 - \boldsymbol{test}$**:** It is based on Chi-squared distribution.

- It is used for comparing a sample variance to a theoretical population variance.

Parametric Tests : $\boldsymbol{F}$ -test

$\boldsymbol{F-test}$**:** It is based on F-distribution.

- It is used to compare the variance of two independent samples.
- This test is also used in the context of analysis of variance (ANOVA) for judging the significance of more than two sample means.

Hypothesis Testing : Assumptions

Case 1: Normal population, population infinite, sample size may be large or small, variance of the population is known.

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma/\sqrt{n}}$$

Case 2: Population normal, population finite, sample size may large or small………variance is known.

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma/\sqrt{n}[\sqrt{(N-n)/(N-1)}]}$$

Case 3: Population normal, population infinite, sample size is small and variance of the population is unknown.

$$t = \frac{\bar{X} - \mu_{H_0}}{S/\sqrt{n}} \qquad with\ degree\ of\ freedom = (n-1)$$

and

$$s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{(n-1)}}$$

Hypothesis Testing

Case 4: Population finite

$$t = \frac{\bar{X} - \mu_{H_0}}{\sigma/\sqrt{n}\left[\sqrt{(N-n)/(N-1)}\right]} \; with\ degree\ of\ freedom = (n-1)$$

Note: If variance of population (σ) is known, replace S by σ. Population normal, population infinite, sample size is small and variance of the population is unknown.

Hypothesis Testing : Non-Parametric Test

- ***Non-Parametric tests***

✔ Does not work under any assumption

✔ Assumes only nominal or ordinal data

Note: Non-parametric tests need entire population (or very large sample size)

REFERENCE

- The detail material related to this lecture can be found in

  Probability and Statistics for Engineers and Scientists (8^{th} Ed.) by Ronald E. Walpole, Sharon L. Myers, Keying Ye (Pearson), 2013.

Any question?

QUESTIONS OF THE DAY…

1. In a hypothesis testing, suppose H_0 is rejected. Does it mean that H_1 is accepted? Justify your answer.

2. Give the expressions for z, t and χ^2 in terms of population and sample parameters, whichever is applicable to each. Signifies these values in terms of the respective distributions.

3. How can you obtain the value say $P(z = a)$? What this values signifies?

4. On what occasion, you should consider z-distribution but not t-distribution and vice-versa?

5. Give a situation when you should consider χ^2 distribution but neither z- nor t-distribution.

Example: Calculating α and β

There are two identically appearing boxes of chocolates. Box A contains 60 red and 40 black chocolates whereas box B contains 40 red and 60 black chocolates. There is no label on the either box. One box is placed on the table. We are to test the hypothesis that "Box B is on the table".

To test the hypothesis an experiment is planned, which is as follows:

- Draw at random five chocolates from the box.
- We replace each chocolates before selecting a new one.
- The number of red chocolates in an experiment is considered as the sample statistics.

Note: Since each draw is independent to each other, we can assume the sample distribution follows binomial probability distribution.

Calculating α

Let us express the population parameter as $p =$ the number of red chocolates in Box B.

The hypotheses of the problem can be stated as:

$$H_0: p = 0.4 \qquad \text{// Box B is on the table}$$

$$H_1: p = 0.6 \qquad \text{// Box A is on the table}$$

Calculating α:

In this example, the null hypothesis (H_0) specifies that the probability of drawing a red chocolate is 0.4. This means that, lower proportion of red chocolates in observations $(i.e., sample)$ favors the null hypothesis. In other words, drawing all red chocolates provides sufficient evidence to reject the null hypothesis. Then, the probability of making a $Type\ I$ error is the probability of getting five red chocolates in a sample of five from Box B. That is,

$$\alpha = P(X = 5 \qquad when\, p = 0.4)$$

Using the binomial distribution

$$= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}\ where\ n = 5, x = 5$$

$$= (0.4)^5 = 0.01024$$

Thus, the probability of rejecting a true null hypothesis is ≈ 0.01. That is, there is approximately 1 in 100 chance that the box B will be mislabeled as box A.

.Calculating $\boldsymbol{\beta}$

The $Type\ II$ error occurs if we fail to reject the null hypothesis when it is not true. For the current illustration, such a situation occurs, if Box A is on the table but we did not get the five red chocolates required to reject the hypothesis that Box B is on the table.

The probability of $Type\ II$ error is then the probability of getting four or fewer red chocolates in a sample of five from Box A.

That is,

$$\beta = P(X \leq 4 \qquad \text{when } p = 0.6)$$

Using the probability rule:

$$P(X \leq 4) + P(X = 5) = 1$$

That is, $P(X \leq 4) = 1 - P(X = 5)$

Now, $P(X = 5) = (0.6)^5$

Hence, $\beta = 1 - (0.6)^5$
$= 1 - 0.07776 = 0.92224$