ACADGILD

Mastering Data Science

Session 17 - Sampling Distribution and Hypothesis Testing

Agenda

1. Sampling Distribution of the Mean
2. Sampling Distribution of Two Dice
3. Difference between 2 Mean
4. Estimation
5. Point and Interval Estimation
6. Confidence Levels
7. Interval Width
8. Selecting Sample Size
9. Sample Size to Estimate Mean Size
10. Hypothesis Testing
11. Concepts of Hypothesis Testing
12. Types of Errors

Sampling Distribution of the Mean

A fair die is thrown infinitely many times, with the random variable X = # of spots on any throw.

The probability distribution of X is:

x	1	2	3	4	5	6
P(x)	1/6	1/6	1/6	1/6	1/6	1/6

and the mean and variance are calculated as well:

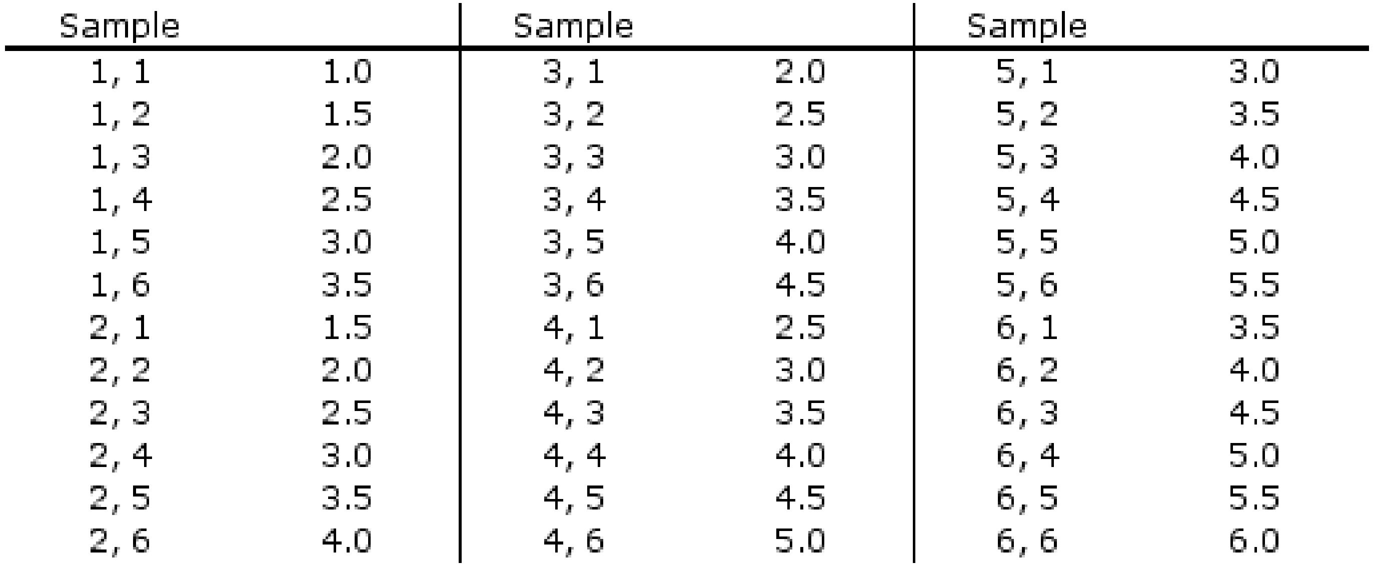
$$\mu = \sum xP(x) = 1(\tfrac{1}{6}) + 2(\tfrac{1}{6}) + \ldots + 6(\tfrac{1}{6}) = 3.5$$

$$\sigma^2 = \sum (x-\mu)^2 P(x) = (1-3.5)^2(\tfrac{1}{6}) + \ldots + (6-3.5)^2(\tfrac{1}{6}) = 2.92$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{2.92} = 1.71$$

Sampling Distribution of Two Dice

A sampling distribution is created by looking at all samples of size n=2 (i.e. two dice) and their means



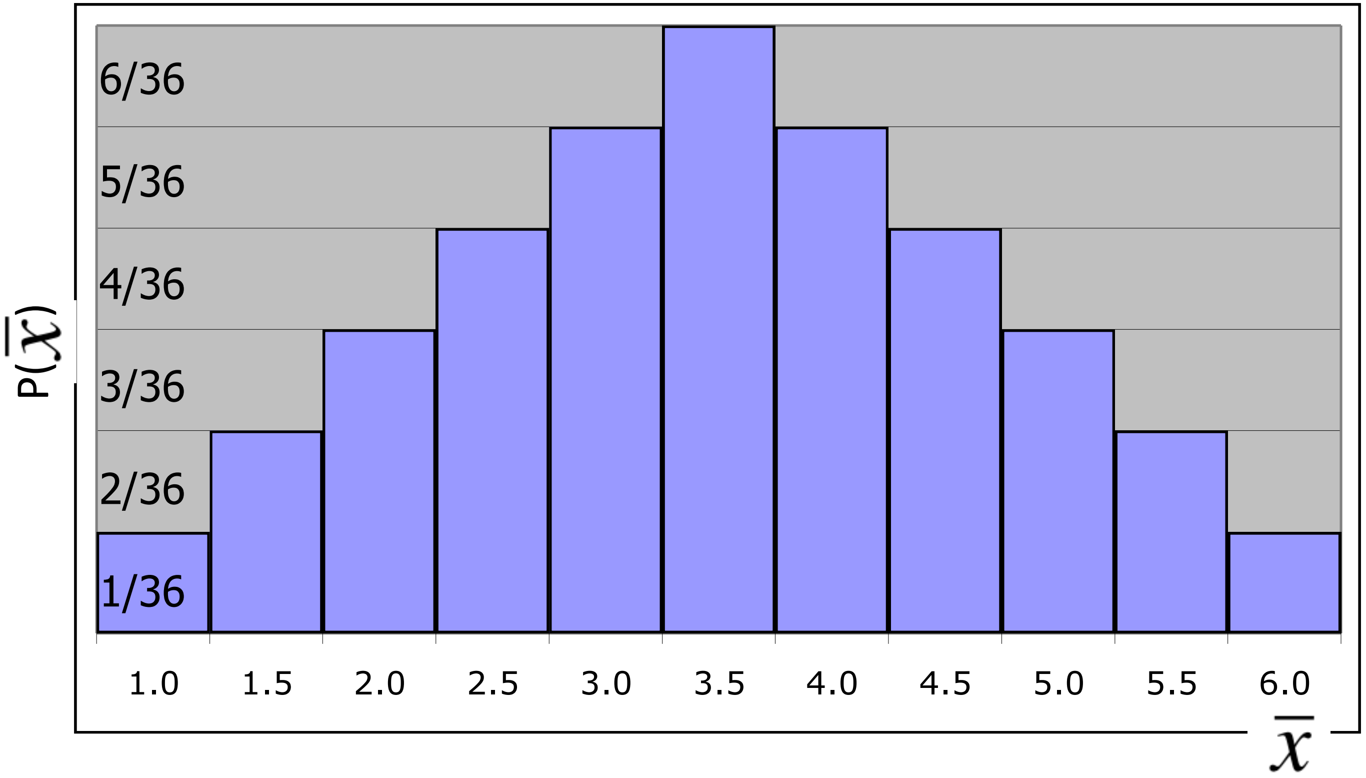
Sample		Sample		Sample	
1, 1	1.0	3, 1	2.0	5, 1	3.0
1, 2	1.5	3, 2	2.5	5, 2	3.5
1, 3	2.0	3, 3	3.0	5, 3	4.0
1, 4	2.5	3, 4	3.5	5, 4	4.5
1, 5	3.0	3, 5	4.0	5, 5	5.0
1, 6	3.5	3, 6	4.5	5, 6	5.5
2, 1	1.5	4, 1	2.5	6, 1	3.5
2, 2	2.0	4, 2	3.0	6, 2	4.0
2, 3	2.5	4, 3	3.5	6, 3	4.5
2, 4	3.0	4, 4	4.0	6, 4	5.0
2, 5	3.5	4, 5	4.5	6, 5	5.5
2, 6	4.0	4, 6	5.0	6, 6	6.0

While there are 36 possible samples of size 2, there are only 11 values for $\overline{x}$, and some (e.g. $\overline{x}$ =3.5) occur more frequently than others. (for example, e.g. $\overline{x}$ =1).

Sampling Distribution of Two Dice

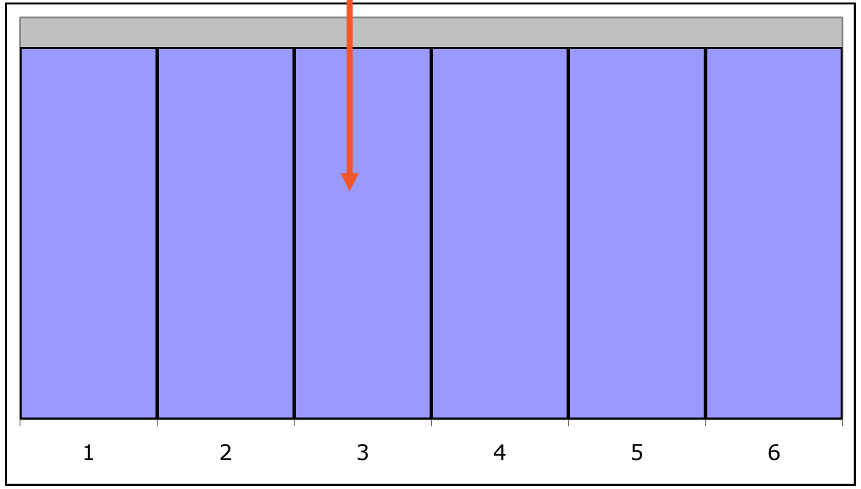
The Sampling Distribution of $\bar{x}$ is shown below:

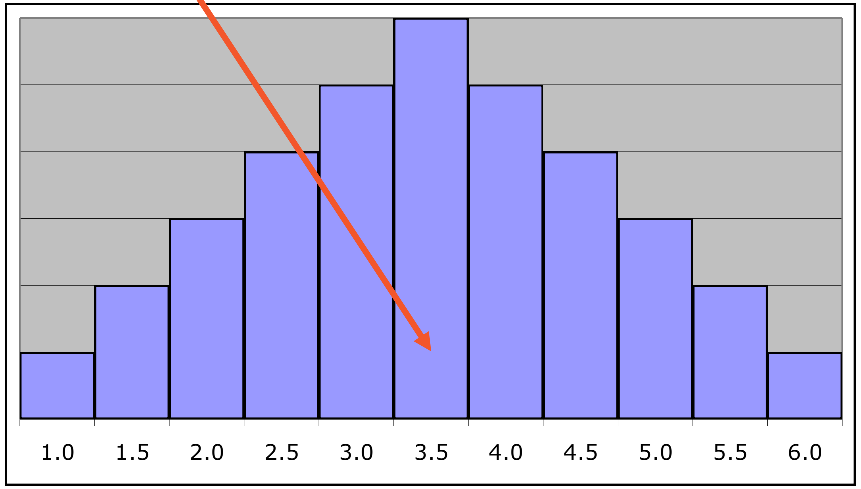
$\bar{x}$	$P(\bar{x})$
1.0	1/36
1.5	2/36
2.0	3/36
2.5	4/36
3.0	5/36
3.5	6/36
4.0	5/36
4.5	4/36
5.0	3/36
5.5	2/36
6.0	1/36



Compare

Compare the distribution of X with the sampling distribution of $\bar{x}$





As well as note that: $\mu_{\bar{x}} = \mu$

$$\sigma^2_{\bar{x}} = \sigma^2 / 2$$

Sampling Distribution – Difference between Two Means

- The final sampling distribution introduced is that of the difference between two sample means. This requires:
  - Independent random samples be drawn from each of two normal populations
- If this condition is met, then the sampling distribution of the difference between the two sample means, i.e. $\bar{X}_1 - \bar{X}_2$ will be normally distributed.

Note: If the two populations are not both normally distributed, but the sample sizes are “large” (>30), the distribution of $\bar{X}_1 - \bar{X}_2$ is approximately normal)

Sampling Distribution – Difference between Two Means

- The expected value and variance of the sampling distribution of $\bar{X}_1 - \bar{X}_2$ are given by:

  Mean, $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$

  Standard deviation = $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Also called the standard error if the difference between two means)

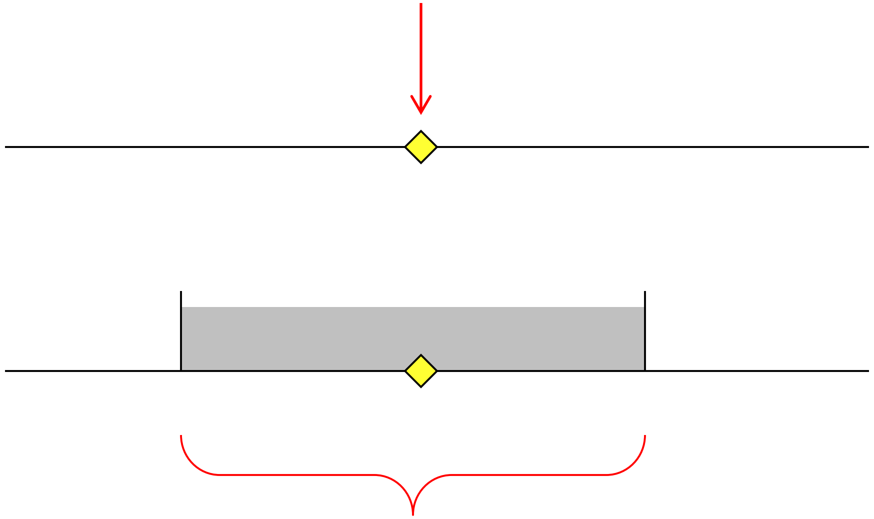
Estimation

- Estimation and Hypothesis Testing are the two types of Inferences. Whereas Estimation is introduced first.
- The objective of Estimation is to determine the approximate value of a population parameter on the basis of a sample statistic.
- E.g., the sample mean ($\overline{x}$) is employed to estimate the population mean (μ).

Estimation

- The objective of estimation is to determine the approximate value of a population parameter on the sample statistic basis.

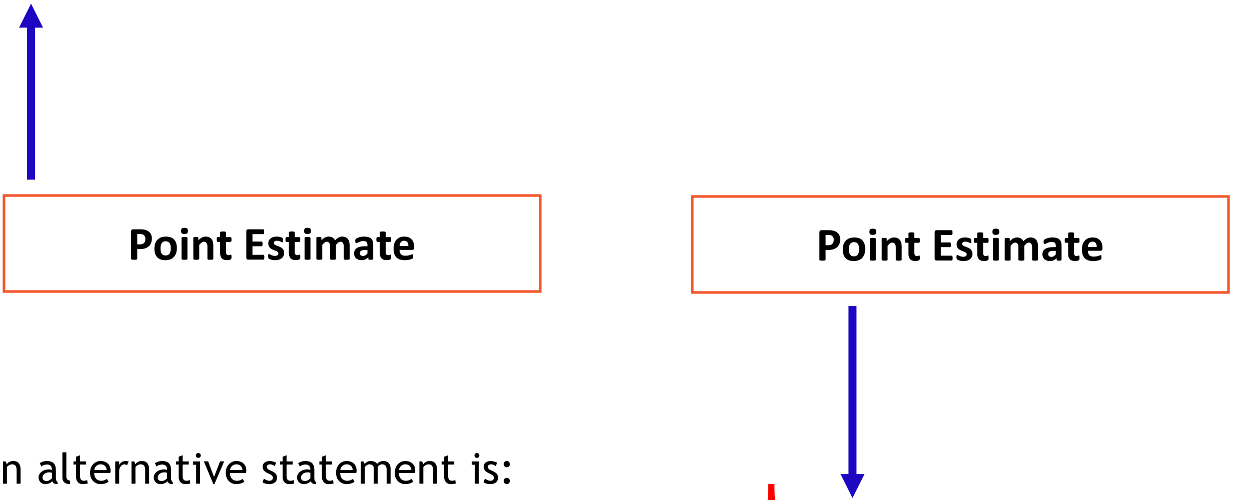
- There are two types of estimators:
  - Point Estimator
  - Interval Estimator



Point and Interval Estimation

- For example, suppose we want to estimate the mean summer income of a class of business students. For n=25 students,

  $\bar{x}$ is calculated to be $400/week.



Point Estimate

Point Estimate

An alternative statement is:

The mean income is ***between*** $\overbrace{\text{380 and 420 \$/week}}$

Estimating μ when σ is known

- We established

The Confidence Interval

$$\overbrace{P(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \bar{x} \le \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha}$$

The Sample Mean is the center of the Interval

- Thus the Probability that the interval is, $\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \left\{\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}$

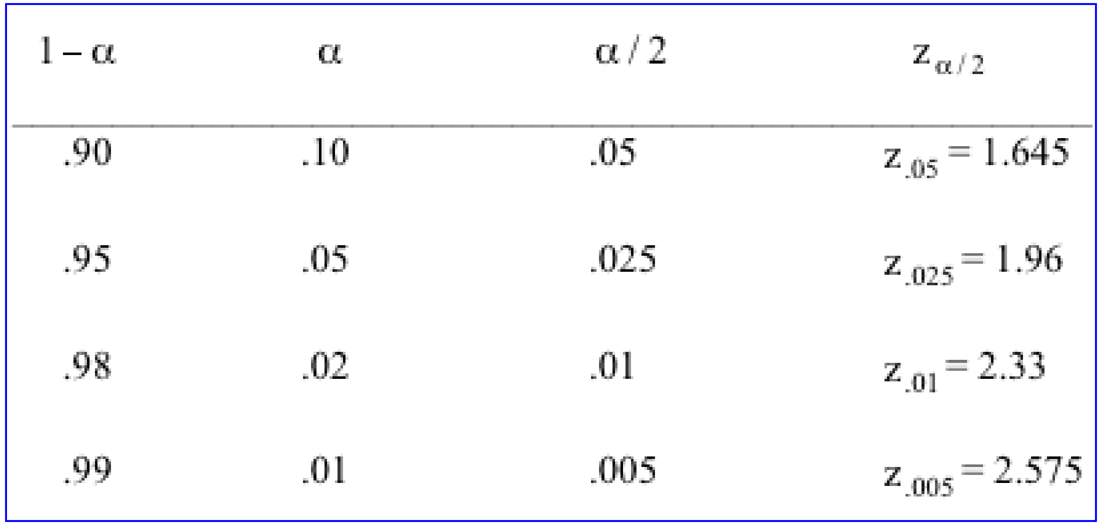
contains the population mean μ is 1– α . This is a ***confidence interval estimator for*** μ .

Commonly Used Confidence Levels

There are four commonly used Confidence Levels.

Confidence Level

↓



$1-\alpha$	α	$\alpha/2$	$z_{\alpha/2}$
.90	.10	.05	$z_{.05} = 1.645$
.95	.05	.025	$z_{.025} = 1.96$
.98	.02	.01	$z_{.01} = 2.33$
.99	.01	.005	$z_{.005} = 2.575$

Example

A computer company samples demand during lead time over 25 time periods:

235	374	309	499	253
421	361	514	462	369
394	439	348	344	330
261	374	302	466	535
386	316	296	332	334

It is known that the standard deviation of demand over lead time is 75 computers. We want to estimate the mean demand over lead time with 95% confidence in order to set inventory levels

Example

In order to use our confidence interval estimator, we need the following pieces of data:

$\bar{x}$	370.16
$z_{\alpha/2}$	1.96
σ	75
n	25

Calculated from the data…

$$1-\alpha = .95, \;\therefore \alpha/2 = .025$$

$$so \;\; z_{\alpha/2} = z_{.025} = 1.96$$

Given

Therefore, the **lower** and **upper** confidence limits are 340.76 and 399.56.

Example - Interpretation

- The estimation for the mean demand during lead time lies between 340.76 and 399.56 — we can use this as input in developing an inventory policy.
- That is, we estimated that the mean demand during lead time falls between 340.76 and 399.56, and this type of estimator is 95% accurate of the time. That also means that 5% of the time the estimator will be incorrect.
- Incidentally, the media often refer to the 95% figure as “19 times out of 20,” which emphasizes the long-run aspect of the confidence level.

Interval Width

A wide interval provides little information.

For example, suppose we estimate with 95% confidence that an accountant’s average starting salary is between $15,000 and $100,000.

In **Contrast** with this: a 95% confidence interval estimate of starting salaries between $42,000 and $45,000.

The second estimate is much narrower, providing accounting students more precise information about starting salaries.

Interval Width

The width of the confidence interval estimate is a function of the **confidence level**, the **population standard deviation**, and the **sample size**…

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Selecting the Sample Size

We can control the width of the interval by determining the sample size necessary to produce narrow intervals.

Suppose we want to estimate the mean demand “to within 5 units”; i.e. we want to the interval estimate to be: $\bar{x} \pm 5$

Since: $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

It follows that $z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 5$

Sample Size to Estimate Mean Size

The general formula for the sample size needed to estimate a population mean with an interval estimate of: $\bar{x} \pm W$

Requires a sample size of at least this large: $n = \left(\dfrac{z_{\alpha/2}\sigma}{W} \right)^2$

Example

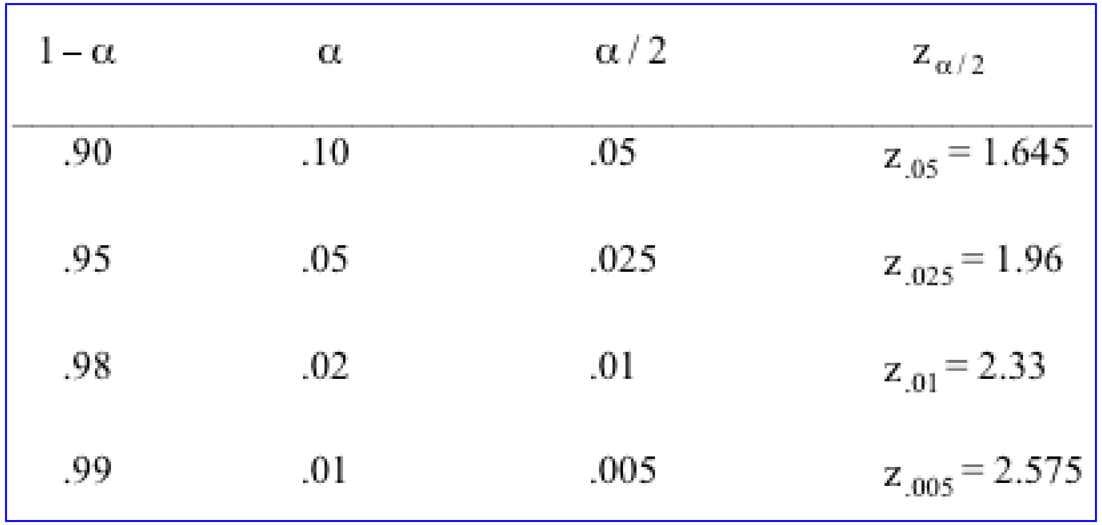
A lumber company must estimate the mean diameter of trees to determine whether or not there is sufficient lumber to harvest an area of forest. They need to estimate this to within 1 inch at a confidence level of 99%. The tree diameters are normally distributed with a standard deviation of 6 inches.

How many trees need to be sampled?

Example

Things we know:

Confidence level = 99%, therefore α =.01



$1-\alpha$	α	$\alpha/2$	$z_{\alpha/2}$
.90	.10	.05	$z_{.05} = 1.645$
.95	.05	.025	$z_{.025} = 1.96$
.98	.02	.01	$z_{.01} = 2.33$
.99	.01	.005	$z_{.005} = 2.575$

$$z_{\alpha/2} = z_{.005} = 2.575$$

We want $\bar{x}$ ± 1, Hence W = 1.

We are given that σ = 6.

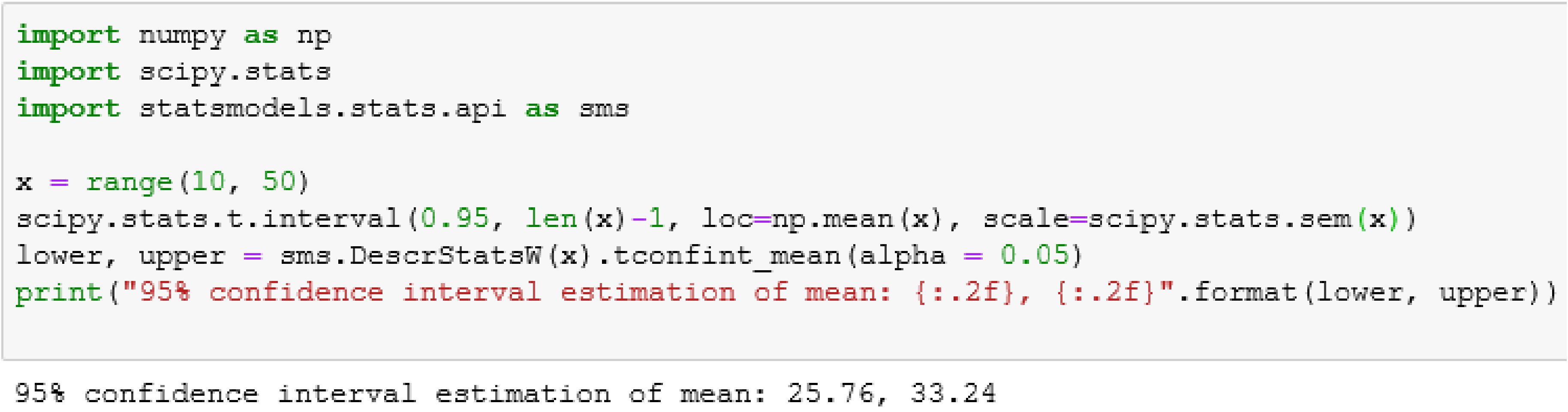
Example

We Compute,

$$n = \left(\frac{z_{\alpha/2}\sigma}{W}\right)^2 = \left(\frac{(2.575)(6)}{1}\right)^2 = 239$$

That is, we will need to sample at least 239 trees to have a 99% confidence interval of $\bar{x} \pm 1$

Mean Interval Estimation using Python



```python
import numpy as np
import scipy.stats
import statsmodels.stats.api as sms

x = range(10, 50)
scipy.stats.t.interval(0.95, len(x)-1, loc=np.mean(x), scale=scipy.stats.sem(x))
lower, upper = sms.DescrStatsW(x).tconfint_mean(alpha = 0.05)
print("95% confidence interval estimation of mean: {:.2f}, {:.2f}".format(lower, upper))
```

```
95% confidence interval estimation of mean: 25.76, 33.24
```

[Practice Code]

Hypothesis Testing

- ➢ A criminal trial is an example of hypothesis testing without the statistics.
- ➢ In a trial a jury must decide between two hypotheses. The null hypothesis is
  - H_0: The defendant is innocent
- ➢ The alternative hypothesis or research hypothesis is
  - H_1: The defendant is guilty

The jury does not know which hypothesis is true. They must make a decision on the basis of evidence presented.

Hypothesis Testing

There are two possible errors:

- Type I Error
  - A Type I error occurs when we reject a true null hypothesis. That is, a Type I error occurs when the jury convicts an innocent person.
  - The probability of Type I error is denoted as α (Greek Letter Alpha)
- Type II Error
  - A Type II error occurs when we don’t reject a false null hypothesis. That occurs when a guilty defendant is acquitted.
  - The probability of Type II error is denoted as β (Greek Letter Beta)

The two probabilities are inversely related. Decreasing one increases the other.

Concepts of Hypothesis Testing

There are two hypotheses:

- Null hypothesis (H_0)
- Alternative or research hypothesis (H_1)

The null hypothesis (H_0) will always states that the parameter equals the value specified in the alternative hypothesis (H_1).

Concepts of Hypothesis Testing

Consider example, Mean demand for computers during assembly lead time again. Rather than estimate the mean demand, our operations manager wants to know whether the mean is different from 350 units. We can rephrase this request into a test of the Hypothesis:

H_0: μ = 350

Thus our research hypothesis becomes:

H_1: μ ≠ 350

Concepts of Hypothesis Testing

There are two possible decisions that can be made:

1. Conclude that there is enough evidence to support the alternative hypothesis
   - Also stated as, rejecting the null hypothesis in favor of the alternative hypothesis
2. Conclude that there is not enough evidence to support the alternative hypothesis
   - Also stated as, not rejecting the null hypothesis in favor of the alternative hypothesis

Note: We do not say that we accept **Null Hypothesis**

Concepts of Hypothesis Testing

Once the null and alternative hypotheses are stated, the next step is to randomly sample the population and calculate a test statistic (In this example, sample mean)

If the test statistic’s value is inconsistent with the null hypothesis we reject the null hypothesis and infer the alternate hypothesis is true.

For example, if we’re trying to decide whether the mean is not equal to 350, a large value of $\bar{x}$ (say, 600) would provide enough evidence. If $\bar{x}$ is close to 350 (say, 355) we could not say that this provides a great deal of evidence to infer that the population mean is different than 350.

Concepts of Hypothesis Testing

A Type I error occurs when we reject a true null hypothesis (i.e. Reject H_0 When it is True)

A Type II error occurs when we don’t reject a false null hypothesis (i.e. Do not Reject H_0 When it is False)

H_0	T	F
Reject	I	
🚫 Reject		II

Email us - support@acadgild.com