Lesson 2: Point and Interval Estimates of the Population Mean, Student's t-distribution, Sample Size and Biases

A **point estimate** involves the use of sample data to calculate a single value (a statistic) that serves as an approximation for an unknown population parameter.

- Example: The sample mean, is a point estimate of the population mean, μ.

The formula used to calculate a point estimate is known as an estimator.

A **confidence interval** uses sample data to calculate a range of possible (or probable) values that an unknown population parameter can take, with a given of probability of (1-α).

- α is called the *level of significance*
- (1 – α) refers to the degree of confidence that the relevant parameter will lie in the computed interval.
- Example: A calculated interval between 100 and 150 at the 5% significance level implies that we can be 95% confident that the population parameter will lie between 100 and 150.

Unbiasedness: An unbiased estimator is one whose expected value is equal to the parameter being estimated.

- The expected value of the sample mean equals the population mean.

Efficiency: An efficient unbiased estimator is the one that has the *lowest* variance among all unbiased estimators of the same parameter.

Consistency: A consistent estimator is one for which the probability of estimates close to the value of the population parameter increases as sample size increases.

- The standard error of the sampling distribution falls as sample size increases, which implies a higher probability of estimates close to the population mean.

Student's t-distribution is a bell-shaped probability distribution that has the following properties:

- It is symmetrical.
- It is defined by a single parameter, the degrees of freedom (df), where degrees of freedom equal sample size minus one (n-1).
- It has a lower peak than the normal curve, but fatter tails.
- As the degrees of freedom increase, the shape of the t-distribution approaches the shape of the standard normal curve.

Important

As the degrees of freedom increase, the t-distribution curve becomes more peaked and its tails become thinner (bringing it closer to a normal curve).

- For a given significance level, the confidence interval for a random variable that follows the t-distribution will become narrower when the degrees of freedom increase.
- We will be *more* confident that the population mean will lie within the calculated interval as more data is concentrated towards the middle (as demonstrated by the higher peak) and less data is in the tails (thinner tails).

The t-distribution is used in the following scenarios:

- It is used to construct confidence intervals for a *normally* (or approximately normally) distributed population whose variance is *unknown* when the sample size is small (n < 30).
- It may also be used for a *non-normally* distributed population whose variance is *unknown* if the sample size is *large.* In this case, the central limit theorem is used to assume that the sampling distribution of the sample mean is approximately normal.

The following reliability factors are used frequently when constructing confidence intervals based on the standard normal distribution:

- For a 90% confidence interval we use $z_{0.05} = 1.65$
- For a 95% confidence interval we use $z_{0.025} = 1.96$
- For a 99% confidence interval we use $z_{0.005} = 2.58$

36 students are taking a mock SAT exam to evaluate their level of preparedness for the actual test. The average score of these students is 1750. The standard deviation of scores of all students (population) who take the actual test is 200 points. Construct and interpret a 99% confidence interval for the average score of all students who take the SAT given the average score of these 36 students.

Example

A sample of the monthly returns of Treptash Ltd. stock over the last two and a half years has a mean return of 3% and a standard deviation of 15%. Compute the 95% confidence interval for the average monthly returns on Treptash stock.

One tail						
df	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0005**
Two tail						
df	**0.2**	**0.1**	**0.05**	**0.02**	**0.01**	**0.001**
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728
16	1.3368	1.7459	2.1199	2.5835	2.9208	4.0150
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.9651
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.9216
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.8834
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.8495
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.8193
22	1.3212	1.7171	2.0739	2.5083	2.8188	3.7921
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.7676
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.7454
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.7251
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.7066
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.6896
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.6739
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.6594
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.6460

When the population is *normally* distributed:

- Use the z-statistic when the population variance is known.
- Use the t-statistic when the population variance is not known.

When the distribution of the population is *nonnormal,* the construction of an appropriate confidence interval depends on the size of the sample.

- If the population variance is *known* and the sample size is large we use the z-statistic. This is because the central limit theorem tells us that the distribution of the sample mean is approximately normal when sample size is large.
- If the population variance is *not known* and sample size is large, we can use the z-statistic or the t-statistic. However, in this scenario the use of the t-statistic is encouraged because it results in a more conservative measure.

This implies that we cannot construct confidence intervals for nonnormal distributions if sample size is less than 30.

Criteria for Selecting Appropriate Test Statistic

When Sampling From a:	**Small Sample**	**Large Sample**
Normal distribution with known variance	z-statistic	z-statistic
Normal distribution with unknown variance	t-statistic	t-statistic*
Nonnormal distribution with known variance	not available	z-statistic
Nonnormal distribution with unknown variance	not available	t-statistic*

* *Use of z-statistic also acceptable*

Data mining is the practice of developing a model by extensively searching through a data set for statistically significant relationships until a pattern "that works" is discovered.

Data-mining bias most commonly occurs when:

- Researchers have not formed a hypothesis in advance, and are therefore open to any hypothesis suggested by the data.
- When researchers narrow the data used in order to reduce the probability of the sample refuting a specific hypothesis.

Warning signs

- *Too much digging warning sign*, which is indicated by a lack of an economic theory that can explain empirical results.
- *No story/ no future warning sign*, which involves testing numerous variables until one that appears to be significant is discovered.

The best way to avoid the data-mining bias is to test the 'apparently statistically significant relationships' on 'out-of-sample' data to check whether they continue to hold.

Sample selection bias results from the exclusion of certain assets (such as bonds, stocks or portfolios) from a study due to the unavailability of data.

- Survivorship bias is present in databases that only list companies or funds currently in existence, which means that those that have failed are not included in the database.

Look-ahead bias arises when a study uses information that was not available on the test date.

Time-period bias arises if a test is based on a certain time period, which may make the results obtained from the study time-period specific.