Basics of Mechanical Engineering-3

Prof. J. Ramkumar

Prof. Amandeep Singh Oberoi

Department of Mechanical Engineering

Indian Institute of Technology, Kanpur

Week

Lecture 52: Basics of Inferential Statistics

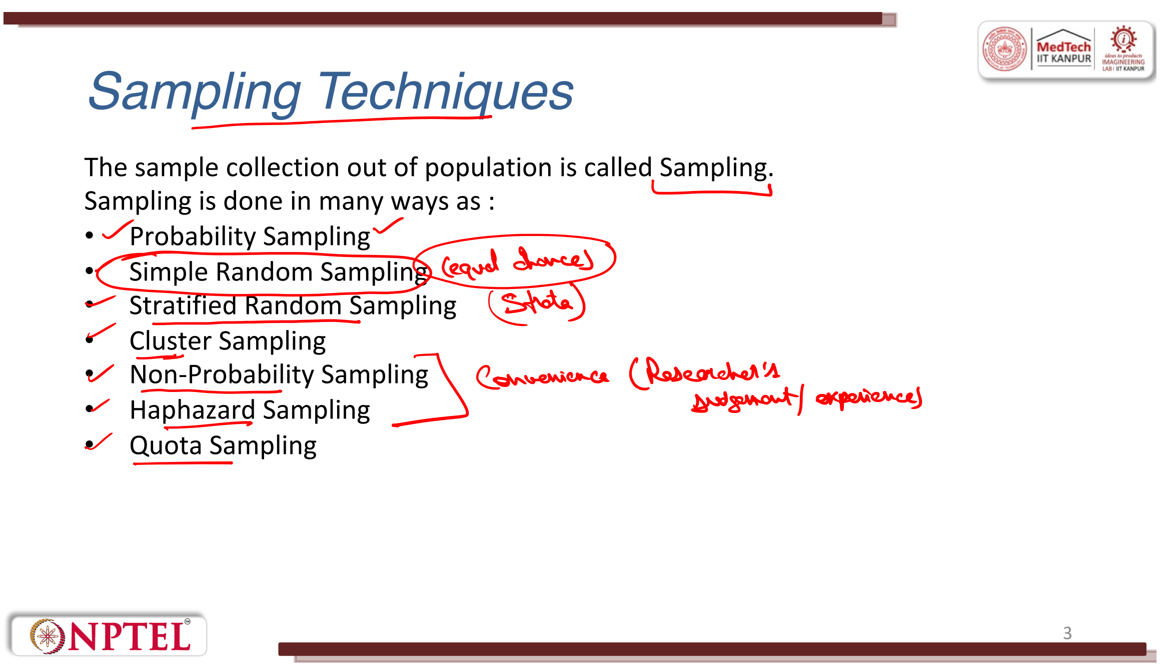
Welcome to the third lecture in week 12 of the course Basics of Mechanical Engineering 3. This week, we are talking about statistics in mechanical engineering. We have talked about the general introduction to statistics, its limitations, and the kinds of statistics. We talked about descriptive statistics and data presentation methods, and how to present pictorial or graphical data in part two. In part three, we'll talk about the basics of inferential statistics. As we discussed, inferential statistics makes assumptions and tests whether those assumptions are true or not. That is, we define and then work on the hypothesis.

Contents

- ✓ Sampling techniques
- ✓ Inferential Statistics
- ✓ Probability Distributions
- ✓ Hypothesis testing
- ✓ Type I and Type II errors

In this lecture, the slides will flow like this. We'll talk about sampling techniques, providing a general discussion about the kinds of sampling techniques we will cover. We talked about random sampling and some other methods. We will also briefly explore other kinds. Inferential statistics, we'll provide more information on that.

This lecture is mainly focused on inferential statistics, what we need to do, and the steps in investigations. For probability distribution, we will only discuss Poisson, binomial, and normal distributions. A quick introduction to Poisson and binomial, as these are also used. The most widely used distribution is the normal distribution, which we will discuss in a little detail. Then we will try to see what hypothesis testing is, and what type 1 and type 2 errors are. Type 1 and type 2 errors—this part we will also see.



First is sampling techniques. Sample collection from a population is called sampling. Sampling is done in many ways. Probability sampling, simple random sampling, stratified, cluster, non-probability, haphazard, or quota sampling. Probability sampling is when each and every unit of the population has a known chance, but not necessarily a non-zero chance, of being selected. For example, there is a big bucket of nozzles that are manufactured. There are maybe 5000 nozzles.

You randomly select 5 out of them. So this is probability sampling. Simple random sampling is very similar to probability sampling. It says it has an equal chance. The example I gave is also an example of simple random sampling. And we pick, for example, 5 pieces out of the 100 sets in a basket. Then comes stratified random sampling. Stratified random sampling means we divide the overall population into strata. For example, there are 1000 pieces in the bucket. We do not keep all 1000 in one group.

We just keep each hour—suppose for the 10 hours the production is happening—for the first hour, we keep the first 100 in one box; for the second hour, we put the second 100 in another box; for the third hour, we keep the third 100 in another box. Total 10 boxes with 100 pieces each.

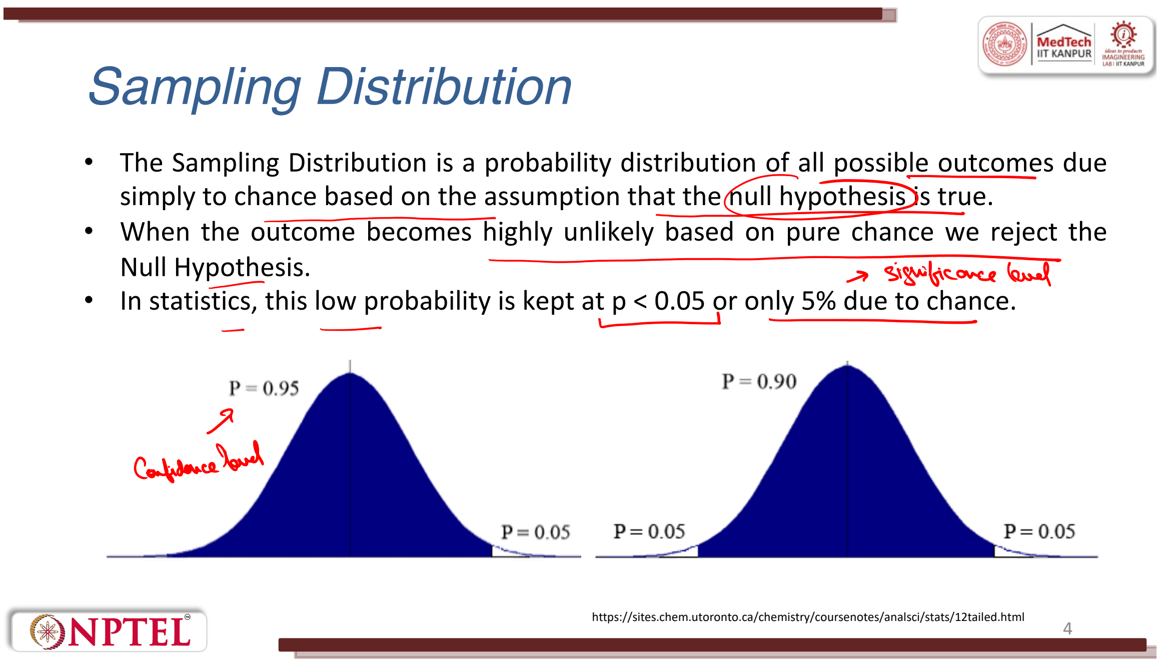
Then we pick, suppose, one sample from each of the boxes. This is known as stratified sampling. From each stratum, it is given. If the box size is bigger or, in a way, if the number of units in that box is more, accordingly or proportionally, we will pick. For example, one box has 200 units. We will pick two out of them. Another box has 100 units. We will pick one out of them as a sample.

This is stratified. Then comes cluster sampling. Clusters are made based on the kinds of similarity between specific units. For instance, there are two different machines manufacturing a specific set of nozzles. For one machine, it could be one cluster. For the second machine, it could be a second cluster. For the third machine, it could be a third cluster.

Or, if it is coming from different factories, it could be in separate clusters. So, that is how a cluster is formed. We will specifically pick from a specific set of units. Non-probability sampling is exactly the opposite of probability sampling. That means not all units have a chance of being selected. It depends on the researcher's judgment or convenience. So, both non-probability and haphazard sampling are known as convenience sampling. It depends on the researcher's judgment or experience.

Then comes quota sampling. Quota sampling could be when a specific quota is given to a specific set. For instance, I talked about cluster sampling given different machines. Now, for instance, there are workers who are working: workers with more than 10 years of experience and workers with less than 5 years of experience. 40% of workers have less than 5 years of experience.

60% are, for example, workers with more than 10 years of experience. Accordingly or proportionally, I will pick the sample size based on which workers have worked on what kind of machines. So, that is quota sampling. Generally, we will talk about simple random sampling only, which is equal chance of occurrence or equal chance of being selected.



Sampling distribution. Sampling distribution is a probability distribution of all possible outcomes due simply to chance, based on the assumption that the null hypothesis is true. I will talk about what the null hypothesis is; sampling distribution is what we are talking about.

When the outcome becomes highly unlikely based on pure chance, we reject the null hypothesis. In statistics, this low probability is generally kept at around 5%, that is, 5% due to chance. Now, if the probability is 5%, this is also known as the significance level. This P, that is, 1 - 5%, is the confidence level.

I'll talk about what the null hypothesis is. I'm talking about confidence and significance levels. While defining a problem or while working on a sample, because we're trying to pick some number of units only from the overall population—out of maybe 10,000 units—we have picked only 20 units to be tested. These 20 units should represent what the overall 10,000 units are behaving like. For that, how confident are we? We keep that confidence level. Whatever we do, there could be a small error. But if we are 95% confident, we set our hypothesis accordingly. Or maybe at a 99% confidence level. This, too—99% confidence means a 1% significance level. I will try to talk about it more, and it will be clearer when I discuss the normal distribution.

Probability

- Probability is the likelihood of the occurrence of some desired result or outcome from the statistical operation.
- If the probability comes out to be low, we reject the possibility of random error.
- A significant result is one that is very unlikely if the null hypothesis is correct.
- α level: probability required for significance.
- Most common α level is $P < 0.05$.
- If there is less than 5% chance that the results were due to random error then the results are considered to be statistically significant.
- There are three types of probabilities:
  1. Poisson Probability
  2. Binomial Probability
  3. Normal Probability

NPTEL

Probability. Just to give a quick definition once again: it is the likelihood of occurrence of some desired result or outcome from the statistical operation. If the probability comes out to be low, we reject the possibility of random error. A significant result is one that is very unlikely if the null hypothesis is correct. An alpha-level probability is required for significance. The most common alpha level is 5%.

If there is less than a 5% chance that the results were due to random error, then the results are considered statistically significant. There are three types of probabilities mainly that we are considering in this lecture: Poisson, binomial, and normal.

Probability

1. Poisson probability

- In Poisson probability distribution the chance of a given number of events occurring in a fixed time interval is calculated, assuming that they occur at a constant average rate and independent of each other.

Examples:

- Number of motor accidents at a specific juncture per month.
- Occurence of powercuts in a week or month.
- Number of calls placed by a call-service employee per hour.

Poisson probability or Poisson probability distribution describes the chance of a given number of events occurring in a fixed time interval. This assumes that they occur at a constant average rate and are independent of each other. See the examples.

The number of motor accidents at a specific junction per month. This could have a Poisson distribution. Occurrences of power cuts in a week or a month. Poisson distribution. The number of calls placed by a call service employee per hour. That is within a specific time interval. The number of occurrences happening. That is Poisson distribution, Poisson probability.

Probability

2. Binomial probability

- It is the probability distribution with n parameters, asking a $\boxed{\text{Yes/No}}$ question and each with its own Boolean-valued outcome: success with probability p or failure with probability q = (1−p).
- The binomial distribution is the basis for the binomial test of statistical significance.

Examples:

- Calculating probability of getting desired successful vs unsuccessful results.
- Measuring the chance of getting one result over another in defined experiments

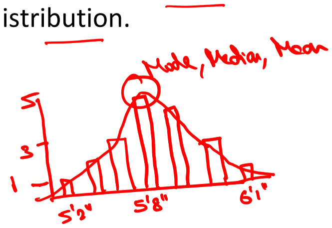
Binomial is generally yes or no. It is a probability distribution with n parameters asking yes or no questions. And each has its own Boolean-valued outcome: success with probability p or failure with probability q = (1 – p). The binomial distribution is the basis for the binomial test of statistical significance.

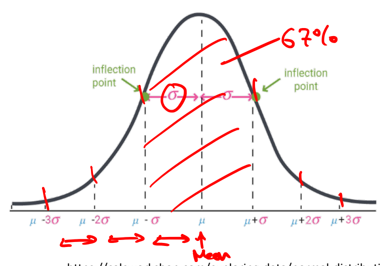
Examples include: calculating the probability of getting desired successful versus unsuccessful results, and measuring the chance of getting one result over another in defined experiments. These are not very detailed. Anyway, we will be given more references to read about binomial distribution. The third one, which is normal distribution, I will give you more details about because we are going to talk about hypothesis testing based on this only.

Probability

3. Normal Probability

- In all statistical models we generally assume that experimental errors are normally distributed.
- This assumption can be verified by plotting the residual errors on a normal probability distribution.
- If all the residuals fall along a straight line drawn, it is said to be a normal distribution.





https://calcworkshop.com/exploring-data/normal-distribution/

8

In all statistical models, we generally assume that experimental data errors are normally distributed. This assumption can be verified by plotting the residual errors on a normal probability distribution. If all the residuals fall along a straight line drawn, it is said to be a normal distribution.

Normal distribution says, for example, the average height of students in a class is 5 feet 8 inches. The maximum height would be something, the minimum height would be something, but most students are 5 feet 8 inches. This is called normal distribution. For instance, if it is like this: most students are 5 feet 8 inches, the minimum is, say, 5 feet 2 inches, and the maximum is, say, 6 feet 1 inch, then most students would be 5 feet 8 inches.

Some students would be 5 feet 9 inches, and some would be this. One or two students could be 6 feet 1 inch. One student could be something like this. This kind of histogram plot gives me a frequency polygon, which when I try to draw, gives me a distribution known as normal distribution. Normal means very common.

Normal, in literal terms, means very common. Commonly, most of the data is distributed like this. If you see average marks in the class, average dimension, or whatever the desired dimension is, most of the pieces will fall into that. Some of them would be larger; some of them would be smaller. So, this is a normal distribution.

And here, this difference is σ, the standard deviation. This is μ, which is the mean, and this is σ, the standard deviation. And $\mu - \sigma$, $\mu - 2\sigma$, $\mu - 3\sigma$, each of them has a specific percentage of area lying within itself. For example, around 67% of the area lies between $\mu - \sigma$ and $\mu + \sigma$. Let me see more about it.

Normal Distribution

- The normal distribution is a continuous probability distribution.
- It is a distribution of continuous variables.
- A **Continuous Variable** is a variable that can assume any value on a continuum (can be assumed as an uncountable number of values).

For example:

- thickness of an item
- time required to complete a process
- temperature of a system etc.

Such parameters can take on any value depending only on the ability to measure them precisely and accurately.

9

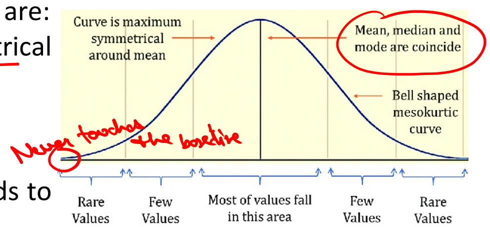
Normal distribution is a continuous probability distribution. It is a distribution of continuous variables. A continuous variable is a variable that can assume any value on a continuum and can take on an uncountable number of values. For instance, height, measurement, diameter, weight—these are all continuous variables. These are the only cases where a normal distribution can be plotted. For the variables for which a continuous variable is not present, for instance, the dimension could be 2.5 micrometers, 2.51 micrometers, 2.512 micrometers, 2.513 micrometers, to any level. This is a continuous distribution.

A non-continuous distribution is when the numbers are discrete. For example, the number of machines, the number of workers, or the number of boxes that are ready. That is the non-continuous distribution. So, normal distribution is for a continuous variable. For example, the thickness of an item, the time required to complete a process, or the temperature of a system, etc. Such parameters can take on any value depending on the ability to measure them precisely and accurately.

Normal Distribution

- The characteristics of Normal Distribution are:

1. The curve is Bell Shaped and symmetrical about the mean.
2. The Mean, Median and Mode are Equal.
3. The total area under the curve is 100%.
4. The tail on either side of the curve extends to infinity.



5. The distribution is defined by two parameters μ and σ. Location is determined by the mean, μ and spread is determined by the standard deviation, σ.
6. The random variable has an infinite theoretical range: +∞ to –∞. The tails of curve extend to infinity, but they do not touch the axis.

NPTEL

https://www.researchgate.net/profile/Atul-Kumar-Goyal/publication/377499413/figure/fig4/AS:11431281218613010@1705652256428/A-normal-distribution-curve.png

Normal distribution; the characteristics of normal distribution are the curve is bell-shaped and symmetrical about the mean. This is a bell-shaped curve. It is symmetrical about the mean. The mean, median, and mode are all equal. This is, as I said, you see in this number of students. For example, five students are there. This is a frequency, which are having a height of 5 feet 8 inches. And there are three students who are having a height of 5 feet 7 inches and maybe 5 feet 10 inches.

And there is only one student with a height of 5 feet 2 inches. So, this is the mode, the maximum number of students. Also, this is the median and mode. And also, this is a continuous variable. This is also the mean. This means the mean, median, and mode are all equal in a normal distribution. Mean, median, and mode coincide. The total area under the curve is 100%. The tail on either side of the curve extends to infinity. It never touches here.

It never touches the baseline or touches the baseline at infinity. The distribution is defined by two parameters, mu and sigma, where mu is the mean, that is, the location is determined by the mean mu, and the spread is determined by the standard deviation sigma. The random variable has an infinite vertical range, that is, plus sigma to - sigma. The tails of the curve extend to infinity, but they do not touch the axes. So, these are the characteristics of a normal distribution.

Normal Distribution

Standard Normal Distribution

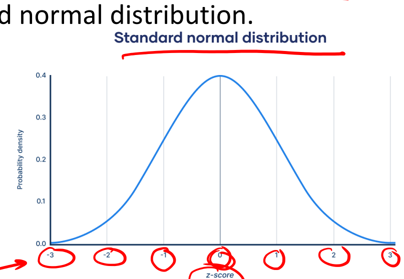
- Standard normal distribution is a special case of normal distribution.
- It is used for easy calculation of area between any two points under the curve. Its mean is 0 and variance is 1.
- Suppose x is a continuous random variable that has a normal distribution:

$N(\mu, \sigma^2)$, then the random variable z follows a standard normal distribution.

$$z = \frac{X - \mu}{\sigma}$$

The horizontal axis is represented by z.
The centre point is the mean, 0.
The z value for a point (x) on the horizontal axis gives the distance between the mean and that point in terms of standard deviation.

(z-score)



https://www.scribbr.com/wp-content/uploads/2023/02/standard-normal-distribution-example.webp

11

Standard normal distribution is a special case of normal distribution where mean is 0 and variance is 1. Suppose x is a continuous variable or random variable that is a normal distribution that is n mu sigma square. This is the way to present the normal distribution. Then random variable z follows a standard normal distribution that is

$$z = \frac{X - \mu}{\sigma}$$

There is z, is -1, -2, -3, 1, 2, 3, 0.

This is z-score. The horizontal axis is represented by z. The center point is the mean 0. The z-value at a point x on the horizontal axis gives the distance between the mean and that point in terms of standard deviation. This is a standard normal distribution with a mean of 0 and a variance of 1.

Normal Distribution

Normal probability plot is generated in following way:

- Arrange the sample data in ascending order of magnitude.
- Compute the cumulative frequency

$$\frac{j - 0.5}{n}$$

Where 'j' is the rank or position of the number in ascending order and n is the total number of observations.

- Transform the cumulative frequency into a standardized normal score Z_j.
- Plot Z_j vs X_j.

So, let me see an example here. A normal probability plot is generally generated in the following way. Arrange the assembled data in ascending order of magnitude. Compute the cumulative frequency, which is $\frac{j-0.5}{n}$. I will show you the example in the next slide. j is the rank or position of the number in ascending order, and n is the total number of observations. Transform the cumulative frequency into a standardized normal score Zj, then plot Zj versus Xj. So, let me see this example.

Normal Distribution

Normal probability plot

Let us consider the hardness of 10 samples of certain alloy steel having undergone a heat treatment process.

290, 248, 235, 228, 256, 223, 263, 195, 272, 237

To obtain a normal probability plot :

1. We tabulate it for normal probability plot.
2. The data arranged in the ascending order of value is shown in the second column.
3. The first column indicates the rank of the ordered data.
4. Third column gives the cumulative frequency of the ordered data.
5. The last column (*Zj)* is the standardized normal score.

j	X_i	(j – 0.5)/10	Z_j
1	195	0.05	–1.64
2	223	0.15	–1.04
3	228	0.25	–0.67
4	235	0.35	–0.39
5	237	0.45	–0.13
6	248	0.55	0.13
7	256	0.65	0.39
8	263	0.75	0.67
9	272	0.85	1.04
10	290	0.95	1.64

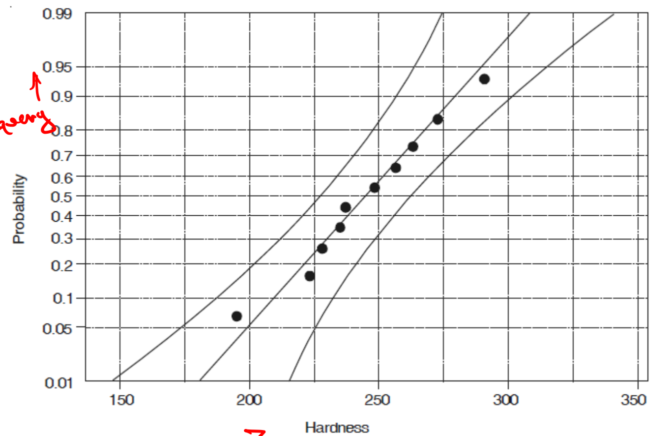
Let us consider the hardness of 10 samples of a certain alloy steel that has undergone heat treatment processes. These are 292, 48, 235, 228, 256, 223, 263, 195, 272, 237. Here, the minimum value is 195, and the maximum value is 290. To obtain a normal probability plot, we tabulate the data for the normal probability plot. We will calculate it. There are 10 numbers. We put the numbers from 1 to 10 here.

Then, we arrange the data in ascending order, which is shown in the second column. That is, the minimum value is 195, and the maximum value is 290. It is put in ascending order. The first column indicates the rank of the ordered data. The third column gives the cumulative frequency of the ordered data. Cumulative frequency is j - 0.5 divided by 10. That is the number of samples; I got it here. Based on this cumulative frequency, I get the value of Zj. And this Zj is now plotted here, you see.

Normal Distribution

- The plot X_j (X-axis) versus (j – 0.5)/10 (Y-axis) is the normal probability plot of the data.
- We can also plot X_j (X-axis) versus Zj (Y-axis) on an ordinary graph sheet to obtain the same result.

Frequency



Zj

This is Zj, and this is probability or frequency. The plot Xj, which is the x-axis, versus (j – 0.5)/10, which is the y-axis, is the normal quantile distribution plot. We can also plot Xj, which is the x-axis, versus Zj, which is the y-axis, on an ordinary graph sheet to obtain the same result. This is plotted here. Hardness here.

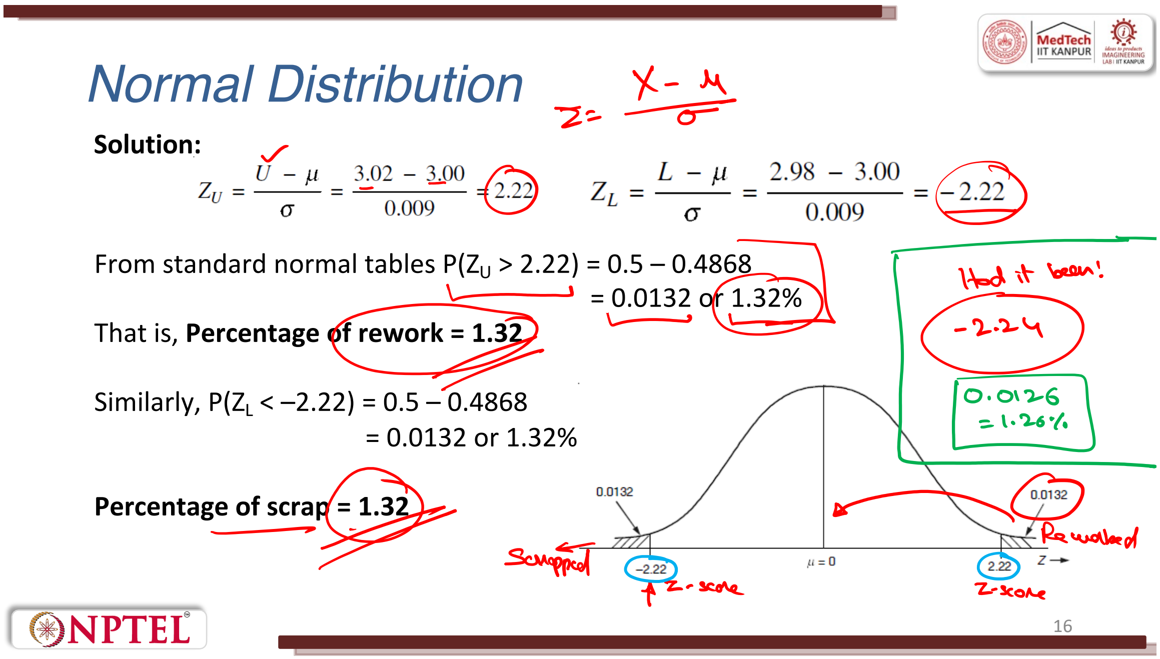
Normal Distribution

Example problem: The diameter of shafts manufactured is normally distributed with a mean of 3.0 cm and a standard deviation of 0.009 cm. The shafts that are with 2.98 cm or less diameter are scrapped and shafts with diameter more than 3.02 cm are reworked. Determine the percentage of shafts scrapped and percentage of rework.

Solution: Mean (μ) = 3.0 cm
Standard deviation (σ) = 0.009 cm
Let upper limit for rework *(U) = 3.02 cm*
Lower limit at which shafts are scrapped *(L) = 2.98*

Now let us determine the Z value corresponding to U and L

https://www.scribbr.com/wp-content/uploads/2023/02/standard-normal-distribution-example.webp



Let me see another example to have a better understanding of the bell-shaped curve of normal distribution. The diameter of shafts manufactured is normally distributed with a mean of 3.0 cm and a standard deviation of 0.009 cm. The shafts that are with 2.98 cm or less diameter are scrapped and shafts with diameter more than 3.02 cm are reworked. Determine the percentage of shafts scrapped and percentage of rework.

Solution: Mean (μ) = 3.0 cm

Standard deviation (σ) = 0.009 cm

Let upper limit for rework *(U) = 3.02 cm*

Lower limit at which shafts are scrapped *(L) = 2.98*

Now let us determine the Z value corresponding to U and L

$$Z_U = \frac{U-\mu}{\sigma} = \frac{3.02-3.00}{0.009} = 2.22 \qquad Z_L = \frac{L-\mu}{\sigma} = \frac{2.98-3.00}{0.009} = -2.22$$

From standard normal tables $P(Z_U > 2.22) = 0.5 - 0.4868$

$= 0.0132$ or 1.32%

That is, Percentage of rework = 1.32

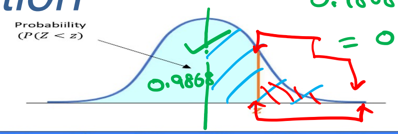
Similarly, $P(Z_L < -2.22) = 0.5 - 0.4868$

$= 0.0132$ or 1.32%

Percentage of scrap = 1.32

Normal Distribution

Solution:



0.9868 − 0.5
= 0.4868

0.5 − 0.4868
= 0.0132

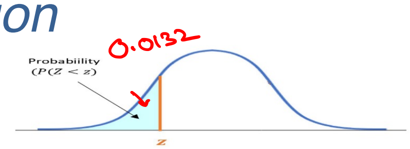
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5754
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7258	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7518	0.7549
0.7	0.7580	0.7612	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7996	0.8023	0.8051	0.8079	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9430	0.9441
1.6	0.9452	0.9463	0.9474	0.9485	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9700	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998	0.9998

This is the area that we have seen in this table. In this table, you see, we can see 2.2 is here. From this 2, 2.20 is this, 2.21 is this, 2.22 is this, 0.0898. So, this is talking about the area on the left-hand side of this, this area. This total area is 0.9868, which means this half of the area would be 0.9868 - 0.5, which is 0.4868.

And from this half area, we are only interested in this region that is on the right-hand side. This region. This region is what 0.5 - 0.4868—I'm talking about this region only. This came out to be 0.0132, you see it here. We're looking at the right side; this area comes out to be 0.0132

Normal Distribution

Solution:



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.5000	0.4960	0.4920	0.4880	0.4841	0.4801	0.4761	0.4721	0.4681	0.4641
-0.10	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.20	0.4207	0.4168	0.4129	0.4091	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.30	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.40	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.50	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.60	0.2743	0.2709	0.2676	0.2644	0.2611	0.2579	0.2546	0.2514	0.2483	0.2451
-0.70	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2207	0.2177	0.2148
-0.80	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.90	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-1.00	0.1587	0.1563	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-1.10	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.20	0.1151	0.1131	0.1112	0.1094	0.1075	0.1057	0.1038	0.1020	0.1003	0.0985
-1.30	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.40	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.50	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.60	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.70	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.80	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.90	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-2.00	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-2.10	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.20	0.0139	0.0136	0.0132	0.0129	0.0126	0.0122	0.0119	0.0116	0.0113	0.0110
-2.30	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.40	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0070	0.0068	0.0066	0.0064
-2.50	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.60	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.70	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.80	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.90	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-3.00	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-3.10	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.20	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.30	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.40	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002	0.0002

https://www.wizeprep.com/online-courses/16904/chapter/10/core/5/1

And on the left-hand side, if I see at - 2.22, another standard normal distribution is here, in which we are showing the area on the left-hand side. This will be directly given because directly we are given. So, we have - 2.22 here. If you wish me to zoom this, I can also zoom this. So, that is more visible.

This is -0.22 on this side. From here 2.20, 2.21 and a little this side 0.0132. So, to put it in the slide mode. So, this is at - 2.22 to this level, at 0.22 this level, this is the value, directly 0.0132. This gives me the percentage of scrap as 1.32 percent, percentage of rework, which is on the right-hand side, as also 1.32 percent. Had it been a different value, for example, had it been a value of maybe - 2.24 on this side. Then we would have taken—I am putting 'had it been,' just to make it more clear, - 2.24. We would have taken the percentage as, let me see in the table, in the green color, this value would have come, 0.0126, this value would have come.

So, in that case, this would have been 0.0126 = 1.26%. This is just a separate note I have given. So, this is how we use standard deviation and normal distribution, and hypothesis testing has a lot of applications in this. Hypothesis, one-tailed or left-hand side, one-tailed or right-hand side, or two-tailed test—these all have different tables. From those tables, we try to see the value and determine the area of the acceptance region, where we reject the hypothesis and accept the alternative hypothesis.

Inferential Statistics

- Inferential statistics is an important tool that allows us to make predictions and conclusions about a population based on sample data.
- Descriptive statistics only summarizes the data while inferential statistics lets us analyze data by testing hypotheses, making estimates and measure the uncertainty about our predictions.
- These tools are essential for evaluating models, testing assumptions and supporting data-driven decision-making.
- Inferential statistics provides the tools to analyze conclusions systematically and mathematically.
- **Inferential Statistics gives the probability that the difference between means reflects random error rather than a real difference.**

NPTEL

Now, having understood the normal distribution, let me talk about inferential statistics—that is, hypothesis testing—in a little more detail. Inferential statistics is an important tool that allows us to make predictions and draw conclusions about a population based on sample data.

Descriptive statistics only summarize the data, while inferential statistics let us analyze data by testing hypotheses, making estimates, and measuring uncertainty in our predictions. These tools are essential for evaluating models, testing assumptions, and supporting data-driven decision-making. Inferential statistics provides the tools to analyze conclusions systematically and mathematically. And this gives the probability that the difference between the means reflects random error rather than a real difference. So, what can inferential statistics do for us?

Inferential Statistics

What Inferential Statistics can do for us?

- Analyzing an entire population is very time consuming and often impossible. Instead, we collect data from a sample population and use inferential statistics tools to:
- Conclude about the whole population.
- Test hypotheses or claims.
- Calculate confidence intervals and p-values to measure uncertainty.
- Make predictions with statistical models.

NPTEL

Because analyzing an entire population is very time-consuming—that is, a census survey requires a lot of resources. We collect data from a sample population and use inferential statistics to draw conclusions about the whole population. Test hypotheses or claims, calculate confidence intervals and p-values (probability values) to measure uncertainty, and make predictions with statistical models.

Inferential Statistics

- The goal of the inferential statistics is to draw conclusions from a sample and generalize them to the population.
- It determines the probability of the characteristics of the sample using probability theory.

The most common methodologies used are:

1. **Hypothesis tests,**
2. **Analysis of variance etc.**

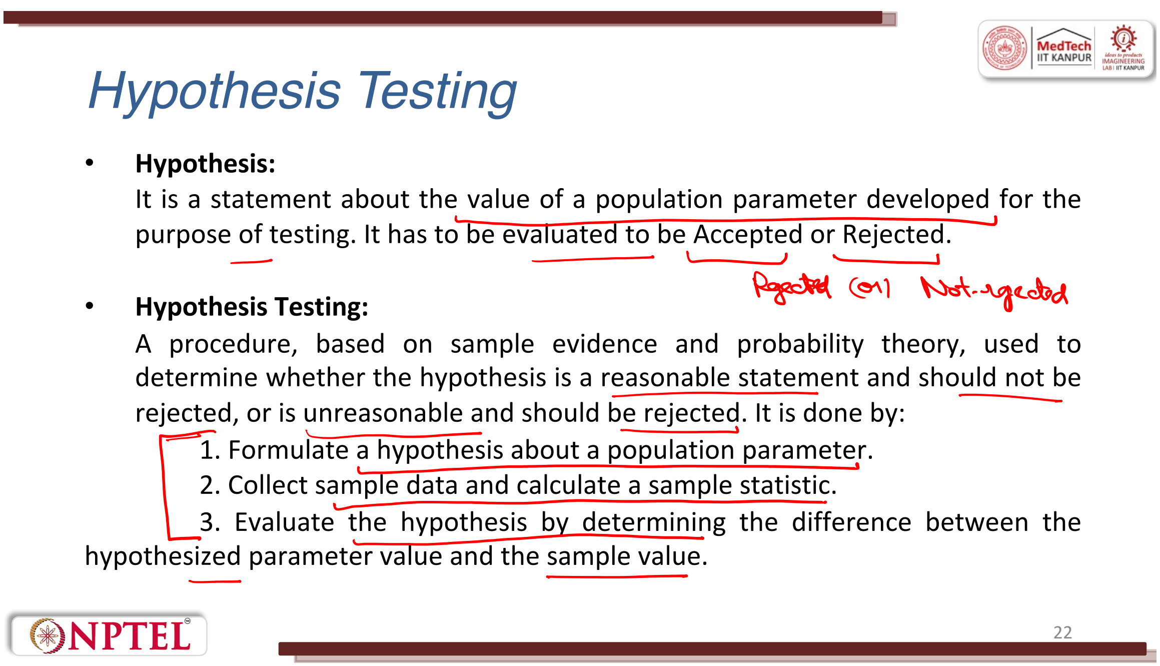
- For it is not feasible to gather data for a large population, we collect information of count.
- We consider this sample for our statistical study of the whole population.

Sampling

NPTEL

Now, the goal of inferential statistics is to draw conclusions from a sample and generalize them to the population. It determines the probability of characteristics of the sample using probability theory. The common methodologies are hypothesis testing and analysis of variance.

I'll talk about hypothesis testing. I'll talk about regression modeling in the next lecture. For it is not feasible to gather data for a large population with limited information. Of course, we consider this sample for our statistical study of the whole population. This is known as sampling. That is the point where we started our lecture.



We specifically select the sample and try to conduct hypothesis testing on it. Now, what is a hypothesis? It is a statement about the value of a population parameter developed for the purpose of testing. It has to be evaluated to be accepted or rejected. Generally, we say it is rejected or not rejected.

A procedure based on sample evidence and probability theory is used to determine whether the hypothesis is a reasonable statement and should not be rejected or is unreasonable and should be rejected. This is done by: Number 1: Formulate a hypothesis about the population parameter. Collect the sample data and calculate the sample statistic.

Evaluate the hypothesis by determining the difference between the hypothesized parameter value and the sample value. We will see this through examples.

Making a statement of hypothesis is very important. How clearly and precisely we make the statement is very important. There could be one-tailed and two-tailed tests. I am talking about tails because the normal distribution has ends which are called tails. A one-tailed test could be, for example, I took the example of a shaft with a 3 cm diameter.

If I say the size of the shaft is less than 3.02 cm, that's all. It's a one-tailed, one-sided test only. Or, if on the other side, I say the size of the shaft is greater than 2.98 cm, that is also a one-tailed test. I am talking about only one side. If I say the size is different from 3 cm, plus or minus this distribution, this plus or - 0.2 value. This is where both tails are considered, which is a two-tailed test. So, accordingly, the area is taken in the normal distribution.

Hypothesis Testing

Types of Hypothesis:

- **The Null Hypothesis H_0**
  It represents a theory or assumption that has not been proven, and is often used as a basis for testing.

For example

- In a clinical trial, the null hypothesis might state that there is no significant difference between a new drug and an existing one, and can be written as:
  H_0: The new drug is no better than the current drug on average.

NPTEL

The type of hypothesis is the null hypothesis. It represents a theory or assumption that has not been proven and is often used as a basis for testing. In a clinical trial, the null hypothesis might state that there is no significant difference between a new drug and an existing one, which can be written as H_0, representing the null hypothesis. The new drug is no better than the current drug on average.

Now, what does this mean? If myelopathy is rejected, that means the 'no better than current drug' statement is rejected. That means there is a significant difference that exists. If it is not rejected, this means it is no better than the current drug.

Hypothesis Testing

- **The Alternative Hypothesis (H_A)**
  It is a statistical test, which aims to prove and is typically the opposite of the null hypothesis.

For example

- In a clinical trial, the alternative hypothesis might be that the new drug has a different effect than the current one, written as:

  H_A: The new drug has a different effect than the current drug on average.

- The result of a hypothesis test can be one of two outcomes:

  Reject H_0 in favour of H_A OR Do not reject H_0 (that is reject H_A).

NPTEL

The alternative hypothesis is there, which aims to prove that there is an alternative way. The alternative hypothesis is when the null hypothesis is rejected; there is an alternative thing that is true. It is a statistical test that aims to prove and is typically the opposite of the null hypothesis.

For example, in a clinical trial, the alternative hypothesis might be that the new drug has a different effect than the current one, which is written as HA (alternative hypothesis). The new drug has a different effect than the current drug, on average. The result of a hypothesis test can be one of these two outcomes. Number one: reject H_0 in favor of HA, or do not reject H_0. That is, we reject HA.

Hypothesis Testing

Null Hypothesis, H_0

- State the hypothesized value of the parameter before sampling.
- The assumptions we wish to test (or the assumption we are trying to reject)
- e.g. population mean $\mu = 10$
- There is no difference of tensile strength between duralumin and mild steel.

Alternative Hypothesis, H_a

- All possible alternatives other than the null hypothesis
- e.g. $\mu \neq 10$
  $\mu > 10$
  $\mu < 10$
- There is a difference of tensile strength between duralumin and mild steel.

NPTEL

Null hypothesis—let me see an example. When we compare it with the alternative hypothesis, state the hypothesized value of the parameter before sampling. The assumptions we wish to test or the assumptions we are trying to reject. The population mean is equal to 10.

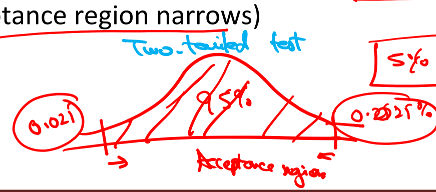
This is the null hypothesis. On the other hand, the alternative hypothesis includes all possible alternatives other than the null hypothesis, that is, either $\mu \neq 10$ (a two-tailed test), or $\mu > 10$, or $\mu < 10$ (both of which are one-tailed tests). In the null hypothesis, there is no difference in tensile strength between duralumin and mild steel. This could be a null hypothesis statement. There is no difference in tensile strength.

The alternative would be that there is a difference in tensile strength. That is what we wish to prove or work upon—the alternative hypothesis. That there is a difference in the tensile strength. There is a lot to be rejected or accepted in the alternative hypothesis.

Hypothesis Testing

Deciding on a criterion for accepting or rejecting the null hypothesis.

- **Significance level** refers to the percentage of sample means that is outside certain prescribed limits. e.g. testing a hypothesis at 5% level of significance means that we reject the null hypothesis if it falls in the two regions of area 0.025.
- Do not reject the null hypothesis if it falls within the region of area 0.95.
- The higher the level of significance, the higher is the probability of rejecting the null hypothesis when it is true. (acceptance region narrows)



5% of s.l.

Deciding on a criterion for accepting or rejecting the null hypothesis is the significance level, which refers to the percentage of sample means that fall outside certain prescribed limits. For example, testing a hypothesis at a 5% level of significance means we reject the null hypothesis if it falls in the two regions of area 0.025. That is in a two-tailed test. When we say a 5% significance level, as I said, the significance level is exactly opposite to the confidence level. We are 95% confident that this is true. And if it is a two-tailed test like this, and if we say 5% of—I will put that s.l., that significance level.

That means 0.025 here and 0.025 on the right side. The total area is 95%, which is left inside. Do not reject the null hypothesis if it falls within the range of the area of 0.95. The higher the level of significance, the higher the probability of rejecting the null hypothesis when it is true, and the acceptance region narrows. That means whenever we move towards this, the acceptance region—this is the acceptance region—narrows.

Hypothesis Testing

- **Example problem:** The null hypothesis is that the battery for a EV has an average life of 1000 days, with the alternative hypothesis that the average life is more than 1000 days, You are the Quality control manager for the battery manufacturer.

  (a) Would you rather make a Type I error (Producer's risk) or a Type II error (Consumer's risk)?

  (b) Based on your answer to part (a), should you use a high or low significance level?

H_0: Average battery life = 1000 days ($H_0: \mu = 1000$)

H_A: Average battery life > 1000 day ($H_A: \mu > 1000$)

a) If H_0 is false, still $\mu > 1000$] we would prefer making Type II error

b) $\alpha \uparrow$ (More chances of Type I error) $\alpha \downarrow$ (Less chances of Type I error)

Let me see an example problem so that we are clear with it. The null hypothesis is that the battery for an electric vehicle has an average life of 1000 days, with an alternative hypothesis that the average life is more than 1000 days. You are the quality control manager for the battery manufacturer. Would you rather make a Type 1 error or a Type 2 error? These are new terms.

I'll talk about what Type 1 and Type 2 errors are. These are the consumer and producer errors. Based on your answer to part A, should you use a high or low significance level? Please try to understand this problem. They say the average battery life is 1000 days. This is what? This is our null hypothesis. Alternatively, that is HA, that average battery life is not given as 1000; it is given as more than 1000. We say the alternative is that average battery life is more than 1000 days, or we can also put it as H naught is mu equal to 1000, or H alternative is mu greater than 1000. Now they say, would you make a type 1 error or a type 2 error? For this, let me first talk about what type 1 and type 2 errors are.

Hypothesis Testing

- A hypothesis test is a simple rule that it is either accepted or rejected.
- Since it is based on sample statistics computed from 'n' observations, the decision is subject to two types of errors.

- **Type I error:** The probability of Type I error is denoted by α. It says that the null hypothesis is true, but rejected.

The value of α represents the significance level of the test.

$$\alpha = P\,(H_0 \text{ is rejected} \mid H_0 \text{ is true})$$ (Producer's error)

(Significance level)

A hypothesis test is a simple rule that is either accepted or rejected. Since it is based on sample statistics computed from n observations, the decision is subject to two types of errors. The point I'm trying to make here is that only a small number of observations represents the whole population. For example, out of 10,000 pieces of the nozzles, we picked only a sample of 20. These 20 will represent the 10,000. Whatever is the outcome of this 20-sample testing data, that would be projected over the 10,000 pieces; the whole lot would be either accepted or rejected based on this. There could be an error that this sample we have selected is good, but the overall lot is not good. Or, on the other side, the sample we have selected is not good, but the overall lot was good. In both cases, there are errors: one is type 1, and the other is type 2. Let us see those.

Hypothesis Testing

- **Type II error:** The probability of Type II error is denoted by **β**. It says that the hypothesis is accepted when it is not true.

  That is, some alternative hypothesis is true.

  β = P (H_0 is accepted | H_0 is false) (Consumer's error)

- **(1 – β)** is called the power of the test. It denotes the probability of not committing the Type II error.

Type 1 error is the probability denoted by alpha. It occurs when the null hypothesis is true but rejected. That is, α = P (H_0 is rejected | H_0 is true). This is known as the producer's error. This means the selected sample is not a good sample. The sample should have been accepted. That is, H_0 should be true here, but it is rejected. That is, the sample is rejected. Based on the sample, the whole lot of 10,000 pieces is rejected. Based on what? A sample size of only 20 pieces.

That means the 10,000 pieces produced by the manufacturer have gone to waste. This is the producer's risk due to the selection of an incorrect sample. This is a type 1 error. On the other hand, there could be a type 2 error, denoted by beta. It occurs when the null hypothesis is accepted when it is not true. This means some alternative hypothesis is true. That is, H_0 is accepted when it should be false. This means the sample that is selected is a good sample, a very wonderful sample, and is accepted. But the lot is not good. Out of the bad lot, a good sample is selected. Because of this sample, the whole lot is now selected.

And that is passed to the consumer. Now, the producer will have its profit. But the consumer will receive the products which are not good. 10,000 pieces which are not good. That means this is a consumer's error. These two errors are possible when we talk about hypothesis testing. So, $1 - \beta$ is called the power of the test and denotes the probability of not committing a type 2 error. So, now let me see the problem statement. They say, would you have a producer's risk or would you have a consumer's risk? Would you rather make which kind of risk would you like to take?

And just to be more clear, this 5% significance level is alpha. Alpha, wherever I have written, this alpha is the significance level. Now, should we increase the significance level or reduce it? This is the question asked. So, whether we will make a Type 1 error or a Type 2 error, this is the first part. Now, Type II error is the consumer risk. Even if this rejects the null hypothesis, still the battery life is more than 1000.

That means, if H_0 is false, still μ is greater than 1000. Therefore, we would prefer making a Type 2 error. This is part A. Part B: Should you use a high or low significance level? Now, if we increase the alpha value, or if we reduce the alpha value, what happens? When we increase alpha, there are more chances of a Type 1 error. And when we reduce it, there are fewer chances of a Type 1 error. And when it comes to alpha, we will try to reduce the alpha value. So, the last topic I would like to discuss is more details on the one-tailed and two-tailed test.

Hypothesis Testing

- A **one-tailed test** is a statistical hypothesis test in which the values for which we can reject the null hypothesis, H_0 are located entirely in one tail of the probability distribution,
- **Lower tailed test** will reject the null hypothesis if the sample mean is significantly lower than the hypothesized mean. Appropriate when:

$$H_o : \mu = \mu_0 \text{ and } H_A : \mu < \mu_o$$

Example: A wholesaler buys light bulbs from the manufacturer in large lots and decides not to accept a lot unless the mean life is at least 500 hours.

$H_0: \mu \geq 500$

$H_A: \mu < 500$

He rejects H_0 only if mean life of bulbs is significantly below 500 hrs. (H_0 is rejected; H_A is accepted)

We talked about the one-tailed and two-tailed test when we were talking about only one direction, whether the acceptance region is only in one direction. For example, if I say the dimension is 2.5 micrometers, and if I say less than 2.7 micrometers is acceptable only, that is one-sided only. Or if I say greater than 3.5 micrometers is acceptable only, something like that. So, this is one-sided only. If I say, like I took an example in the previous plot here, this was half of the area on this side, half of the area on this side at 5% significance level had it been a one-tailed test.

It could have been, this is the acceptance area, the acceptance region. And this would have been 0.05, that is five percent on one side itself, so this first one is a one-tailed test, this is a two-tailed test. Let me try to see this with very small and easy examples.

Let me try to talk about first the lower tail test, then I will try to talk about the upper tail test. A lower tail test will reject the null hypothesis if the sample mean is lower than the hypothesized mean, that is H_0 is $\mu = \mu_0$ and H_1 is when μ is less than μ_0. So, we will reject if the alternative hypothesis is true.

For example, a wholesaler buys light bulbs from the manufacturer in large lots and decides not to accept a lot unless the mean life is at least 500 hours. When I say at least 500 hours, it looks like we are talking about the upper tail, but this is a lower tail test where we can design my hypothesis as H_0, where $\mu \geq 500$ because my focus is upon the lower tail test here. Alternative hypothesis is what? That has to be the lower tail, that μ is less than 500.

This is the alternative hypothesis. So, here the phrase 'lower tail' indicates the alternative hypothesis should have a less-than sign. So, the wholesaler is looking for evidence against the null hypothesis here. The word 'at least' means greater than or equal to 500. This is the null hypothesis. This means he rejects H naught only if the mean life of bulbs is significantly below Just watch my words here: significantly below 500 hours. That is, H naught is rejected, which means HA is accepted. This is a one-sided test.

Hypothesis Testing

- **Upper tailed test** will reject the null hypothesis if the sample mean is significantly higher than the hypothesized mean. Appropriate
- when H_0: $\mu = \mu_0$ and H_A: $\mu > \mu_0$

Example: A highway safety engineer decides to test the load bearing capacity of a 20 year old bridge. The minimum load-bearing capacity of the bridge must be at least 10 tons.

H_0: $\mu = 10$ tons H_A: $\mu > 10$ tons   (Upper-tailed test)

He rejects H_0 only if mean load bearing capacity of the bridge is significantly higher than 10 tons

Then we have an upper-tail test, which will reject the null hypothesis if the sample mean is significantly higher than the hypothesized mean. Here, H_0: $\mu = \mu_0$, and H_A: $\mu > \mu_0$. A highway safety engineer decides to test the load-bearing capacity of a 20-year-old bridge. The minimum load-bearing capacity of the bridge must be at least 10 tons. You see the design of the two examples I am taking. The previous one was that the bulb life should be more than 500 hours.

If it should be more than 500 hours, less than 500 hours is the alternative. Here, it is the bridge design. If it is 10 tons minimum load-bearing capacity, the alternative hypothesis based upon the upper-tail test would be H_A: $\mu > 10$ tons. So, this means this uses an upper-tail test. The null hypothesis is then the design. The null hypothesis is $\mu = 10$ tons, or I could have even put $\mu <= 10$ tons, any of those. But the alternative hypothesis is $\mu > 10$ tons.

So, here he rejects H_0 only, If the mean load-bearing capacity of the bridge is significantly higher than 10 tons. See these words. He rejects H_0 only if it is significantly higher than 10 tons. This is for the upper tail. For lower tails, he rejects only if it is significantly below \$500. This is the lower tail.

Hypothesis Testing

- Two tailed test will reject the null hypothesis if the sample mean is significantly higher or lower than the hypothesized mean.
- Appropriate when $H_0 : \mu = \mu_0$ and $H_A : \mu \neq \mu_0$

Example: The manufacturer of light bulbs wants to produce light bulbs with a mean life of 500 hours.

- If the lifetime is shorter he will lose customers to the competition and if it is longer then he will incur a high cost of production.
- He does not want to deviate significantly from 500 hours in either direction.

$H_0: \mu = 500$ hrs $\quad H_A: \mu \neq 500$ hours

Two tailed test

32

Let us see an example of the two-tailed test. A two-tailed test will reject the null hypothesis if the sample mean is significantly higher or lower than the hypothesized mean. In this, the null hypothesis is given as $H_0 : \mu = \mu_0$ and $H_A : \mu \neq \mu_0$.

It could be lower, higher, or anything. The manufacturer of light bulbs wants to produce light bulbs with a mean life of 500 hours. If the lifetime is shorter, they will lose customers to the competition. If it is longer, they will incur a higher cost of production.

They do not want to deviate significantly from 500 hours in either direction. This is simpler here. H naught is designed as mu equals 500 hours, and H alternative is mu is not equal to 500 hours. And this is a two-tailed test. With this, I'm closing this lecture.

To Recapitulate

- What are main types of sampling?
- Why is sampling needed in research?
- Define inferential statistics in simple words.
- What is a probability distribution used for?
- Give an example of binomial and normal distributions.
- What is the purpose of hypothesis testing?
- Difference between null and alternative hypothesis?

To recapitulate, we talked about the main types of sampling. The major types we discussed include random sampling as the most used type, but we also talked about stratified sampling, cluster sampling, and other types. What is sampling, and why is sampling needed in research?

Define inferential statistics in simple words. What is probability distribution used for? We talked about binomial, Poisson, and normal distributions, hypothesis testing and its purpose, and the difference between null and alternative hypotheses. With this, I am closing the third lecture. The last lecture of this course will discuss the design of experiments, where I will also talk about regression analysis in some detail.

Thank you.