ACADGILD

LEARN. DO. EARN

DATA ANALYTICS

ACADGILD

Statistics (Contd.)

Session 3

LEARN. DO. EARN

Agenda

ACADGILD

Sl. No.	Agenda Topics
1	Types Of Distributions
2	Binomial Distribution
3	Binomial Example
4	Geometric Distribution
5	The Normal Distribution
6	Importance of Normal Distribution
7	The Standard Normal Distribution
8	Skewness And Kurtosis
9	Skewness Interpretation
10	Kurtosis
11	Kurtosis: Interpretation
12	Quantile-Quantile (q-q) Plots
13	Central Limit Theorem
14	CLT Facts
15	Practical Application 1 Of CLT

Sl. No.	Agenda Topics
16	Confidence Interval & Probability
18	(Mis)interpreting The Confidence Interval
19	Interpreting Confidence Interval
20	P-value , Z-score
21	Unknown Mean And Known Standard Deviation
22	T-Distribution
23	Hypothesis Testing
24	Null Hypothesis Vs Alternative Hypothesis
25	Type I Vs Type II Errors
26	P-value And Significance Level
27	**a Vs b**
28	Steps of Hypothesis Testing
29	Determine a p-Value - Testing A Null Hypothesis
30	Framing Hypothesis Question
31	Upper-Tailed, Lower-Tailed, Two-Tailed Tests

Types Of Distributions

ACADGILD

- Discrete probability distributions
  - Binomial Distribution
  - Geometric Distribution

- Continuous probability distributions
  - Normal Distribution

Binomial Distribution

ACADGILD

- A binomial random variable is the number of successes in a series of trials, for example, the number of 'tails' occurring when a coin is tossed 200 times.

- A discrete random variable, X is said to follow a Binomial distribution with parameters n and p, if it has probability distribution

$$P(X=x) = (nCx)*p^x* (1-p)* (n-x)$$

where

x = 0, 1, … , n

n = 1, 2, …

p = success probability; 0 < p < 1

Binomial Distribution (Contd.)

- The trials must have the following characteristics:
  - the total number of trials is fixed in advance
  - there are just two possible outcomes of each trial: success or failure
  - the outcomes of all the trials are statistically independent
  - all the trials have the same probability of success

- Mean (μ) of the Binomial Distribution:

$\mu = np$

- Variance (σ^2) of the Binomial Distribution:

$\sigma^2 = np(1-p)$

Binomial Example

ACADGILD

- Take the example of 5 coin tosses. What's the probability that you flip exactly 3 heads in 5 coin tosses?

Solution:

$\therefore P(\text{3 heads and 2 tails}) = P(\text{heads})^3 * P(\text{tails})^2 =$

$10 \times (½)^5 = 31.25\%$

Geometric Distribution

ACADGILD

- A Geometric random variable is the number of trials required to obtain the first success.

- A discrete random variable X is said to follow a Geometric distribution with parameter p, if it has probability distribution:

  $P(X=x) = p(1-p)x-1$

  where

  $x = 1, 2, 3, \ldots$

  p = success probability; $0 < p < 1$

Geometric Distribution (Contd.)

ACADGILD

- The trials must meet the following requirements:
  - the total number of trials is potentially infinite
  - there are just two outcomes of each trial; success and failure
  - the outcomes of all the trials are statistically independent
  - all the trials have the same probability of success

Mean (μ) of the Geometric Distribution:

$$\mu = 1/p$$

Variance (σ2) of the Geometric Distribution:

$$\sigma 2 = (1-p)/p2$$

The Normal Distribution

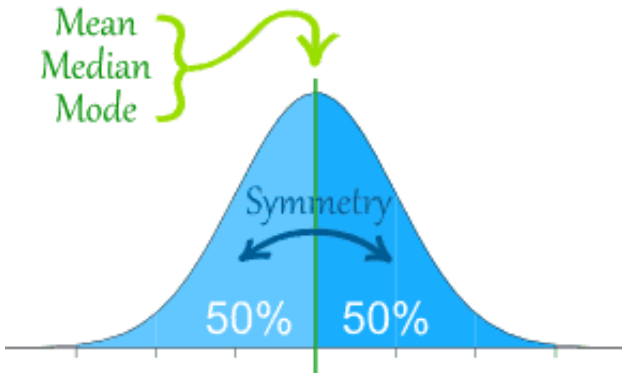
ACADGILD

- A continuous random variable X follows a normal distribution if it has the following probability density function (p.d.f.)

- The parameters of the distribution are m and s2, where m is the mean (expectation) of the distribution and s2 is the variance. We write X ~ N(m, s2) to mean that the random variable X has a normal distribution with parameters m and s2.

- The normal distribution is symmetrical about its mean:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$



Importance of Normal Distribution

ACADGILD

- One reason why the normal distribution is important is because many psychological and educational variables are distributed approximately, normally

- The second reason why the normal distribution is so important is because it is easy for mathematical statisticians to work with

- If a random variable X follows the normal distribution, then we write:

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

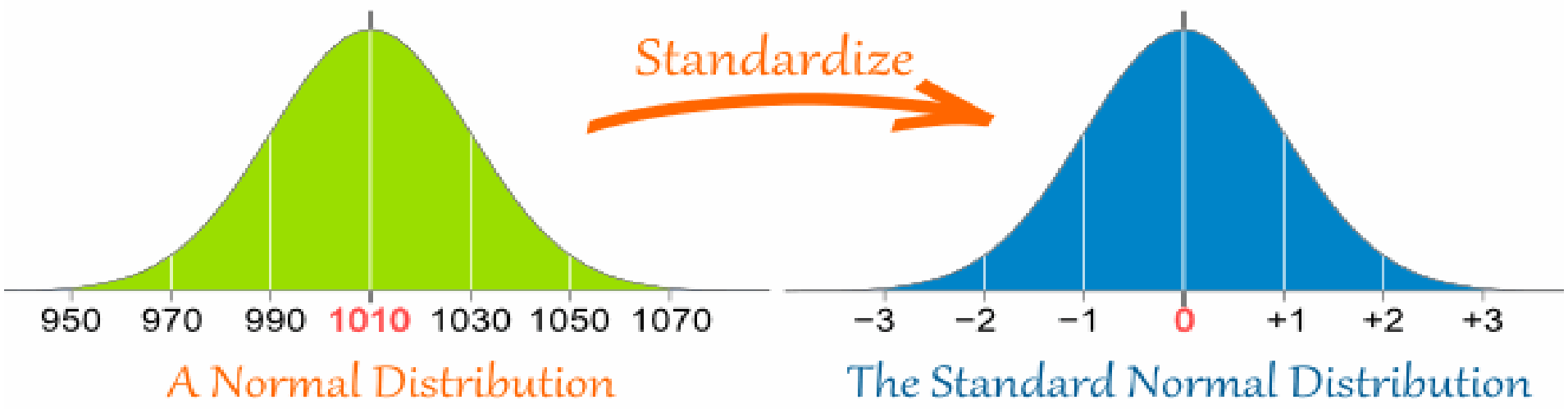
The Standard Normal Distribution

ACADGILD

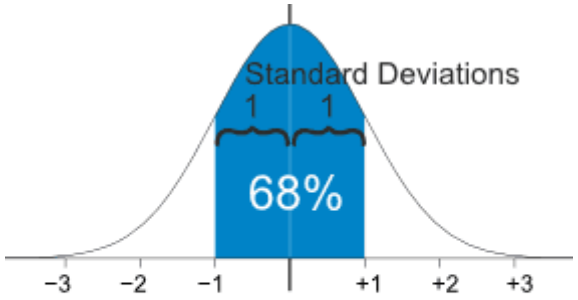
Z ~ N(0, 1), then Z is said to follow a standard normal distribution

$$Z = (x-\mu)/\sigma$$

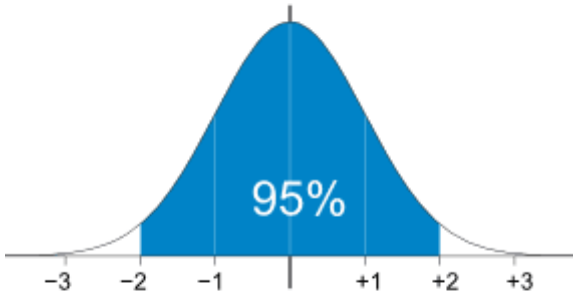
P(Z < z) is known as the cumulative distribution function of the random variable Z



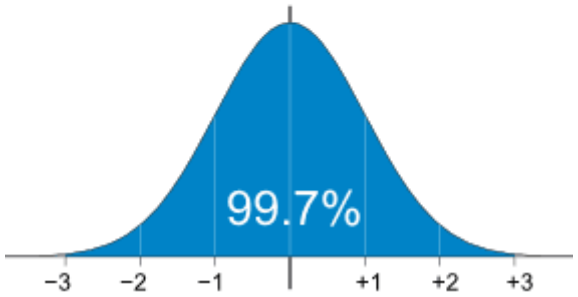
The Standard Normal Distribution (Contd.)



- **68%** of values are within **1 standard deviation** of the mean



- **95%** of values are within **2 standard deviations** of the mean

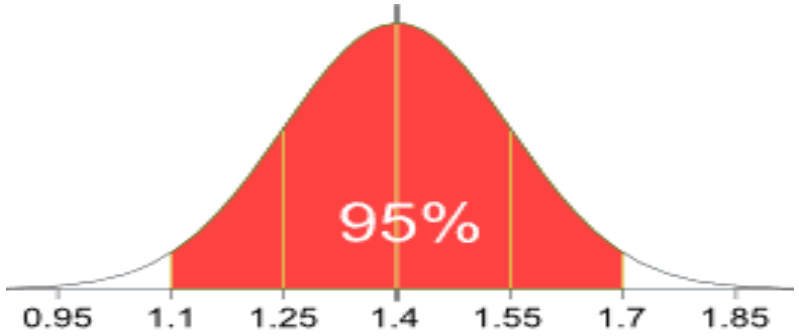


- **99.7%** of values are within **3 standard deviations** of the mean

Example

ACADGILD

- 95% of students at school are between 1.1m and 1.7m tall.
- Assuming this data is normally distributed can you calculate the mean and standard deviation?
- The mean is halfway between 1.1m and 1.7m:
- Mean = (1.1m + 1.7m) / 2 = 1.4m
- 95% is 2 standard deviations either side of the mean (a total of 4 standard deviations) so:
- An 1 standard deviation= (1.7m-1.1m) / 4 = 0.6m / 4 = 0.15md this is the result:



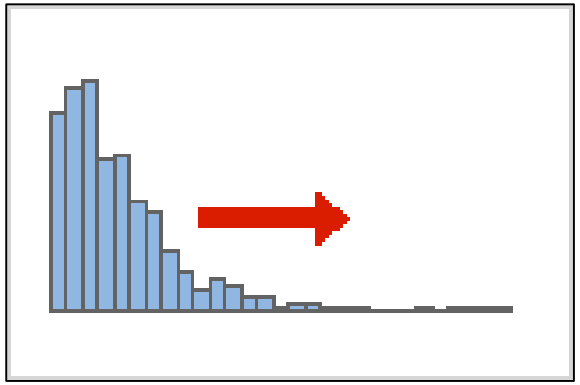
Skewness And Kurtosis

ACADGILD

- Skewness is the extent to which the data is non-symmetrical
- Whether the skewness value is 0, positive, or negative reveals information about the shape of the data. As data becomes more symmetrical, its skewness value approaches zero

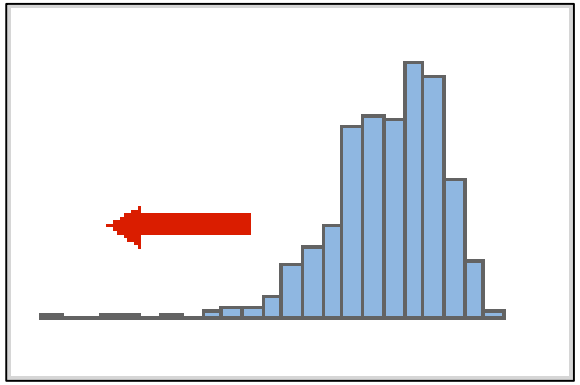
Positive or right skewed distributions

Positive skewed or right skewed data is so named because the "tail" of the distribution points to the right and its skewness value will be greater than 0 (or positive).



Negative or left skewed distributions

Left skewed or negative skewed data is so named because the "tail" of the distribution points to the left, and it produces a negative skewness value.



Skewness Interpretation

ACADGILD

- **Skewness > 0:** Right skewed distribution - most values are concentrated on the left of the mean, with extreme values to the right

- **Skewness < 0:** Left skewed distribution - most values are concentrated on the right of the mean, with extreme values to the left

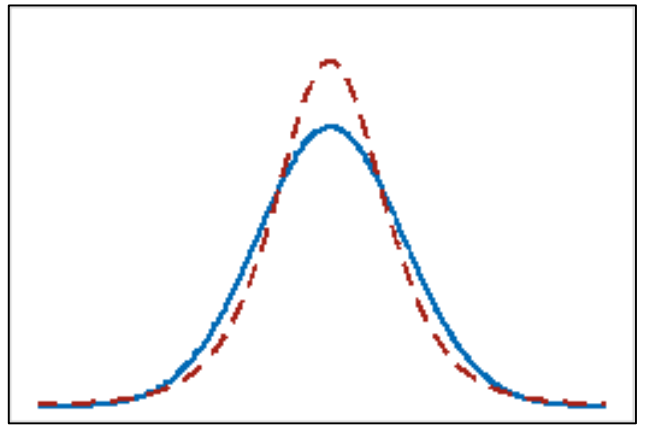
- **Skewness = 0:** mean = median, the distribution is symmetrical around the mean.

Kurtosis

- Kurtosis indicates how the peak and tails of a distribution differ from the normal distribution. Use kurtosis to help you initially understand general characteristics about the distribution of your data.

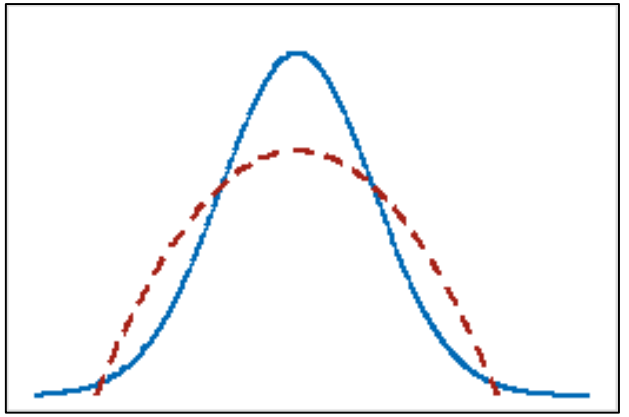
Positive kurtosis

A distribution with a positive kurtosis value indicates that the distribution has heavier tails and a sharper peak than the normal distribution.



Negative kurtosis

A distribution with a negative kurtosis value indicates that the distribution has lighter tails and a flatter peak than the normal distribution.



Kurtosis: Interpretation

ACADGILD

- Kurtosis > 3 - Leptokurtic distribution, sharper than a normal distribution, with values concentrated around the mean and thicker tails. This means high probability for extreme values

- Kurtosis < 3 - Platykurtic distribution, flatter than a normal distribution with a wider peak. The probability for extreme value is less than for a normal distribution, and the values are wider spread around the mean

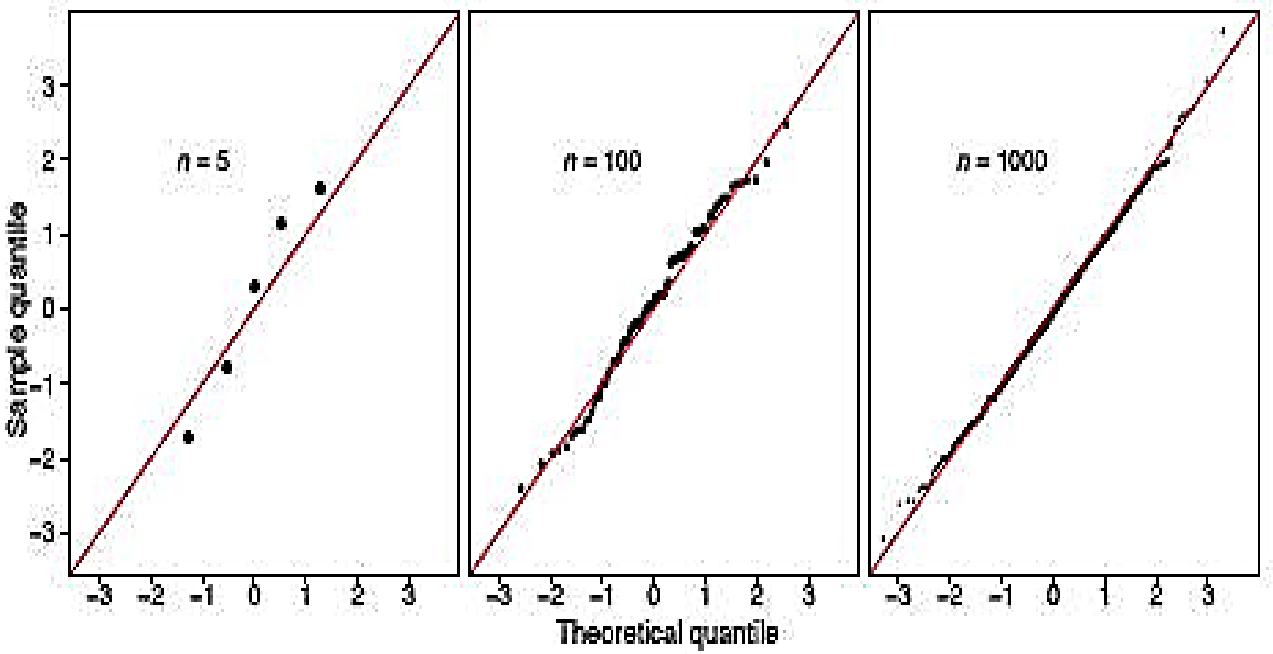
- Kurtosis = 3 - Mesokurtic distribution has normal distribution

Quantile-Quantile (q-q) Plots

ACADGILD

- The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

- The advantages of the q-q plot are:
  - The sample size need not equal
  - Many distributional aspects can be simultaneously tested
- 3rd Graph shows QQ plot of a normal distribution



Central Limit Theorem

ACADGILD

- Given a distribution with a mean μ and variance σ2, the sampling distribution of the mean approaches a normal distribution with a mean (μ) and a variance σ2/n as n, the sample size, increases

- The amazing and counter-intuitive thing about the central limit theorem is that no matter what the shape of the original (parent) distribution, the sampling distribution of the mean approaches a normal distribution

- Three different components of the central limit theorem
  - (1) successive sampling from a population
  - (2) increasing sample size
  - (3) population distribution

- Remember that this theorem applies only to the mean and not to other statistics

CLT Facts

ACADGILD

- If you draw samples from a normal distribution, then the distribution of sample means is also normal

- The mean of the distribution of sample means is identical to the mean of the "parent population", the population from which the samples are drawn

- The higher the sample size that is drawn, the "narrower" will be the spread of the distribution of sample means

Practical Application 1 Of CLT

ACADGILD

- The mean salary of the 9,000 employees at Holley.com is μ = 26,000 with a standard deviation of σ = 2420. A pollster samples 400 randomly selected employees and finds that the mean salary of the sample is 26 650. Is it likely that the pollster would get these results by chance, or does the discrepancy suggest that the pollster's results are fake?

Practical Application 1 Of CLT (Contd.)

- The question deals with the mean of a group of 400 individuals, which is a case for the Central Limit Theorem. The theorem tells us that if we select many groups of 400 individuals and compute the mean of each group, the distribution of means will be close to normal with a mean of $\mu = 26,400$ and a standard deviation of = $\sigma/\sqrt{n} = 2400 / \sqrt{400} = 121$. Within the distribution of means, a mean salary of 26,650 has a z-score of

Z = data value – mean / standard deviation = 26,650-26400 / 121 = 2.07

Practical Application 1 Of CLT (Contd.)

ACADGILD

- In other words, if we assume that the sample is randomly selected, its mean salary is more than 2 standard deviations above the mean salary of the entire company. According to the z-score chart, a z-score of 2.07 lies near the 98th percentile. Thus, the mean salary of this sample is greater than the mean salary we would find in 98% of the possible samples of 400 workers. That is, the likelihood of selecting a group of 400 workers with a mean salary above 26,650 is about 2 % or 0.02. The mean salary of the sample is surprisingly high; perhaps the survey was flawed.

Confidence Interval & Probability

ACADGILD

- A confidence interval is expressed in terms of a range of values and a probability (e.g. my lectures are between 60 and 70 minutes long 95% of the time)

- For this example, the confidence level that is used is the 95% level, which is the most commonly used confidence level

- Other commonly selected confidence levels are 90% and 99%, and the choice of confidence level to be used when constructing an interval often depends on the application

(Mis)interpreting The Confidence Interval

ACADGILD

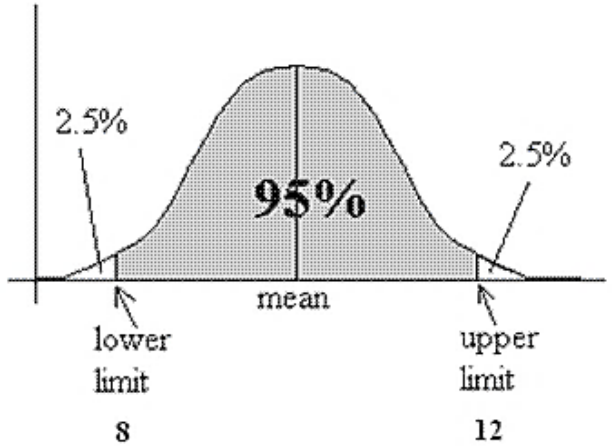
- 90% Confidence Interval for the mean score of the students is [75-80]

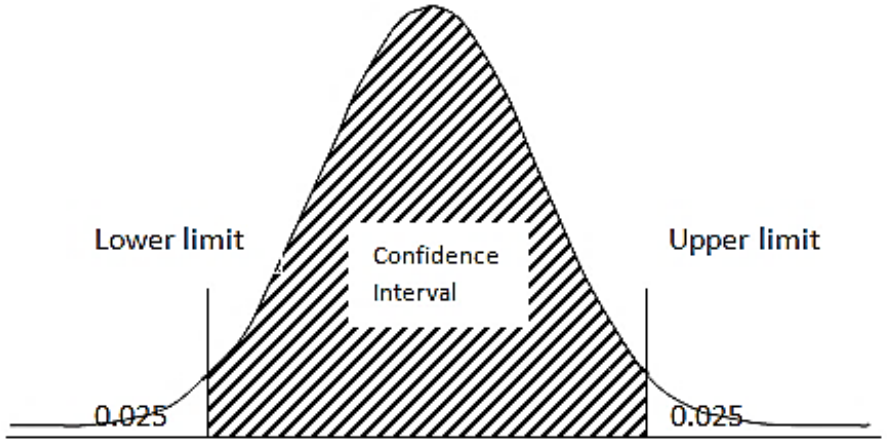
- Does this mean that
  - 90% of the students have a score in this range?
  - The mean score of the class lies in this range?
  - The mean score is in this range, 90% of the time?

Interpreting Confidence Interval

ACADGILD

- A 95% confidence interval is a range of values that you can be 95% certain and contains the true mean of the population. This is not the same as a range that contains 95% of the values.

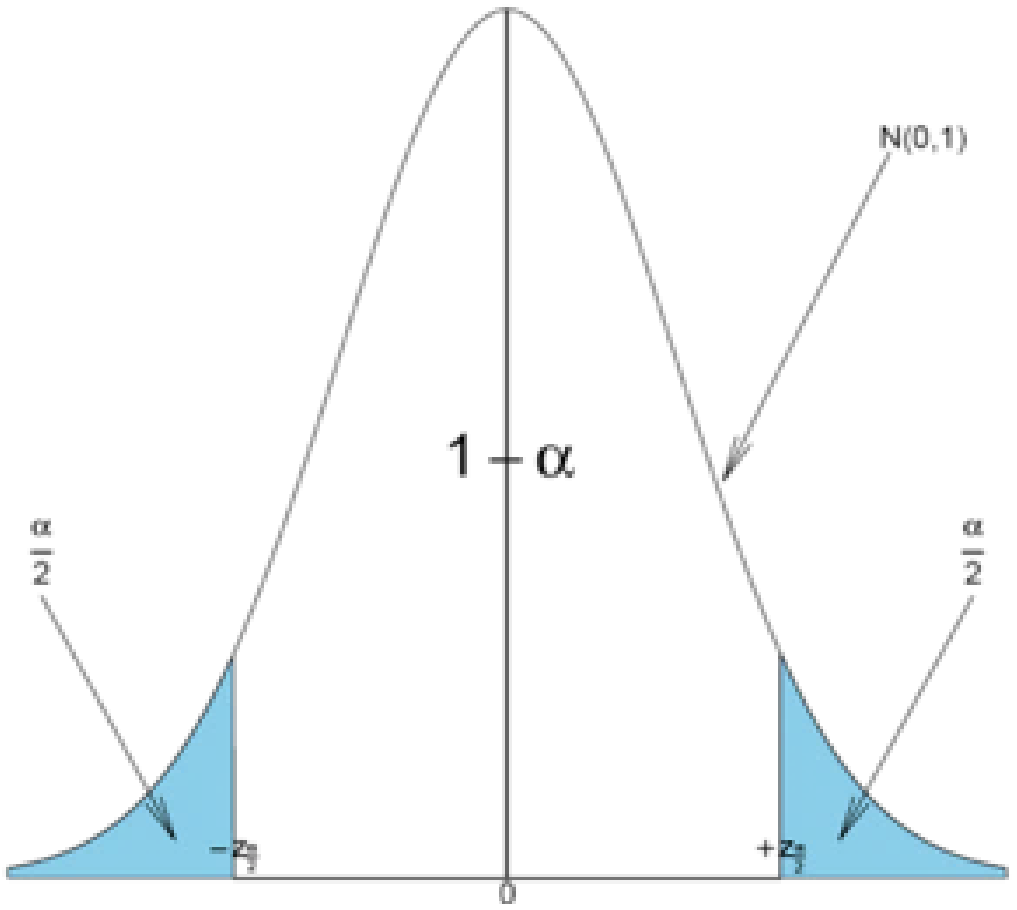




P-value , Z-score

ACADGILD

- The central region on this graph is the acceptance area and the tail is the rejection region, or regions. In this particular graph of a two tailed test, the rejection region is shaded blue. The tail is referred to as "alpha", or p-value (probability value). The area in the tail can be described with z-scores. For example, if the area of the tails was 5% (2.5% each side), the z-score would be 1.96 (from the z-table), which represents 1.96 standard deviations from the mean. The null hypothesis will be rejected if z is less than -1.96 or greater than 1.96.



Unknown Mean And Known Standard Deviation

- For a population with unknown mean(μ) and known standard deviation(σ), a confidence interval for the population mean, based on a simple random sample (SRS) of size n is:

$$\overline{x} \pm z^* \frac{\sigma}{\sqrt{n}},$$

where z* is the upper (1-C)/2 critical value for the standard normal distribution.

- An increase in sample size will decrease the length of the confidence interval without reducing the level of confidence. This is because the standard deviation decreases as *n* increases.
- As the level of confidence decreases, the size of the corresponding interval will decrease.

Unknown Mean And Known Standard Deviation (Contd.)

- Mostly, the standard deviation for the population of interest is not known. In this case, the standard deviation σ is replaced by the estimated standard deviation s, also known as the standard error

- Since the standard error is an estimate for the true value of the standard deviation, the distribution of the sample mean is no longer normal with mean μ and standard deviation σ/√n. Instead, the sample mean follows the t distribution with mean and standard deviation. Instead, the sample mean follows the ***t distribution*** with mean μ, and standard deviation s/√n

T-Distribution

- The use of a t-distribution is precluded by the standard deviation of the population parameter being unknown and allows the analyst to approximate probabilities, based on the mean of the sample, the population, the standard deviation of the sample and the sample's degrees of freedom.
- The t-distribution is also described by its degrees of freedom. For a sample of size n, the t distribution will have n-1 degrees of freedom.
- As the sample size n increases, the t-distribution becomes closer to the normal distribution, since the standard error σ approaches the true standard deviation for large n.
- For a population with unknown mean μ and unknown standard deviation, a confidence interval for the population mean, based on a simple random sample (SRS) of size n, is , where t* is the upper (1-C)/2 critical value for the t-distribution with n-1 degrees of freedom, t(n-1).

Hypothesis Testing

- What is Hypothesis?
  - An educated guess
  - A claim or statement about a property of a population

- What is the goal of Hypothesis Testing?
  - To analyze a sample in an attempt to distinguish between population characteristics, that are likely to occur and population characteristics that are unlikely to occur.

Null Hypothesis Vs Alternative Hypothesis

Null Hypothesis

- Statement about the value of a population parameter
- Represented by H0
- Always stated as an Equality

Alternative Hypothesis

- Statement about the value of a population parameter that must be true if the null hypothesis is false
- Represented by H1
- Stated in on of three forms

  >

  <

  ≠

Type I Vs Type II Errors

- A type I error is the incorrect rejection of a true null hypothesis (a "false positive"), while a type II error is the failure to reject a false null hypothesis (a "false negative")

Examples of type I errors include a test that shows a patient to have a disease when in fact the patient does not have the disease, a fire alarm going off indicating a fire when in fact there is no fire.

- A type II error (or error of the second kind) is the failure to reject a false null hypothesis.

Examples of type II errors would be a blood test failing to detect the disease it was designed to detect, in a patient who really has the disease, a fire breaking out and the fire alarm does not ring.

Type I Vs Type II Errors (Contd.)

ACADGILD

	Condition of null hypothesis	
Possible action	**True**	**False**
Fail to reject H_0	**Correct** $(1-\alpha)$	**Type II error** β
Reject H_0	**Type I error** α	**Correct** $(1-\beta)$

P-value And Significance Level

ACADGILD

P-Value

In statistical significance testing, the p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.

Significance level

The probability of rejecting the null hypothesis is when it is called the significance level α

Note: If the *p*-value is equal to or smaller than the significance level (*α*), it suggests that the observed data is inconsistent with the assumption that the null hypothesis is true and thus this hypothesis must be rejected (but this does not automatically mean the alternative hypothesis can be accepted as true).

α Vs β

- α is the probability of Type I error
- β is the probability of Type II error
- The experimenters (you and I) have the freedom to set the α-level for a particular hypothesis test.
- That level is called the level of significance for the test. Changing α can (and often does) affect the results of the test—whether you reject or fail to reject H_0.
- As α increases, β decreases and vice versa.
- The only way to decrease both α and β is to increase the sample size. To make both quantities equal zero, the sample size would have to be infinite- you would have to sample the entire population.

Steps of Hypothesis Testing

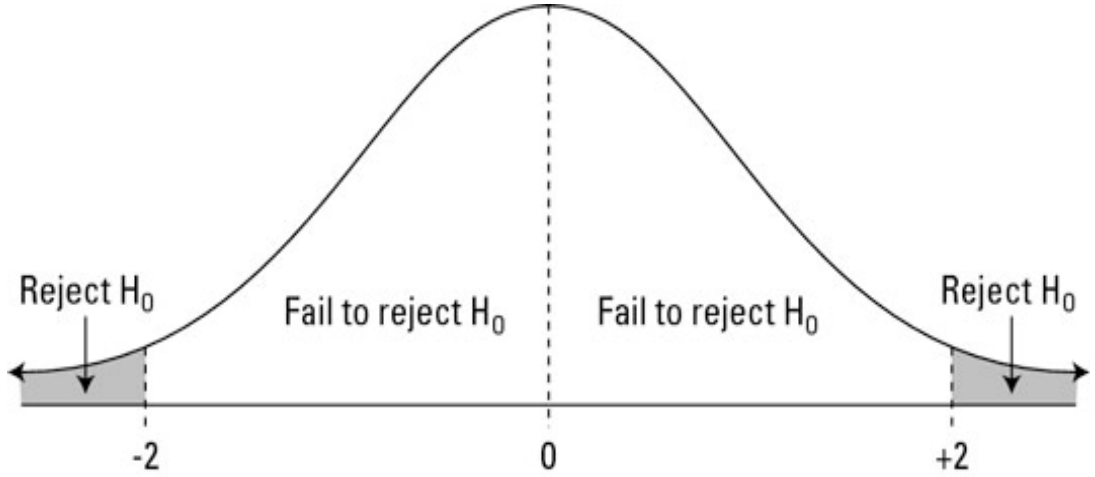
ACADGILD

1. Describe the population characteristic about which hypotheses are to be tested
2. State the null hypothesis, Ho
3. State the alternative hypothesis, H1 or Ha
4. Display the test statistic to be used
5. Identify the rejection region
6. Is it an upper, lower, or two-tailed test?
7. Determine the critical value associated with α, the level of significance of the test
8. Compute all the quantities in the test statistic, and compute the test statistic itself
9. State the conclusion. That is, decide whether to reject the null hypothesis, Ho, or not to reject the null hypothesis. The conclusion depends on the level of significance of the test. Also, remember to state your result in the context of the specific problem

Determine a p-Value - Testing A Null Hypothesis

ACADGILD

- If the alternative hypothesis is the less-than alternative, you reject H0 only if the test statistic falls in the left tail of the distribution (below –2). Similarly, if Ha is the greater-than alternative, you reject H0 only if the test statistic falls in the right tail (above 2)



Determine a p-Value - Testing A Null Hypothesis (Contd.)

- If Ha contains a less-than alternative, find the probability that Z is less than your test statistic
- If Ha contains a greater-than alternative, find the probability that Z is greater than your test statistic (look up your test statistic on the Z-table, find its corresponding probability, and subtract it from one). The result is your p-value.
- If Ha contains a not-equal-to alternative, find the probability that Z is beyond your test statistic and double it.
  - If your test statistic is negative, first find the probability that Z is less than your test statistic (look up your test statistic on the Z-table and find its corresponding probability). Then double this probability to get the p-value.
  - If your test statistic is positive, first find the probability that Z is greater than your test statistic (look up your test statistic on the Z-table, find its corresponding probability, and subtract it from one). Then double this result to get the p-value.

Framing Hypothesis Question

ACADGILD

- **True/false:** First decide on the null hypothesis. Then analyze the data and calculate the probability value. Look at this probability value, and depending on what it is, choose an appropriate alpha level. Then you decide whether you can reject the null hypothesis.

Framing Hypothesis Question (Contd.)

ACADGILD

- Ans – False
- You have to select the alpha level before you calculate the probability value. You compare your probability value to your previously selected alpha level when deciding whether or not you can reject the null hypothesis.

Upper-Tailed, Lower-Tailed, Two-Tailed Tests

- H1: μ > μ 0 , where μ0 is the comparator or null value (e.g., μ0 =191 in the example about weight in men in 2006) and an increase is hypothesized - this type of test is called an upper-tailed test

- H1: μ < μ0 , where a decrease is hypothesized and this is called a lower-tailed test

- H1: μ ≠ μ 0, where a difference is hypothesized and this is called a two-tailed test.

Problem

ACADGILD

- Suppose the seller says that the mean life of a fan is more than 15,000 hours. In a sample of 40 fans, it was found that they only last 14,900 hours on average. Assume the population standard deviation is 110 hours. At .05 significance level, can we reject the claim by the seller?

The null hypothesis is μ ≥ 15000. Begin with computing the test statistic.

```
> xbar = 14900            # sample mean
> mu0 = 15000             # hypothesized value
> sigma = 110           # population standard deviation
> n = 40                # sample size
> z = (xbar−mu0)/(sigma/sqrt(n))
> z                     # test statistic
[1] -5.749091
```

Answer

- The critical value at .05 significance level

```
> alpha = .05
>> –z.alpha            # critical value
[1] –1.6449
```

- The test statistic - 5.749091 is less than the critical value of -1.6449. Hence, at .05 significance level, we reject the claim that mean life of a fan is above 15,000 hours

Problem

ACADGILD

- Suppose the chocolate wrapper states that there is at most 4 grams of saturated fat in a single chocolate. In a sample of 70 chocolates , it is found that the mean amount of saturated fat per chocolate is 4.2 grams. Assume that the population standard deviation is 0.50 grams. At .05 significance level, can we reject the claim on wrapper?

Answer

- The null hypothesis is that μ ≤ 4. We begin with computing the test statistic.

```
> xbar = 4.2              # sample mean
> mu0 = 4                 # hypothesized value
> sigma = 0.50            # population standard deviation
> n = 70                  # sample size
> z = (xbar−mu0)/(sigma/sqrt(n))
> z                       # test statistic
[1] 3.344
```

- We then compute the critical value at .05 significance level.

```
> alpha = .05
> z.alpha                 # critical value
[1] 1.6449
```

- The test statistic 3.344 is greater than the critical value of 1.6449. Hence, at .05 significance level, we reject the claim that there is at most 4 grams of saturated fat in a chocolate.

ACADGILD

THANK YOU

Email us at - support@acadgild.com

LEARN. DO. EARN