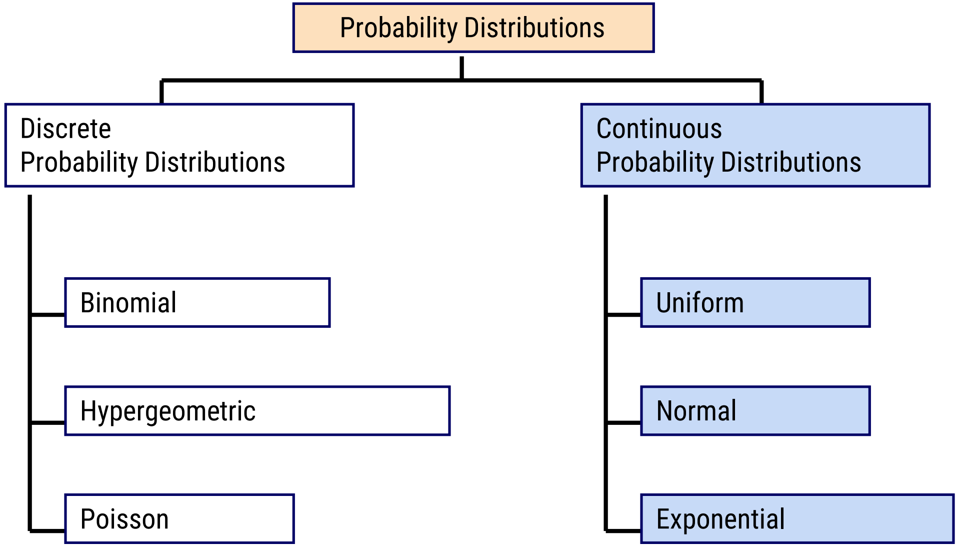
Machine Learning
Session-1

Normal Distribution, Hypothesis Testing, Chi-Sq

Digital Vidya

Probability Distributions

(continued)



Probability Distributions

Discrete Probability Distributions

- Binomial
- Hypergeometric
- Poisson

Continuous Probability Distributions

- Uniform
- Normal
- Exponential

Continuous Probability Distributions

- A continuous random variable is a variable that can assume any value in an interval
  - thickness of an item
  - time required to complete a task
  - temperature of a solution
  - height, in inches
- These can potentially take on any value, depending only on the ability to measure accurately.

Cumulative Distribution Function

- The cumulative distribution function, F(x), for a continuous random variable X expresses the probability that X does not exceed the value of x

$$\boxed{F(x) = P(X \le x)}$$

- Let a and b be two possible values of X, with a < b. The probability that X lies between a and b is

$$\boxed{P(a < X < b) = F(b) - F(a)}$$

Probability Density Function

The probability density function, f(x), of random variable X has the following properties:

- □ $f(x) > 0$ for all values of x
- □ The area under the probability density function f(x) over all values of the random variable X is equal to 1.0
- □ The probability that X lies between two values is the area under the density function graph between the two values
- □ The cumulative density function $F(x_0)$ is the area under the probability density function $f(x)$ from the minimum x value up to x_0

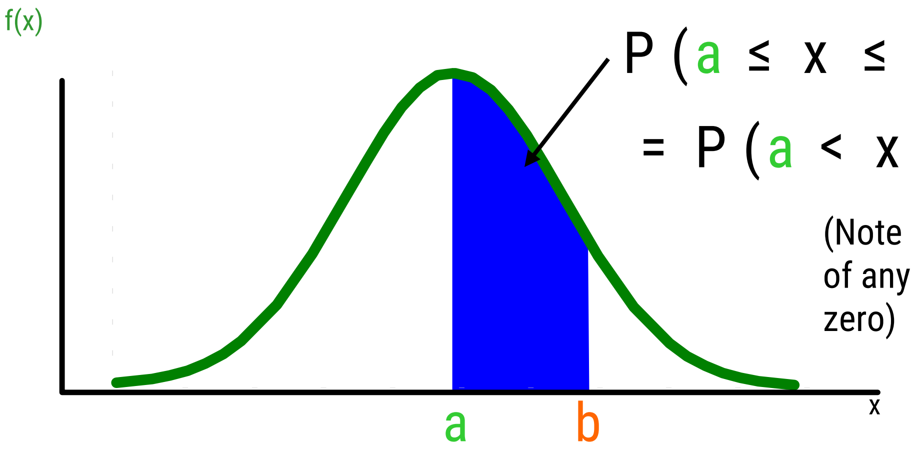
$$\boxed{f(x_0) = \int_{x_m}^{x_0} f(x)dx}$$

where x_m is the minimum value of the random variable x

Digital Vidya

Probability as an Area

Shaded area under the curve is the probability that X is between a and b



$$P(a \le x \le b)$$

$$= P(a < x < b)$$

(Note that the probability of any individual value is zero)

Linear Functions of Variables

(continued)

- An important special case of the previous results is the standardized random variable

$$\boxed{Z = \frac{X - \mu_X}{\sigma_X}}$$

- which has a mean 0 and variance 1

The Normal Distribution

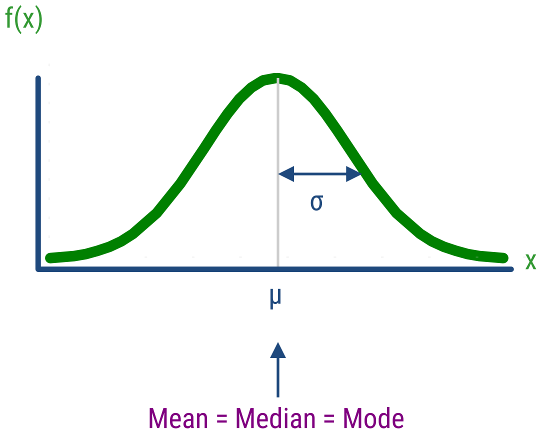
(continued)

- ‘Bell Shaped’
- Symmetrical
- Mean, Median and Mode are Equal

Location is determined by the mean, μ

Spread is determined by the standard deviation, σ

The random variable has an infinite theoretical range:
$+\infty$ to $-\infty$

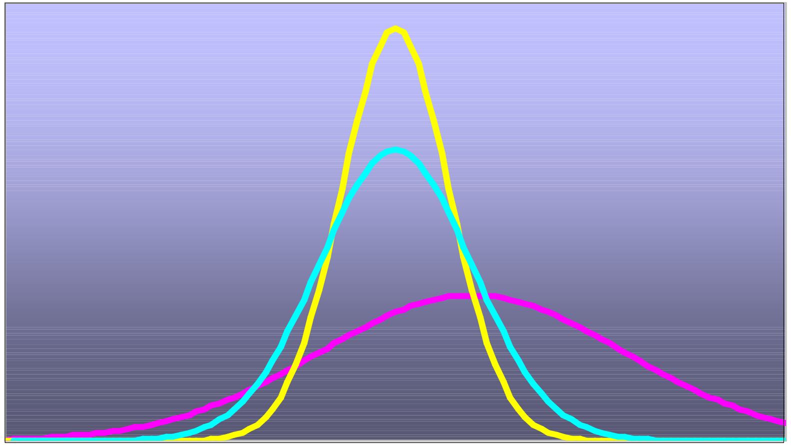


The Normal Distribution

(continued)

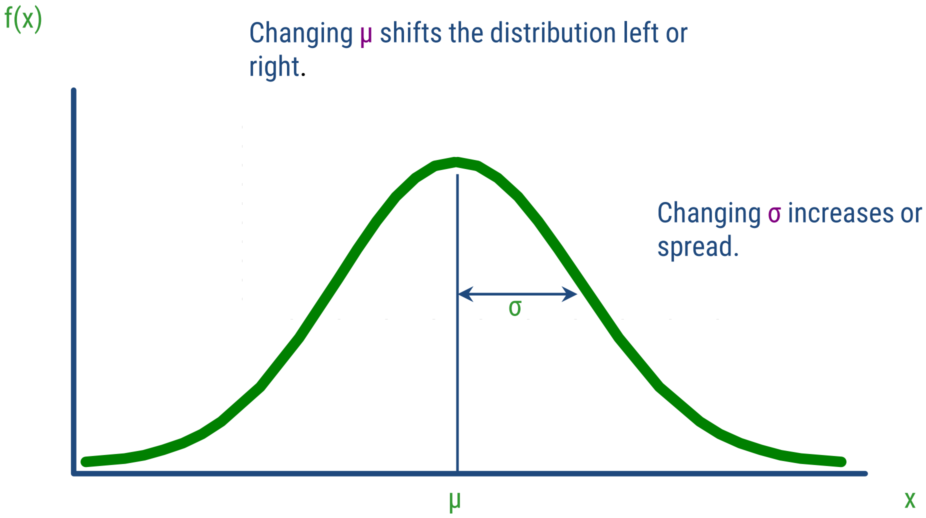
- The normal distribution closely approximates the probability distributions of a wide range of random variables
- Distributions of sample means approach a normal distribution given a “large” sample size
- Computations of probabilities are direct and elegant
- The normal probability distribution has led to good business decisions for a number of applications

Many Normal Distributions



By varying the parameters μ and σ, we obtain different normal distributions

The Normal Distribution Shape



Changing μ shifts the distribution left or right.

Changing σ increases or decreases the spread.

Given the mean μ and variance σ we define the normal distribution using the notation

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

Normal Probability Density Function

- The formula for the normal probability density function is

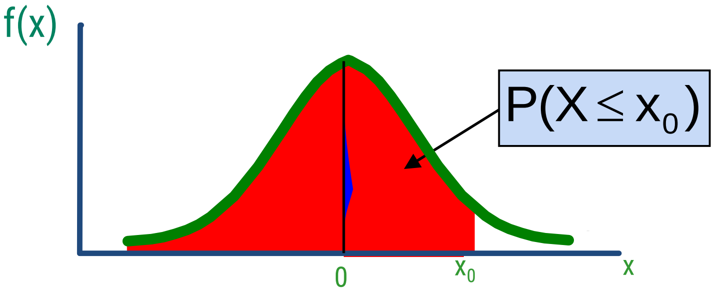
$$\boxed{f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}}$$

Where e = the mathematical constant approximated by 2.71828
π = the mathematical constant approximated by 3.14159
μ = the population mean
σ = the population standard deviation
x = any value of the continuous variable, $-\infty < x < \infty$

Cumulative Normal Distribution

- For a normal random variable X with mean μ and variance σ^2 , i.e., $X \sim N(\mu, \sigma^2)$, the cumulative distribution function is

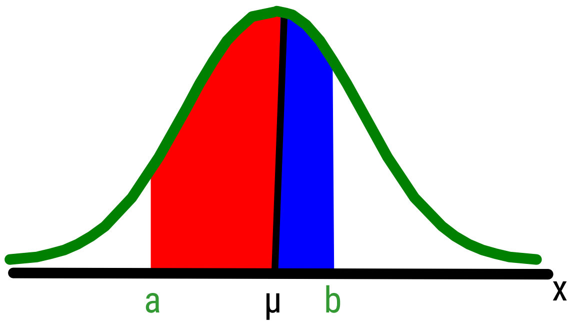
$$\boxed{F(x_0) = P(X \le x_0)}$$



Finding Normal Probabilities

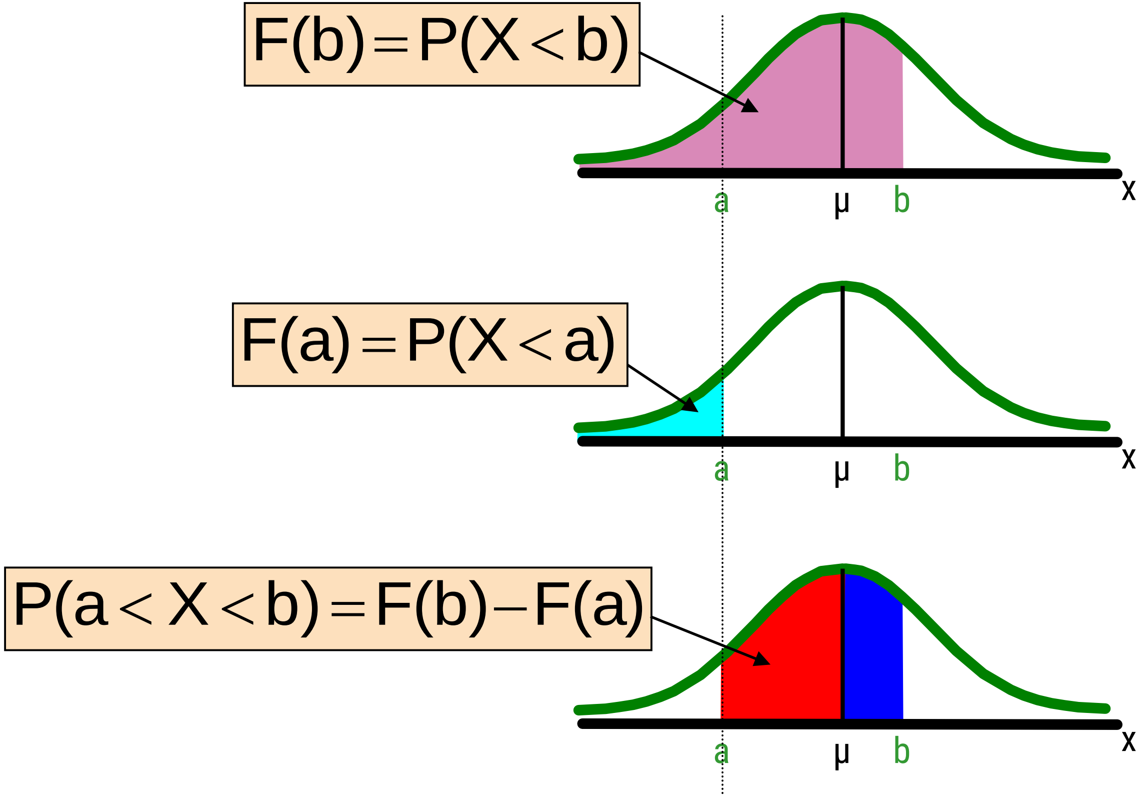
The probability for a range of values is measured by the area under the curve

$$\boxed{P(a < X < b) = F(b) - F(a)}$$



Finding Normal Probabilities

(continued)



$F(b) = P(X < b)$

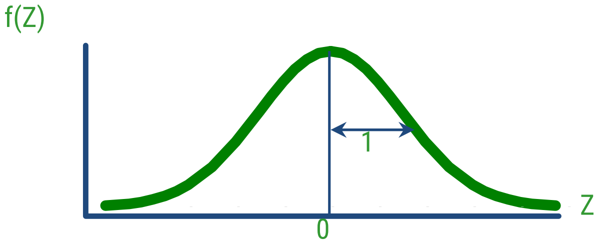
$F(a) = P(X < a)$

$P(a < X < b) = F(b) - F(a)$

The Standardized Normal

- Any normal distribution (with any mean and variance combination) can be transformed into the standardized normal distribution (Z), with mean 0 and variance 1

$$\boxed{Z \sim N(0,1)}$$



- Need to transform X units into Z units by subtracting the mean of X and dividing by its standard deviation

$$\boxed{Z = \frac{X - \mu}{\sigma}}$$

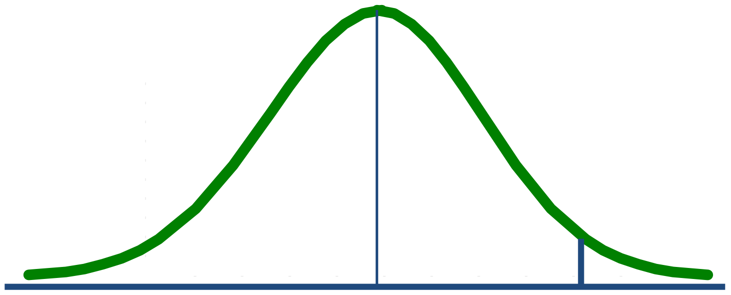
Example

- If X is distributed normally with mean of 100 and standard deviation of 50, the Z value for X = 200 is

$$\boxed{Z = \frac{X - \mu}{\sigma} = \frac{200 - 100}{50} = 2.0}$$

- This says that X = 200 is two standard deviations (2 increments of 50 units) above the mean of 100.

Comparing X and Z units



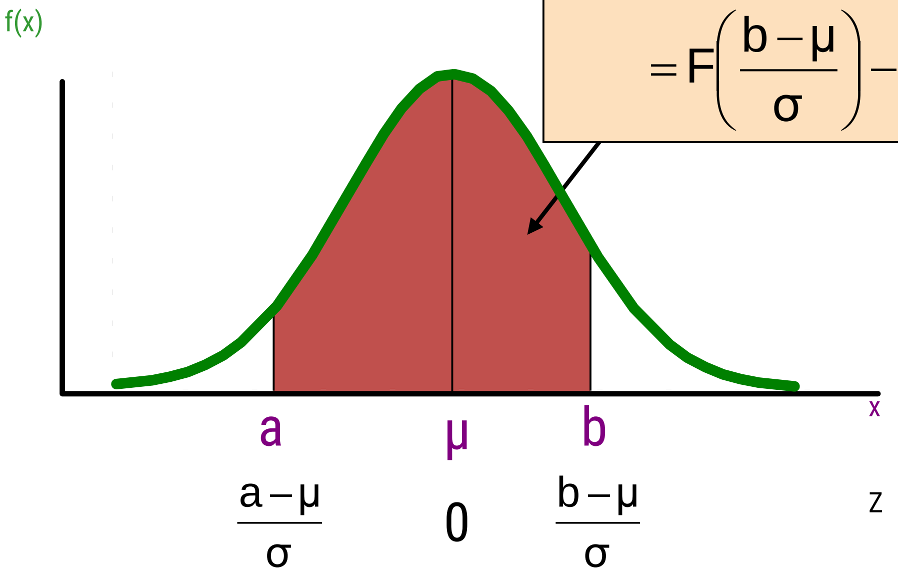
100	200	X	(μ = 100, σ = 50)
0	2.0	Z	(μ = 0, σ = 1)

Note that the distribution is the same, only the scale has changed. We can express the problem in original units (X) or in standardized units (Z)

Finding Normal Probabilities

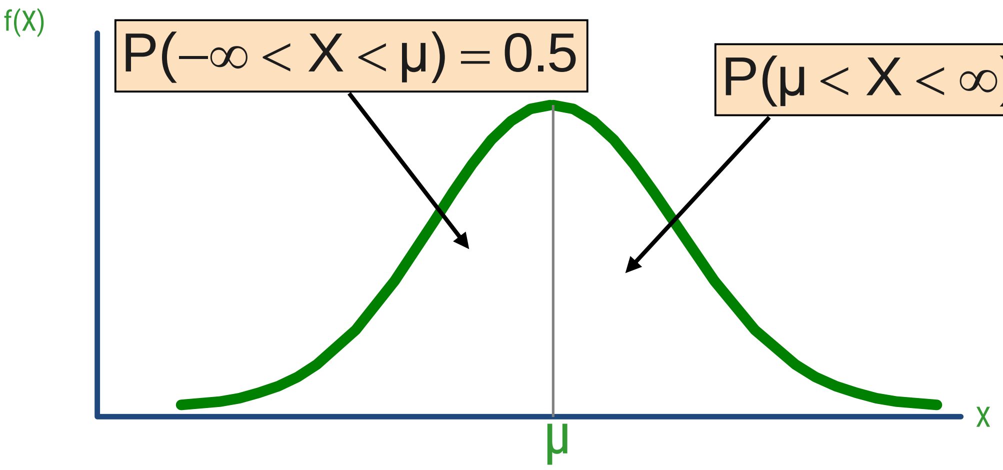
$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$

$$= F\left(\frac{b-\mu}{\sigma}\right) - F\left(\frac{a-\mu}{\sigma}\right)$$



Probability as Area Under the Curve

The total area under the curve is 1.0, and the curve is symmetric, so half is above the mean, half is below



$P(-\infty < X < \mu) = 0.5$

$P(\mu < X < \infty) = 0.5$

$P(-\infty < X < \infty) = 1.0$

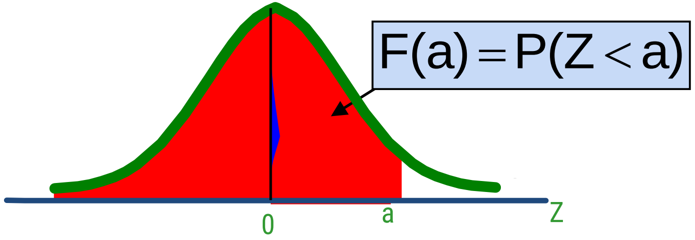
Digital Vidya

Standard Normal Table

- The Standardized Normal table shows values of the cumulative normal distribution function

- For a given Z-value a , the table shows F(a)

  (the area under the curve from negative infinity to a)

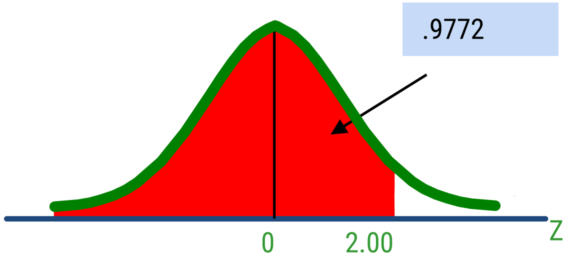


The Standardized Normal Table

- Standard Normal Table gives the probability F(a) for any value a

Example:

P(Z < 2.00) = .9772



The Standardized Normal Table

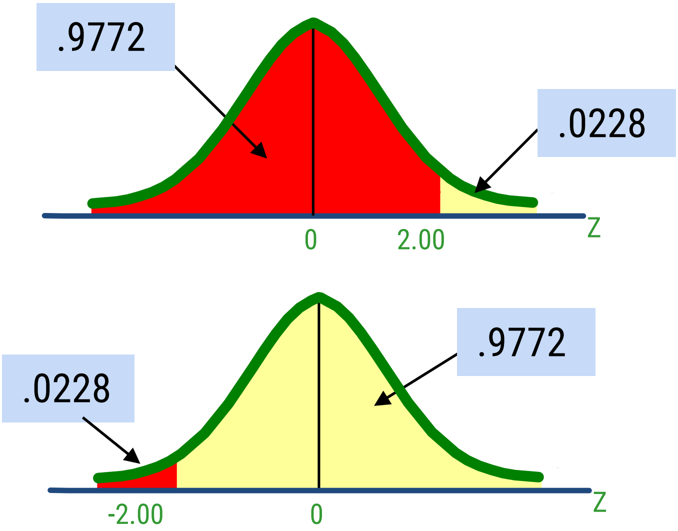
(continued)

- For negative Z-values, use the fact that the distribution is symmetric to find the needed probability:

Example:

P(Z < -2.00) = 1 – 0.9772

= 0.0228



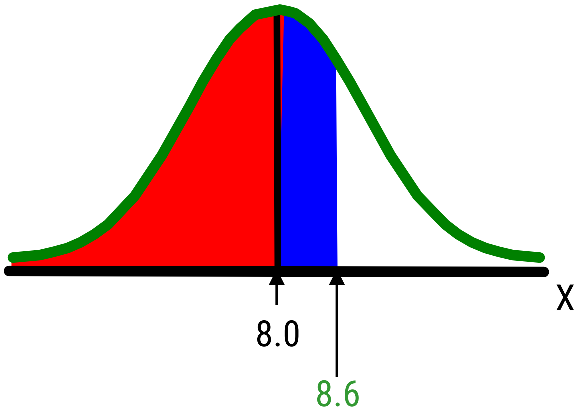
General Procedure for Finding Probabilities

To find $P(a < X < b)$ when X is distributed normally:

- Draw the normal curve for the problem in terms of X
- Translate X-values to Z-values
- Use the Cumulative Normal Table

Finding Normal Probabilities

- Suppose X is normal with mean 8.0 and standard deviation 5.0
- Find P(X < 8.6)

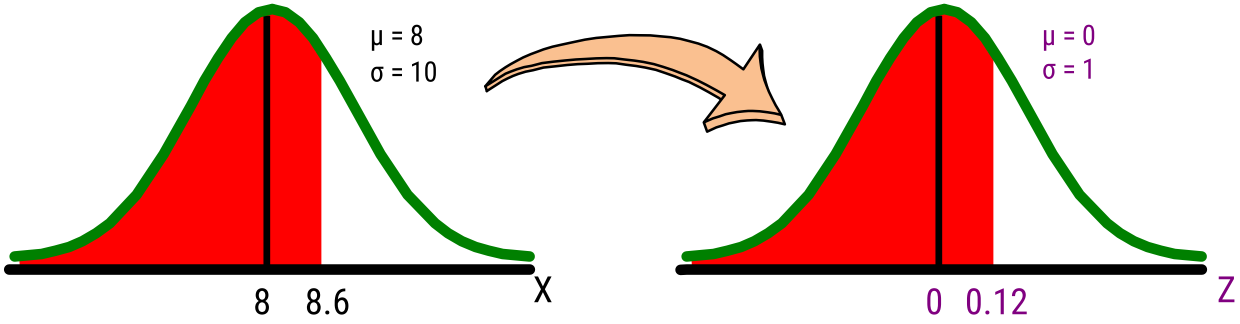


Finding Normal Probabilities

(continued)

- Suppose X is normal with mean 8.0 and standard deviation 5.0. Find P(X < 8.6)

$$\boxed{Z = \frac{X - \mu}{\sigma} = \frac{8.6 - 8.0}{5.0} = 0.12}$$



P(X < 8.6)

P(Z < 0.12)

Solution: Finding P(Z < 0.12)

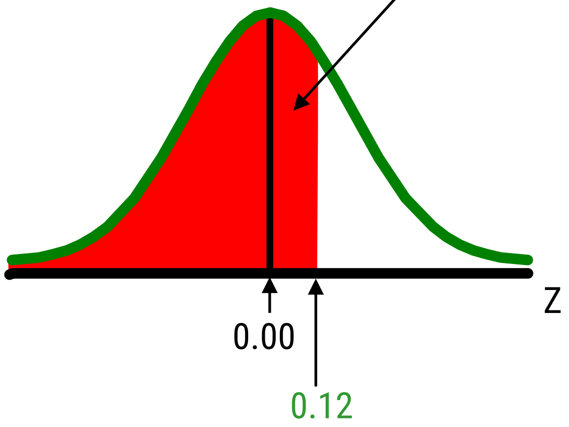
Standardized Normal Probability Table (Portion)

z	F(z)
.10	.5398
.11	.5438
.12	.5478
.13	.5517

P(X < 8.6)

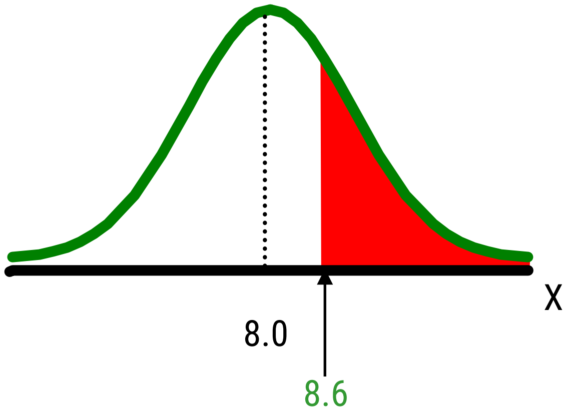
= P(Z < 0.12)

F(0.12) = 0.5478



Upper Tail Probabilities

- Suppose X is normal with mean 8.0 and standard deviation 5.0.
- Now Find P(X > 8.6)



Upper Tail Probabilities

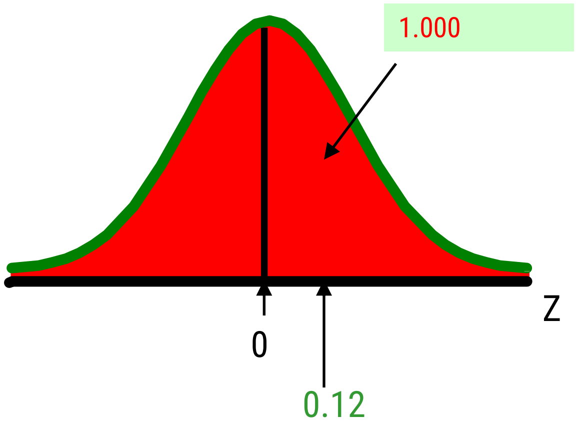
(continued)

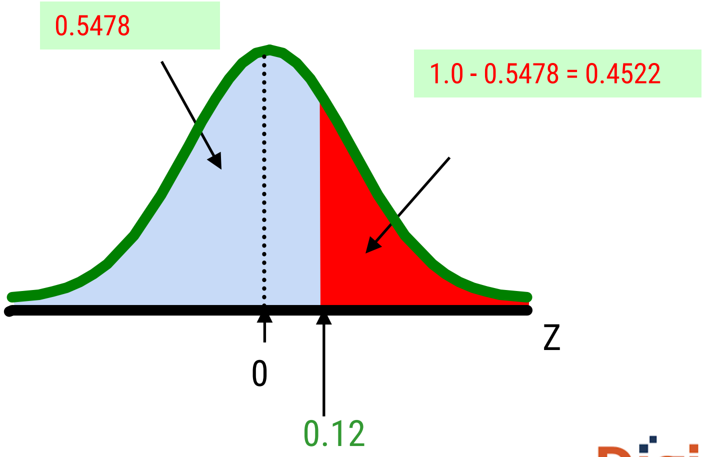
- Now Find P(X > 8.6)…

$P(X > 8.6) = P(Z > 0.12)$

$= 1.0 - P(Z \le 0.12)$

$= 1.0 - 0.5478 = \boxed{0.4522}$





Finding the X value for a Known Probability

- Steps to find the X value for a known probability:

1. Find the Z value for the known probability

2. Convert to X units using the formula:

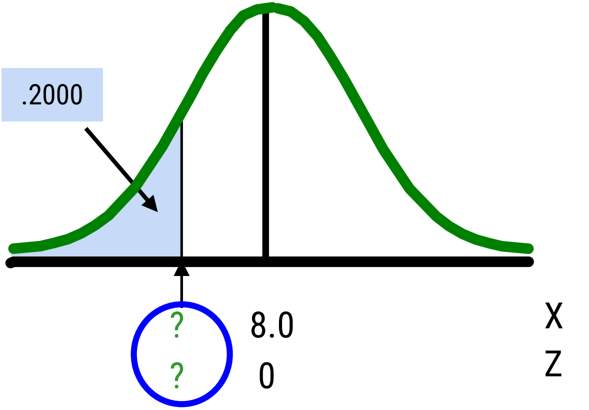
$$\boxed{X = \mu + Z\sigma}$$

Finding the X value for a Known Probability

(continued)

Example:

- Suppose X is normal with mean 8.0 and standard deviation 5.0.
- Now find the X value so that only 20% of all values are below this X



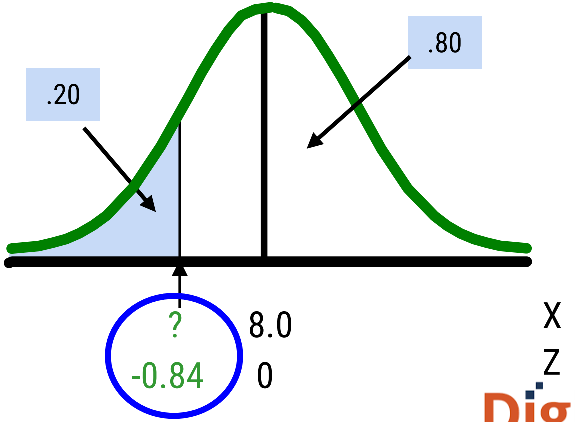
Find the Z value for 20% in the Lower Tail

1. Find the Z value for the known probability

Standardized Normal Probability Table (Portion)

z	F(z)
.82	.7939
.83	.7967
.84	.7995
.85	.8023

20% area in the lower tail is consistent with a Z value of -0.84



Finding the X value

2. Convert to X units using the formula:

$$\begin{aligned} X &= \mu + Z\sigma \\ &= 8.0 + (-0.84)5.0 \\ &= 3.80 \end{aligned}$$

So 20% of the values from a distribution with mean 8.0 and standard deviation 5.0 are less than 3.80

Assessing Normality

- Not all continuous random variables are normally distributed

- It is important to evaluate how well the data is approximated by a normal distribution

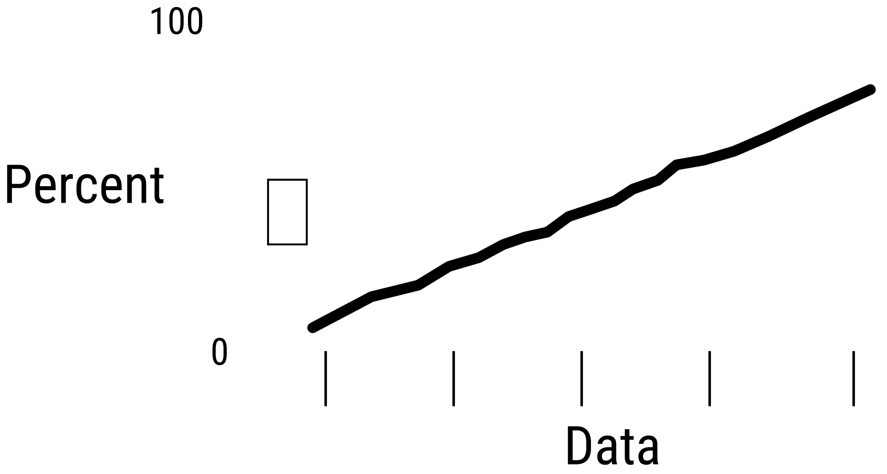
The Normal Probability Plot

- Normal probability plot
  - Arrange data from low to high values
  - Find cumulative normal probabilities for all values
  - Examine a plot of the observed values vs. cumulative probabilities (with the cumulative normal probability on the vertical axis and the observed data values on the horizontal axis)
  - Evaluate the plot for evidence of linearity

The Normal Probability Plot

(continued)

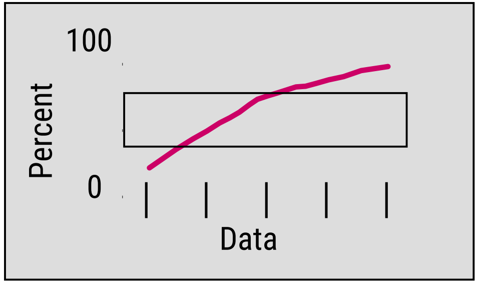
A normal probability plot for data from a normal distribution will be approximately linear:



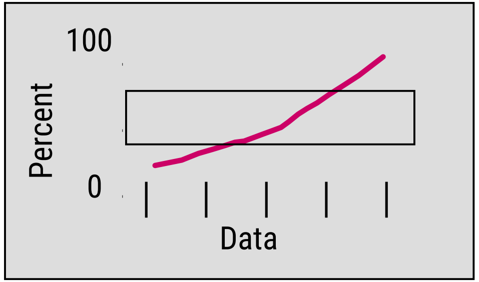
The Normal Probability Plot

(continued)

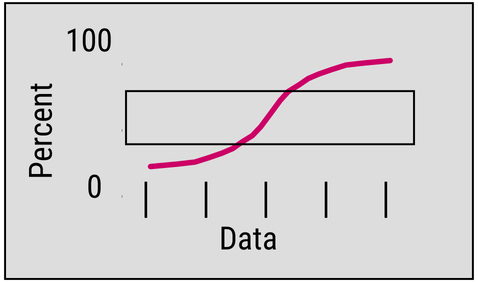
Left-Skewed



Right-Skewed



Uniform

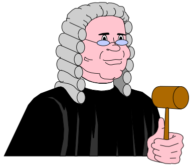


Nonlinear plots indicate a deviation from normality

The Null Hypothesis, H_0

(continued)

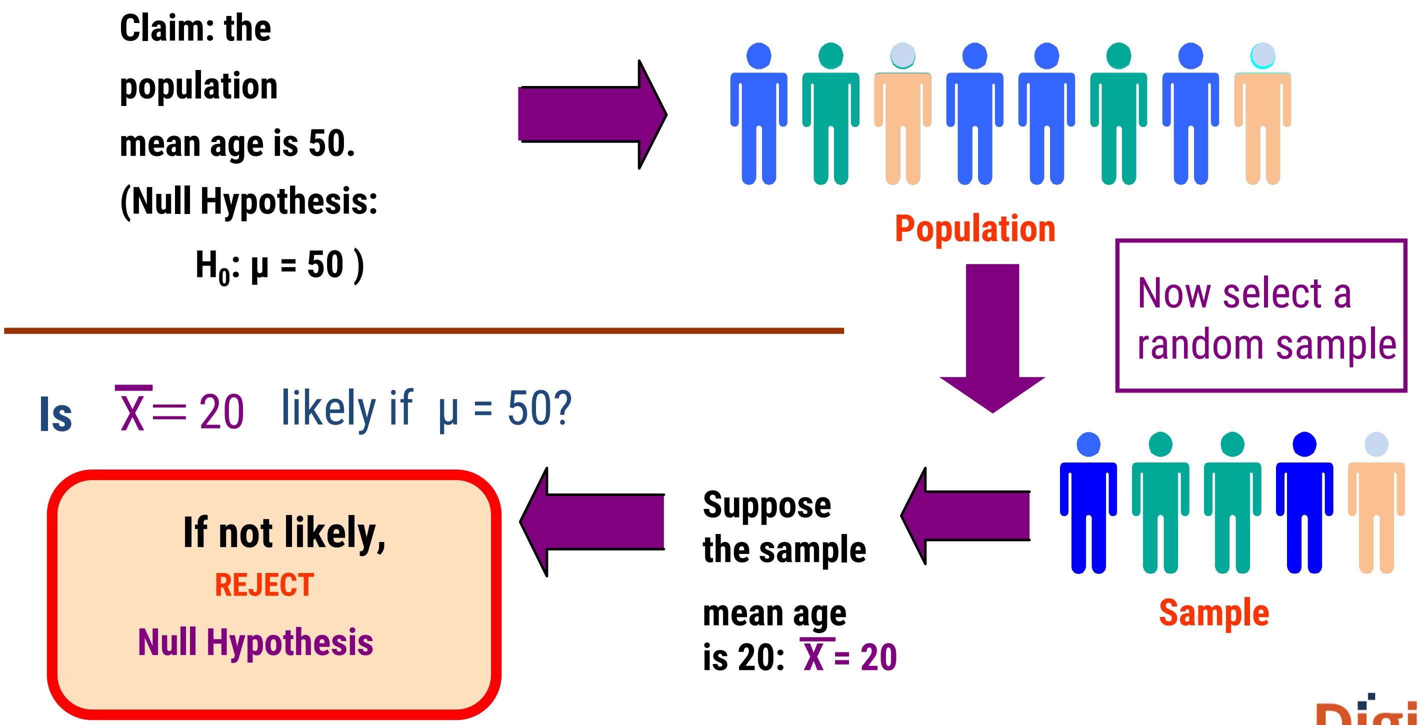
- Begin with the assumption that the null hypothesis is true
  - Similar to the notion of innocent until proven guilty
- Refers to the status quo
- Always contains "=" , "≤" or "≥" sign
- May or may not be rejected



The Alternative Hypothesis, H_1

- Is the opposite of the null hypothesis
  - e.g., The average number of TV sets in U.S. homes is not equal to 3 ($H_1: \mu \neq 3$)
- Challenges the status quo
- Never contains the “=” , “≤” or “≥” sign
- May or may not be supported
- Is generally the hypothesis that the researcher is trying to support

Hypothesis Testing Process



Claim: the population mean age is 50. (Null Hypothesis:

H_0: $\mu = 50$)

Population

Now select a random sample

Sample

Suppose the sample mean age is 20: $\bar{X} = 20$

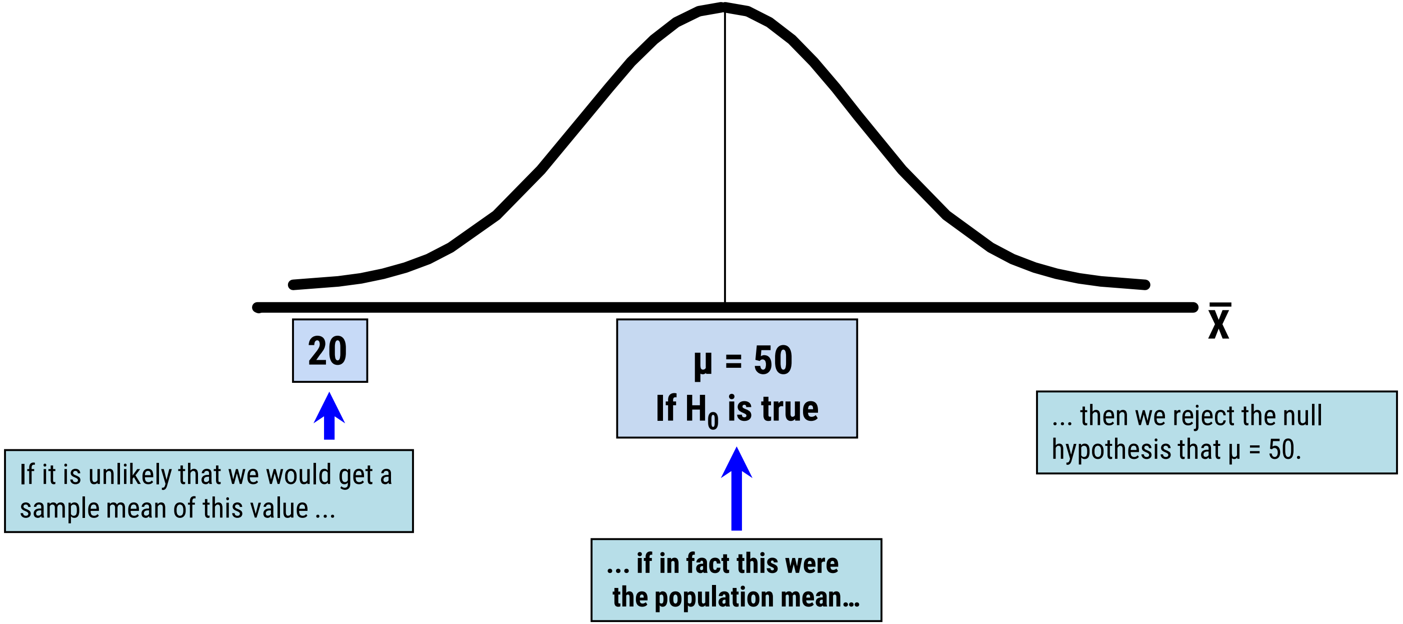
Is $\bar{X} = 20$ likely if $\mu = 50$?

If not likely, REJECT Null Hypothesis

Digital Vidya

Reason for Rejecting H_0

Sampling Distribution of $\overline{X}$



If it is unlikely that we would get a sample mean of this value ...

... if in fact this were the population mean…

... then we reject the null hypothesis that μ = 50.

Level of Significance, α

- **Defines the unlikely values of the sample statistic if the null hypothesis is true**
  - Defines rejection region of the sampling distribution
- Is designated by α , (level of significance)
  - Typical values are .01, .05, or .10
- Is selected by the researcher at the beginning
- Provides the critical value(s) of the test

Level of Significance and the Rejection Region

Level of significance = α

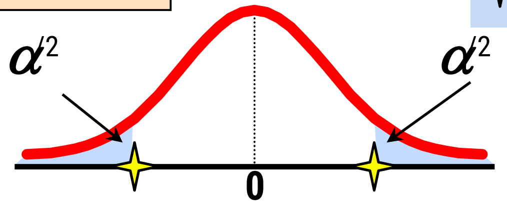
✦ Represents critical value

Rejection region is shaded

H_0: μ = 3

H_1: μ ≠ 3

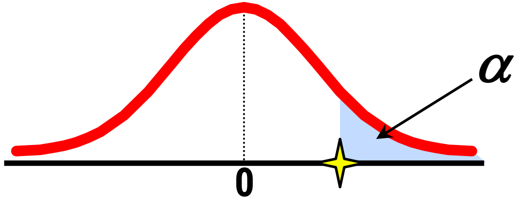
Two-tail test



H_0: μ ≤ 3

H_1: μ > 3

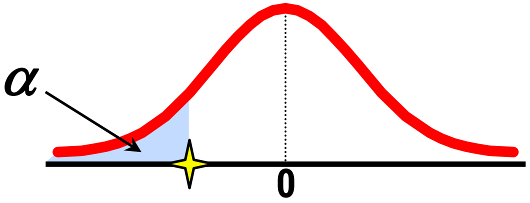
Upper-tail test



H_0: μ ≥ 3

H_1: μ < 3

Lower-tail test



Errors in Making Decisions

- **Type I Error**
  - Reject a true null hypothesis
  - Considered a serious type of error

The probability of Type I Error is α

  - Called level of significance of the test
  - Set by researcher in advance

Errors in Making Decisions

(continued)

- **Type II Error**
  - Fail to reject a false null hypothesis

The probability of Type II Error is β

Outcomes and Probabilities

Possible Hypothesis Test Outcomes

	Actual Situation	
Decision	**H_0 True**	**H_0 False**
Do Not Reject H_0	No error (1 - α)	Type II Error (β)
Reject H_0	Type I Error (α)	No Error (1 - β)

Key:
Outcome
(Probability)

Digital Vidya

Type I & II Error Relationship

- Type I and Type II errors can not happen at the same time
- Type I error can only occur if H_0 is true
- Type II error can only occur if H_0 is false

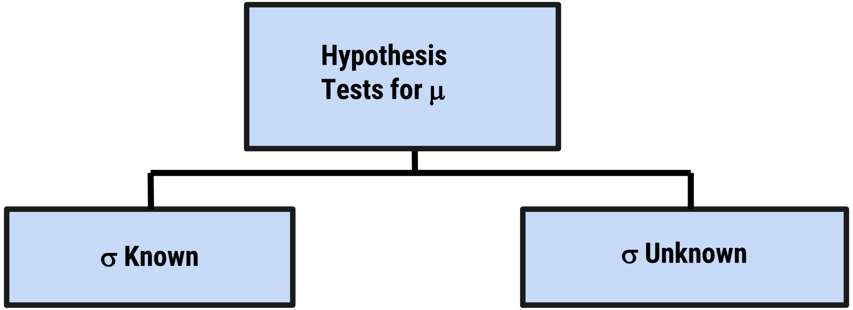
If Type I error probability (α) ⬆, then

Type II error probability (β) ⬇

Factors Affecting Type II Error

- All else equal,
  - β ⬆ when the difference between hypothesized parameter and its true ⬇ value
  - β ⬆ when α ⬇
  - β ⬆ when σ ⬆
  - β ⬆ when *n* ⬇

Hypothesis Tests for the Mean



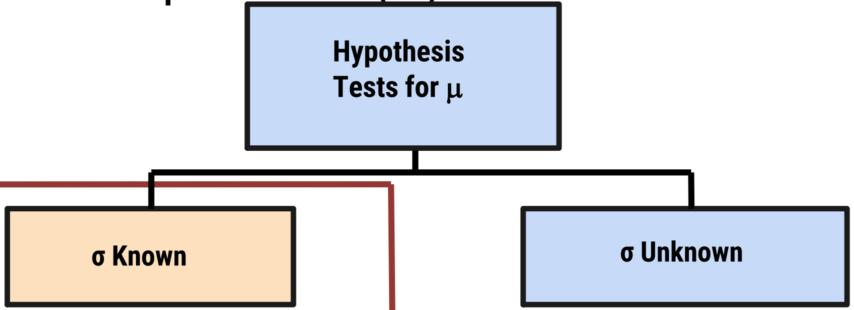
Hypothesis Tests for μ

σ Known

σ Unknown

Test of Hypothesis for the Mean (σ Known)

- Convert sample result ($\bar{x}$) to a z value



Hypothesis Tests for μ

σ Known

σ Unknown

Consider the test

$$\boxed{\begin{array}{l} H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 \end{array}}$$

(Assume the population is normal)

The decision rule is:

$$\boxed{\text{Reject } H_0 \text{ if } \; z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_\alpha}$$

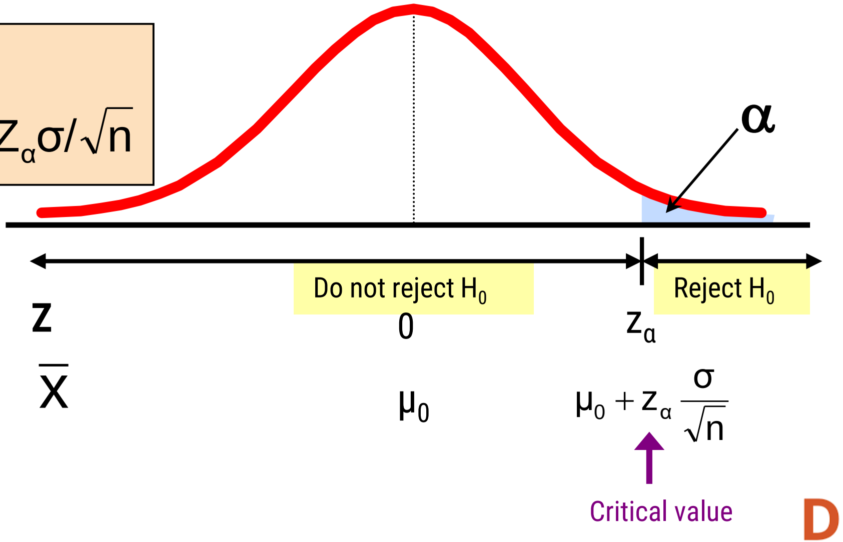
Decision Rule

$$\text{Reject } H_0 \text{ if } \; z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_\alpha$$

H_0: $\mu = \mu_0$
H_1: $\mu > \mu_0$

Alternate rule:

$$\text{Reject } H_0 \text{ if } \; \bar{X} > \mu_0 + Z_\alpha \sigma/\sqrt{n}$$



Digital Vidya

p-Value Approach to Testing

- p-value: Probability of obtaining a test statistic more extreme (≤ or ≥) than the observed sample value given H_0 is true
  - Also called observed level of significance
  - Smallest value of α for which H_0 can be rejected

p-Value Approach to Testing

(continued)

- Convert sample result (e.g., $\bar{X}$) to test statistic (e.g., z statistic)
- Obtain the p-value
  - For an upper tail test:

$$\text{p-value} = P\left(Z > \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}, \text{ given that } H_0 \text{ is true}\right)$$

$$= P\left(Z > \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \mid \mu = \mu_0\right)$$

- Decision rule: compare the p-value to α
  - If p-value < α , reject H_0
  - If p-value $\geq \alpha$, do not reject H_0

Digital Vidya

Example: Upper-Tail Z Test for Mean (σ Known)

A phone industry manager thinks that customer monthly cell phone bill have increased, and now average over $52 per month. The company wishes to test this claim.

(Assume $\sigma = 10$ is known)

Form hypothesis test:

H_0: $\mu \le 52$ the average is not over $52 per month

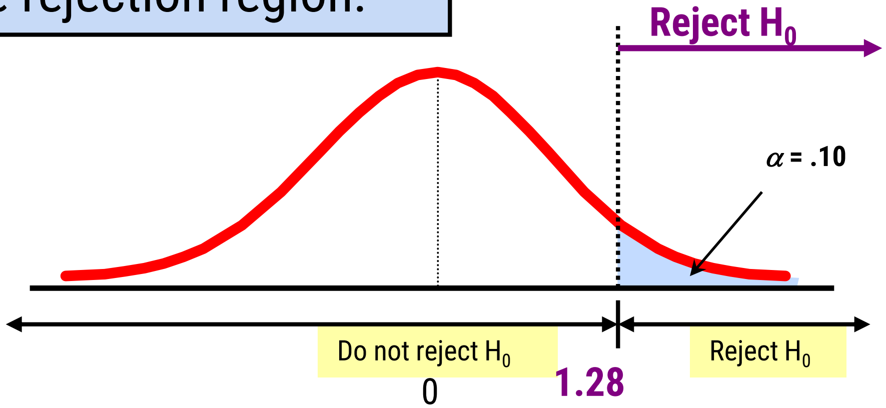
H_1: $\mu > 52$ the average **is** greater than $52 per month (i.e., sufficient evidence exists to support the manager's claim)

Example: Find Rejection Region

(continued)

- Suppose that α = .10 is chosen for this test

Find the rejection region:



$$\text{Reject } H_0 \text{ if } z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > 1.28$$

Example: Sample Results

(continued)

Obtain sample and compute the test statistic

Suppose a sample is taken with the following results:

$n = 64$, $\bar{x} = 53.1$ (σ=10 was assumed known)

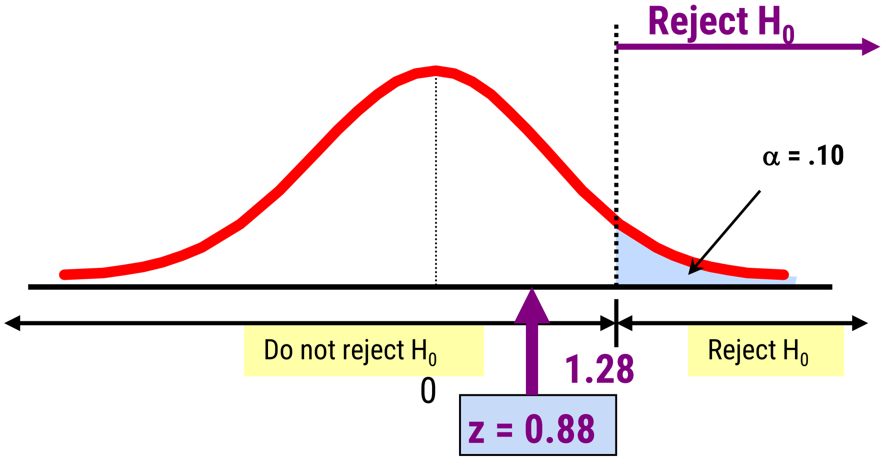
- Using the sample results,

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{53.1 - 52}{\frac{10}{\sqrt{64}}} = 0.88$$

Example: Decision

(continued)

Reach a decision and interpret the result:



Do not reject H_0 since z = 0.88 < 1.28

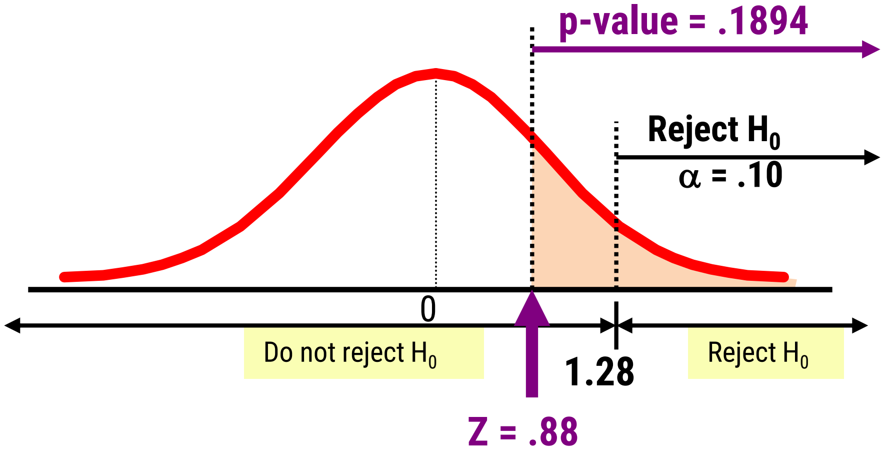
i.e.: there is not sufficient evidence that the mean bill is over $52

Example: p-Value Solution

(continued)

Calculate the p-value and compare to α

(assuming that $\mu = 52.0$)



$$P(\bar{x} \geq 53.1 | \mu = 52.0)$$

$$= P\left(z \geq \frac{53.1 - 52.0}{10/\sqrt{64}} \right)$$

$$= P(z \geq 0.88) = 1 - .8106$$

$$= .1894$$

Do not reject H_0 since p-value = .1894 > α = .10

Digital Vidya

One-Tail Tests

- In many cases, the alternative hypothesis focuses on one particular direction

$H_0: \mu \le 3$
$H_1: \mu > 3$

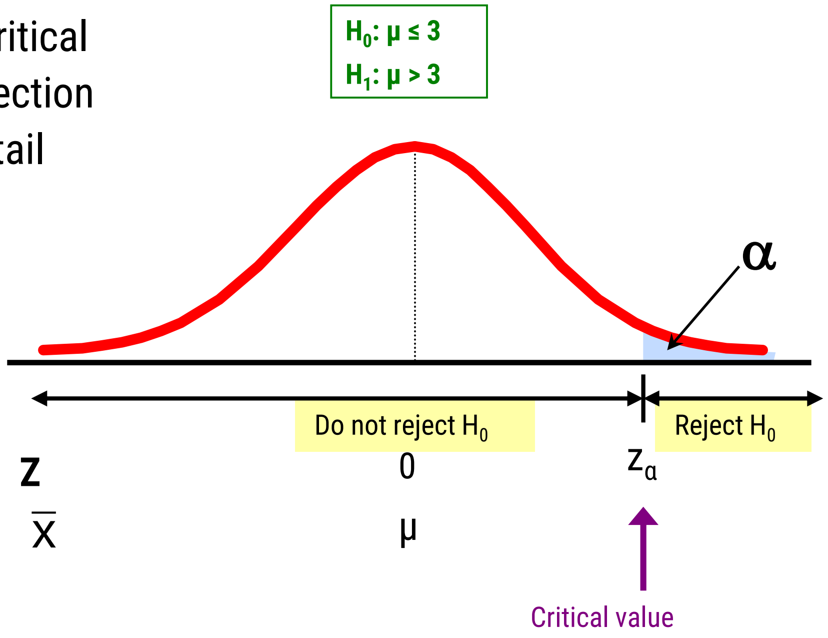
→ This is an upper-tail test since the alternative hypothesis is focused on the upper tail above the mean of 3

$H_0: \mu \ge 3$
$H_1: \mu < 3$

→ This is a lower-tail test since the alternative hypothesis is focused on the lower tail below the mean of 3

Upper-Tail Tests

- There is only one critical value, since the rejection area is in only one tail



H_0: $\mu \leq 3$

H_1: $\mu > 3$

α

Do not reject H_0

Reject H_0

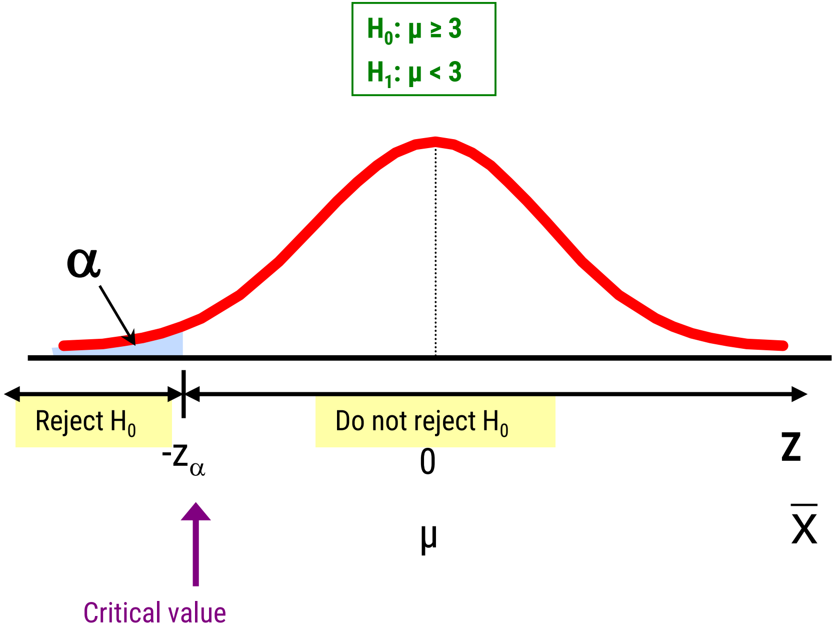
Z 0 z_α

$\bar{X}$ μ

Critical value

Lower-Tail Tests

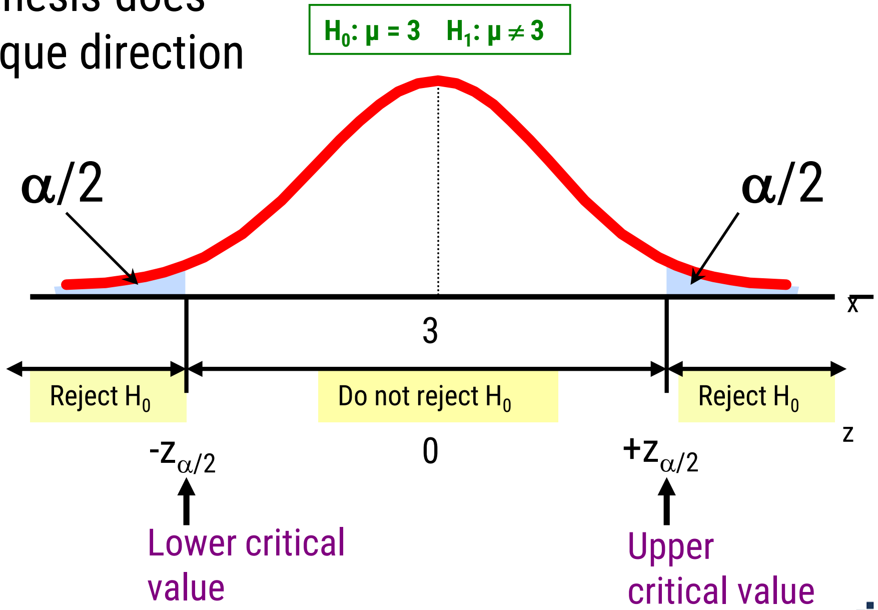
- There is only one critical value, since the rejection area is in only one tail



Two-Tail Tests

- In some settings, the alternative hypothesis does not specify a unique direction

- There are two critical values, defining the two regions of rejection

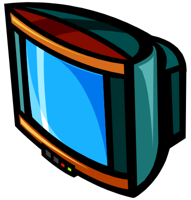


Hypothesis Testing Example

Test the claim that the true mean # of TV sets in US homes is equal to 3.

(Assume σ = 0.8)

- State the appropriate null and alternative hypotheses
  - H_0: μ = 3 , H_1: μ ≠ 3 (This is a two tailed test)
- Specify the desired level of significance
  - Suppose that α = .05 is chosen for this test
- Choose a sample size
  - Suppose a sample of size n = 100 is selected

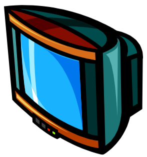


Hypothesis Testing Example

(continued)

- Determine the appropriate technique
  - σ is known so this is a z test
- Set up the critical values
  - For α = .05 the critical z values are ±1.96
- Collect the data and compute the test statistic
  - Suppose the sample results are

    n = 100, $\overline{x}$ = 2.84 (σ = 0.8 is assumed known)



So the test statistic is:

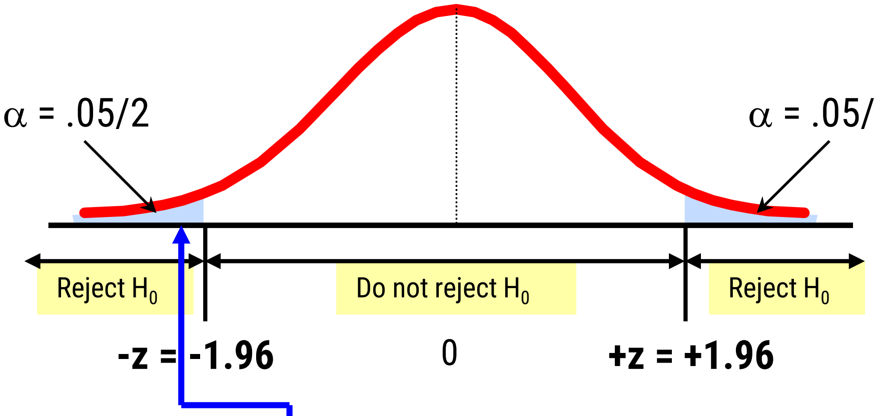
$$\boxed{z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{2.84 - 3}{\frac{0.8}{\sqrt{100}}} = \frac{-.16}{.08} = -2.0}$$

Hypothesis Testing Example

(continued)

- Is the test statistic in the rejection region?

Reject H_0 if $z < -1.96$ or $z > 1.96$; otherwise do not reject H_0

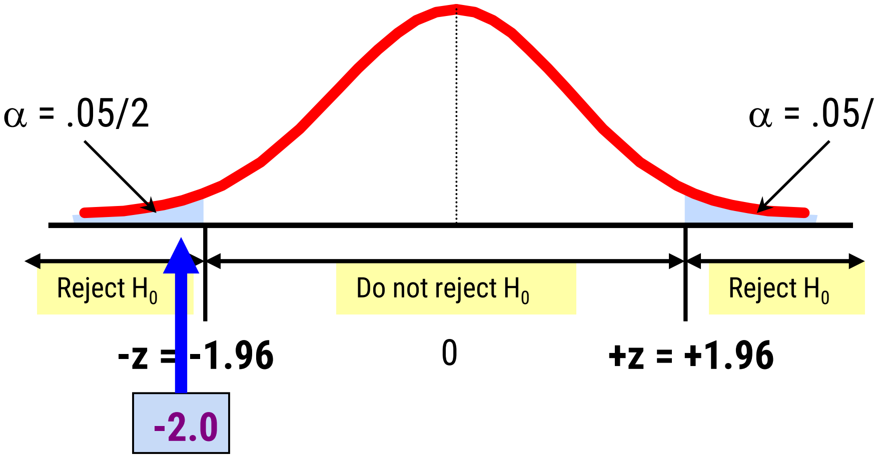


Here, $z = -2.0 < -1.96$, so the test statistic is in the rejection region

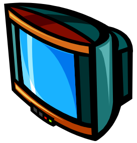
Hypothesis Testing Example

(continued)

- Reach a decision and interpret the result



Since z = -2.0 < -1.96, we reject the null hypothesis and conclude that there is sufficient evidence that the mean number of TVs in US homes is not equal to 3



Example: p-Value

- Example: How likely is it to see a sample mean of 2.84 (or something further from the mean, in either direction) if the true mean is μ = 3.0?

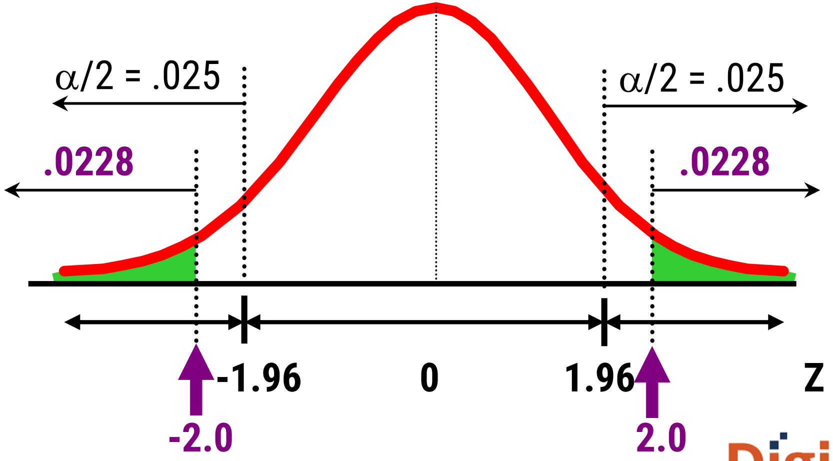
$\bar{x}$ = 2.84 is translated to a z score of z = -2.0

$$P(z < -2.0) = .0228$$

$$P(z > 2.0) = .0228$$

p-value

= .0228 + .0228 = .0456



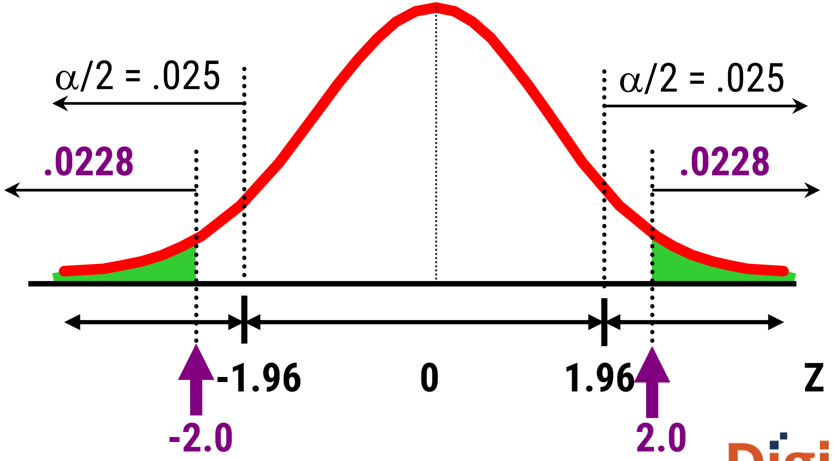
Example: p-Value

(continued)

- Compare the p-value with α
  - If p-value $< \alpha$, reject H_0
  - If p-value $\geq \alpha$, do not reject H_0

Here: p-value = .0456
α = .05

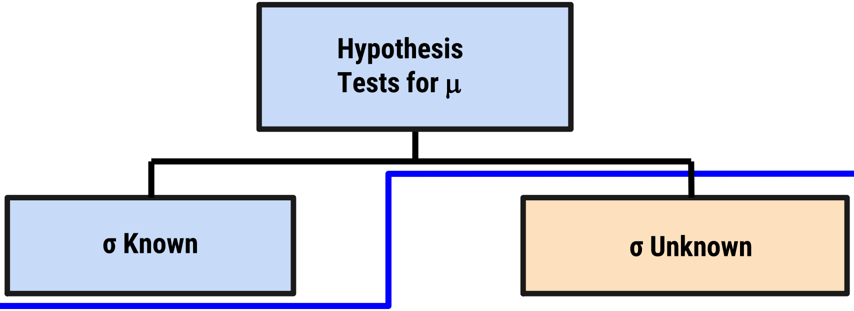
Since .0456 < .05, we reject the null hypothesis



Digital Vidya

t Test of Hypothesis for the Mean (σ Unknown)

- Convert sample result ($\bar{x}$) to a t test statistic



Consider the test

$$\boxed{\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu > \mu_0 \end{aligned}}$$

(Assume the population is normal)

The decision rule is:

$$\text{Reject } H_0 \text{ if } \quad t = \frac{\bar{x} - \mu_0}{\dfrac{s}{\sqrt{n}}} > t_{n-1,\,\alpha}$$

t Test of Hypothesis for the Mean (σ Unknown)

(continued)

- For a two-tailed test:

Consider the test

$$\boxed{\begin{array}{l} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{array}}$$

(Assume the population is normal, and the population variance is unknown)

The decision rule is:

$$\text{Reject } H_0 \text{ if } \boxed{t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} < -t_{n\text{-}1,\, \alpha/2}} \text{ or if } \boxed{t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} > t_{n\text{-}1,\, \alpha/2}}$$

Example: Two-Tail Test (σ Unknown)

The average cost of a hotel room in New York is said to be $168 per night. A random sample of 25 hotels resulted in

$\overline{x}$ = $172.50 and

s = $15.40. Test at the

α = 0.05 level.

(Assume the population distribution is normal)

H_0: μ = 168

H_1: $\mu \neq 168$

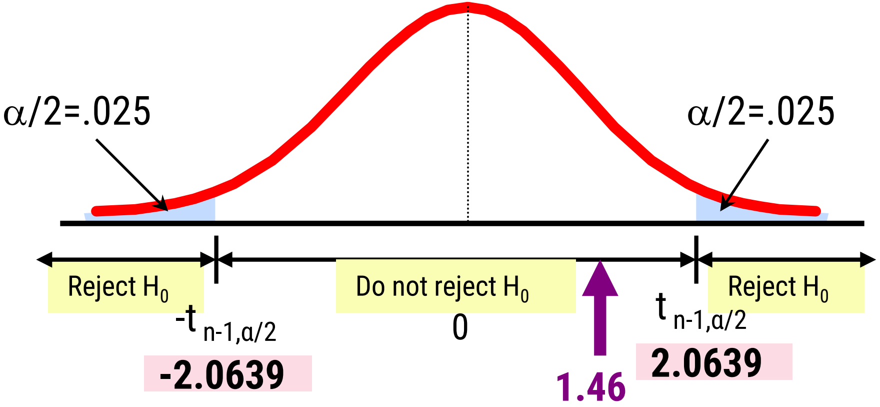
Digital Vidya

Example Solution: Two-Tail Test

H_0: μ = 168
H_1: μ ≠ 168

- α = 0.05
- n = 25
- σ is unknown, so use a **t statistic**
- Critical Value:

$t_{24\,,\,.025} = \pm\, 2.0639$



$$t_{n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{172.50 - 168}{\frac{15.40}{\sqrt{25}}} = 1.46$$

Do not reject H_0: not sufficient evidence that true mean cost is different than $168

Chi-Square as a Statistical Test

- ***Chi-square test:*** an **inferential statistics** technique designed to test for **significant relationships** between two variables organized in a bivariate table.

- Chi-square requires **no assumptions** about the shape of the population distribution from which a sample is drawn.

- It can be applied to **nominally** or **ordinally** measured variables.

Hypothesis Testing with Chi-Square

Chi-square follows five steps:

1. Making assumptions (**random sampling**)
2. Stating the research and null hypotheses and selecting alpha
3. Selecting the sampling distribution and specifying the test statistic
4. Computing the test statistic
5. Making a decision and interpreting the results

The Assumptions

- The chi-square test requires **no assumptions** about the **shape of the population distribution** from which the sample was drawn.

- However, like all inferential techniques it assumes **random sampling**.

- It can be applied to variables measured at a **nominal** and/or an **ordinal** level of measurement.

Stating Research and Null Hypotheses

- The **research hypothesis** (H_1) proposes that the two variables are **related** in the population.

- The **null hypothesis** (H_0) states that **no association exists** between the two cross-tabulated variables in the population, and therefore the variables are **statistically independent**.

H_1: The two variables are **related** in the population.

Gender and fear of walking alone at night are ***statistically dependent***.

Afraid	Men	Women	Total
No	71.1%	83.3%	57.2%
Yes	28.9%	16.7%	42.8%
Total	100%	100%	100%

H_0: There is **no association** between the two variables.

Gender and fear of walking alone at night are statistically independent.

Afraid	Men	Women	Total
No	71.1%	71.1%	71.1%
Yes	28.9%	28.9%	28.9%
Total	100%	100%	100%

The Concept of Expected Frequencies

Expected frequencies f_e : the cell frequencies that would be **expected** in a bivariate table **if** the two tables were **statistically independent**.

Observed frequencies f_o: the cell frequencies **actually observed** in a bivariate table.

Calculating Expected Frequencies

$$f_e = \frac{\text{(column marginal)(row marginal)}}{N}$$

To obtain the expected frequencies for any cell in any cross-tabulation in which the two variables are assumed independent, **multiply** the row and column totals for that cell and **divide** the product by the total number of cases in the table.

Chi-Square (obtained)

- The test statistic that **summarizes** the differences between the **observed** (*fo*) and the **expected** (*fe*) frequencies in a bivariate table.

Calculating the Obtained Chi-Square

$$\chi^2 = \sum \frac{(f_e - f_o)^2}{f_e}$$

f_e = expected frequencies

f_o = observed frequencies

The Sampling Distribution of Chi-Square

- The sampling distribution of chi-square tells the **probability** of getting values of chi-square, **assuming no relationship** exists in the population.

- The chi-square sampling distributions depend on the **degrees of freedom**.

- The χ^2 sampling distribution is not one distribution, but is **a family of distributions**.

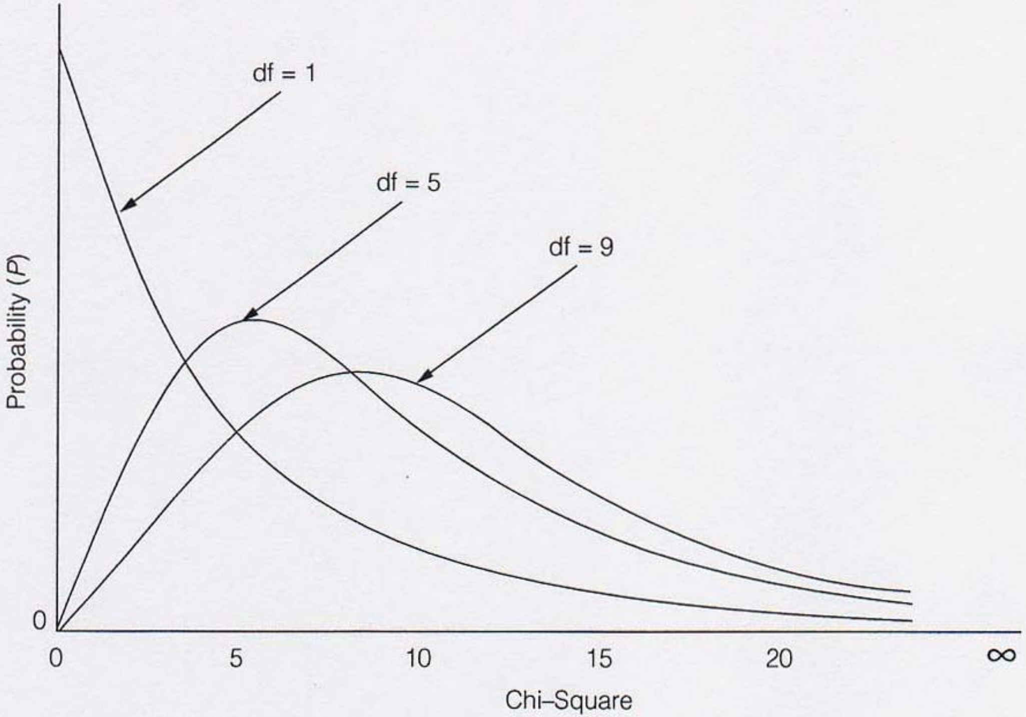
The Sampling Distribution of Chi-Square

- The distributions are **positively skewed**. The research hypothesis for the chi-square is **always** a **one-tailed test**.

- Chi-square values are **always positive**. The minimum possible value is zero, with **no upper limit** to its maximum value.

- As the number of degrees of freedom increases, the χ^2 distribution becomes **more symmetrical**.

Chi-Square Distributions for 1, 5, and 9 Degrees of Freedom



Determining the Degrees of Freedom

$$\mathbf{df = (r - 1)(c - 1)}$$

where

r = the number of rows

c = the number of columns

Calculating Degrees of Freedom

How many degrees of freedom would a table with 3 rows and 2 columns have?

$(3 - 1)(2 - 1) = 2$

2 degrees of freedom

Award Preference & SAT

The data in **StudentSurvey** includes two categorical variables:

Award = Academy, Nobel, or Olympic
HigherSAT = Math or Verbal

Do you think there is a relationship between the award preference and which SAT is higher? If so, in what way?

Award Preference & SAT

HigherSAT	**Academy**	**Nobel**	**Olympic**	Total
Math	21	68	116	205
Verbal	10	79	61	150
Total	31	147	177	355

Data are summarized with a 2×3 table for a sample of size *n*=355.

H_0 : Award preference is not associated with which SAT is higher

H_a : Award preference is associated with which SAT is higher

If H_0 is true $\Longrightarrow$ The award distribution is expected to be the same in each row.

Expected Counts

$$\text{Expected Count} = \frac{\text{row total ´ column total}}{n}$$

HigherSAT	**Academy**	**Nobel**	**Olympic**	Total
Math				205
Verbal				150
Total	31	147	177	355

Note: The expected counts maintain row and column totals, but redistribute the counts as if there were *no* association.

Chi-Square Statistic

HigherSAT	**Academy**	**Nobel**	**Olympic**	Total
Math	21 (17.9)	68 (84.9)	116 (102.2)	205
Verbal	10 (13.1)	79 (62.1)	61 (74.8)	150
Total	31	147	177	355

HigherSAT	**Academy**	**Nobel**	**Olympic**
Math			
Verbal			

$$\chi^2 = \sum \frac{\left(\text{observed - expected}\right)^2}{\text{expected}}$$

Chi-Square (χ2) Distribution

- If each of the expected counts are at least 5, AND if the null hypothesis is true, then the χ^2 statistic follows a χ^2 –distribution, with degrees of freedom equal to

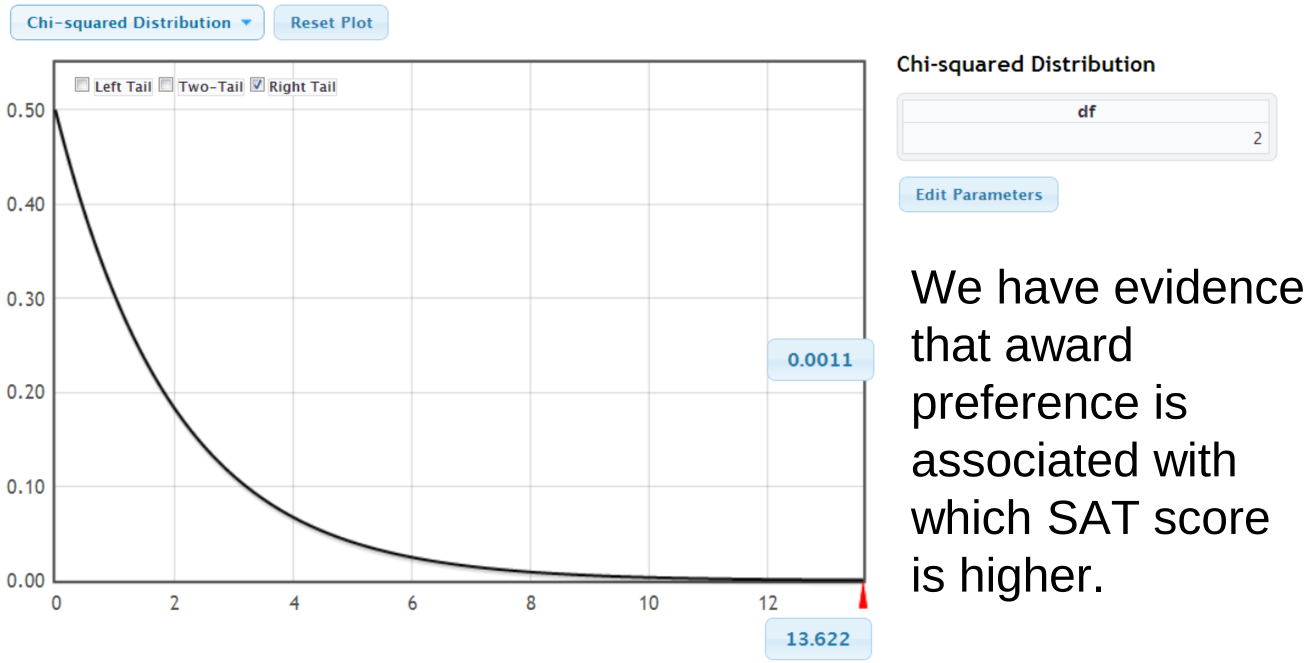
$$df = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

- Award by HigherSAT:

$$df = (2 - 1)(3 - 1) = 2$$

Chi-Square Distribution

For Higher SAT vs. Award: df = (2 – 1)(3 – 1) = 2



□ Left Tail □ Two-Tail ☑ Right Tail

0.0011

13.622

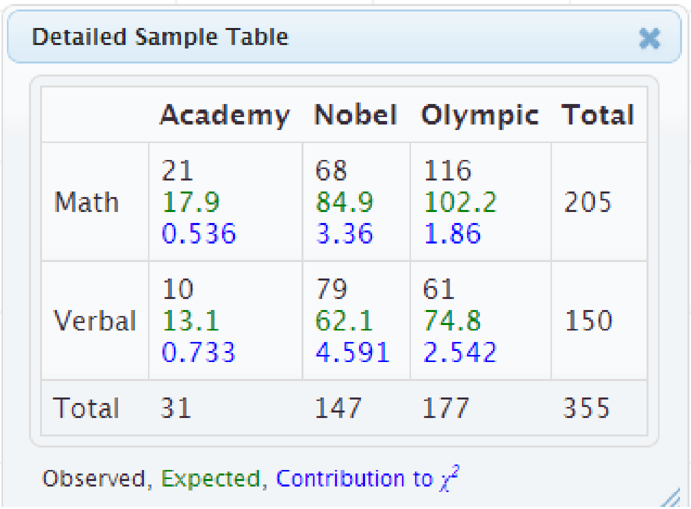
Chi-squared Distribution

df
2

We have evidence that award preference is associated with which SAT score is higher.

Chi-Square Test for Association

Note: The χ^2-test for two categorical variables only indicates **if** the variables are associated. Look at the contribution in each cell for the possible nature of the relationship.



Detailed Sample Table

	Academy	Nobel	Olympic	Total
Math	21 17.9 0.536	68 84.9 3.36	116 102.2 1.86	205
Verbal	10 13.1 0.733	79 62.1 4.591	61 74.8 2.542	150
Total	31	147	177	355

Observed, Expected, Contribution to χ^2

Limitations of the Chi-Square Test

- The chi-square test does **not** give us much information about the ***strength*** of the relationship or its ***substantive significance*** in the population.

- The chi-square test is **sensitive** to ***sample size***. The size of the calculated chi-square is **directly proportional** to the size of the sample, independent of the strength of the relationship between the variables.

- The chi-square test is also **sensitive** to **small expected frequencies** in one or more of the cells in the table.

Thank You

Digital Vidya