The Normal Distribution and Normality

What is Normal?

1

The Normal Distribution and Normality

Ask Webster

- Sane
- Even and Balanced
- Conforming to a type, standard, or regular pattern
- Containing neither basic hydroxyl nor acid hydrogen
- Having the property of commutativity under multiplication by the transpose of the matrix each of whose elements is a conjugate complex number with respect to the corresponding element of the given matrix

Skewness and Kurtosis

The upsides, downsides, peaks and valleys of data analysis

3

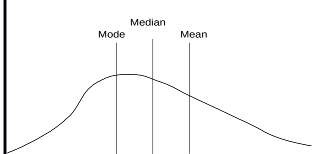
Skewness

The degree of departure from symmetry of a distribution. A positively skewed distribution has a "tail" in the positive direction. A negatively skewed distribution has a "tail" in the negative direction. A zero value indicates that the values are relatively evenly distributed on both sides of the mean, typically but not necessarily implying a symmetric distribution.

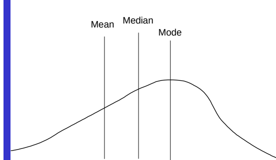
Why important? Dependent variable usually has to be normally distributed and not skewed. Else, you must transform the variable to normality or use non-parametric tests.

4

Common Shapes of Plotted Data (Skewness)



Right Skewed



Left Skewed

Examples: Right Skewed - Blood Pressure Drugs, Weight Loss Medications

Left Skewed - Azathioprene (Platelets), Ginko Balbo (Memory)

Population Skewness

- skewness: $g_1 = m_3 / m_2^{3/2}$
- where
- $m_3 = \sum(x-\bar{x})^3 / n$ and $m_2 = \sum(x-\bar{x})^2 / n$
- $\bar{x}$ is the mean and n is the sample size, as usual. m_3 is called the **third moment** of the data set. m_2 is the **variance**, the square of the standard deviation

6

Sample Skewness

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$$

7

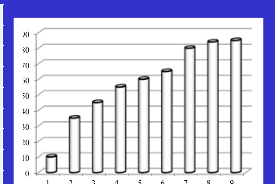
Interpretation of Skewness

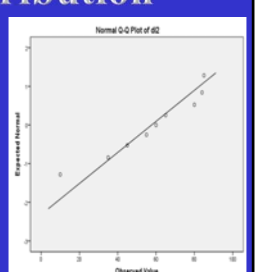
- If <−1 or >+1: **Highly Skewed**.
- If between −1 and −0.5 or +0.5 and +1: **Moderately Skewed**.
- If between −0.5 and +0.5: **Approximately Symmetric**

8

Skewed Distribution

di2
10
35
45
55
60
65
80
84
85





Descriptive Statistics							
	N	Mean	Std Dev.	Skewness	Skew SEM	Kurtosis	Kurt SEM
Left Skewed	9	57.6667	24.87971	-.749	.717	.142	1.400

9

Transformation For Skewness

Distribution Skewness	Transformation
Moderately positive	NEWX = SQRT(X)
Moderately negative	NEWX = SQRT(K – X)
Substantially positive	NEWX = LG10(X)
Substantially negative	NEWX = LG10(K – X)

K = a constant from which each score is subtracted so that the smallest score is 1; usually equal to the largest score + 1.

Work Around is to use Non-Parametric Tests

10

Class Example

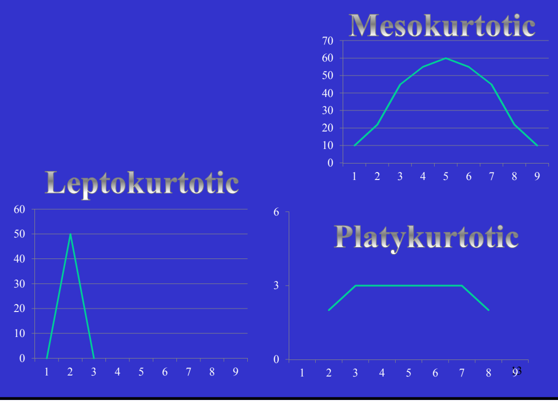
11

Kurtosis

The degree of peakedness of a distribution

12

Kurtosis Profiles

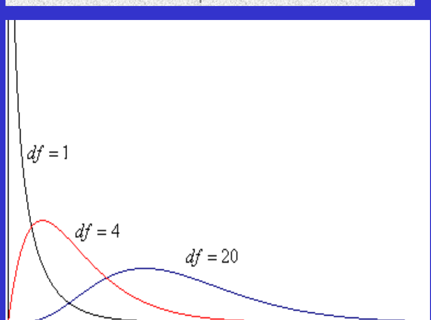


Central Limit Theorem

The larger the sample, the more likely the data will be normally distributed.

14

Influence of Sample Size on Variance



Larger the Sample, more normal the distribution and the less variance to account for.

15

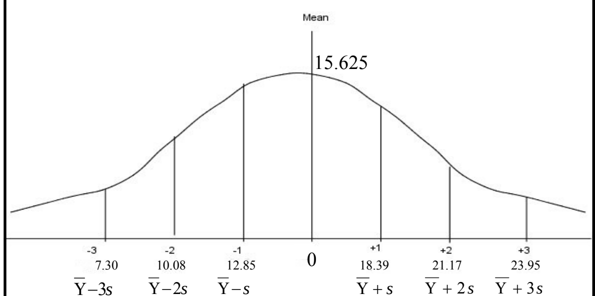
Application of the Standard Deviation To The Normal Curve

16

Application of the Standard Deviation

Mean = 15.623±2.77

S = 2.77424



Chebyshev's Theorem

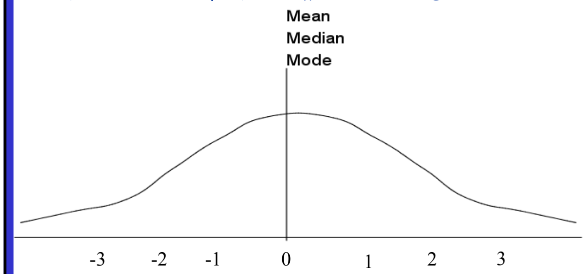
- Values will fall within k standard deviations of the mean can be determined by $1-1/k^2$, when $k > 1$.

For 3 standard deviations: $1-1/3^2 = 88.89\%$

For 2 standard deviations: $2 = 1-1/2^2 = 75\%$

18

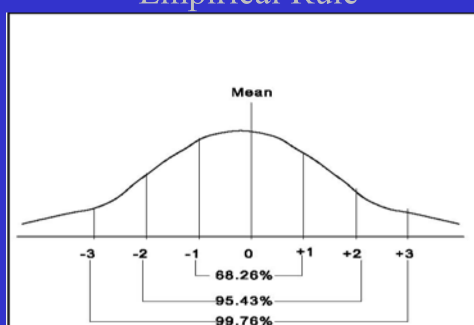
Characteristics of the Normal Curve and Distribution



*2 Halves that are mirror images
*Generated by Random Sampling
*Tails go to Infinity
*Negative Values lie Below Mean

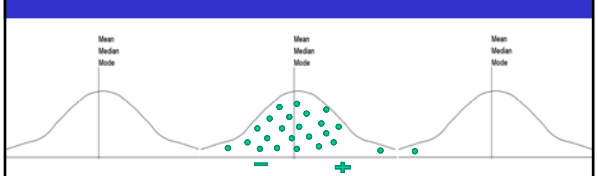
* Total Value of Curve = 1.0
* Mean = 0 and S = 1
* Opposite Objects Equal Distance From Mean

Empirical Rule



Units under curve are standard deviations and major divisions are critical values.

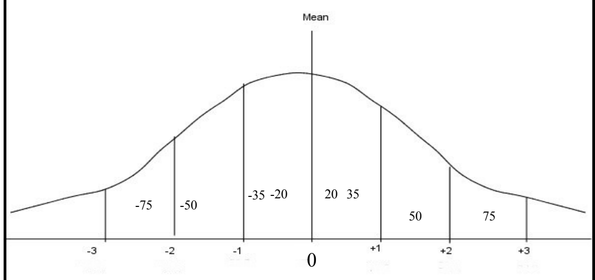
Where the Tails Actually Go



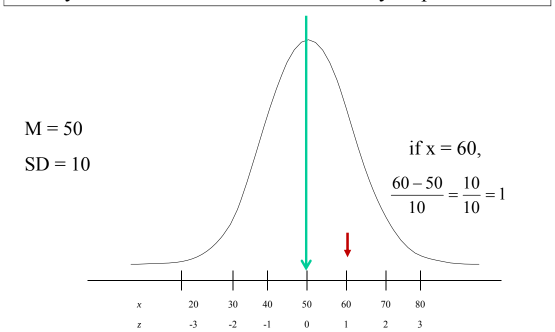
21

Why Normal Distribution Has Mean of "0"

Observations below mean have corresponding value above mean



Why is the SD of Normal Curve always equal to 1.0?



Determining if Data are Normally Distributed

- Relationship between Mean and Median
- Observing data with histogram
  (bar chart with polygon)
- Software, eg Komogorov-Smirnov 1 Sample Test

Uses of the Normal Distribution

- Determining if parametric tests can be used
- Comparison of observations with normalized and standardized data.

24

Definitions

- Critical Value - The value that sets the boundary for the acceptance region.
- Critical Region - The tail of the curve beyond the acceptance region. It is also called the "Rejection Region".
- α or "Significance Level" - The size of the critical region. The fixed probability of wrongly rejecting the null hypothesis (Type I Error). Example: α = .05
- p - Value - Probability of getting a value as extreme or more extreme by chance alone. Example: Difference between 2 means obtained by chance from the same population.

25

Significance Level or "α"

- "α" is the percent times that a researcher allows the rejection of a true null.
- "α" is set when the research project is being developed.
- Common α are .01, .05, .1
- The smaller "α", the less likely a Type I error will be made. Therefore, α = .01 is less risky than α = .05
- "α" is located in the tails of the curve.

26

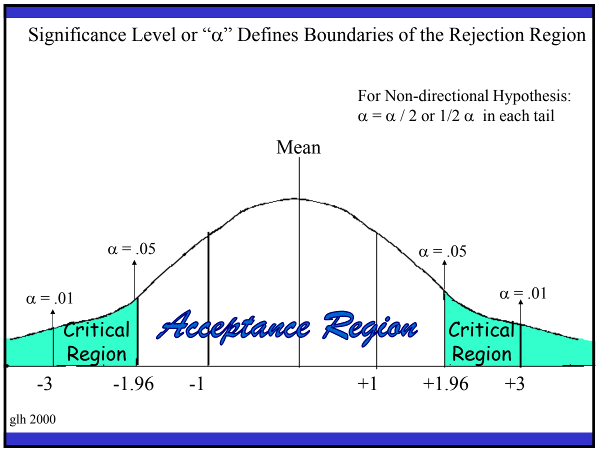
Critical Values of Normal Curve

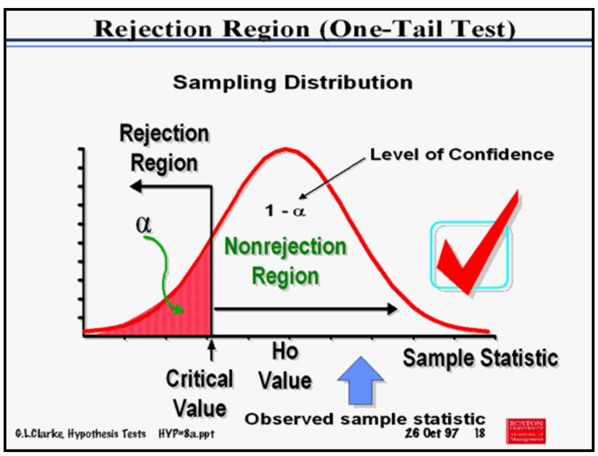
2 Tails

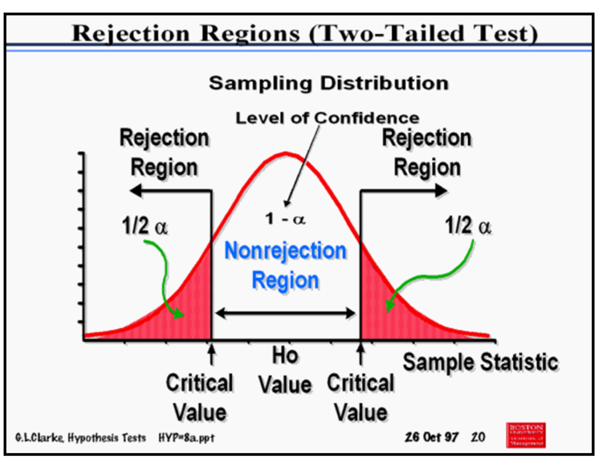
Confidence Level	Alpha = 1-CL	Critical Value (SD)
90	.10	1.64
95	.05	1.96
98	.02	2.33
99	.01	2.58

If n is large, then t-distribution and binomial distributions approximate the normal distribution. Therefore these values can be used in calculations that are binomial or are derived from the normal distribution, to include power analyses.

27


Significance Level or "α" Defines Boundaries of the Rejection Region
For Non-directional Hypothesis:
α = α / 2 or 1/2 α in each tail
Mean
α = .05
α = .05
α = .01
α = .01
Critical Region
Acceptance Region
Critical Region
-3
-1.96
-1
+1
+1.96
+3
glh 2000


Rejection Region (One-Tail Test)
Sampling Distribution
Rejection Region
Level of Confidence
1 - α
α
Nonrejection Region
Ho Value
Sample Statistic
Critical Value
Observed sample statistic
G.L.Clarke, Hypothesis Tests HYP=8a.ppt
26 Oct 97 18


Rejection Regions (Two-Tailed Test)
Sampling Distribution
Level of Confidence
Rejection Region
Rejection Region
1 - α
1/2 α
1/2 α
Nonrejection Region
Ho Value
Critical Value
Critical Value
Sample Statistic
G.L.Clarke, Hypothesis Tests HYP=8a.ppt
26 Oct 97 20

Statistical Significance

- A rare event that deviates dramatically from the normal or from what is expected.

Probability of event occurring is less than lower limit of chance error.

Something statistically significant might not be clinically significant or even important.

31

Chance Error

- Outcome of statistical test is to accept or to not accept the null hypothesis
- Chance Error is set at beginning of study, called α error
- Typical error values are .01, .05, .1
- Signifies acceptable times that a null outcome can be wrongly rejected and algorithm is robust enough to handle.
- Alpha = .05 means null 5% time is ok

32

Relationship Between P-Value and α

If α = .05, an outcome can occur 5% of the time by chance and it is acceptable, even if it is an error. Therefore, outcomes that occur at a probability of <.05 are rare and are statistically significant when alpha is set at .05.

Likewise, if α = .01, an outcome would be considered to be rare and significant if it occurred with a probability < .01.

33

Statistical Significance and P-Value

- Is determined by a probability called "P", that ranges from 0 to 1
- Smaller the "P" (e.g. $p < .05$) the less likely the difference is due to chance.
- "P" = probability of seeing a difference of the given magnitude if the Null were true.

Males	Females	
$\bar{y} = 25.67 \pm 3.81$	$\bar{y} = 26.92 \pm 4.79$	$P = .661$

66.1% chance of seeing this difference "by chance"

So the Null is "accepted"

Relationship Between P-Value, Significance, CI and α

Significant Outcome if $p < \alpha$

$CI = 1 - \alpha$

CI = Acceptance Region

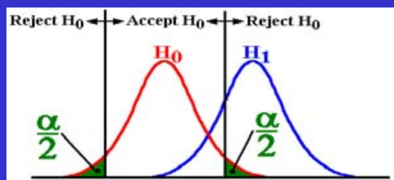
Statistical Error

- Type I Error - Rejecting the Null Hypothesis when it is true. This is an "α" error.
- Type II Error - Accepting a False Null Hypothesis. This is a "β" error.
- Statistical Power - The ability of a test to reject the null when it is truly false (or to make the right decision).

Type I Error

To Reject a Null That Should be Accepted

- One Directional Hypothesis
- Data in Tails. α Error



Decision: Effect of intervention greater than placebo. Results from sample generalized to population. Example: Medication that lowers BP.

Classic Examples When Decision Based on Fraud:

Snake Oil

Some Holistic Remedies

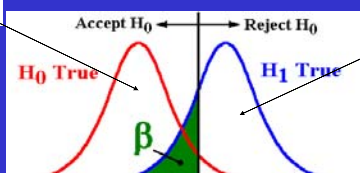
Fake Cancer Cures

Type II Error

To Accept a Null that Should be Rejected

Sophomores

Sampling Problem



Juniors

Hypothesis: No Difference in Obturation Times

Decision: Effect of intervention same as competitor or has no effect. Results from flawed sampling.

Classic Examples When Decision Based on Fraud:

Effect of Tobacco

Cable vs Satellite

Vivoxx

Standardization

$$Z = \frac{\mu - y}{s}$$

Magnitude of difference between y and the mean are controlled by the error that is associated with the difference.

39

Standard Scores

1. Standardizing a distribution of scores to itself as standard deviates of the mean
2. Standardizing a distribution of scores to a mean with a preset level of error

40

Z Scores: Standardizing a distribution to itself

$$Z = \frac{\mu - y}{s}$$

41

Class Demonstration

- Creating a "Z" or Standardized Distribution
- Use of the Normal Table

42

Z Scores to a Standard Score

Standard Score = Z * S + Desired Mean

To standardize distribution to a mean of 75 with a sd of 6.1:

Standard Score Mean of 75 = Z * 6.1 + 75

43

Z-Score and Standard Score Calculations

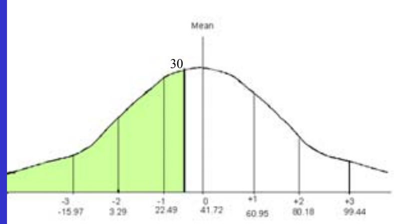
CLASS EXERCISE

44

Problem Solving the Normal Distribution

How many were 30 years old and younger?

N = 50



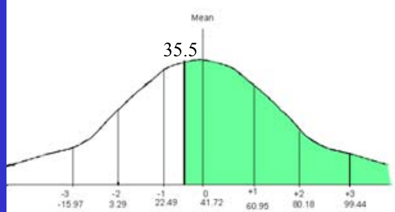
$Z = y - \bar{y} / s = 30 - 41.72/19.23 = .6095$

Now to the tables.

45

Problem Solving the Normal Distribution

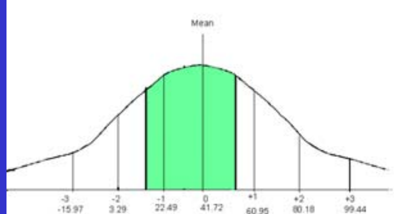
How many were over 35 years of age ?



46

Problem Solving the Normal Distribution

How many were between 15 and 52 years of age?



47

Relationship Between Z and SEM

48