Assessment of Children with Special Needs

Part I. Screening & Assessment

1.1. Screening

- Definition: “Screening is the process of collecting data to decide whether more intensive assessment is necessary.”
- Screening is an initial stage during which students who may evidence a particular problem, disorder, disability, or disease are sorted out from the general population.

a. True Positive (TP)

- Individuals who perform poorly on screening measures are called **“at risk,”**
- When individuals are identified as having the problem or disability using the screening measure and also demonstrate the same problem or disability during the follow up assessment, they are called **“True Positive,”**
- **True Positive (TP):** Screened children as having some problems, who are also correctly diagnosed as having the problem
- **Hit Rate** is the proportion of accurate positive decisions.

b. False Positive (FP)

- Individuals who perform poorly on screening measure but do well on later follow up assessment are **“false positives,”** which shows that they do not have the condition for which they were screened,
- **False Positive (FP):** Screened children as having some problems but later turn out to be normal or healthy; they are incorrectly identified as having the problem,

c. False Negative (FN)

- Individuals who were considered ‘normal’ or without any problems using the screening measure but later they evidence the very problem for which the screening was conducted and they are said to be **“false negatives”**.
- **False Negative (FN):** Children screened as normal or without the target problem but found out as having the problem during the assessment; such kind of children are wrongly left behind as if they had no problem.

d. True Negative (TN)

- **“True Negatives”** are individuals who were considered normal or without any problems using the screening measure and later on found out to be, or confirmed, without the problem.
- **True Negative (TN):** Children screened as normal or without the target problem are also proven to have no problem in the follow up assessment; i.e., such children are correctly identified as having no problem which the assessment is trying

Sensitivity, Specificity, PPV & NPV

- **Sensitivity:** measures the proportion of actual positives correctly identified by a screening measure from the total actual positives.
- **Specificity:** measures the proportion of actual negatives correctly identified by a screening measure from the total actual negatives.
- **Positive predictive value (PPV):** is the probability of hitting true positives from the total referred positives.
- **Negative Predictive Value (NPV):** Is the probability of correctly identifying true

Examples

		Assessment Result		
Screening Result		YES	NO	TOTAL
	YES	a	b	(a+b)
	NO	c	d	(c+d)
	TOTAL	(a+c)	(b+d)	(a+b+c+d)

		Assessment Result		
Screening Result		**POSITIVE**	**NEGATIVE**	**TOTAL**
	POSITIVE	**TP**	**FP**	**(TP+FP)**
	NEGATIVE	**FN**	**TN**	**(FN+TN)**
	TOTAL	**(TP+FN)**	**(FP+TN)**	**(TP+FP+FN+TN)**

- Sensitivity = a/(a+c)

$$Sensitivity = \frac{Number of True Positives (TP)}{Number of True Positives (TP) + Number of False Negatives (FN)}$$

- Specificity = d/(b+d)

$$Specificity = \frac{Number of True Negatives (TN)}{Number of False Positive + Number of True Negatives (TN)}$$

- Positive Predictive Value (PPV) = a/(a+c)

$$PPV = \frac{Numberof True Positives(TP)}{Numberof True Positives(TP) + Numberof False Positive(FP)}$$

- Negative Predictive Value (NPV) = d/(b+d)

$$NPV = \frac{Numberof True Negatives(TN)}{Numberof True True Negatives(TN) + Numberof False Negative(FN)}$$

Example 1.		Assessment Results		
		Positive	*Negative*	
Screening Results (Item 3)	*Positive*	**TP = 11**	**FP = 1**	**→ Positive predictive value** = TP / (TP + FP) = 11 / (11 + 1) = 11 / 12 **= 91.7%**
	Negative	**FN = 3**	**TN = 15**	**→ Negative predictive value** = TN / (FN + TN) = 15 / (3 + 15) = 15 / 18 **= 83.3%**
		↓ **Sensitivity** = TP / (TP + FN) = 11 / (11 + 3) = 11 / 14 **= 78.6%**	↓ **Specificity** = TN / (FP + TN) = 182 / (18 + 182) = 182 / 200 **= 91%**	

Example 1.			Assessment Results		Total
			No	Yes	
Item3	No	Count	15	3	18
		% within Item3	83.3%	16.7%	100.0%
		% within Assessment Re	93.8%	21.4%	60.0%
	Yes	Count	1	11	12
		% within Item3	8.3%	91.7%	100.0%
		% within Assessment Re	6.3%	78.6%	40.0%
Total		Count	16	14	30
		% within Item3	53.3%	46.7%	100.0%
		% within Assessment Re	100.0%	100.0%	100.0%

Example 2.		Assessment Results		
		Positive	*Negative*	
Screening Results (Item 8)	*Positive*	**TP = 12**	**FP = 4**	→ **Positive predictive value** = TP / (TP + FP) = 12 / (12 + 4) = 12 / 16 **= 75%**
	Negative	**FN = 2**	**TN = 12**	→ **Negative predictive value** = TN / (FN + TN) = 12 / (2 + 12) = 12 / 14 **= 85.7%**
		↓ **Sensitivity** = TP / (TP + FN) = 12 / (12 + 2) = 12 / 14 **= 85.7%**	↓ **Specificity** = TN / (FP + TN) = 12 / (4 + 12) = 12 / 16 **= 75%**	

Example 2.			Assessment Results		Total
			No	Yes	
Item4	No	Count	12	2	14
		% within Item4	85.7%	14.3%	100.0%
		% within Assessment Re	75.0%	14.3%	46.7%
	Yes	Count	4	12	16
		% within Item4	25.0%	75.0%	100.0%
		% within Assessment Re	25.0%	85.7%	53.3%
Total		Count	16	14	30
		% within Item4	53.3%	46.7%	100.0%
		% within Assessment Re	100.0%	100.0%	100.0%

Selecting a Good Screening Measure

- A good screening measure should have 0.70 and above sensitivity and specificity indexes and reasonable high predictive values.

Use of Screening

- Screening takes place at all levels of education.
  - Children are screened before they enter kindergarten or first grade to determine their academic readiness in terms of:
    - Language, cognitive, and motor development,
    - Social and emotional functioning,
  - Children may also be given vision and hearing screening tests.

Use of Screening (Cont’d)

- Screening is also used throughout the school years to identify students who need extra attention because their performance is markedly different from the average performance.
- Cut off scores are based on the average performance of students at various age or grade levels.

Use of Screening (Cont'd)

- Referral for further assessment or evaluation: such as psychological and educational assessment,
  - Two kinds of decisions:
    - Decisions about exceptionality: Whether the child is disabled or gifted
    - Decisions about special learning needs

- Referral is a formal process involving the completion of a referral form to be submitted to a team of professionals, sometimes called child study team. However, referral may result from a teachers' observations, a parent's request, or the student's own request.

Use of Screening (Cont'd)

- Depending on the nature of the case, the child study team may consist of:
  - General education teachers,
  - Special education teachers,
  - Administrators,
  - Parent(s) of the student, and
  - Related services personnel, such as the school psychologist, nurse, social worker, and counselor,.

Assessment

- Assessment is the process of collecting data for the purpose of making decisions about students.
- Students who fail to perform as expected are referred to the child study team for assessment.
- Assessment is dynamic and ongoing. After the declaration of the eligibility for services and specification of learning needs for a student, teachers continue to observe how the student performs under different circumstances and to

Assessment (Cont’d)

- Note: It takes longer to decide to declare students with mild disabilities eligible for special education services than it does to make the same decision for students with severe disabilities.
- Assessment personnel need to also be engaged in extensive assessment to ascertain whether a child has developmental disorder.

Assessment Domains

1)Academic Problems

- Parents & teachers are concerned about a student not performing well in school and initiate referral.
- Examples of contents of academic problems include:
  - Reading, Mathematics, and written language and communication
- Academic competence is assessed in making exceptionality and eligibility decisions
  - E.g., a student referred for reading problems might be given a reading test to provide a comparison of his or her reading skills with those of other students in the school or even the nation.
  - The result of the test would be used to verify or

- Academic performance is also usually assessed in making instructional-planning decisions. In deciding what to teach a student, the teacher or team must specify which academic competencies the student already has.

Assessment Domains (Cont'd)

2) Behavioral Problems

- Students are often referred for psychological or educational assessment because they demonstrate behavior problems.
  - Students for whom severe behavior problems can be specified and verified are often declared eligible for special education services.

- Behavior problems may include:
  - Failure to get along with peers,
  - Delinquent activities,
  - Excessive withdrawal,
  - Unregulated and maladaptive behavior
  - Disruptive behavior and
  - non-compliant behavior
- Assessment results are compared to the typical child’s behavior.

Assessment Domains (Cont'd)

3) Physical Problems

- Physical problems include sensory disabilities (such as in vision or hearing), problems of physical structure (for example, spina bifida or cerebral palsy) , and chronic health problems (such as diabetes or asthma)
- Severe physical problems are brought to the attention of parents by physicians

- Milder, but nonetheless important, problems that have not been noticed by the parents are often discovered during routine screening.
- E.g., when a school nurse notices that a first grader child has 65-decibel loss in his more sensitive ear, she could refer him for an audiologist for further assessment and further decisions.

Part II. Basic Concepts of Measurement

Contents

2.1. Descriptive Statistics

a) Scales of (or Levels of) Measurement

- **Nominal Scales:** Nominal measures offer names or labels for certain characteristics. E.g., sex (male, or female), other categories and qualitative descriptions. Categorical variables are assessed on nominal scale.
  - Adjacent values have no inherent relationship
  - Nominal scales name values on the scale
  - Permissible statistics: mode, chi-square
  - We cannot add orange and banana and seek for an average!! We simply ask “how many or more orange or banana are there?”

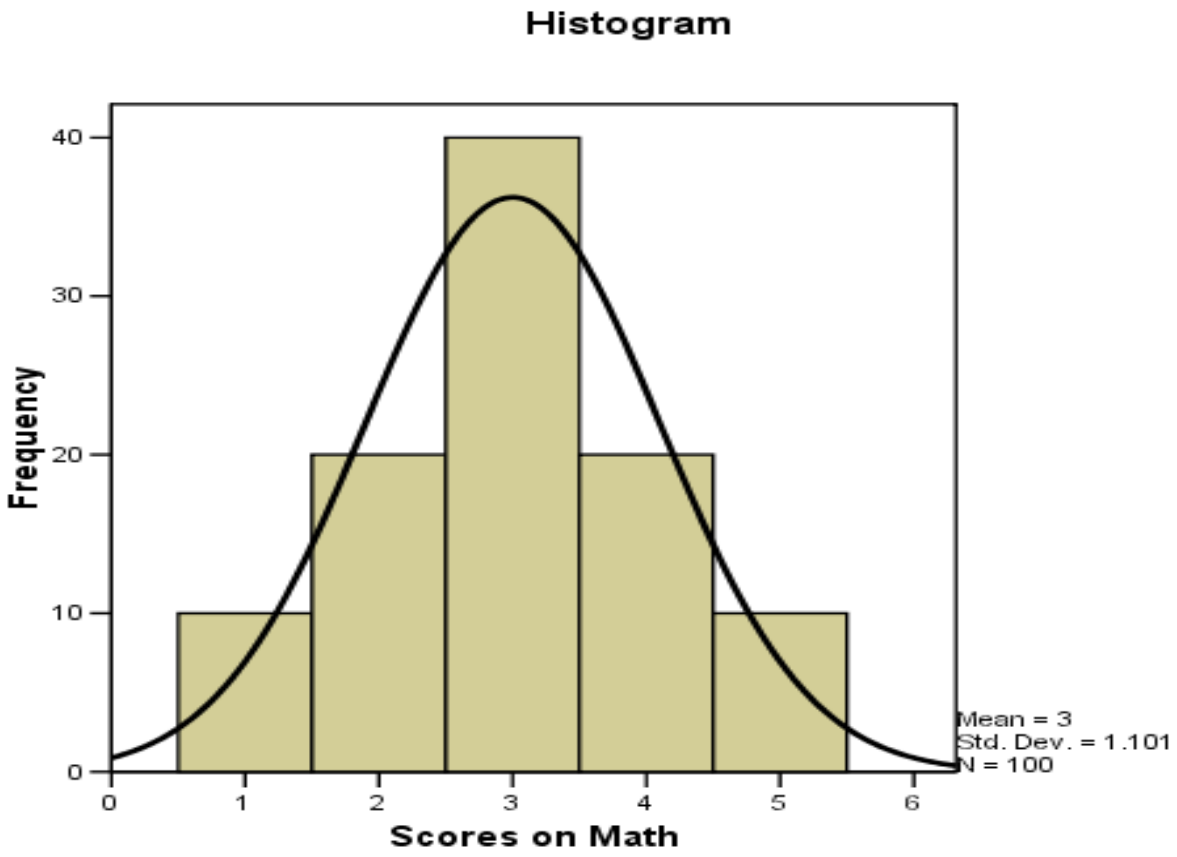
Ordinal Scales: In this scale type, the numbers assigned to objects or events represent the rank order (1st, 2nd, 3rd etc.) of the entities assessed. Ordinal scales show relative position or rank such as first, second, third, etc,

- Adjectival comparisons like “low", "medium", “high“; “good”, “better”, “best”; “poor”, “worse”, “worst”.
- Adjacent values are either in increasing or decreasing order
- Differences in adjacent ranks do not reflect the magnitude of differences in the raw scores.
- Ordinal values can be assigned all along the continuum such as the ranks in class OR only in some parts of the continuum such as the top 20 (i.e., achievers) or the bottom 20 (i.e., losers).

- **Interval Scales:**
- It is more powerful than nominal and ordinal as it not only orders or ranks or rates but also shows exact distances in between. But it does not start from zero; no logical or absolute zero.
  - If there is zero like zero temperature it is not natural but arbitrary as 0 degree does not mean no temperature.
  - If a student gets zero or all incorrect in an arithmetic test, it does not mean that the student knows nothing about arithmetic.
- Interval scale is used in addition or subtraction of scale value to calculate mean, range, variance, standard deviation, correlation and regression.
- Ordinal scale only ranks but does not measure difference between the two ranks like “satisfactory” and “not-satisfactory”. Interval scale not only ranks but also give exact distance between them by assigning a value.
- Difference in temperature of 20 degree and 40 degree is 20 but 40 is not double hot than 20.
- Ratio comparisons cannot be made.

- **Ratio Scales:**
- This scale can perform all functions. It can show all mathematical and geographical indicators. It is useful when exact figures are required in objective matters are required.
- All properties in nominal, ordinal and interval scales PLUS an absolute zero. Hence, ratio comparisons are possible.
- If a person is drawing a salary of $20,000 and another $ 40,000, it can be said that the latter is getting double the salary of the former.

Score Distributions



- Normal Distribution
- Skewed Distribution
  - Positively or negatively skewed distribution
- Kurtosis
  - Platykurtic curve and Leptokurtic curve

Measures of Central Tendency and Dispersion

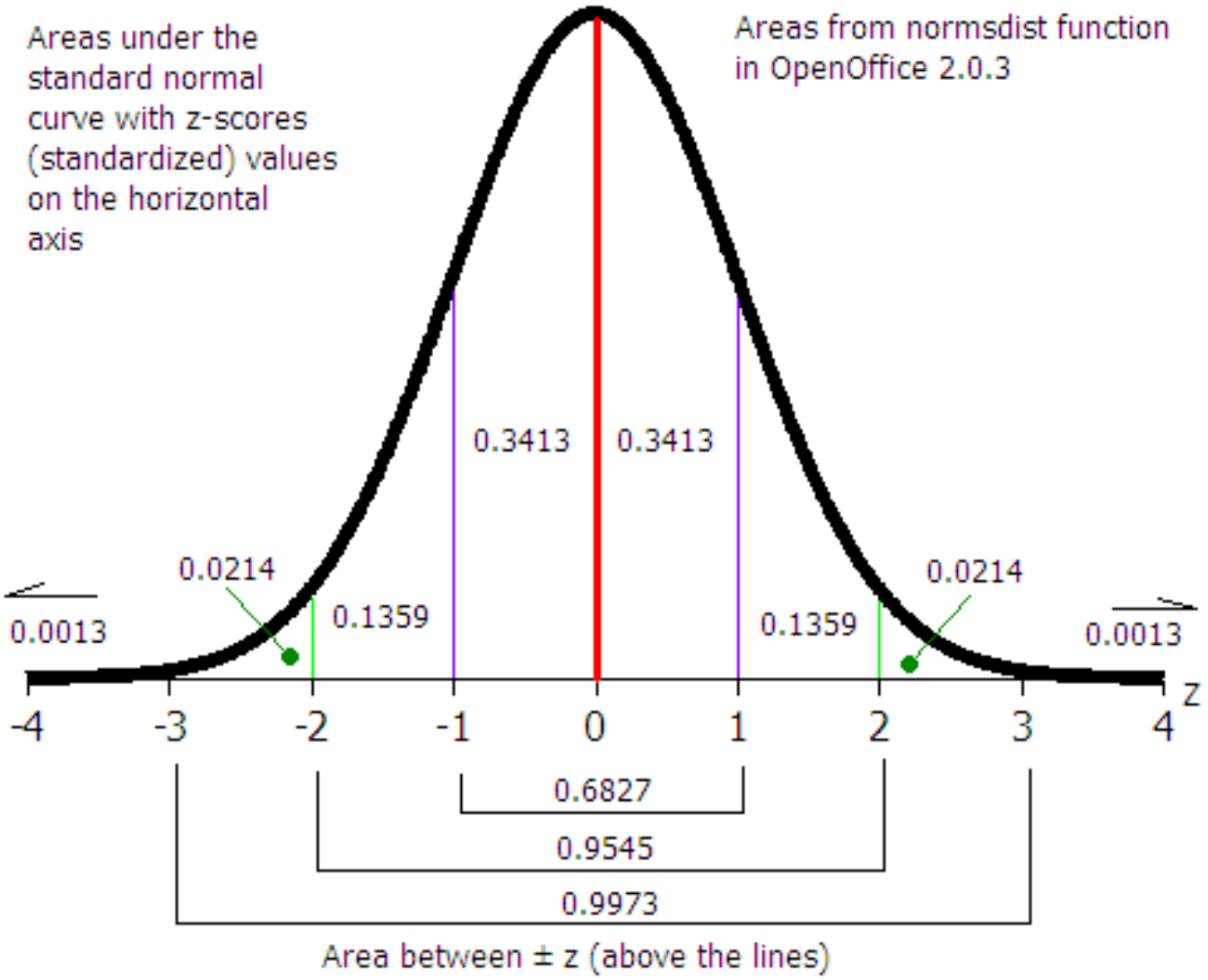
- Measures of Central Tendency
  - Arithmetic Mean
  - Median
  - Mode
- Measures of Dispersion
  - Variance
  - Standard Deviation
  - Range
  - Inter-quartile range

Measures of Dispersion

- Variance = S^2
- The standard deviation S is the square root of the sample variance:

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

- It can be used to make inferences about the population variance.


Areas under the standard normal curve with z-scores (standardized) values on the horizontal axis
Areas from normsdist function in OpenOffice 2.0.3
0.3413
0.3413
0.0214
0.0214
0.1359
0.1359
0.0013
0.0013
z
-4
-3
-2
-1
0
1
2
3
4
0.6827
0.9545
0.9973
Area between ± z (above the lines)

- **Normal Curve Areas:**
- Mean, Median and Mode are at the center in a normally distributed score.
- 68.26% of the scores are 1 standard deviation below and above the mean.
- 95.44% of the scores are 2 standard deviations below and above the mean.
- 99.72% of the scores are within 3 standard deviations below and above

Correlation

- Correlation quantifies the relationships between variables.
- Correlation coefficients are numerical indexes of the relationship between variables.
- Values of correlation coefficients :
  - 0 to +1
  - 0 to -1
- Correlation coefficients are used in measurement to estimate the both reliability and validity

- Pearson Product Moment Correlation Coefficient, r_{xy} .

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

Individuals	X	Y	XY	X2	Y2
1	1	2	2	1	4
2	2	2	4	4	4
3	3	4	12	9	16
4	4	6	24	16	36
5	5	4	20	25	16
Total	15	18	52	55	76

$r_{xy} = 40/52.915 = 0.7559$

- Types of relationship
  - Linear vs. curvilinear relationship
  - Positive vs. negative relationship
  - No relationship at all,
  - No relationship when one variable is constant
- Pearson Family Correlation Coefficients
  - Spearman rho, ρ, correlation of variables on ordinal scale such as ranks of scores.
  - Phi coefficient,Φ, correlation of dichotomous variable with continuous variable
- Point biserial correlation coefficients, r

- General rule of thumb for correlation coefficients and you can use the same rule for the Phi coefficient.
  - -1.0 to -0.7 strong negative association.
  - -0.7 to -0.3 weak negative association.
  - -0.3 to +0.3 little or no association.
  - +0.3 to +0.7 weak positive association.
  - +0.7 to +1.0 strong positive association.

Correlation and Causality

i) Correlation can occur by chance

ii) X can cause Y (‘burning building cause firefighter presence’)

iii) Y can cause X (sometimes ‘firefighters cause fire’)

iv) A third variable, Z, can cause both X and Y (there is a positive relationship between shoe size and mental age):

- Big feet (X) do not cause mental development (Y)
- Mental development (Y) does not cause big feet (X)
- Maturation (Z) causes both X & Y. As children grow older, they develop both mentally and

- The Spearman correlation coefficient is often thought of as being the Pearson correlation coefficient between the ranked variables.
- In practice, however, a simpler procedure is normally used to calculate ρ. The n raw scores X_i, Y_i are converted to ranks x_i, y_i, and the differences $d_i = x_i - y_i$ between the ranks of each observation on the two variables are calculated.
- If there are no tied ranks, then ρ is given by:
- If tied ranks exist, Pearson's correlation coefficient between ranks should be used

Individuals	IQ, X_i	Hours of TV per week, Y_i	rank x_i	rank y_i	d_i	d^2_i
1.	86	0	1	1	0	0
2.	97	20	2	6	−4	16
3.	99	28	3	8	−5	25
4.	100	27	4	7	−3	9
5.	101	50	5	10	−5	25
6.	103	29	6	9	−3	9
7.	106	7	7	3	4	16
8.	110	17	8	5	3	9
9.	112	6	9	2	7	49
10.	113	12	10	4	6	36

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

= 1-[6(194)/10(100-1)]

= 1- 1.17575

= 0.175

- **Interpretation:** This low value shows that the correlation between IQ and hours spent watching TV is very low.

		Exclusive breast feeding three days after discharge		
		Yes	No	Total
Exclusive breast feeding at discharge	Yes	A=45	B=7	E=52
	No	C=2	D=34	F=36
	Total	G=47	H=41	88

- Phi coefficient, Φ

$$\phi = \frac{A.B - C.D}{\sqrt{E.F.G.H}}$$

= 1523/1899.3 = 0.80

- **Interpretation:** There is a strong association between breast feeding status at discharge and breast feeding status 3 days after discharge.

- **What is a point biserial correlation?**
- The point biserial correlation is a measure of association between a continuous variable and a binary variable. It is constrained to be between -1 and +1.

- This is mathematically equivalent to the traditional correlation formula. The interpretation is similar. The point biserial correlation is positive when large values of X are associated with Y=1 and small values of X are associated with Y=0.

- **Calculation of the point biserial correlation**
- Assume that X is a continuous variable and Y is categorical with values 0 and 1. Compute the point biserial correlation using the formula

$$r_{ptbis} = \frac{(\overline{X_1} - \overline{X_0})\sqrt{p(1-p)}}{S_x}$$

$\overline{X_0}$ = the mean of X when Y = 0

$\overline{X_1}$ = the mean of X when Y = 1

S_x = the standard deviation of X

p = the proportion of the values where Y = 1

2.2.Quantification of Test Performance

a) Scores Used in Norm Referenced Assessment

- When a large domain is assessed, a student's performance is typically interpreted by comparing it with the performances of a group of subjects of known demographic characteristics (age, gender, grade in school, and so on). This group is called a *normative sample, or norm group*.
- The comparison scores are called *derived scores* and are of two types: ***developmental scores*** and ***scores of relative standing***

Developmental Scores

i) Developmental equivalents

- Developmental equivalents are one type of derived or transformed score.
- The most common types of developmental equivalents are **age equivalents** (mental age, for example) and **grade equivalents**.
- Example:
  - Suppose the average performance of 10 year old children on an intelligence test was 27 correct answers.
  - If Abebe, who is a 10 years old boy also scored 27 correct on the same test, then he earns a mental age of 10 years. It means Abebe answered as many questions correctly as the average 10 year-year-old childre.
- **Age equivalents** are expressed in years and months; a hyphen is used in age scores (for example, 7-1 for 7 years, 1 month old).

- **Grade Equivalent** means the child's raw score is the average (the median or mean) performance for a particular grade.
- Example:
  - Suppose the average performance of grade 5 pupils on a reading test was 50 correct answers.
  - If Abebe, who is a 4th grader and scored 50 correct on the same test, then he earns a grade equivalent of grade 5. It means that although Abebe a 4th grader, he answered as many questions correctly as the average grade 5 pupils do.
- **Grade equivalents** are expressed in grades and tenths of grades; a **decimal point** is used in grade scores (for example

- Problems with developmental scores
  - *Systematic misinterpretation:* students who earn the typical average score does not necessarily and actually perform as the child in the typical age.
  - *Need for interpolation and extrapolation:* there may be no children tested in the norm group.
  - *Promotion of typological thinking:* The average pupil is a mere abstraction sometimes non-existent (such as age 12-2 OR grade 2.3)
  - Implication of a false standard of performance: Since half of the test takers earn scores below the median, the expectation that a third grader would performer at a third grade level makes the standard false.
  - Tendency for scales to be *ordinal, not equal-interval* and scores should not be added or multiplied in any computation.

ii) Developmental quotients

- IQ ratio = $\frac{\text{Mental Age, in months (MA)}}{\text{Chronological Age, in months (CA)}} X\ 100$

- Suppose Almaz earns a mental age of 120 months while her chronological age is 8-6 or 102 months, her IQ quotient would be 117.65, which is above average.
- Interpretation:
  - Developmental age: is often interpreted as the level of functioning,
  - Developmental quotient: is interpreted as the rate of

- **Problems with Developmental Quotients**
  - All problems that apply to developmental levels also apply to developmental quotients.
  - Additional problem: the variance of developmental scores within different chronological age or grade groups may not be the same.
    - The same quotient may mean different things at different ages.
    - Different quotients at different ages can mean the same thing.
  - Thus, different variances at different ages and grades render it impossible to interpret scores without knowing the variances

Scores of Relative Standing

- Unlike developmental scores, scores of relative standing use more information than the mean or median to interpret a person's test score.
- When the same type of relative standing score is used, the units of measurement are exactly the same.
- Thus, we can compare the performances of different people even when they differ in age, and we can compare one person's scores on several different tests.
- Scores of relative standing put raw scores into comparable units, such as **percentiles**

Percentiles

- A percentile rank indicates a student's relative position in terms of the percentage of students scoring lower.
- Definition:
  - A percentile rank is the **percentage of people whose scores are at or below a given raw score**.
- If we consult a table of norms and find that a student's raw score is 29 which equals a percentile rank of 70, we know that 70 percent of the students in the reference group obtained a raw score lower than 29.
- Or, the student's performance surpasses that of 70% of the group.

- Steps:
  - Sort and order scores from highest to lowest,
  - Compute the percentage of people with **scores below** the score to which you wish to assign a percentile rank,
  - Compute the percentage of people with scores **at the score** to which to which you wish to assign a percentile rank,
  - Add the percentage of people with scores **below the score** to one half the percentage of people with scores **at the score** to obtain the percentile rank.

- **Example of Percentile Rank:**
  - Scores: 15,13, 11, 9, 8, 7, 5, 3.
  - Corresponding Ranks: 1, 2, 3, 4, 5, 6, 7, 8.
  - Steps to find out the Percentile Rank of a student getting a score of 9:
    - S_{bs} = % of people with scores below 9 = 4/8 = 50%
    - S_{as} = % of people with scores at the score 9 = 1/8 = 12.5%
    - X = Raw score for which percentile rank is required = 9
    - L = The lower real limit of the score X = 8.5
    - Percentile Rank = 50% + ½ (12.5%) = 56%
  - Formula: PR= S_{bs} + (X-L)(S_{as}).

- Because percentile ranks are computed using one half the percentage of those obtaining a particular score, it is **not possible** to have a percentile rank of either 0 or 100 (it could be 99.9 or 0.1).
- **Deciles** and **quartiles** are bands of percentiles that are 10 and 25 percentile ranks in width in the norm group.
- First decile contains percentile ranks 0.1 through 9.9, second decile 10-19.9; the tenth decile 90-99.9.
- First quartile contains percentile ranks 0.1 through 24.9, and the fouth quartile contains the ranks 75-99.9.

Standard Scores

- A **standardized distribution** is a set of scores that have been transformed so that the mean and standard deviation of the test take predetermined (standard) values.

- The most basic standardized distribution is the **z-distribution**. A z-distribution has a predetermined mean of zero and a predetermined standard deviation of one.

- ***Standard score*** is the general name of any derived score that has been standardized.

- Five commonly used standard-score distributions: z-scores, deviation IQs, normal curve equivalents, and stanines.

Z-scores

- Z-scores are standard scores, with a mean of zero and standard deviation of one.
- Formula: $z = \frac{(X - \overline{X})}{S}$

- Z-scores are often transformed to other standard scores for practical reasons:

$$SS = \overline{X_{ss}} + (S_{ss})(z)$$

- Where, SS is the standard score, $\overline{X_{ss}}$ is the mean of the distribution of the standard scores, and S_{ss} is the standard deviation of the distribution of the standard score and z is the z-score.

	Test 1. English	Test2. Mathematics
Mean	72	30
Standard Deviation	6	5
Abebe’s raw score	70	36
Yilma’s raw score	75	28
Abebe’s Z-score	-0.333	+1.20
Yilma’s Z-score	+0.50	-0.40
Abebe’s T-Score (=50+10Z)	46.67	62
Yilma’s T-Score (=50+10Z)	55	46

Z-score	T-Score (mean = 50 and Sd = 10) T-score = 50+10(z)	IQ deviations (mean = 100 and Sd = 15) = 100+15(z)
Z=+1	60	115
Z=+2	70	130
Z=+3	80	145
Z=0	50	100
Z=-1	40	85
Z=-2	30	70
Z=-3	20	55

- Each point on the base line of the normal curve could be equated to percentile ranks:

  -2 SD =2%

  -1 SD =16% (OR 2% and 14% of the area)

  0 SD =Mean = 50% (or 16%+34%)

  +1SD = 84% (OR 50% + 34%)

  +2SD = 98% (OR 84% + 14%)
- The relationship between standard deviation units and percentile ranks enables us to interpret standard scores in simple and familiar terms

- Advantages & disadvantages of standard scores
- Percentiles are better understood than standard scores by pupils and parents,
- Standard scores have all the advantages of percentiles plus an additional merit: they can be combined (i.e., added or averaged).

Scores Used in Criterion-Referenced Assessment

- Unlike norm-referenced scores, criterion referenced scores compare a student's performance against an objective and absolute standard or criterion of performance.
- Criterion-referenced measures are of two types:
  - scores on single skills and
  - scores on multiple skills

	Type of Skill	Score
Single Ski	Melody voice	pass-fail;
	Completion of requ	complete-incomplete;
	Answer to a questi	Right-wrong;
	Adequacy of self h	•Frequently drinks from a cup without assista •Seldom drinks from a cup without assistance •Never drinks from a cup without assistance
	Assistance needed	•Drinks from a cup with out physical assistanc •Drinks from a cup with verbal prompts, •Drinks from a cup with physical guidance.

- **Multiple- Skill Scores** involves complex skills (such as oral reading) and multiple observations of single skills. The scores could be correct percentages or rate.
  - **Percentages:** Number of correct or incorrect responses as a function of the total number of responses (%correct); or
  - **Rate:** Number of correct or incorrect responses as a function of time (the number of correct responses per minute)

  - Verbal labels for percentages that facilitate instruction are: **“mastery”** and **“instructional level”**

- **Mastery Level:**
  - *Mastery* = 90% and more correct
  - *Non-mastery* = Less than 90% correct
  - In reality, sometimes 100% is the mastery (like crossing the road safely)
- **Instructional Level:**
  - *Frustration level*, less than 85% correct,
  - *Instructional level,* between 85% & 95% correct,
  - *Independent level*, above 95% correct.
- **Example:**
  - In reading, students who decode more than 95% correct of the words should be able to read a passage without assistance,
  - Students who decode between 85 and 95 percent of the words in a passage should be able to read and comprehend that passage with assistance,
  - Students who cannot decode 85% of the words in a passage will probably have difficulty comprehending the material, even with assitance.

2.3. Norms

- To understand a student’s performance, we must also know the characteristics and abilities of the people with whom we compare a test taker.
- In norm referenced testing, we compare test taker with a group of students tested by the test author. This comparison group is called the norm or standardization sample.

Norm (Cont'd)

A) Representativeness:
   - Does the norm sample contain individuals with relevant characteristics and experiences?
   - Are the characteristics and experiences present in the sample in the same proportion as they are in the population of reference?

- Most commonly considered developmental and socio-cultural characteristics include:
   - Developmental characteristics (gender, age, grade)
   - Socio-cultural characteristics (gender, ethnicity, socioeconomic status, acculturation of parents, race and cultural identity of parents)
   - Geographical locations
   - Intelligence

B) Technical Considerations

- **Planning:** Planning to get representative sample of people and locate potential participants with the needed characteristics (cluster sampling of urban, semi-urban, and rural).
- **Proportional representation:** test authors adjust norms to make them representative.
  - Norm samples can be manipulated to confirm with population characteristics.
  - If a test has covers Kindergarten through twelfth grade, 13 (1 for each grade) representative nor groups should be adequately represented. If this test included males and females, then there are 26 norm groups. Hence, representativeness should be demonstrated for each comparison group.
- **Number of participants:** it should be large enough to guarantee stability and calculate a full range of derived scores. In practice, 100 participants in each age or grade is considered the minimum.
- **Normative updates:**

2.4. Reliability

- **Definition:** Reliability means "repeatability" or "consistency". A measure is considered reliable if it would give us the same result over and over again, assuming that what we are measuring isn't changing.
- **Types of Reliability**
  - There are four *general classes of reliability estimates*, each of which estimates reliability in a different way. They are:
  - **Inter-Rater or Inter-Observer Reliability**
    Used to assess the degree to which different raters/observers give consistent estimates of the same phenomenon.
  - **Test-Retest Reliability**
    Used to assess the consistency of a measure from one time to another.
  - **Parallel-Forms Reliability**
    Used to assess the consistency of the results of two tests constructed in the same way from the same content domain.
  - **Internal Consistency Reliability**
    Used to assess the consistency of results across items within a test.

2.5. Validity

- **Definition:** Validity is the extent to which a test measures what it claims to measure. It is vital for a test to be valid in order for the results to be accurately applied and interpreted.
- **Types of validity:**
  - **Content Validity:** When a test has content validity, the items on the test represent the entire range of possible items the test should cover. Individual test questions may be drawn from a large pool of items that cover a broad range of topics.
    - Appropriateness if included items (you might use table of specification),
    - Completeness of content,
    - Appropriateness of the measurement of the content,

- **Criterion-related Validity:** A test is said to have criterion-related validity when the test is demonstrated to be effective in predicting criterion or indicators of a construct. Two different types:
  - **Concurrent Validity** occurs when the criterion measures are obtained at the same time as the test scores. This indicates the extent to which the test scores accurately estimate an individual's current state with regards to the criterion. For example, on a test that measures levels of depression, the test would be said to have concurrent validity if it measured the current levels of depression experienced by the test taker.

- **Predictive Validity** occurs when the criterion measures are obtained at a time after the test. Examples of test with predictive validity are career or aptitude tests, which are helpful in determining who is likely to succeed or fail in certain subjects or occupations.

- **Construct Validity:** A test has construct validity if it demonstrates an association between the test scores and the prediction of a theoretical trait. Intelligence tests are one example of measurement instruments that should have construct validity.

Relationship between reliability and validity

2.6. Adapting Tests to Accommodate Students with Disabilities

Part III. Assessing Instructional Ecology

Part IV. Portfolio Assessment