STAT 530/J530
November 22nd, 2005

Instructor: Brian Habing

Department of Statistics

LeConte 203

Telephone: 803-777-3578

E-mail: habing@stat.sc.edu

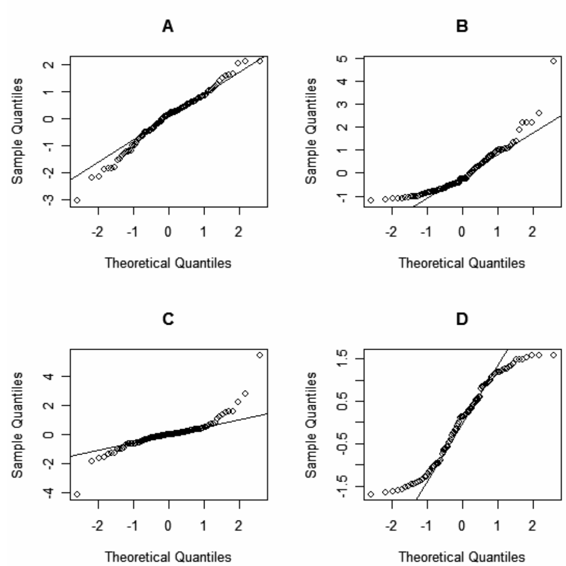
Homework 9

1)	Distribution	Q-Q plot	Boxplot
Heavy Tailed	__________	___	___
Light Tailed	__________	___	___
Normal	__________	___	___
Skewed Right	__________	___	___

chi-square distribution,
normal distribution,
t-distribution, &
uniform distribution

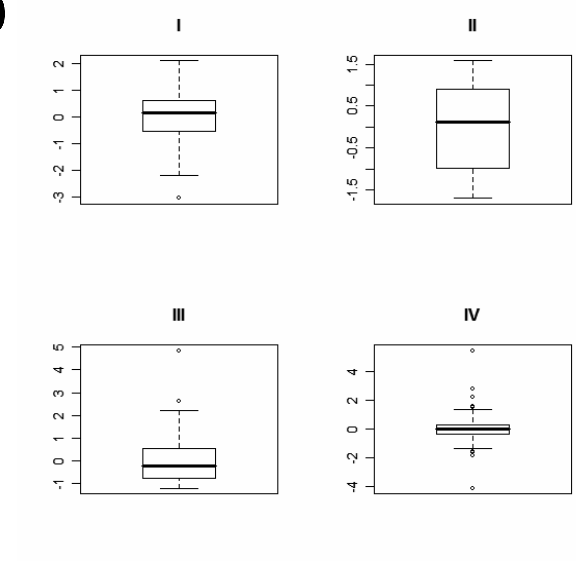
Homework 9

Q-Q plots



Homework 9

Box Plots



Homework 9

2) Assume we are trying to verify that a data set is multivariate normal. Briefly explain why can we stop checking if we find just one q-q plot, bivariate box-plot, or chi-square plot is extremely far from where it should be? For MANOVA why should we continue the checking if we find one that is only slightly off?

Homework 9

3) For the county data on the exam the total county income was divided by the population. Similarly on the homework the crab measurements were divided by the total crab size. Briefly explain why scaling in this manner is desirable.

Homework 9

4) Which of the following of the following are NOT reasons to standardize the data:

- standardization reduces differences between the individual observations
- it alleviates the problem of different variables using possibly different units of measurement
- it adjusts for variables having different variances
- it causes different groups of observations to separate more on each variable
- it makes it so that one variable doesn't dominate all of the others simply because of a larger range of values.

Homework 9

5) ______ Using Factor Scores

______ Using a Representative Question

______ Using a Sum Score

a. There is only one score for each observation

b. Each observation receives five scores

c. The results for each individual observation will heavily depend on the other observations used to fit the model

d. It can only be performed if the different questions have similar sorts of scales

e. It keeps less information than the other two methods

Next

Thursday 24th: Thanksgiving – No Class

Tuesday 29th: Structural Equation Modeling

Thursday 1st: Homework 10 is due, ice cream field trip as penance for Homework 6 grade being late! With time for questions while we eat.

5:30pm Tuesday, December 6th – Final Exam is Due

Item Response Theory

IRT is a class of methods for modeling the relationship between the <u>responses</u> to the items or questions on a test, scale, or questionnaire and the underlying <u>latent</u> <u>trait</u>(s) that the test is designed to measure.

The Data

Items→

Subjects↓

```
110011010110100001010110
000111001111111000001010
111000111110000000000100
111101111111111100101000
010000101000000000000000
110111111100000100001000
011101111111100000001111
111111111110000000000110
011111111111101000000001
111100110011100000011010
011100111010011000101000
101010011111111101000010
011000001111110000010010
111111111011100010000100
111111111111111111111111
001010110110101000010111
```

U_{ij} is the response of examinee *j* to item *i*.

Sample Data Sets

- The GRE
- A test to measure clinical depression
- A survey to measure abortion attitudes
- A survey to measure illegal behaviors and social support
- An instrument to measure pain

The Goals

- Estimating the properties of the items
- Choosing the most effective items
- Estimating subjects' "ability" levels
- Determining the dimensional structure of the latent construct.
- Detecting "bias" in the questions
- Equating/Linking different forms

Three Alternatives

- Classical Test Theory
- Classical Item Analysis
- Factor Analysis

Classical Test Theory

The basis of classical test theory is the formula

$$X = T + E$$

observed = true + error
score score

Where E(X)=T and Cor(E,T)=0

Using CTT

Reliability of the test is $\rho_{XT}^2 = \rho_{TT'}$

The standard error of measurement is

$$\hat{\sigma}_E = \hat{\sigma}_X \sqrt{1 - \hat{\rho}_{XT}^2}$$

Can get an estimate of T using regression $T = \rho_{XT}^2 (X - \mu_X) + \mu_T$

Weaknesses in CTT

- Entirely test and population dependent
- Does not describe individual items
- Has a constant standard error of measurement across all true score levels

Classical Item Analysis

Item Difficulty measured by Item p-value (percentage correct)

Item Discrimination measured by the Biserial or Point-Biserial Correlation between the item score and total observed score

Both depend on the particular examinees, and discrimination depends on the remaining test items.

Factor Analysis

$$X_1 = a_{11}F_1 + \cdots a_{1m}F_m + e_1$$
$$X_2 = a_{21}F_1 + \cdots a_{2m}F_m + e_2$$
$$\vdots$$
$$X_p = a_{p1}F_1 + \cdots a_{pm}F_m + e_p$$

The X_i are the observed variables, the F_j are the m common factors, the e_i are the specific errors, and the a_{ij} are the factor pxm factor loadings.

Weaknesses in Factor Analysis

- Designed for continuous observed responses
- Hypothesis testing assumes multivariate normality of the underlying latent traits

Item Response Theory

$\underline{U}=(U_1,\ldots U_i,\ldots U_n)$ is the vector of item responses
u_i is a particular possible response to item *i*
Θ are the latent traits measured by the test
$\underline{\theta}$ is a particular level of the latent traits

The goal is to model:

$$P[\underline{U} = \underline{u} \mid \underline{\Theta} = \underline{\theta}]$$

Monotone Homogeneity Model

Three commonly made assumptions are:

- Local Independence = Item responses are conditionally independent given θ
- Unidimensionality = θ is scalar
- Monotonicity = $P[U_i \geq k|\Theta=\theta]$ is increasing in θ.

Item Response Functions

The $P[U_i \geq k|\Theta=\theta]$ are the item response functions (IRFs) for the test. They are sometimes called item characteristic curves (ICCs).

For dichotomous items these are

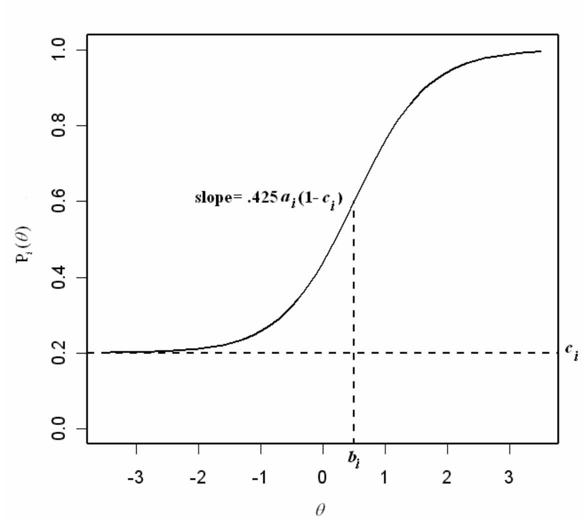
$$P_i(\theta)= P[U_i =1|\Theta=\theta]$$

Birnbaum's 3PL Model

One of the common models for dichotomous item tests is Birnbaum's (1968) three parameter logistic model.

$$P_i(\theta) = c_i + \frac{1-c_i}{1+\exp(-1.7a_i(\theta-b_i))}$$

3PL Model a_i=discrimination, b_i=difficulty, c_i=guessing



Invariance

Because the models are based on an underlying latent trait they have an invariance property.

When the IRT model fits the data the same IRFs are obtained regardless of the ability distribution of the sample of examinees or which other items are used on the test.

1PL or Rasch Model

Rasch's (1960) model is the 3PL model with the guessing set to 0 and the discrimination constant across all items.

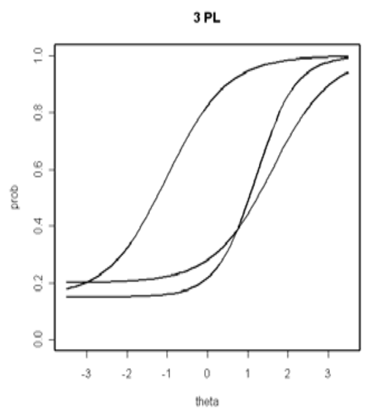
$$P_i(\theta) = \frac{1}{1 + \exp(-a(\theta - b_i))}$$

Benefits of Rasch

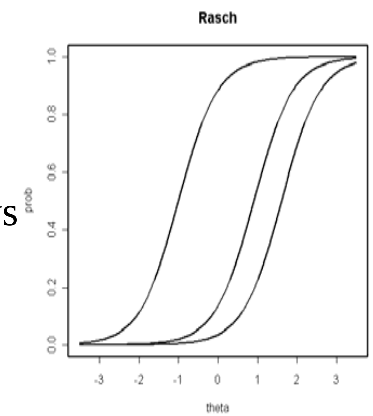
The item response functions do not cross, so that items can be ranked on difficulty regardless of examinee ability level.

The marginal sums are sufficient statistics for the examinee ability and item difficulty respectively.

Objective Measurement



vs



Problem with Rasch

It often just doesn't fit the data!

Other Models

- Polytomous/Likert Scale items (cumulative probability, adjacent category, or continuation ratio)

Widely used and studied.

- Multidimensional Abilities

Difficult to fit and not widely used.

Can be re-parameterized as non-linear factor analysis.

Other Models

- Non-Monotone Response Functions (unfolding models)

Very recently developed

- Locally Dependent Items (testlet models or conjunctive IRT)

There are several strong proponents of testlet models. Conjunctive IRT model has received little attention.

Estimation

The item parameters and q's are commonly estimated by marginal maximum likelihood using the EM algorithm. (First items, then examinees.) MCMC is used for some more complicated models.

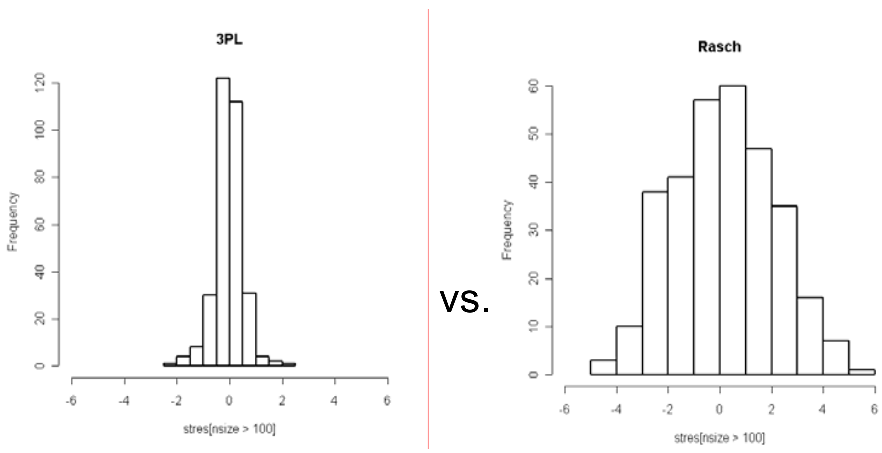
For most procedures, 20 dichotomous items and 400 examinees is considered a small sample size.

Goodness of Fit

There are no widely accepted test statistics for testing goodness of model fit (although most IRT packages will produce some)

Residual plots are often used to determine whether the IRF has the correct parametric form.

Standardized Residual Plots



Checking LI and d=1

The twin assumptions of local independence and unidimensionality are often tested by fitting a unidimensional model and then testing for lack of local independence (for example observing correlations between the residuals).

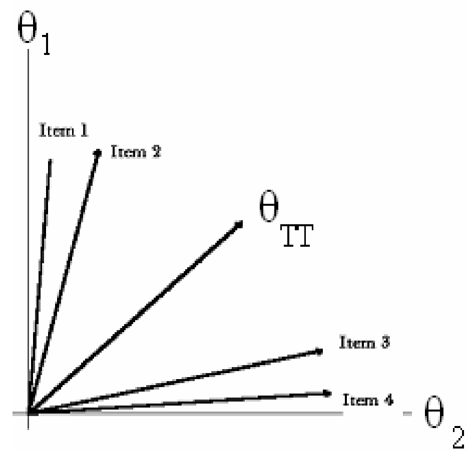
Conditional Covariances

Zhang and Stout (1999) demonstrated the relationship between the covariance of an item pair conditioned on the “best” unidimensional sub-trait,

$$Cov(U_i,U_j|\Theta_{TT}=\theta)$$

is directly related to the underlying dimensional structure.

Geometric Representation



Items on opposite sides of Θ_{TT} have negative CCOVs

those on the same have positive

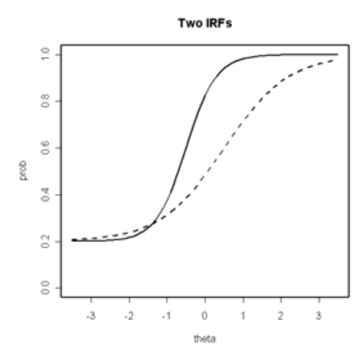
CCOV Based Procedures

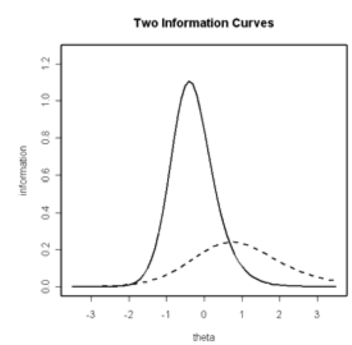
The CCOVs can be estimated using d=1 parametric models, or by using the observed test score as a proxy for Θ_{TT}.

They can then be used to construct hypothesis tests, or be converted into a distance for clustering or scaling.

Selecting Items

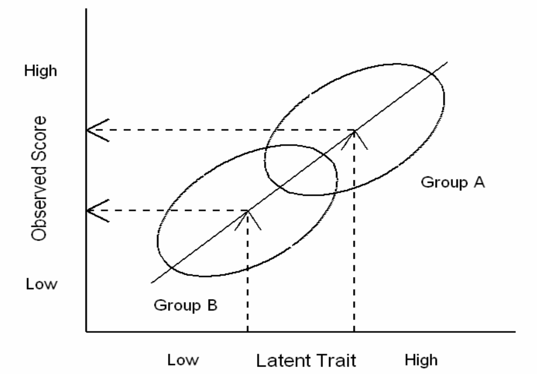
If a parametric model is fit, then each item produces an item information curve.





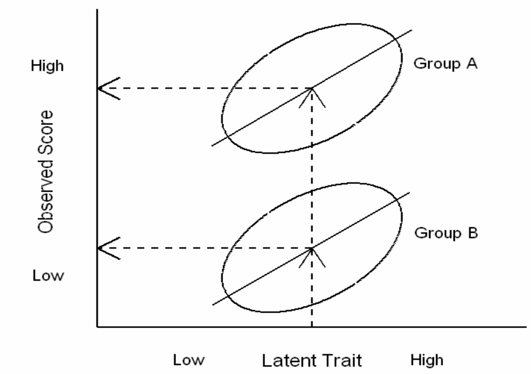
Test Equity

Impact – An item demonstrates impact if it has different statistical properties for members in different groups.



Test Equity

Differential Item Functioning – DIF is when the item has different properties for members of different groups after "controlling" for the ability it is supposed to measure.



Test Equity

An item is **biased** if it has different statistical properties for members of different groups that are due to factors in the test beyond what the construct is designed to measure.

Equating / Linking

If tests share at least a few items in common then the estimated scores can be placed on a common metric by fitting the two exams simultaneously.

If the tests don't share items in common it is necessary to make assumptions about the underlying populations.