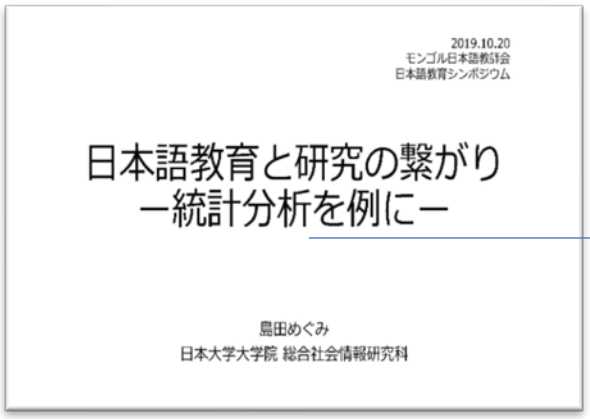
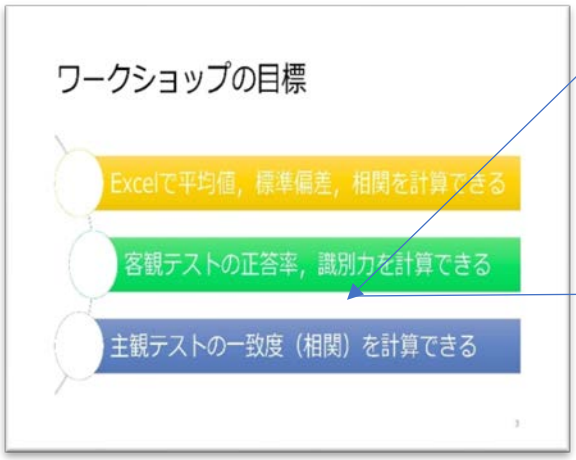
基調講演

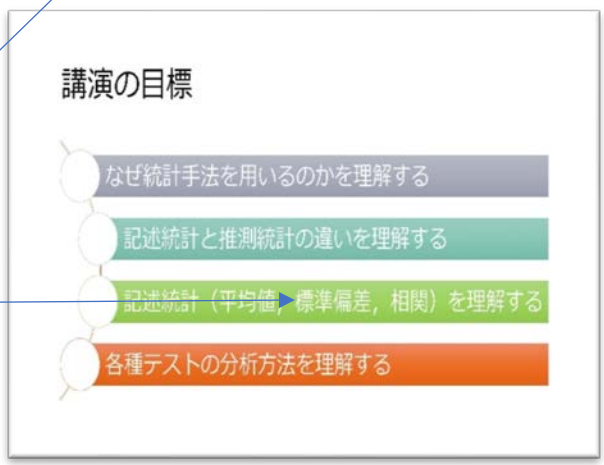
「日本語教育と研究のつながり―統計分析を例に―」

島田めぐみ先生

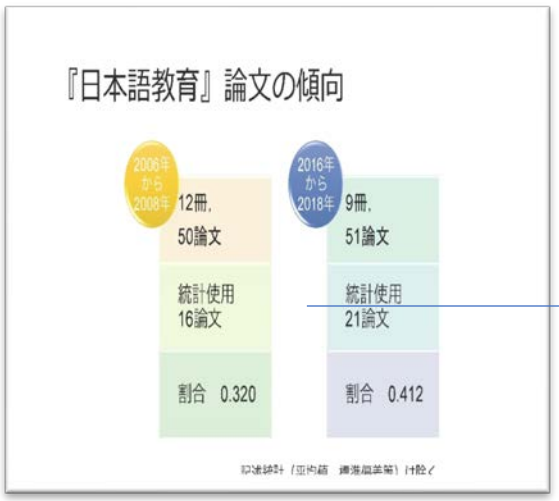
日本大学大学院総合社会情報研究科　教授

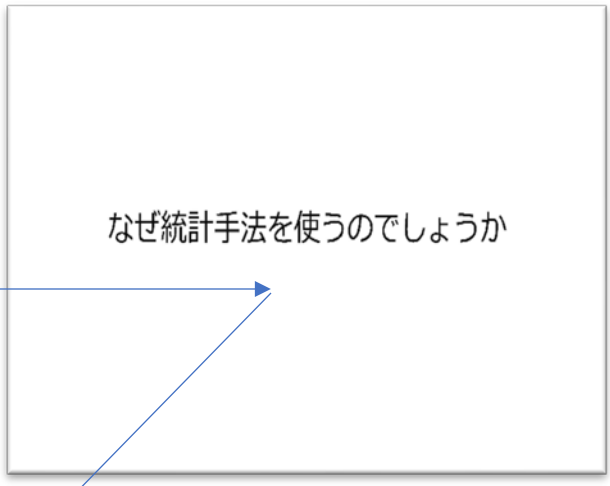


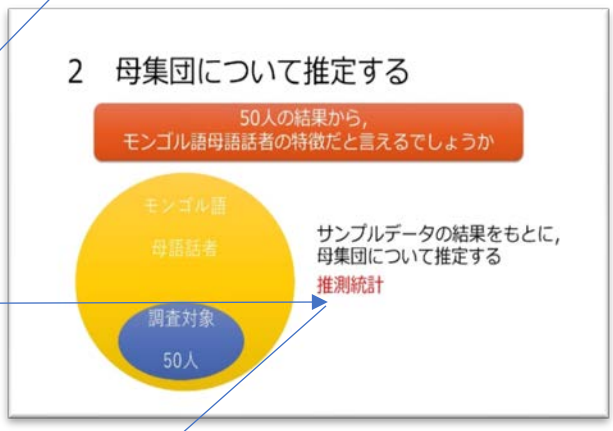


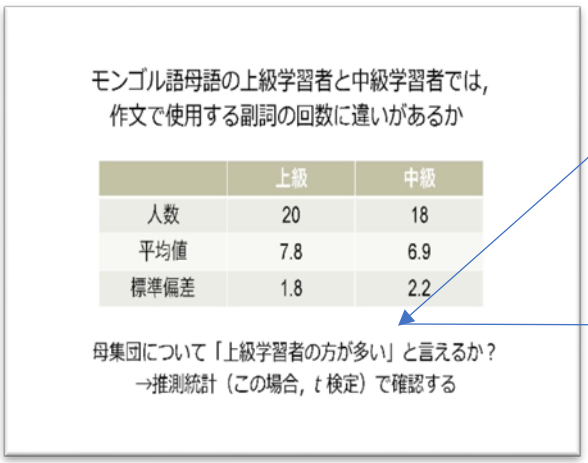


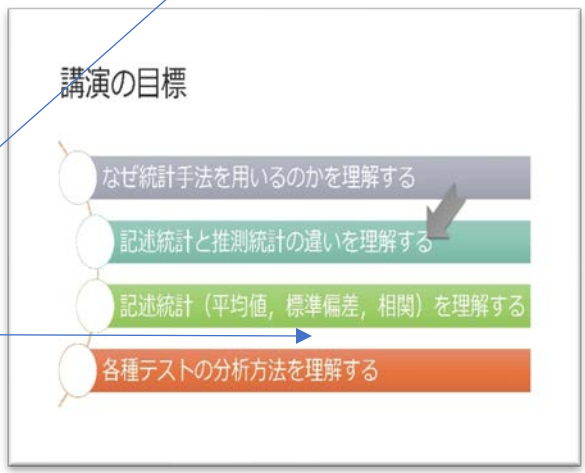
日本語教育と研究の繋がり
-統計分析を例に-









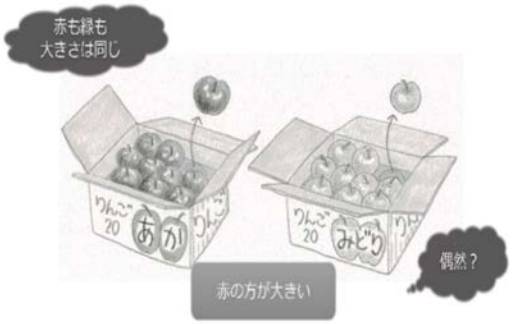


統計的記述と統計的推測の違い

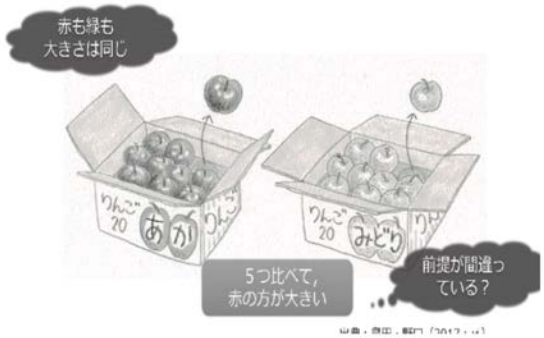
- 統計の種類：統計的記述（記述統計）と統計的推測（推測統計）
- 統計的記述：収集したデータを要約し，わかりやすい値で表すこと（例：平均値，標準偏差など）
- 統計的推測：収集したデータ（標本）を元に母集団の特徴を推測すること（例：t検定，χ^2検定など）

統計
- 統計的記述（記述統計）
- 統計的推測（推測統計）

母集団の推定の考え方



母集団の推定の考え方



母集団の推定の考え方

記述統計

記述統計

例1：クラスで実施した期末テストの結果を知りたい

→統計的記述→平均値，標準偏差，ヒストグラムなど

例2：学校で実施した授業評価アンケートの結果を報告したい

→統計的記述→平均値，標準偏差，相関など

例3：学校で実施した自己評価の結果が、学年で異なるか、知りたい

→統計的記述→平均値，標準偏差などの比較

日本語教育と研究の繋がり
-統計分析を例に-

推測統計

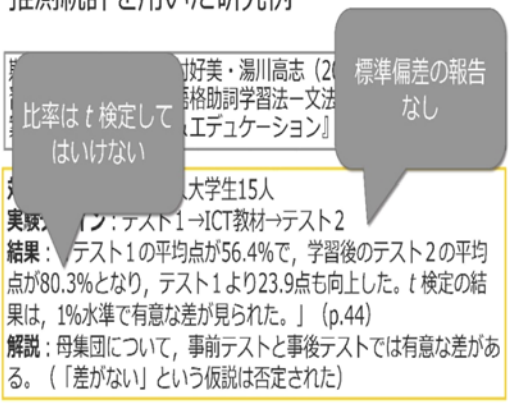
例1
［目的］モンゴル語母語話者の日本語の発音の特徴を検討する
［母集団］モンゴル語母語話者の日本語学習者
［標本］モンゴル語母語話者50人と日本語母語話者50人
→統計的推測：*t* 検定などにより、標本を元に母集団について推測
例2
［目的］中等教育学習者と高等教育学習者で学習動機に違いがあるかを調べる
［母集団］中等教育日本語学習者と高等教育日本語学習者（日本国外）
［標本］合計100人のデータを収集
→統計的推測：χ^2検定などにより、標本を元に母集団について推測

推測統計を用いた研究例

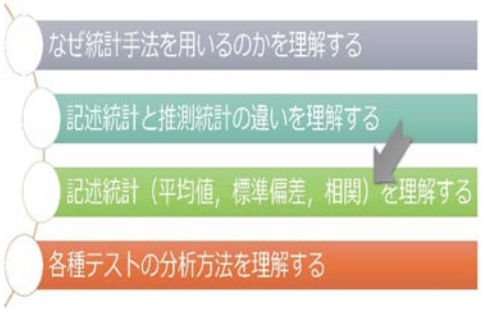
斯日貢・中平勝子・福村好美・湯川高志（2016）「モンゴル人学習者を対象とした日本語格助詞学習法ー文法対比型学習法の提案ー」『コンピュータ&エデュケーション』41，40-45

対象：在日内モンゴル人大学生15人
実験デザイン：テスト1→ICT教材→テスト2
結果：「テスト1の平均点が56.4%で，学習後のテスト2の平均点が80.3%となり，テスト1より23.9点も向上した。*t* 検定の結果は，1%水準で有意な差が見られた。」（p.44）
解説：母集団について，事前テストと事後テストでは有意な差がある。（「差がない」という仮説は否定された）

推測統計を用いた研究例



講演の目標



記述統計の報告例

	N	*M*	*SD*
英語母語話者	16	78.8	10.6
中国語母語話者	20	79.8	11.6

データの特徴を表す

・データ（例えばテストの得点）が得られたら，要約して，全体的な特徴を見る

1. データの分布の状況は？（何点が多い？高得点は何人？）
→度数分布表，ヒストグラム

度数の分布

データの羅列→傾向わからない

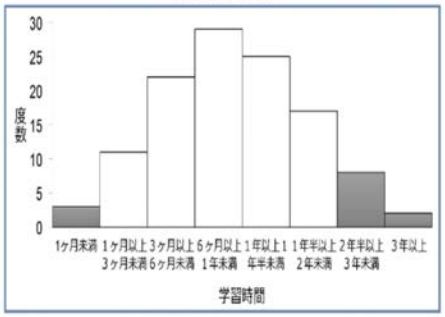
学習期間	学習期間
1年以上2年未満	1年以上2年未満
1ヶ月以上3ヶ月未満	1ヶ月以上3ヶ月未満
3ヶ月以上6ヶ月未満	3ヶ月以上6ヶ月未満
6ヶ月以上1年未満	2年以上3年未満
6ヶ月以上1ヶ月未満	6ヶ月以上1ヶ月未満
3ヶ月以上6ヶ月未満	1年以上2年未満
1年以上2年未満	3ヶ月以上6ヶ月未満
2年以上3年未満	6ヶ月以上1ヶ月未満

度数分布表→傾向わかりやすい

学習期間	度数
1ヶ月未満	3
1ヶ月以上3ヶ月未満	11
3ヶ月以上6ヶ月未満	22
6ヶ月以上1年未満	29
1年以上1年半未満	25
1年以上2年未満	17
2年以上3年未満	8
3年以上	2

度数の分布

ヒストグラム



データの特徴を表す

- データ（例えばテストの得点）が得られたら，要約して，全体的な特徴を見る

1. データの分布の状況は？（何点が多い？高得点は何人？）
   →度数分布表，ヒストグラム
2. データの中心は（代表値）？
   →平均値，中央値

代表値

例1　20 30 40 50 60 70 80

平均値 50　　中央値 50

例2　3 30 40 50 60 70 80

平均値 47.6　　中央値 50

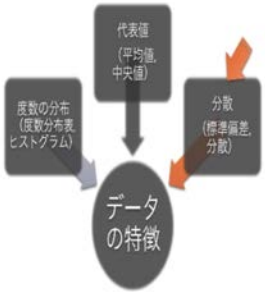
平均値はすべての値が反映される

中央値は変化なし

データの特徴を表す

- データ（例えばテストの得点）が得られたら，要約して，全体的な特徴を見る

1. データの分布の状況は？（何点が多い？高得点は何人？）
   →度数分布表，ヒストグラム
2. データの中心は（代表値）？
   →平均値，中央値
3. データの散らばりの具合は？
   →標準偏差，分散



データの散らばり

Aさんは聴解テストと読解テストの両方で70点を取りました。
どちらのテストも平均点は50点。
聴解テストも読解テストも，受験者内の順位は同じでしょうか。

データの散らばり

Aさんは聴解テストと読解テストの両方で70点を取りました。
どちらのテストも平均点は50点。
聴解テストも読解テストも，受験者内の順位は同じでしょうか。

受験者9人の聴解テストの得点
10 20 30 40 50 60 70 80 90
Aさん
散らばりが大きい

受験者9人の読解テストの得点
30 35 40 45 50 55 60 65 70
Aさん
散らばりが小さい

平均値だけではテストの特徴を表せない
→散らばり（標準偏差，分散）の確認が必要

標準偏差

標準偏差とは
各データ（この場合得点）が平均値から平均してどのぐらい離れているか

得点	20	30	40	50	60	70	80
❶	-30	-20	-10	0	10	20	30
❷	900	400	100	0	100	400	900
❸				400			
❹				20			

❶得点-平均値
各データについて平均値との差を**計算する。**

❷（平均値-得点）を二乗
プラスマイナスは関係なく，平均値からの距離を　得るために，❶を二乗してマイナスを取る。

❸❷の平均

❹平方根
❸の値は二乗した値なので，平方根に戻す

平均値だけではなく
標準偏差も報告することが必要

2変量のデータの関係を見る

漢字テストの結果と読解テストの結果は関連があるでしょうか。
つまり，漢字テストの得点が高い人は読解テストの得点も高いでしょうか。

2変量のデータの関係を見る

漢字テストの結果と読解テストの結果は関連があるでしょうか。
つまり，漢字テストの得点が高い人は読解テストの得点も高いでしょうか。

図で表すと

学生	A	B	C	D	E	F
漢字	46	70	65	87	58	56
読解	32	54	86	92	68	45

散布図
漢字得点が高い人は読解得点も高い傾向
読解テスト
漢字テスト

2変量のデータの関係を見る

漢字テストの結果と読解テストの結果は関連があるでしょうか。
つまり，漢字テストの得点が高い人は読解テストの得点も高いでしょうか。

図で表すと

学生	A	B	C	D	E	F
漢字	46	70	65	87	58	56
読解	32	54	86	92	68	45

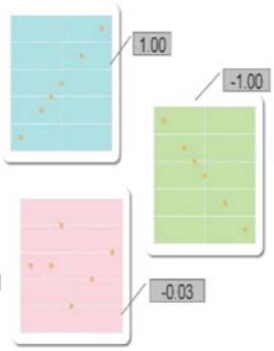
散布図
相関係数：0.791
数値で表すと
読解テスト
漢字テスト

相関係数と散布図

相関係数
- 2変量の関係がぴったり一致すると1.0
- 全く逆だと-1.0
- 相互に関連がないとほぼ0.0

解釈の目安

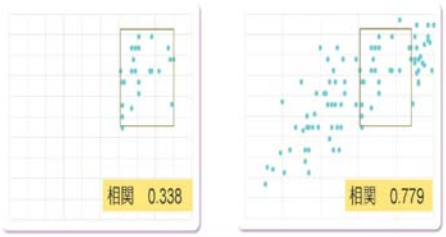
-0.2~0.2	ほとんど相関がない
-0.4~-0.2, 0.2~0.4	弱い相関
-0.7~-0.4, 0.4~0.7	中程度の相関
-1.0~-0.7, 0.7~1.0	強い相関



集団の範囲の影響

自己評価と日本語テストの関係をクラスで調査。あまり相関が高くない？

複数のレベルのクラスで実施した結果。自己評価と日本語テストは強い相関。



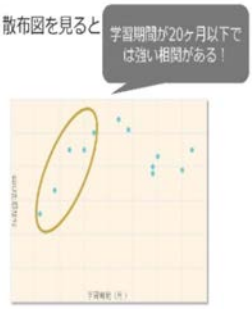
クラスのように集団の範囲が狭いと相関は低くなりがち

散布図を描いてわかること

学生の学習期間と自己評価得点の相関を計算

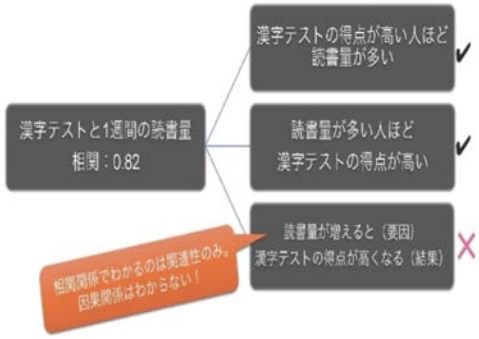
- 学習期間と自己評価得点の相関 →0.306
- 学習期間と自己評価得点の間の相関は弱い？？
- 結論づける前に散布図を確認！

散布図を見ると



散布図で確認すると得られる情報は多い

相関関係でわかること



相関を用いた研究例

島田めぐみ・三枝令子・野口裕之（2016）「日本語Can-do-statementsを利用した言語行動記述の試み」『世界の日本語教育』16，75-88

対象：T大学留学生91名
データ：JLPT問題からなるテスト（PT），Cds自己評価の結果
結果：「Cds調査総点とPT総点との相関は0.804と高い値を示し，各類についても高い値が示されている。（中略）したがって，Cdsは，JLPTとも高い相関関係になり，日本語能力を反映する尺度としての有効性が示されたと言える。」（p.80）
解説：JLPTの得点が高い者は，Cds自己評価の得点も高いので，Cdsが日本語能力を反映するものだと言える。

講演の目標

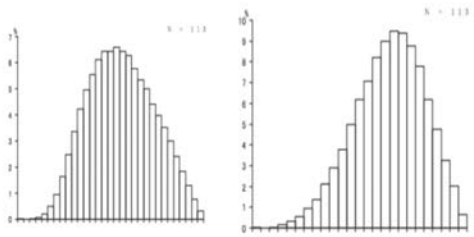
各種テストの分析

- 客観テスト
  - 日本語能力試験（2009年まで）
  - 日本語能力試験（2010年以降）
  - 教育機関のテスト→ワークショップ
- 主観テスト
  - 作文テスト→ワークショップ

客観テストの分析

- 総合点について全体的傾向を計算する
  - →平均値，標準偏差を計算，ヒストグラムを作成
- 各項目の分析をする
  - →困難度，識別力を計算

ヒストグラム（度数の分布）
日本語能力試験 2009年度第1回1級 読解・聴解



ヒストグラム（度数の分布）
日本語能力試験 2009年度第1回1級 読解・聴解

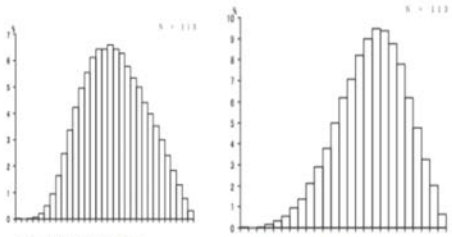


図2-4 2009年度第1回1級聴解
得点分布（30点満点 1点刻み）

図2-6 2009年度第1回1級読解
得点分布（120点満点 5点刻み）

平均値（代表値），標準偏差（分散）
日本語能力試験 2009年度第1回1級

文字・語彙 標準偏差 14.41

聴解 標準偏差 18.31

	文字・語彙	聴解		総点
	71.04	55.98	131.36	258.45
標準偏差	14.41	18.31	29.42	50.45
満点	100	100	200	400
受験者数	113,641	113,618	113,517	113,482

平均値からの距離が
平均して14.41離れている

	語彙	読解	文法
	37.30	81.35	50.01
	8.22	21.53	12.28
	56	26	74
受験者数	113,641	113,641	113,517	113,517

客観テストの分析

- 総合点について全体的傾向を計算する
  - →平均値，標準偏差を計算
- 各項目の分析をする
  - →困難度，識別力を計算

項目の難しさ（正答率）

受験者の能力を識別する程度
（合計点の高い人ほど正答する項目は識別力高い）

日本語能力試験 2009年度第1回 1級 文字語彙

お世話になった恩人を心を込めて_____。

1 もらした　　2 もてなした
3 もたらした　　4 もがいた

国際交流基金・日本国際教育支援協会（2011）
『平成21年度日本語能力試験（第1回・第2回）分析評価に関する報告書』p.215より

日本語能力試験 2009年度第1回 1級 文字語彙

お世話になった恩人を心を込めて_____。

1 もらした(11.5%)　　*2 もてなした(44.2%)
3 もたらした(33.7%)　　4 もがいた(10.3%)

正答率 0.442　識別力 0.451

国際交流基金・日本国際教育支援協会（2011）
『平成21年度日本語能力試験（第1回・第2回）分析評価に関する報告書』p.215より

得点段階別選択率

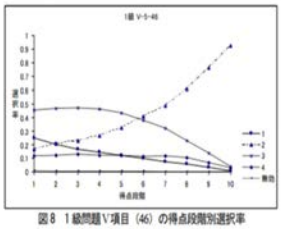


図8　1級問題V項目（46）の得点段階別選択率

国際交流基金・日本国際教育支援協会（2011）
『平成21年度日本語能力試験（第1回・第2回）分析評価に関する報告書』p.216より

得点段階別選択率

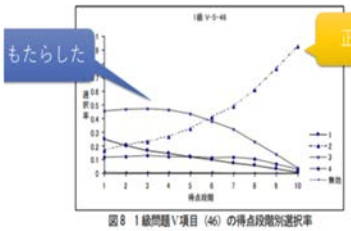


図8　1級問題V項目（46）の得点段階別選択率

国際交流基金・日本国際教育支援協会（2011）
『平成21年度日本語能力試験（第1回・第2回）分析評価に関する報告書』p.216より

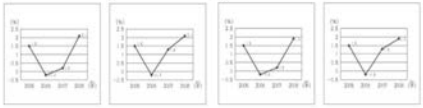
日本語能力試験 2009年度第1回 1級 聴解

男の人が国の経済成長率について話しています。この国の経済成長率を示す正しいグラフはどれですか。

低成長期の経済成長率は一般に1.3%程度であると言われていますが，今日発表された2008年の経済成長率は2%に迫る伸びとなりました。2006年の成長率はマイナス，2007年もわずかな伸びにとどまっていたことを考えると，ここへきて，景気の回復傾向がはっきりしてきたと言えるでしょう。

日本語能力試験 2009年度第1回 1級 聴解

1　2　3　4



1 (29.4%)　2 (37.1%)　3 (18.7%)　4 (14.7%)

正答率 0.187　識別力 0.070

国際交流基金・日本国際教育支援協会（2011）
『平成21年度日本語能力試験（第1回・第2回）分析評価に関する報告書』p.294 より

日本語教育と研究の繋がり
-統計分析を例に-

得点段階別選択率

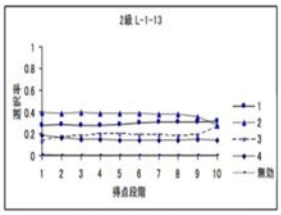
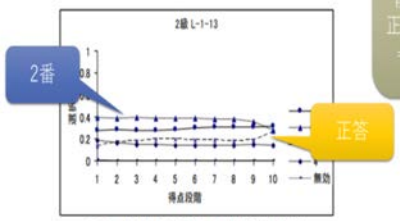


図17 2級問題1項目（13）の得点段階別選択率

正答率 0.187 識別力 0.070

国際交流基金・日本国際教育支援協会（2011）
『平成21年度日本語能力試験（第1回・第2回）分析評価に関する報告書』p.295より

得点段階別選択率



2番

正答

能力の高い人と
能力の低い人で
正答率に差がない
=識別力が低い

図17 2級問題1項目（13）の得点段階別選択率

正答率 0.187 識別力 0.070

国際交流基金・日本国際教育支援協会（2011）
『平成21年度日本語能力試験（第1回・第2回）分析評価に関する報告書』p.295より

特点段階別選択率

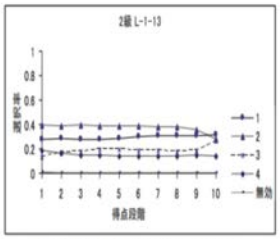
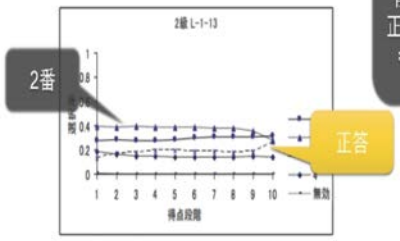


図17 2級問題1項目（13）の得点段階別選択率

正答率 0.187 識別力 0.070

国際交流基金・日本国際教育支援協会（2011）
『平成21年度日本語能力試験（第1回・第2回）分析評価に関する報告書』p.295より

特点段階別選択率



2番

正答

能力の高い人と
能力の低い人で
正答率に差がない
=識別力が低い

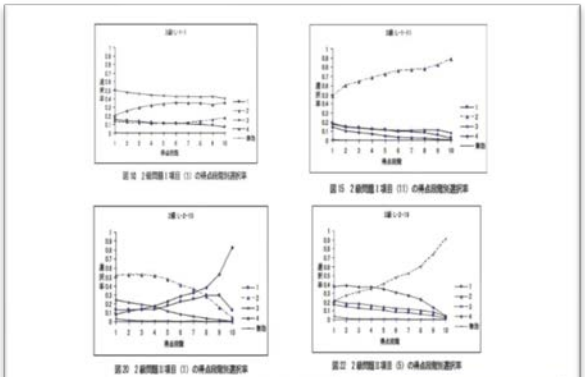
図17 2級問題1項目（13）の得点段階別選択率

正答率 0.187 識別力 0.070

国際交流基金・日本国際教育支援協会（2011）
『平成21年度日本語能力試験（第1回・第2回）分析評価に関する報告書』p.295より

識別力の考え方

- -1.0から1.0の間の値
- プラス：能力の高い人が正答する傾向
- マイナス：能力の低い人が正答する傾向
- 0.3以上 高め（識別する力が強い）
- 0.3以下 低め
- 0.0 程度 識別する力なし（正答したかどうかは能力と関係ない）
- 正答率が高すぎたり低すぎたりする場合は識別力は低めとなる



正答率 0.717 識別力 0.235

正答率 0.317 識別力 0.078

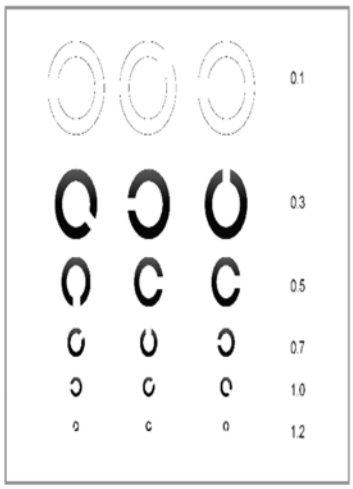
正答率 0.311 識別力 0.460

正答率 0.481 識別力 0.418

国際交流基金・日本国際教育支援協会（2011）

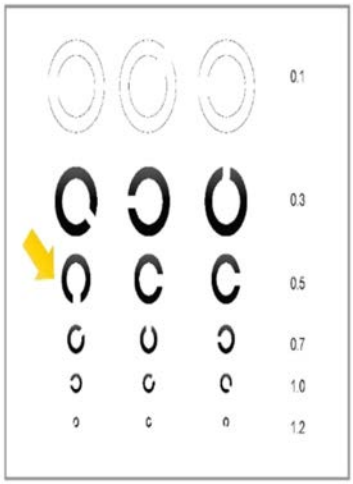
項目応答理論（IRT）の導入
項目の難しさ

古典的テスト理論	IRT
「学校」の読み方 正答率 中級のクラスでは0.9 初級のクラスでは0.5 →この項目の難しさは？ →正答率は受験者集団に影響される	項目の難しさは受験者集団とは独立して決定 受験者の能力に左右されない、その項目の難しさを特定。 ＝視力検査の環のよう



項目応答理論（IRT）の導入
得点の計算

古典的テスト理論	IRT
• 正答数に基づいて得点が決まる受験者 • 同一のでも，難しい問題項目が多いテストでは得点が低く，易しい問題項目が多いテストでは得点が高くなる • →この受験者の能力は？	• 各項目の難しさは最初に計算されていて，易しい項目から難しい項目まで1本の尺度 • 受験者自身も尺度上に位置付けられる • ＝視力検査のように • （難しい検査表，易しい検査表はない）

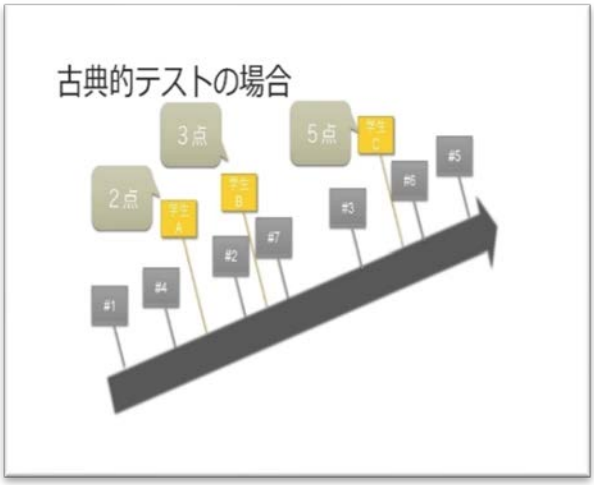


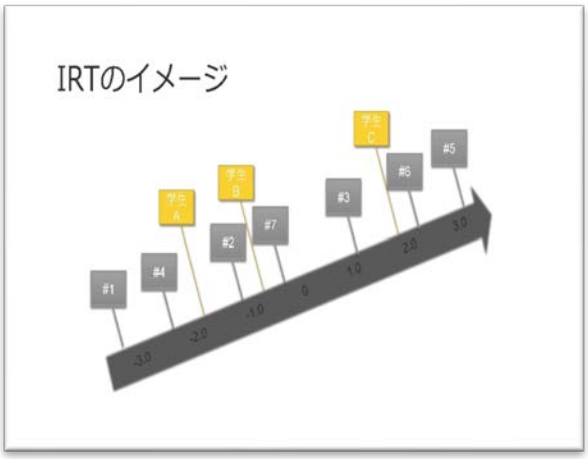
IRTのイメージ

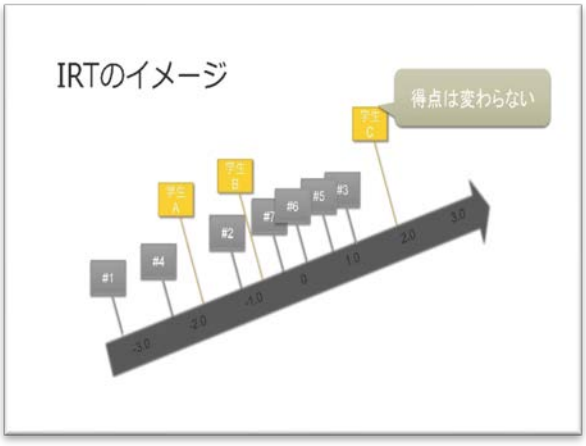
日本語能力試験

旧試験	新試験
素点（正答した項目数）に基づいて計算→難しいテストだと得点が低くなる。テストの得点は問題の難しさに影響される。 →違う回のテストの結果を比較できない。	尺度得点（IRT）に基づいて計算。 →いつでも同じ基準なので、違う回のテストの結果を比較できる。

日本語教育と研究の繋がり
-統計分析を例に-

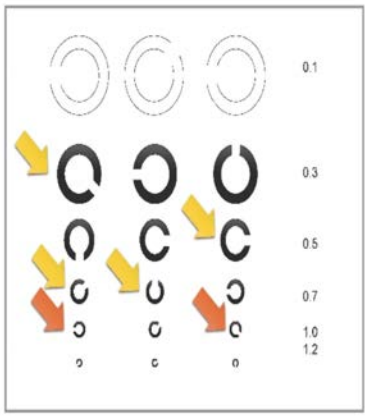






項目応答理論（IRT）の導入
CAT

- CAT（Computer Adaptive Testing，コンピュータ適応型テスト）は IRTによって難しさが推定されている項目が出題される
- 最初に中程度の困難度の項目が出題され，受験者がその問題項目に正答すると，それよりも難しい項目が出題され，逆に間違えると，その項目よりも易しい項目が出題される
- 受験者によって項目数も所要時間も異なる
- CBT（Computer Based Testing）とは違う



テストの種類による正答率と識別力の考え方

熟達度テスト（プレースメントテスト）	到達度テスト（定期テスト）
• いろいろな難しさがほしい。 • 識別力が高い項目が望ましい。	• 全員にできてほしい • 正答率の低すぎる項目は要注意 • 正答率が高い（ほぼ全員正答）と，識別力は低くなる。その場合は識別力が低くても可。

教育機関でテスト結果を分析する

テストの種類により分析の観点は異なります

学習内容に関係なく，受験者の能力を測定する
→いろいろな難しさの項目を出題
JLPT など

学習（教授）内容をどの程度習得したか測定する
→教授した内容を出題
定期テスト

教育機関でテスト結果を分析する

テストの種類により分析の観点は異なります

プレースメントテスト
JLPT など
小テスト
定期テスト

テストを分析する意義

① 学生の習熟度がわかる

② 教授について反省点が得られる（正答率が低い項目から）

・教授内容や教授方法に問題がある可能性がある

・教材に問題がある可能性がある

→テスト分析は教授内容について振り返るいい機会となる

主観テストの分析

- 総合点について全体的傾向を計算する
  →平均値，標準偏差を計算
- 信頼性を計算する
- →評価者間の信頼性，評価者内の信頼性

評定者間信頼性

学生番号	評定者A	評定者B	評定者C
1	8	9	9
2	4	5	5
3	3	2	2
4	8	6	6
5	5	4	9
6	10	10	9
7	3	4	7
8	6	6	3
9	2	1	4
10	9	7	7

評定者間信頼性

相関係数を計算

学生番号	評定者A	評定者B	評定者C
1	8	9	9
2	4	5	5
3	3	2	2
4	8	6	6
5	5	4	9
6	10	10	9
7	3	4	7
8	6	6	3
9	2	1	4
10	9	7	7

評定者間信頼性

学生番号	評定者A	評定者B	評定者C
1	8	9	9
2	4	5	5
3	3	2	2
4	8	6	6
5	5	4	9
6	10	10	9
7	3	4	7
8	6	6	3
9	2	1	4
10	9	7	7

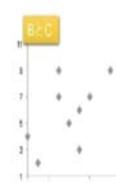
評定者間信頼性

学生番号	評定者A	評定者B	評定者C
1	8	9	9
2	4	5	5
3	3	2	2
4	8	6	6
5	5	4	9
6	10	10	9
7	3	4	7
8	6	6	3
9	2	1	4
10	9	7	7

評定者間信頼性

相関係数をまとめると…

	評定者A	評定者B	評定者C
評定者A	1.000		
評定者B	0.914	1.000	
評定者C	0.573	0.637	1.000



評定者間信頼性

相関係数をまとめると…

	評定者A	評定者B	評定者C
評定者A	1.000		
評定者B	0.914	1.000	
評定者C	0.573	0.637	1.000

評定者間信頼性が低い理由：
評価項目がない，
評価項目がわかりにくい
（解釈が評定者によって違う），
評定者の訓練が十分ではない，
など

評定者内信頼性

相関係数を計算

学生番号	評定者A1	評定者A2
1	8	9
2	4	5
3	3	3
4	8	8
5		
6		
7	3	4
8	6	6
9	2	1
10	9	9

0.970

講演の目標

- なぜ統計手法を用いるのかを理解する
- 記述統計と推測統計の違いを理解する
- 記述統計（平均値，標準偏差，相関）を理解する
- 各種テストの分析方法を理解する

引用文献

- 斯日貢・中平勝子・福村好美・湯川高志（2016）「モンゴル人学習者を対象とした日本語格助詞学習法ー文法対比型学習法の提案ー」『コンピュータ&エデュケーション』41，40-45
- 国際交流基金・日本国際教育支援協会（2011）『平成21年度日本語能力試験（第1回・第2回）分析評価に関する報告書』
- 島田めぐみ・三枝令子・野口裕之（2016）「日本語Can-do-statementsを利用した言語行動記述の試み」『世界の日本語教育』16，75-88
- 島田めぐみ・野口裕之（2017）『日本語教育のためのはじめての統計分析』ひつじ書房

講演の目標

- なぜ統計手法を用いるのかを理解する
- 記述統計と推測統計の違いを理解する
- 記述統計（平均値，標準偏差，相関）を理解する
- 各種テストの分析方法を理解する

ありがとうございました

shimada.megumi@nihon-u.ac.jp

◆配布資料の補足

・統計についての講演依頼をいただいたのは、日本国外ではモンゴルが初めてだったので、驚きつつも、モンゴル日本語教師会の熱心な様子が推察できました。

・数字に苦手意識を持つ方が多いですが、テストをすれば必ず統計が必要になるので、統計は避けては通ることができません。

目標（1）なぜ統計手法を用いるのか理解する

・日本語教育学会発行『日本語教育』の中の論文の傾向を現在と 10 年前とで比べてみると、専門的な統計を使用した論文の割合が増えています（0.320→0.412）。40%超という数字は少なくなく、統計手法に関する知識がないと論文が読みにくいことになります。統計手法は使うだけではなく、知っておくといいものです。

・記述統計：得られたデータをまとめ、要約して、全体的な特徴を見るものです。たとえば、みなさんが普段もしている平均の計算（平均値を知ること）も、この 1 つです。

・推測統計：記述統計より少し高度ですが、母集団について推定するものです。たとえば、モンゴル人日本語学習者の特徴を知りたいとき、モンゴルの全ての学習者（1万人以上）を調査することはできません。そこで調査対象を 50 人とし、結果を出しますが、その結果のみから「これはモンゴル人学習者全てについての特徴である」と言うことはできません。偶然やサンプルが偏っている可能性があるからです。そのようなとき、推測統計を用いることで、50 人から母集団（1 万人以上）について推定することができます。

目標（2）記述統計と推測統計の違いを理解する

・推測統計：赤と緑、それぞれ 20 個ずつのりんごが入った箱があり、「赤も緑も大きさは同じ」と言われていたとします。箱から赤と緑、1つずつりんごを取り出したところ、赤のほうが大きかったです。でも、1 つだけなら、「偶然かな」と考えます。では、5 つずつ取って、5 つとも赤のほうが大きかったら、どうでしょうか。言われていた前提（仮説）が間違っていたのでは、と思いませんか。このように、サンプル（5 つのりんご）の特徴を見て母集団（20 個のりんご）を推定するのが、推測統計です。

・記述統計：全てのりんごについて量ったら、推測する必要はありません。そのときは記述統計を使用します。たとえば、自分のクラスの学習者全員にテストをし、全員のデータが得られるような場合です。

・推測統計を用いた研究例：モンゴル人を扱ったものはあまりありませんでしたが、1 つ見つけました。調査対象は 15 人ですが、ここから母集団（内モンゴル人学習者）を推測しています。ただ、この論文には間違いがあります。1つ目は、標準偏差の報告がないことです。平均値を出したら必

ず標準偏差も報告しなければなりません。2 つ目は、比率（%）は *t* 検定してはいけないことです。このように、統計を使った論文の中には間違いが見られます。このような間違いをしないように、しっかり知識を持っておく必要があります。

目標（3）記述統計（平均値、標準偏差、相関）を理解する

＜記述統計によってわかるデータの特徴①　度数の分布（度数分布表、ヒストグラム）＞

・たとえば、学習期間についてアンケートをした場合、結果を羅列しても、どの答えが多いかわかりません。これを度数分布表にまとめれば、傾向がわかりやすくなります。これは皆さんもしたことがある基本的なことですが、これも記述統計の 1 つです。

・度数分布表をヒストグラムにすると、さらにわかりやすくなります。最近の Excel ではヒストグラムを作れることが多いです。

・ヒストグラムと棒グラフの違いは次の通りです。

ヒストグラム①横軸の項目を並べ替えてはいけない
②隣り合ったバーを離してはいけない（連続したデータを扱っているため）

棒グラフ　①横軸の項目を並べ替えてもいい
②隣り合ったバーを離さなければならない（連続したデータではないため）

＜記述統計によってわかるデータの特徴②　代表値（平均値、中央値）＞

・平均値とは全てのデータの平均で、中央値とは全体の真ん中の値のことです。

・クラスの中には、少し特殊な事情がある場合があります。たとえば、日本語母語話者や両親のどちらかが日本人の学習者がいる場合、一般の学習者と比べたらテスト結果は異なり、高得点を取るでしょう。反対に、日本であればビザの関係で来日が遅れた学習者や、遅刻、体調不良などでテストがほとんどできなかった学習者がいる場合もあります。そこでクラスのデータを考えるとき、その特殊な学習者や、本当の実力とは違う結果になった学習者を外して考えたいですが、平均値では外せません。一方、中央値であればあまり影響はありません。このように、通常は平均値を使いますが、場合によっては中央値を使うこともあります。

＜記述統計によってわかるデータの特徴③　分散（標準偏差、分散）＞

・A さんは、平均点がどちらも 50 点だった聴解テスト・読解テストで、どちらも 70 点を取りました。しかし PPT のように、クラス内での順位は違い、聴解は 3 位、読解は 1 位でした。これはク

ラス内での点数の散らばりが違うからで、ここから、平均値だけではその人の順位やテストの特徴を表せないことがわかります。その散らばりを表すのが、標準偏差と分散です。

・標準偏差：各データと平均値との距離の平均（各データが平均値からどのくらい離れているかを計算し、その平均を求めたもの）です。標準偏差の算出は Excel で自動的にできますが、計算の仕組みを理解しておけば意味への理解が深まります。

・分散：標準偏差を算出する途中で出た値で、高度な統計手法を用いるとき使われることがあります。

・レポートや報告書、論文などで平均値を計算することは多いですが、その際は必ず標準偏差も計算し報告してください。

＜記述統計によって 2 変量のデータの関係を見る（相関）＞

・たとえば、作文や会話テストを二人の先生が採点したとき、同じ得点にはなかなかなりません。そこでどのくらい一致しているかを表すために、相関という統計手法を用います。

・漢字テストと読解テストの点数には関連があると思いますか。そのとき、散布図を作成します。両テストの点数が交わる部分に点を打っていった結果、右肩上がりの形になりました。また、これを数値で表すと 0.791 になりました。これは両方のテストの関連性が高いことを表します。

＜相関をみるときの注意＞

①集団の範囲の影響

：たとえば 1 つのクラスの中で自己評価と日本語テストの関係を調査した結果、相関係数は 0.338 となり、相関はあまり高くありませんでした。しかし、複数のクラスで実施してみた結果、先ほどと同じデータを一部に含んでいるのにも関わらず、相関係数は 0.779 になりました。このように、1つのクラスのように集団の範囲が狭いと相関は低くても、学年や日本語レベルなど集団を広げてみると、実は相関が高くなることも多いです。

②散布図を描いてわかること

：学習期間と自己評価得点の相関を計算したところ、相関係数は 0.306 でした。しかし、これだけで両者の相関は低いと判断しないでください。散布図を描いてみると、学習期間が短い 20 か月以下では右肩上がりの分布になっており、この期間は関連性があることがわかります。このように、散布図で確認すると、相関係数だけでは得られない情報が得られるかもしれません。

③相関関係でわかること

：漢字テストと一週間の読書量の相関が 0.82 だったとき、ここからわかることは以下の 2 点です。

・「漢字テストの得点が高い人ほど読書量が多い」

・「読書量が多い人ほど漢字テストの得点が高い」

しかし、「読書量が増えると（原因）、漢字テストの得点が高くなる（結果）」とは言えないので、注意してください。相関係数からは関連性は言えますが、原因・結果はわかりません。「漢字テストの得点が高くなると（原因）、読書量が増える（結果）」という逆の関係かもしれません。論文でもこのような間違いは時々見られます。

目標（4）各種テストの分析方法を理解する

テスト作成・採点が大変なので、さらにその結果を分析・報告する方は少ないと思います。しかし、分析によっていろいろなことがわかります。

＜テストの種類＞

・客観テスト：だれが採点しても同じ結果になる、答えがはっきりしているものです。

・主観テスト：会話や作文のテストで、採点者の主観が入るため、
ほかの採点者と結果が 100%一致することはありません。

＜客観テストの分析：旧日本語能力試験を例に＞

分析①総合点（合計点）について全体的傾向を計算する

・平均値と標準偏差を計算します。ヒストグラムも作れたらいいですが、人数が少ない場合はうまく作れません。

・日本語能力試験の旧試験では、テスト問題や結果についての様々な細かいデータが分析され、報告書が出されていました。

・2009 年度第 1 回 1 級の聴解・読解のヒストグラムを比べると、読解のほうが得点が高く、山のピークも右側にあります。これは 1 級の受験者は中国人が多いためでしょう。モンゴル人のヒストグラムを作ったならば、聴解のほうが得意ということなので、別の形になるでしょう。

・同回の標準偏差を見ると、文字語彙：14.41、聴解：18.31 で、ここから聴解はできる人・できない人のばらつきが大きいことがわかります。（読解は 200 点満点なので、単純には比べられません。）

分析②各項目の分析をする

・問題を作成する場合、1 つ 1 つがいい問題でなければなりません。自分ではいい問題だと思って作っても、答えが 2 つあったり、答えが選べなかったりすることは時々あります。そのため各問題の分析をすることは非常に重要です。

・正答率（困難度）：難しさの度合いです。
・識別力：受験者の能力を区別する力です。
・「得点段階別選択率」の図の見方
：縦軸＝選択率
横軸＝全受験者を得点別に 10 段階に分けたもの
（得点が上位約 10%に入る人のグループが得点段階 10 になる）
正答は右肩上がりになり、逆に誤答は右肩下がりになる（＝能力が高い人ほど正答率が高く、能力が低い人は選べない）問題は、識別力が高い、いい問題であると言えます。
一方、能力が高い人と低い人で正答率に差がない問題は、識別力が低く、これによって能力をはかることはできないため、出題する意味はありません。
・識別力が低かった場合、必ず問題を見直し、場合によっては分析対象から外したほうがいいです。権威がある日本語能力試験でさえ時折このような問題が出題されるので、私たちが問題のある問題項目を作ってしまうことがあるのは仕方がないことです。その時は問題を変えればいいです。

＜項目応答理論（IRT）：新日本語能力試験を例に＞

「古典的テスト理論」

：個人のクラスや旧日本語能力試験で使われている理論です。「古典」と言いますが過去のものという意味ではなく、私も使っています。

①たとえば、「がっこう」を漢字で書け、というテストをしたとします。中級クラスでやれば正答率は高いでしょうが、初級クラスでは正答率が 0.5 でした。この問題は易しいでしょうか、難しいでしょうか。それはクラス（受験者）に影響されます。このように、古典的テスト理論では、難しさは絶対的に決められません。

②いくつ正答したか、によって合計得点が決まります。よって、難しい問題が多いテストでは得点が低くなり、易しい問題が多いテストでは得点が上がる、ということがおきます。そうすると、この受験者の能力はよくわからず、また、違う回のテストの結果を比較できません。

③たとえばテストの中に難しさが違う問題が 5 問あったとします。受験者 A は易しめの問題を、受験者 B はより難しい問題を間違えたとしても、いくつ正答したかによって得点が決まるので、得点は AB どちらも 4 点です。

「IRT」

：データ数が多く必要なので個人のクラスで使うのは無理ですが、新日本語能力試験、日本留学試験、TOEFL、TOEIC など大きい試験に使われている理論として知っておくといいです。

①難しさは受験者の能力に影響されず、絶対的です。これは視力検査のマークのようだと言われます。このマークの難しさは相手によって変わることはなく、誰にとっても共通です。

②テスト問題は、視力検査のように、一つ一つの問題の難しさはもう決まっており、易しい問題から難しい問題まで並んで一本の尺度ができています。受験者はどの問題までできたか・できなかったかによって、その尺度の中に位置づけられ、能力がはかれます。このように、問題項目の難しさがわかっており、どの回のテストも同じ基準となるので、違う回のテストの結果を比較できます。

③たとえばテストの中に難しさが違う問題が 5 問あり、受験者 A は易しめの問題を、受験者 B はより難しい問題を間違えたとします。このとき、A は難しい問題ができているので、B より能力が高い可能性があり、それを考慮して得点が決まるので、AB の得点は変わります。よって、新試験を受けた受験者から、「絶対 1 問間違えたはずなのに、満点だった」という話を聞くことがありますが、これは正答した問題数で得点を計算しているのではなく、このように IRT で計算しているからです。

④CAT（コンピュータ適応型テスト）は全て IRT によって計算されて問題が出題されています。

＜教育機関でテスト結果を分析する場合：主観テストの分析＞

・評定者間信頼性：複数の評定者がいる場合、それぞれの評定が一致しているか調べます。

・評定者内信頼性：同じ人物が別の機会に評定した場合、それぞれの評定が一致しているかを調べるものです。これは一致度が高い場合が多いです。

日本語教育と研究の繋がり
-統計分析を例に-

島田めぐみ先生のご講演の様子