

## 第1章 測定と評価 pp1-14

### § 1-1 統計分析を行う前に

#### 1-1-1 測定と評価のための統計分析

○**仮説検証型** (hypothesis testing) の研究

仮説を立てる→その仮説が正しいかどうかを検証

例) 社交性とスピーキング力の関連性

**仮説**

「社交的な学生ほど海外留学による英語のスピーキング力の伸びが大きい」

**検証**

- ・留学前に「社交性」に関する質問紙を用いて各学生の社交性の度合いを調査
- ・留学前後に英語のスピーキングテストを実施し、伸びを測定

→測定した社交性とスピーキング力の関係を分析し、仮説が成立するかを**統計分析**を用いて検証

社交性、スピーキング力：目に見えない人間の特性や能力

→正しく**評価** (evaluate) するために質問紙やテストなどの**尺度** (scale) を使わなければならない

#### 1-1-2 測定の二大要素：妥当性と信頼性

**妥当性**：その尺度が測定すべきものを測定しているかの概念

**信頼性**：測定が安定していて正確であるかの概念

測定においては妥当性 (validity) のある尺度を使って、信頼性 (reliability) の高いデータを得ることが大切

例) 英作文の能力を測定するテスト

妥当性を高めるために：

- ・実際に英語を書かせる
- ・明確な採点基準を設ける

→信頼性の向上

○ダーツのアナロジー

的の中心を射ているか=妥当性

誤差、ぶれがないか=信頼性

### § 1-2 妥当性の捉え方と検証法

#### 1-2-1 伝統的観点から見た妥当性の分類

妥当性の種類	定義	検証法
--------	----	-----

内容的妥当性 (content validity)	項目が、どの程度全体を偏りなく代表しているか。	その分野の専門家に判断してもらう。
基準関連妥当性 (criterion-related validity)		
1. 併存的妥当性 (concurrent validity)	外部の基準とどの程度関連があるか。	同時期に得た基準との相関を検証する。
2. 予測的妥当性 (predictive validity)	その尺度の実施以降の変化などをどの程度適切に予測しているか。	尺度の得点とその後の基準との相関を検証する。
構成概念妥当性 (construct validity)	測定しようとする構成概念をどの程度適切に反映しているか。尺度の得点とモデルや理論との整合性がどの程度あるか。	信頼できる基準との相関や因子分析、MTMM（収束的・弁別的妥当性）などによって検証する。

構成概念妥当性 = 基準関連妥当性 + 内容的妥当性  
 妥当性の検証 = 構成概念妥当性の検証



単一概念

### 1-2-2 新しい妥当性の定義 : Messick の分類

妥当性…テスト得点を用いたある特定の推論が、適切であるか、意味があるか、有用であることを示すこと。

テストの妥当化…その推論を裏付けるための推論を累積するプロセス

#### 構成概念妥当性の 6 つの側面

妥当性の側面	妥当性の証拠	具体的な情報の収集法
内容的側面 (content aspect)	項目の内容が設定したドメインの内容に対応しているか、十分にドメインを代表しているかを示す証拠。	カリキュラム分析、タスク分析、談話分析など。 専門家や尺度のユーザーによる判断。
本質的側面 (substantive aspect)	項目やタスクなどの反応(回答)プロセスが心理学的に理論的根拠があるという証拠。	タスク中の発話プロトコルやタスク後の再生。 目の動きや反応時間などの観察。 質問紙・インタビュー。
構造的側面 (structural aspect)	尺度内の項目間の関係が理論的構造に一致していることを示す証拠。尺度の次元性や応答パターンが心理測定モデルに合うという証拠。	得点化手続きの適切性の診断。 相関分析。 因子分析や構造方程式モデリング。 項目応答理論。
一般化可能性の側面	当該データの理論的特性(平均や標準	一般化可能性理論。

(generalizability aspect)	偏差、項目間の相関構造など)が一貫しており、且つ他の集団、実施場面、実施時期、同様の項目セットに対しても不変であるという証拠。	再検査信頼性 (再テスト信頼性)。代替検査信頼性 (同等フォーム)。アルファ係数。
外的側面 (external aspect)	当該尺度と他の尺度に理論上想定される相関パターンが示される問う証拠。	相関分析や MTMM。構造方程式モデリング。
結果的側面 (consequential aspect)	その尺度を使用することの適切さ。短期的、長期的な悪影響が理論的、経験的に生じない、予期されないという証拠。	観察、質問紙、インタビューによる波及効果の分析。事前、事後テストの分析。

### 1-2-3 多特性・多方法行列分析法(MTMM)

#### MTMM (multi-trait multi-method analysis)

収束的妥当性(convergent validity)と弁別的妥当性(discriminant/divergent validity)を使って構成概念妥当性を検証する方法。

#### 収束的妥当性(convergent validity)

同じ特性を異なる方法で測定した尺度間の相関 (同一特性異方法相関 : monotrait-heteromethod correction, MH) で表す。

#### 弁別的妥当性(discriminant/divergent validity)

異なる特性を同じ方法で測定したテスト間の相関 (異特性同一方法相関 : heterotrait-monomethod correlation, HM) と、異なる特性を異なる方法で測定した尺度間の相関 (異特性異方法相関 : heterotrait-heteromethod correlation, HH) の 2 つによって表す。

### 1-2-4 波及効果

波及効果(washback/ backwash effect)・・・結果的側面の中でテストなどが学習に与える影響。

インパクト(impact)・・・テストが教育組織や社会に与える影響。

## § 1-3 テストの信頼性

### 1-3-1 信頼性の定義

信頼性 : 尺度が測定している構成概念をどの程度高い精度で測定しているかという、得点の安定性あるいは一貫性

## ○信頼性係数

古典的テスト理論(classical test theory)

＜観測値＝真値＋誤差＞

**真値**＝対象とする能力を示す真の得点

**誤差**＝真の得点からのズレ

\*仮定\*

①誤差はランダムに生ずる。つまり、真値との相関はゼロである

②誤差はプラス、マイナスどちらの値をとることもありその平均値はゼロである。

③誤差どうしの相関はゼロである。

①の場合、

＜観測値の分散＝真値の分散＋誤差分散＞

分散 (variance) : それぞれのデータの、平均値からの距離を2乗した値

＜信頼性係数  $\rho$  (ロー)＝真値の分散/観測値の分散＝真値の分散/真値の分散＋誤差の分散＞

→観測値の誤差分散が大きいほど信頼性が低くなる

→信頼性係数は0から1の間の値をとる (信頼性係数0.80以上…信頼性が高いデータ)

### 1-3-2 信頼性係数の推定方法

(1)再テスト法 (2)同等フォーム法 (3)内的一貫性 (4)評価者信頼性

(1)再テスト法

同内容のテストや質問紙を同一被験者に二回実施し、得点の相関係数を求める

—特性がすぐに変わらない心理検査などで実施可能

(2)同等フォーム法 (平行フォーム法)

同一被験者に同じ形式の二つのテストを実施し、得点の相関で信頼性を推定

—同等のテスト作成が難しく、実用的ではない

(3)内的一貫性 (内的整合性)

同じ構成概念を測定する尺度内で受験者の項目得点の一貫性をみる

①折半法

テスト項目を半分に分け、それぞれの合計点の相関を示したのち、全体の信頼性係数をスピアマン・ブラウン公式 (Spearman-Brown prophecy formula) を用いて算出

—高い信頼性係数を算出するにあたっての項目の増減を検討可

—問題点：テスト項目の分け方によって結果が変動する

②アルファ係数 ( $\alpha$ 係数：クロンバック・アルファ)

—折半法の問題点を克服、すべての折半方法によって推定した信頼性の平均値を統計的に算出  
—最もよく用いられる

③キューダー・リチャードソン20 (KR20)

—2値データをアルファ係数で求めた場合

④キューダー・リチャードソン21 (KR21)

—内的一貫法の中でも信頼性係数が低く推定され、過小評価される傾向

(4) 評価者信頼性

被験者のパフォーマンス評価において評価側の判断への主観性の介入を考慮 (評価者の採点の一致度)

・複数人で被験者を評価した場合

→評価者間で採点が一貫していたかという評価者間信頼性 (inter-rater reliability)

・複数の評価を一人で行った場合

→同一評価者が一貫して評価したかどうかという評価者内信頼性 (intra-rater reliability)

…2名の評価者の採点の一致度を求める場合は相関分析 (7章) やカッパ係数 (7章) を、3名以上の場合はアルファ係数を用いて算出

その他にも一般化可能性理論 (高信頼性を得るために必要な評価者数や項目数を検討可能) や、項目応答理論 (それぞれの被験者の能力及び評価者の厳しさや一貫性を同時に分析可能) を用いる方法もある。