構成概念の測定

構成概念 (Construct)

- 性格，感情，気持ち，知能，能力，学力などの特性は，頭の中で考えた（心理的）構成概念
- 構成概念を用いると，何らかの現象を上手に説明することができる
- 構成概念は抽象的なものであり，物理的な存在ではない
- 物理的に存在しないので，物差しや秤などで直接測定することができない
- しかし，私達は「あの人は周囲からの人望が厚い」などと言っている

構成概念

- 我々は，物理的に存在しない構成概念についても，その程度を何らかの方法で測定している
- 構成概念の程度が，思考，行動，知識，成果物などに表れていると，暗黙に仮定している
- 思考，行動，知識，成果物などをみることにより，構成概念の程度を間接的に測定している
- 間接測定しているもの（測っているもの）が，測りたいものと一致しているか確認が必要

構成概念の測定の難しさ

- 間接測定なので誤差が大きい
- 観測値がゼロでもその特性が「無い」とは限らない
- 値がいくつならどの程度というスケールがない
- 構成概念は頭の中で考えた抽象的なものであり，人によって定義や捉え方が異なる

- 測りたい構成概念を，確かに測っていると多くの人が納得できる証拠を提示することが必要

  尺度・テストの評価：測定の妥当性・信頼性

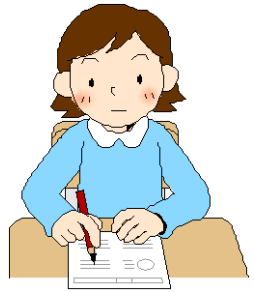
  項目・問題の評価　：項目分析

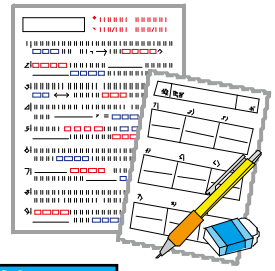
測定の妥当性・信頼性

構成概念の測定のイメージ

どの程度，精度良く測れているか？

信頼性





質問紙

項目１

項目２

・

・

項目X

測りたいもの

＝

？

測っているもの

間接測定

回答

どの程度，測りたいものを測れているか？

妥当性

測定の妥当性・信頼性

- 妥当性 (Validity)

  測定したい構成概念を，どの程度適切に測定しているか

  測定したい特性を捉えていると言える証拠の強さで評価

- 信頼性 (Reliability)

  測定している構成概念を，どの程度精度良く（小さい誤差で）測定しているか

  各回答者において回答が一貫している程度で評価

妥当性の確認

- いろいろな複数の結果を証拠として提示
  専門家のお墨付き，他の尺度との相関係数　など
- 尺度自体が「妥当性」という性質を持つのではない
- その尺度を用いた測定によって導かれる解釈や結果の運用の適切性が妥当性
- よくできたアメリカの社会科のテストでは，日本で社会科のテストとして妥当な測定はできない

内容的妥当性 (Content Validity)

- 尺度の内容や見た感じのそれらしさ

- 論理的妥当性 (Logical Validity)

　専門家の目からみて，それを使って適切な測定ができると言えるか

- 表面的妥当性 (Face Validity)

　一般 (回答者) の目からみて，それを使って適切な測定ができると言えるか

　もっともらしい物と見えなければ，きちんと回答してくれない

基準関連妥当性(Criterion-Referenced Validity)

- 客観的な外的基準との関連の強さ

- 併存的妥当性 (Concurrent Validity)

　尺度を使った測定と同時に，外的基準となる変数の測定が行われ，測定値と基準値との関連 (相関) を検討

- 予測的妥当性 (Predictive Validity)

　外的基準となる変数の値が後日得られ，測定値と基準値とのの関連 (相関) を検討

- 外的基準の妥当性が問題になることもある。妥当性の堂々巡り

構成概念妥当性 (Construct Validity)

- 測りたいものを測定していると解釈できる証拠全般

- 収束的妥当性 (Convergent Validity)
  関連のありそうなものとは相関がある

- 弁別的的妥当性 (Discriminant Validity)
  関連のなさそうなものとは相関がない

- 内容的妥当性も基準関連妥当性も，大きくは構成概念妥当性の枠組みで捉えることが可能

統計分析力尺度の例

- この尺度を用いた測定の妥当

	n	M	SD	統計分析力との相関係数
統計能力テスト	365	64.83	12.47	.51
数学	365	54.17	16.07	.33
批判的思考力	365	30.11	6.14	.39
国語(現代文)	365	65.09	9.28	.08
自己効力感	365	51.99	9.76	.13

- 統計能力，数学，批判的思考力とは弱～中程度の相関
- 現代文や自己効力感とはほぼ相関なし
- 理にかなっている
- 妥当性が確保されていると判断する

医歯薬大学式　統計分析力尺度

開発責任者　　医歯薬大学教授　成上 崇命

以下の各項目は統計分析力に関する項目です．各項目の内容を読んで，あてはまる程度を，1～5の数値から選んで○をつけて下さい．

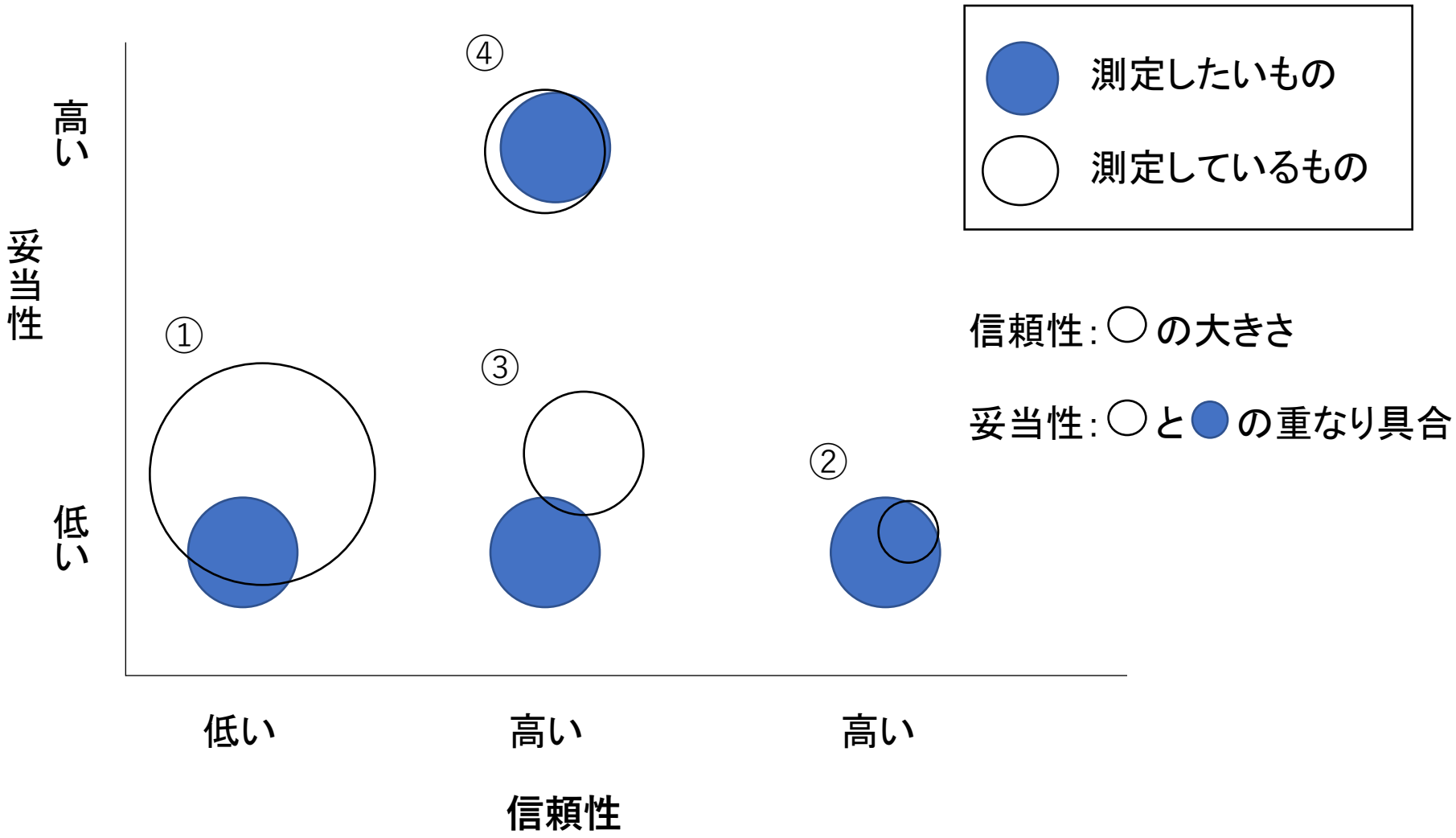
	まったくあてはまらない	あまりあてはまらない	どちらともいえない	まああてはまる	よくあてはまる
1) 分析結果全体を説明できるような解釈を導くことができる	1	2	3	4	5
2) 分析結果を見ていて思わぬ発見をすることがある	1	2	3	4	5
3) 分析結果でよく分からない出力があったら，何を意味するものか調べる	1	2	3	4	5
4) 結果から自分の仮説が支持されなかったとき，仮説は誤りだったと素直に認めることができる	1	2	3	4	5
5) 研究目的を見失わずに分析を進めることができる	1	2	3	4	5
6) 1つの分析法がうまく適用できないとき，他の方法を考えることができる	1	2	3	4	5
7) いろいろな分析法を用いたことがある	1	2	3	4	5
8) いろいろな分析法を知っている	1	2	3	4	5
9) パソコンの扱いは得意なほうだ	1	2	3	4	5
10) 数値で考えるのは得意なほうだ	1	2	3	4	5

測定の信頼性

- 尺度が測定している構成概念を，どの程度精度良く測定しているか
- 測定誤差が小さい（少ない）ほど信頼性の高い測定と言える
- 誤差が小さければ各回答者において測定値（回答）は一貫するはず
- 各回答者において回答が一貫性している程度を表す信頼性係数を算出し，信頼性の高低を評価
- 信頼性において，何を測っているかは問題ではない

  それは妥当性の話

信頼性と妥当性の関係


④
①
③
②
高い
妥当性
低い
低い
高い
高い
信頼性
測定したいもの
測定しているもの
信頼性：○の大きさ
妥当性：○と●の重なり具合

信頼性と妥当性の関係

① 信頼性が低い ⇒ 妥当性は低い

信頼性が低ければ，測定誤差が大きくなるので，測りたいものを適切に測ることはできない

② 信頼性が高すぎる ⇒ 妥当性は低い

信頼性が高すぎれば，測定している概念はごく狭いものになるので，測りたいものを適切に測ることはできない

③ 信頼性は適度に高いが的が外れている ⇒ 妥当性は低い

信頼性が高ければ何らかの構成概念を適切に測っているが，それが測りたいものかは別の話

信頼性と妥当性の関係

④　信頼性が適度に高く，かつ，的を射ている ⇒ 妥当性は高い

　信頼性が高ければ何らかの構成概念を適切に測っており，それが測りたいものであって初めて，妥当性の高い測定となる

- 信頼性は妥当性の必要条件（一部）

  信頼性も，測りたいものを測っていると言える証拠の1つ

- いわゆる「短縮版尺度」には注意が必要

  項目数が少なくて信頼性が高い ＝ 狭いことしか捉えていない

  多くの場合，5〜10項目を使って1つの心理的構成概念を測定

古典的テスト理論・信頼性係数

顕在変数・潜在変数

- 顕在変数 (観測変数)

  Manifest Variable, Observed or Observable Variable

  実際に観測可能な変数

  質問紙への回答，尺度得点，テスト得点　など

- 潜在変数

  Latent Variable

  実際に観測はできないが，現象をうまく説明するものとして用いられる構成概念もしくは変数

  性格特性，情動，学力，

  因子得点，真の得点，能力値　など

古典的テスト理論(Classical Test Theory : CTT)

- モデル

  観測得点 X は，真の得点 T と，誤差得点 E の和で構成される

  $$X = T + E$$

- 得点

  観測得点 X： ある1回の測定で得られる得点 (顕在変数)

  真の得点 T： 同じ測定を繰り返し行ったときの得点の期待値
  (潜在変数)

  誤差得点E： あて推量，記入ミスなど，非系統的に得点に影響するもの
  (潜在変数)

古典的テスト理論(Classical Test Theory : CTT)

- 仮定

  誤差Eの平均は 0 とする

  　系統誤差は真の得点に含まれる

  誤差は互いに無相関である

  　系統誤差は真の得点に含まれる

  誤差Eと真の得点 T は無相関 である

  　真の得点が高いほど，あて推量がよくあたる訳ではない

- 注意点！

  無相関であればよく，関連がない (独立) とまでは言っていない

古典的テスト理論(Classical Test Theory : CTT)

- 帰結

観測得点の平均 ＝ 真の得点の平均

$$\bar{X} = \bar{T} \qquad (\bar{X} = \bar{T} + \bar{E} = \bar{T} + 0)$$

潜在変数である真の得点は，個々の値は分からないが，その平均は，顕在変数である観測得点の平均に等しい

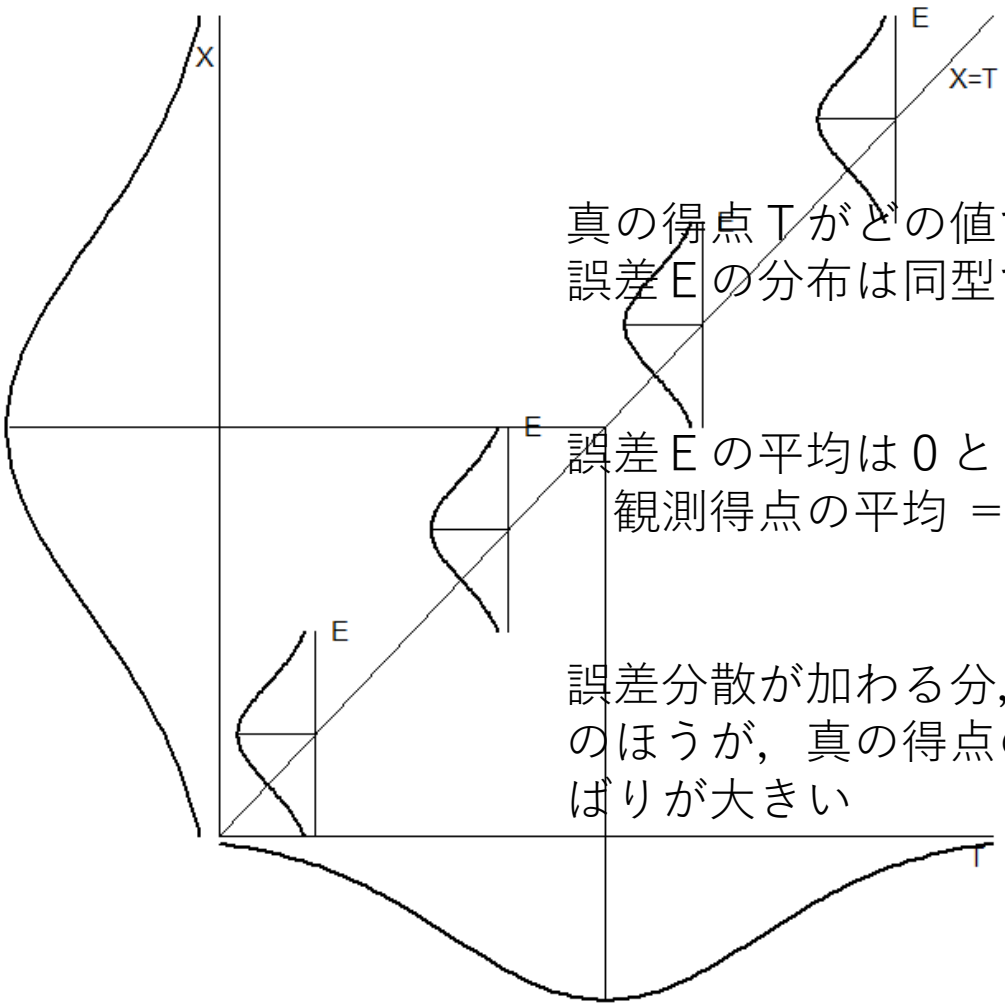
観測得点 X の分散 ＝ 真の得点 T の分散 ＋ 誤差 E の分散

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$$

観測得点の散らばりは，真の得点の散らばりに，誤差の散らばりが加わったもの

古典的テスト理論の概念図

観測得点 X の分布



真の得点 T がどの値であっても，誤差 E の分布は同型で，分散は σ_E^2

誤差 E の平均は 0 としているので，観測得点の平均 ＝ 真の得点の平均

誤差分散が加わる分，観測得点の分布のほうが，真の得点の分布より，散らばりが大きい

真の得点 T の分布

信頼性係数 (Reliability Coefficient)

- 信頼性は各回答者において回答が一貫している程度

- 誤差 E が小さいほど，回答は一貫するはず
- 誤差 E が小さいほど，誤差分散は小さくなる

- 信頼性係数の定義

  観測得点の分散に占める真の得点の分散の割合

$$\rho_X^2 = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}$$

- 信頼性係数は 0～1の値を取り，1に近いほど精度高い

信頼性係数の推定

- 観測得点 X（データ）は入手可能
- 真の得点 T や誤差 E は構成概念であり，実際にはこれらの変数のデータは無い（潜在変数）
- 真の得点の分散の値は分からない
- 定義式に従って信頼性係数を求めることはできない
- 何らかの方法で信頼性係数の値を推定する

信頼性係数の推定

- 一貫性をどのように捉えるかにより，いくつかの推定法がある
- 再現性：同じテストを2回やったら，だいたい同じ結果になる
- 安定性：同じ特性を測るテストを2つやったら，だいたい同じ結果になる
- 内的整合性：同じ特性を測る各項目に対しては，同じような回答をする

再検査信頼性係数(Test-Retest Reliability Coefficient)

- 同一集団に対し，同じ尺度（テスト）を，時間間隔をあけて2回実施したときの，1回目のデータと2回目のデータの相関係数 (ex. r= .82)

- 時間間隔は，記憶効果がなくなり，かつ当該特性が変化しないと考えられる期間で，２週間～1ヵ月程度とするのが一般的

- 特性がすぐに変化してしまう尺度やテストの信頼性の評価には適さない

- 医療系では，2回の測定値の級内相関係数（ICC(2,1)）を求めることが推奨されている

	n	M	SD	相関係数		α
1回目	365	30.68	6.08	1	.82	.84
2回目	156	29.79	5.90	.82	1	.83

平行検査信頼性係数

- 同一集団に対し，同じような尺度（テスト）を2つ実施したときの，1つめのデータと2つめのデータの相関係数

- 記憶効果が影響せず，時間をあけなくても良いのはいいが，同じような検査を作るのが大変

- 簡単な計算問題や漢字テストなどには使えるかもしれないが，実際にはあまり実用的な方法ではない

α係数（内的整合性信頼性係数）

- ある集団に対し，当該尺度を1回実施したときの，各項目のデータと，合計点のデータから算出（ex. α=0.84, α=0.83）

$$\alpha = \frac{p}{p-1}\left[1 - \frac{s_1^2 + s_2^2 + \cdots + s_p^2}{s_X^2}\right]$$

p

s_i^2：項目得点の分散

s_x^2：テスト得点の分散

	n	M	SD	相関係数		α
1回目	365	30.68	6.08	1	.82	.84
2回目	156	29.79	5.90	.82	1	.83

- クロンバックのα（Cronbach's alpha）と言われることも多い
- １回の測定で推定できるので，単に信頼性係数を確認する場合や，特性がすぐに変化してしまう尺度やテストの信頼性の評価に適している

経験的な信頼性係数の大きさ

- 信頼性が高いと認められる範囲

  0.95 以上： 英語，数学

  0.90 以上： 物理，化学

  0.80 以上： 生物，地学，国語，社会

  0.70 以上： 性格検査

- 英語や数学の試験でも測定誤差は５％くらいある

  偏差値±2ポイント程度の変動は誤差の範囲

- 性格検査の点数は30％程度の測定誤差を含んでいる

経験的な信頼性係数の大きさ

- 信頼性係数が0.50 以下となる尺度は，真の得点より誤差の分散のほうが大きいので，使用すべきでない
- 多くの場合，信頼性係数が0.6以下となる尺度は，信頼性が低いと見なされる
- 信頼性係数が0.6～0.7だと悩ましいところ（よくある）
- 物理的なものに比べ，試験や性格検査の測定誤差は，非常に大きい
- （１回の）テストや性格検査の結果を過信してはいけない

信頼性に関する議論

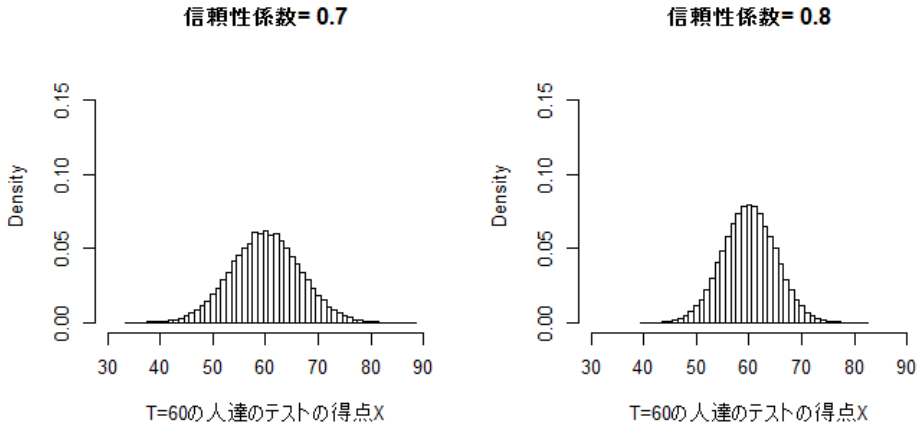
テスト得点と信頼性の関係

テスト得点と真の得点の関係のシミュレーション

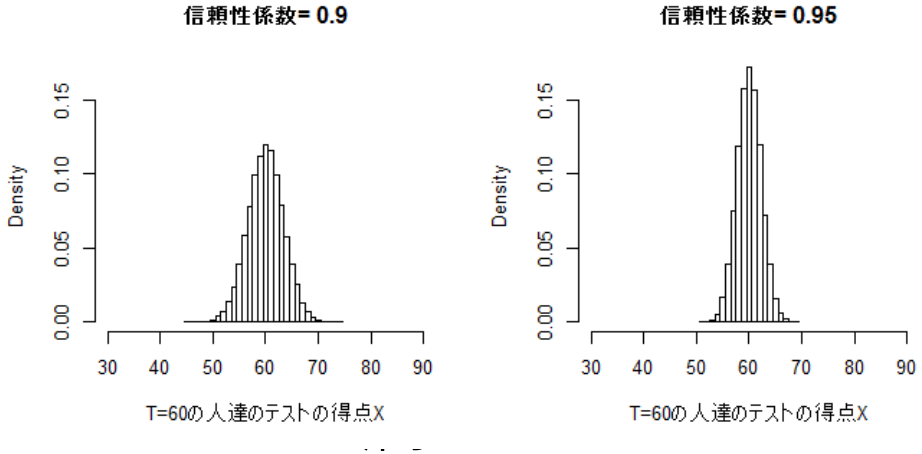
- 100点満点のテスト
- 真の得点の平均点を60点，標準偏差を10点と設定
- 信頼性係数として 0.7, 0.8, 0.9, 0.95 の4通りの値を設定
- 真の得点 T と 誤差 E を正規乱数を用いて発生
- テスト得点 X が同じ値の受検者の真の得点 T の分布を観察

真の得点が60点の受検者のテスト得点の分布

- 真の得点Tが同じでも，実際のテスト得点Xはか

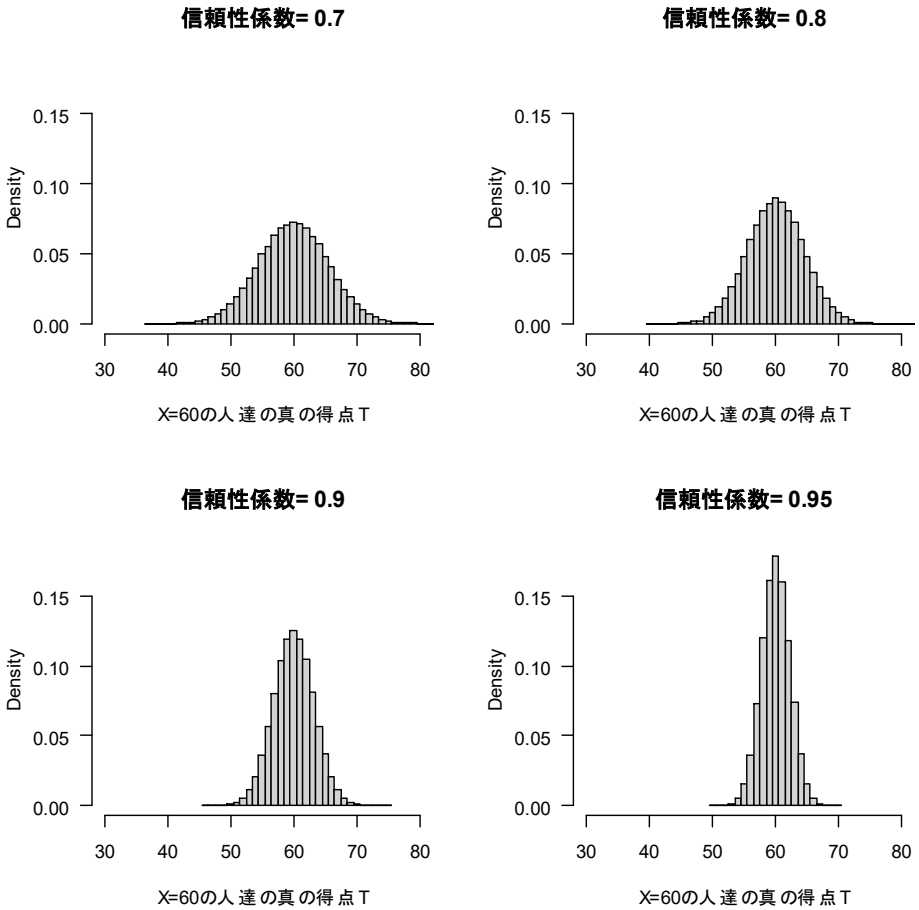


ご，そのばらつきは大き



ようなテストでは，真の
得られる確率は ４分の

テスト得点が60点の受検者の真の得点の分布



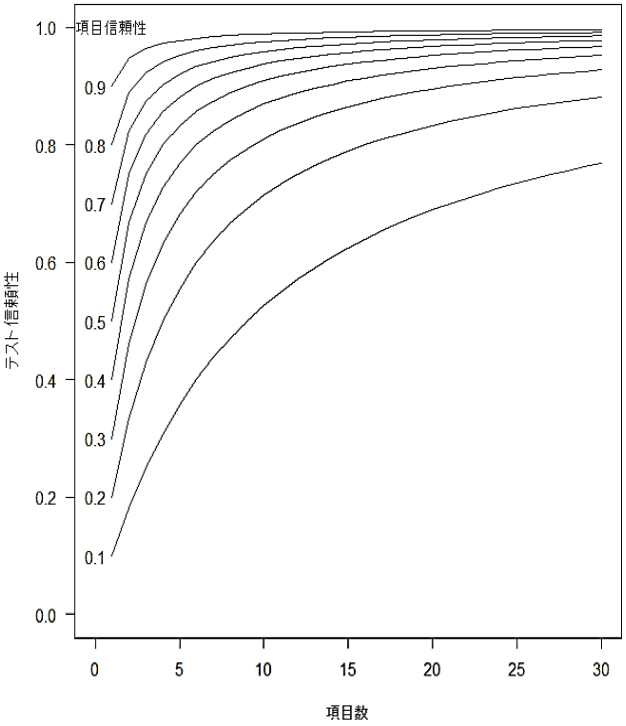
- 得点が60点の受検者真の得点はかなり広く分布している
- 信頼性係数が0.7 (誤差30%) の試験では，真の得点は40～80 点 (±20) くらいに分布する
- 信頼性係数が0.9 (誤差10%) の試験でも，真の得点は50～70 点 (±10) くらいに分布する

テスト得点の扱いについて

- テスト得点は大きな誤差を含む
- 物理的な測定値と同列に考えてはいけない
- １回の試験成績だけでなく，小テストやレポート，出席回数など，複数の指標を組み合わせて多面的に成績を評価するのがよい

アルファ係数の性質

項目数とα係数の関係



- **項目数が多ければ，α係数は大きくなる**
- 1項目あたりの信頼性が低いほどその傾向は顕著
- 項目を増やしてもα係数の上昇には陰りがある
- 不必要に多量の項目を入れると，回答者は疲れてくる
- 多くの場合，１構成概念あたり，5～10項目の尺度で測定
- 逆に，2項目や3項目で信頼性が高いときは妥当性を疑う必要あり

多くの尺度を使いたいとき

- 2項目や3項目からなる心理尺度は妥当性が低いのでお勧めしない

- 全回答者にすべての尺度に答えてもらうのではなく，部分的に尺度を入れ替えた冊子をいくつか作成し，各回答者にはいずれかの冊子に回答して頂く

- 冊子Aは尺度1, 2, 3，冊子Bは尺度1, 2, 4，冊子Cは尺度1, 3, 4 で構成する　など

相関係数の希薄化

相関係数の希薄化

- 信頼性の低い尺度を使うと，標本相関係数が，真の相関係数よりも小さい (0に近い) 値になる

- 標本相関係数 ＝ 真の相関係数 $\times \sqrt{\text{各尺度の信頼性係数の積}}$

  |標本相関係数| ≦ |真の相関係数|

- 複数の尺度を用いて変数間の関連を検討する場合には，各尺度の信頼性が同等であることが望ましい

- 各尺度の信頼性が同等でないと，変数間の関連が崩れた結果が得られてしまう

相関係数の希薄化

	真の相関係数			観測される相関係数					
	信頼性係数＝1			信頼性係数が揃っている場合			信頼性係数が不揃いの場合		
	国語	社会	物理	国語	社会	物理	国語	社会	物理
国語	1	0.8	0.6	0.7	0.56	0.42	0.7	0.47	0.48
社会		1	0.4		0.7	0.28		0.5	0.27
物理			1			0.7			0.9

対角要素：そのテストの信頼性係数，　非対角要素：相関係数

- 真の状態

  国と社の相関　＞　国と物の相関　＞　社と物の相関

- 各教科の信頼性が同等のとき（信頼性係数は高くはないが揃っている）

  国と社の相関　＞　国と物の相関　＞　社と物の相関

- 物理の信頼性が高く，社会の信頼性が低いとき（信頼性係数不揃い）

  国と物の相関　≧　国と社の相関　＞　社と物の相関

相関係数の希薄化

- 各テストの信頼性が同等であれば，標本相関係数の値自体は小さくなるものの，関係性は保存される

- 各テストの信頼性に大小があると，標本相関係数の大小関係が真の相関係数の大小関係と不一致となり，本来とは異なった推論をしてしまう

- 信頼性の低い尺度では精度が低いので，原理的に，本来見たい現象を捉えることができない

- 研究に用いる複数の尺度の信頼性係数は，ほど良く高く，同等であることが望ましい

相関に基づく分析と相関係数の希薄化

- 相関に基づく分析：回帰分析，構造方程式モデリングなど
- 相関係数を利用する分析なので，相関係数の希薄化の問題が生じる

- 説明変数の信頼性係数に大小があるとき，

  信頼性係数が高い説明変数ほど偏回帰係数が大きくなり，

  　基準変数に対する説明力が強いと判断される

  信頼性係数が低い説明変数ほど偏回帰係数が小さくなり，

  　基準変数に対する説明力が弱いと判断される

- 分析結果が説明変数に用いた尺度の信頼性の評価になってしまう可能性がある