BBS 2005/6/22　　発表: 中西

American Psychologist, Vol.60, No.2, pp.170-180.

Inference by eye: Confidence Intervals and how to read pictures of data

Geoff Cumming & Sue Finch

平均値とエラーバーを示す図を読むときの７つの目則

第１の目則： 平均値とエラーバーが**表しているものは何なのか**を見極めよ.

(a) 従属変数は何か？それはもともとの単位か、標準化されているか？

(b) バーは信頼区間か、標準誤差か、その他か？ 信頼区間ならば信頼水準は何%か？

(c) 実験デザインはどうなっているか？

(d) どの効果や比較が主たる関心なのか？ 示された平均値や信頼区間はそれに対してどのように関係しているのか？ 推論についての注意をどこに集中すべきか？

第２の目則： 平均値を**実質的に**解釈せよ.

(a) その効果がどれくらい重要なのか、関心が高いか

(b) その効果がどれくらい大きいか

第３の目則： 信頼区間や他のエラーバーを**実質的に**解釈せよ.

第４の目則: 独立２群の比較に関して、95%信頼区間の重なり比率が **1/2 以下なら** $p \leqq .05$ である．さらに、**重なっていないなら、** $p \leqq .01$ である．（両群のサンプルサイズが 10 以上でエラーマージンが２倍以上異ならない場合）

第５の目則： 対応のあるデータに関しては、**差の平均値**とそのエラーバーを解釈せよ．一般的に、反復測定独立変数での個別の平均値のバーは**差についての推論とは関係ない**ので、用心せよ.

第６の目則： 標準誤差のバーは 95%信頼区間のバーの**およそ半分の大きさ**で、**68%信頼区間**のそれと近似する（サンプルサイズが 10 以上の場合）

第７の目則： 独立２群の比較に関して、標準誤差バーが少なくとも**バー１つ分離れていれば、** $p \leqq .05$ である．さらに、少なくとも**２つ分離れていれば、** $p \leqq .01$ である．（両群のサンプルサイズが 10 以上で標準誤差が２倍以上異ならない場合）

信頼区間とは何か？

仮想例）

ある都市の子供の言語能力を測りたいと仮定しよう。広く認められた言語能力のテストを選び、その得点は母集団において正規分布すると想定する。

$n = 36$ の無作為標本をテストし、標本平均 $M = 62$、標本標準偏差 $SD = 30$、だとわかった。この M は母平均の**点推定値**である。

また、95%信頼区間も求める。これは**区間推定値**である。この 95%は**信頼水準**と呼ばれる。$C = 95$ 以外でも、各 C に応じた区間を求められる。95%にするのは単なる慣例である。

この信頼区間は M を中心とし、両側に同じ長さ w で広がる。w は**エラーマージン**(margin of error)と呼ばれる。エラーマージンは標準誤差 SE に基づいている。

この場合の SE は、$SE = SD / \sqrt{n} = 30 / \sqrt{36} = 5$。

エラーマージンは $w = SE \times t_{(n-1),C}$ であり、$t_{(n-1),C}$ は C に応じた t 統計量の臨界値。この場合（$C = 95$, $d.f. = n - 1 = 35$）臨界値は 2.03 なので、$w = 5 \times 2.03 = 10.15$。よって信頼区間の下限は $M - w = 51.85$、上限は $M + w = 72.15$ となり、95%信頼区間は(51.85, 72.15)である。

前述の式から、一般に、区間の長さは C と n に依存し、95%信頼区間よりも 99%信頼区間のほうが広く、90%信頼区間のほうが狭い。

信頼区間と確率とのつながりを考えるときはいつも注意せよ。

Hays (1973) によれば信頼区間とは「ある高い確率でもって真なる母集団の値をカバーするような推定された値の範囲」

「$M - w \leqq \mu \leqq M + w$ である確率は 95%」は正しい言明だが、これは標本毎に変動する下限と上限についての確率的言明である。「求めた信頼区間（51.85, 72.15）が 95%の確率で μ を含んでいる」は正しくない。変動するのは信頼区間であって、μ は未知だが固定した値である。

なぜ信頼区間を使うか？

信頼区間を使うことの４つの主な利点

(a) 測定単位についての点推定値と区間推定値を得られ、研究においてとてもわかりやすい。

(b) 信頼区間は p 値と（従って帰無仮説検定とも）つながりがある。

(c) 推定に焦点を置いた実験間の証拠の結合、すなわちメタ分析とメタ分析的思考を助ける。

(d) 精度についての情報をもたらし、検出力の計算よりも有用である。

適切な信頼区間が求められない状況もまだあるが（多変量解析やモデル適合の評価など）、信頼区間の計算できる分野はどんどん広がっている。効果サイズの信頼区間や回帰分析のパラメータの信頼区間も求められるようになってきた。

信頼区間は *p* 値とつながりがあるだけでなく、さらなる重要なアドバンテージを提供する。信頼区間がある値をカバーしているかどうかを見ることは帰無仮説検定と同じだが、信頼区間をそれだけに使っていては意味がない。
伝統的な帰無仮説検定では**二分法的**決定が行われる。APA の publication manual には二分法的決定も記載されているが、正確な *p* 値を報告すべきだ(p.25)とも記載されている。**正確な *p* 値の報告**は、二分法的決定としての帰無仮説検定から、**解釈への有益なインプットとしての *p* 値の考慮**への移行を助長するものだ。
信頼区間を提示し議論するためのよりよい方法の開発と、信頼区間の幅広い使用によって、心理学者が推定の観点からアイディアを定式化し実験をデザインするようになるだろう。そしてこれは、より量的な理論、より informative な実験、そして一般的により強く経験的な学問へとつながる。
そして、「**目による推論**」を行うことが、この動向を強く支援する。

図の多義性と実験デザイン

第１の多義性： **エラーバー**が何を示しているのかわからない.

APA のマニュアルでは、キャプションでエラーバーを説明するように(pp.180, 182)、とされている。

第２の多義性： **実験デザイン**がどうなっているのかわからない.

Fig.3 を参照

左 a：独立２群（対応なし）

２つの平均値の差の 95%信頼区間は、エラーマージンを w_D とすると区間は[$(M_B - M_A) - w_D$, $(M_B - M_A) + w_D$]。

中央 b：対応のあるデータ

２回の反復測定の差の平均値を M_d とし、その信頼区間のエラーマージンを w_d とすると、区間は($M_d - w_d$, $M_d + w_d$)。

右 c：実験間

Fig.3

実験デザインは平均値と信頼区間の解釈に関して決定的な意味を持っている。

Gardner & Altman (1986) 「個々の平均値をただ漠然と示すのではなく、その研究における**主要な対比を直接に示す**ほうがよい」→ Fig.3 のほうがよい。
単純な差に関する話だけでなく、もっと複雑なデザインでも同様である。**主効果、単純主効果、交互作用、その他いかなる対比についても、その効果や対比の信頼区間を図に示すことができる**（信頼区間を計算する為の適切な誤差項の選択は簡単でないかもしれないが）。

すぐれた図のデザインは適切なデータ解釈に不可欠である (p.174)

複雑な実験デザインの場合

Fig.7 は被験者間要因１つ、被験者内要因１つの２要因計画。被験者間要因の単純主効果については第４の目則が使われるかもしれないが、それ以外（例えば被験者内要因の単純主効果、主効果、交互作用など）についてはここからは判断できない。
APA のマニュアルの p.181 にある図も、あのような一般的な注釈をつけるのは誤り。

この話の制限ポイント

平均値の両側信頼区間
正規分布母集団
第４、第７の目則はサンプルサイズ 10 以上、エラーバーが２倍以上違わない、が前提
多重比較に関する調整は考えていない

信頼区間の実質的解釈のための４つのアプローチ

自分の得た信頼区間は無限の連続の中からのたった１つ

また、同様に、ある研究者が 95%信頼区間を人生を通して幾度となく報告した場合、そのうちの 95%は推定されたパラメータを捉えていると期待できる。

区間と、区間の中の値を考える

一般的に、信頼区間の中の値はμについての有力候補で、その外側の値はそうではないのだが、これを確率についての不適切な言明をほのめかさずに表現するためにどんな言葉を使うのがよいかに関して、学者間で意見が割れている。

- 私たちの信頼区間はμについてのまことしやかな値の範囲である。信頼区間の外側の値は比較的信じがたい。（著者のおすすめ）
- 私たちは信頼区間がμを含むことを 95%確信できる。
- 私たちのデータは信頼区間の中ではμについてのいかなる値とも適合するが、外側の値とは比較的適合しない。
- 区間の下限はパラメータについてのありそうな最小限の推定値で、区間の上限はありそうな最大限の推定値だ。
- 区間内のどの値についても値の実質的解釈を考えよ。例えば下限と上限の解釈を考え、それらを平均値の解釈と比較せよ。

信頼区間、帰無仮説検定、p 値

帰無仮説検定として考えれば、区間の外側の値は両側検定で $p < .05$。内側の値は $p > .05$。信頼区間の上限と下限はちょうど $p = .05$。.05 は 95%信頼区間の場合で、一般的には $1 - C / 100$。Cを変化させれば信頼区間も変化するので、ある値に下限（または上限）が一致するように C を調節すれば、信頼区間は $p < .05$ かどうかを示すだけでなく正確な p 値も与えることができる。

精度の指標：w

エラーマージン w は研究の精度の指標である。w の実質的解釈をせよ。

w は、実験のプランニング（ある w を得るのにどれくらいの n が必要か）に関しても、結果の報告に関しても、**統計的検出力推定よりも有用**である。

帰無仮説検定が帰無仮説を採択しようとするときはいつでも用心が必要であり、信頼区間でもこの用心が減ぜられることはないが、w は強力なガイドになる。大きな w は実験にほとんど価値がないことを示すだろう。片や、信頼区間が狭く、帰無仮説の値μに近く（あるいは含んでいて）、信頼区間内のどの値も実践的な目的においてμと同等だと判断される場合、**信頼区間は実践的目的に関してμを真の値だとみなすことを正当化する**。

第４および第７の目則に関して

Fig. 4

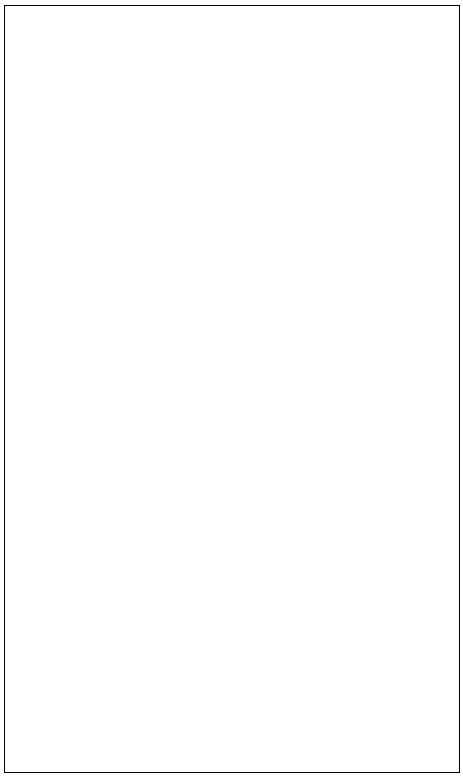

Fig. 5

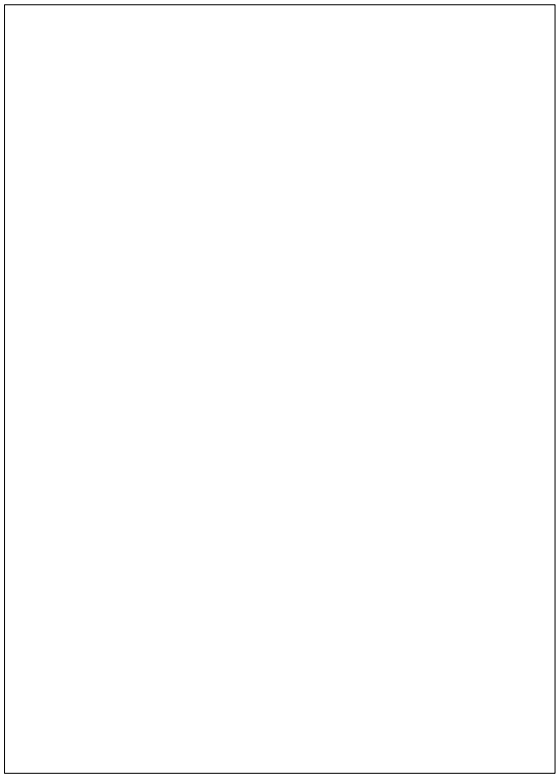

Fig. 7

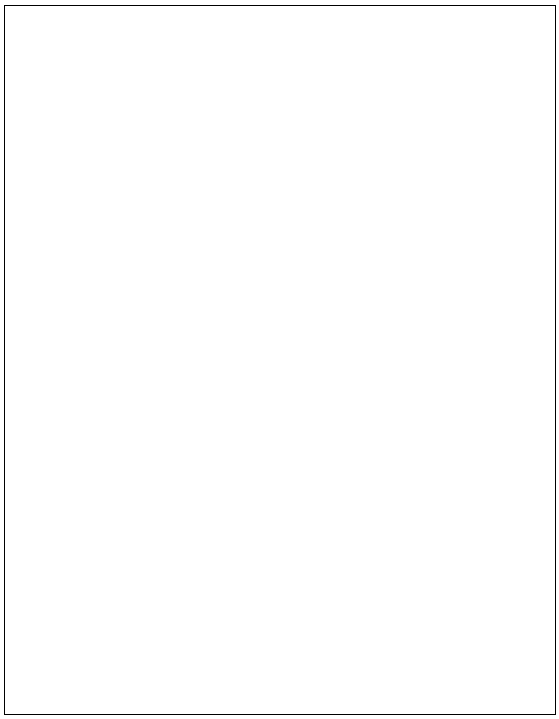

Fig. 6