

今日からできる！音響心理実験

a dummy's guide to psychoacoustical experiments

東京芸術大学
音楽環境創造科
丸井淳史

MARUI Atsushi
Tokyo University of the Arts
2009年7月23日



音の評価法 / Audio Evaluation Methods

物理評価 / Physical

- 音圧 / SPL
- スペクトル分布
spectral distribution
- 生理的反応 / physiological
 - 心拍数 / heart beat rate
 - 血糖消費 など
blood sugar consumption
level, etc.

心理評価 / Psychological

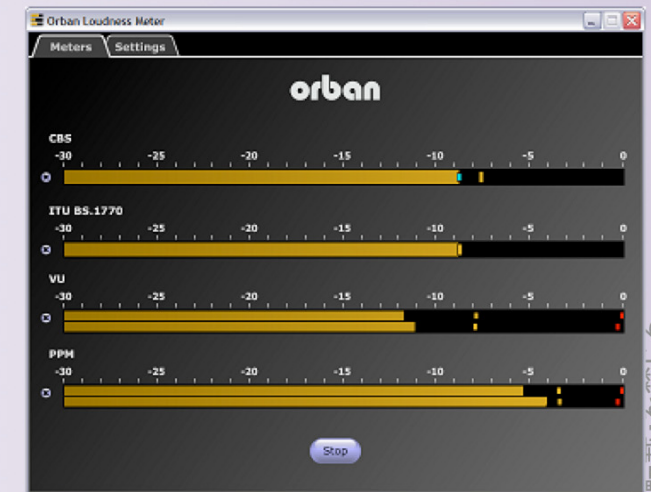
- 直接測定
direct measurement
- 評価実験
evaluation experiment
- モデル測定
predictive models

心理評価 / perceptual evaluation

- 直接計測 / direct measurement
 - 脳に電極を付けるなどして、音に対する反応を計測する
obtain perceptual data directly from brain wave?
 - 現在のところ非常に困難
very difficult at this point
- 評価実験 / evaluation experiment
 - 知覚測定 perceptual measurement
 - 評価すべき音刺激に対する、心理的個別属性の強さを数値化する
objective quantification of individual perceptual attributes for given sound stimuli
 - 感情測定 affective measurement
 - 評価すべき音刺激に対する、総合的な印象を数値化する
objective quantification of overall impression for given sound stimuli
- モデル計測 / predictive models
 - 人の反応を元に構築された予測モデルを用いて計測を行う
use predictive models made based on human responses

予測モデル / predictive models

- 予測モデルの種類 / two types
 - 特定の感覚属性を評価するもの
measures individual attributes
 - loudness、sharpness、roughness, etc.
 - 全体的な印象を評価するもの
measures overall impression
 - 通話やコーデックの音声品質
 - PEAQ、PESQ, etc.
- 予測モデルの利点 / benefits
 - 速い / fast
 - 安い / cheap
 - 安定している (評価結果に変動がない) / stable (little variability over time)
 - 長時間働かせても文句を言わない / works overtime without extra payment
 - 実験者に要求される知識量が少ない / comparably less knowledge necessary



Orban's loudness meter

評価実験でわかること / what you can expect from listening tests

- 複数の音刺激に聴感上の差違があるかどうか
identify whether two stimuli are perceptually identical
- 二つの音刺激をある感覚において比較したときの
大小関係
whether a stimulus is superior, equal, or inferior with regard to an attribute
- 二つの音刺激をある感覚において比較したとき、
一方がもう一方と比べて**どのくらい**大きい
quantify how much superior a stimulus is against one another
- ある音声システムが用途に適しているかどうか
whether an audio system is suitable for a purpose
- 様々な感覚属性を用いた音声システムの性能評価
performance rating of audio systems in many attributes

評価実験ではわからないこと / what you can't expect from listening tests

- アルゴリズムの問題点の指摘
tell what was the problem with the algorithm
- どのシステムがオーディオ雑誌によるランキング上位に入るかを予測すること
which system will be the winner in Hi-Fi magazine shoot-out
- ライバル社の音声システムが優れているのは、どのような技術のおかげかを調べること
what technology in a rival company's system make it better
- 製品の音声品質を向上する指針の提供
how a developer can improve their systems

ただし、評価実験の綿密なデータ分析と、その他の技術的情報を組み合わせることで、これらの疑問に答えられることもある

It's sometime possible to answer some of the issues with advanced statistical techniques combined with technical information.

評価実験前の確認事項 / to note before experiment

- 音刺激の物理測定は十分な情報を与えてくれるだろうか？
do physical measurements give enough information?
- 直接計測ができる方法・設備はあるだろうか？
is there a way to directly measure it?
- 利用できる予測モデルはあるだろうか？
is there a suitable predictive model?
- 自分が知りたいことは評価実験から得られるだろうか？
can information I need obtained from listening tests?

主観評価実験は / Subjective evaluation tests

- **高価** / takes money
- **時間がかかる** / takes time
- **努力を要する** / takes effort
- **やりなおしできない** / no repeats
- **有用な情報が得られる保証もない** / no guarantee of good data

可能な限り主観評価実験より低コストの方法を用いるべし

Use lower cost ways whenever possible!

まとめ / summary

- 不要な実験はやらない

don't do experiments that's unnecessary

- 参考文献、業界標準、予測モデル、耳 などを利用

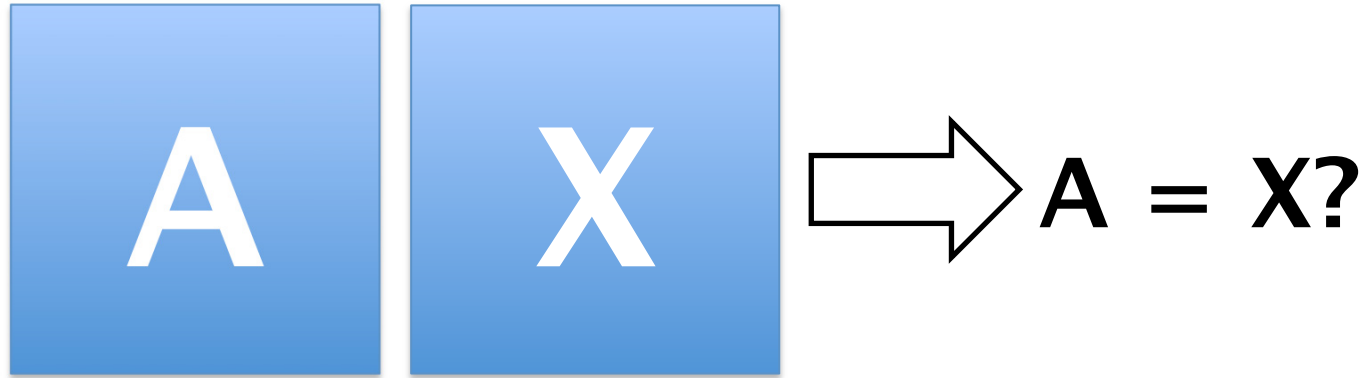
much can be learned from literatures, standards, predictive models, and your ears.

心理実験の方法

1. 違いが分かるか / are they different?
2. どう違うのか / different in what aspect?
3. どのくらい違うのか / how much different?

違いがあるかどうかを確かめる実験

experiment to check if the two are different



- 二つの刺激が提示される two stimuli are presented
 - 1つめ=A、2つめ=X / first A, second X
 - AとXが等しいかどうかをyes/noで回答する
Answer whether A and X are identical
- 問題点：バイアス (判断や回答の偏り) bias
 - yesと答えやすい人、noと答えやすい人がある
tendency to answer yes or no

違いがあるかどうかを確かめる実験 (ABX法)

experiment to check if the two are different



- 三つの刺激が提示される
 - 1つめ=A、2つめ=B、3つめ=AとBのどちらか
Three stimuli are presented:
First A, second B, third either A or B
 - XがAとBのどちらと同じかを回答する
Asked to answer which stimulus X was (A or B?)
- 三刺激二者強制選択
Three-interval Two-alternative Forced Choice (3I2AFC)

例：音声品質聞き比べ / Example

- コーデック1と2で符号化された音ファイル2つ

Sound files from two codecs

- ABXの組み合わせは4通り (順序効果を考慮)

Four possible combinations for ABX method

- 121, 122, 211, 212



原音

ABXが鳴ります。

XはAとBのどちらと同じでしょうか？

(今回は順序効果がないものとして一回のみ行います)

Now, let us try the ABX method once.



3I2AFCの分析法 / analysis

- 「試行数のうち何%が正解だったか」が違いの分かりやすさに関係している

Ratio of correct answer is related to the difference

- まったくでたらめに選んだら正答数は50%

Completely random selection gives 50%.

- 50%をどれだけ上回ったら、はっきりとした違いだと自信を持って良いのか?

How much higher should we get to say it *for sure*?

- 60%では? 80%では?

二項分布 (binomial distribution) を使う

二項分布 / binomial distribution (in Excel)

=BINOMDIST(成功数, 試行回数, 母比率, 関数形式)

#correct #trials pop.ratio

- n回行う試行のうち何回正答したか
 - 二者強制選択では正答の母比率は50%
- population probability is 50% in 2AFC

関数形式

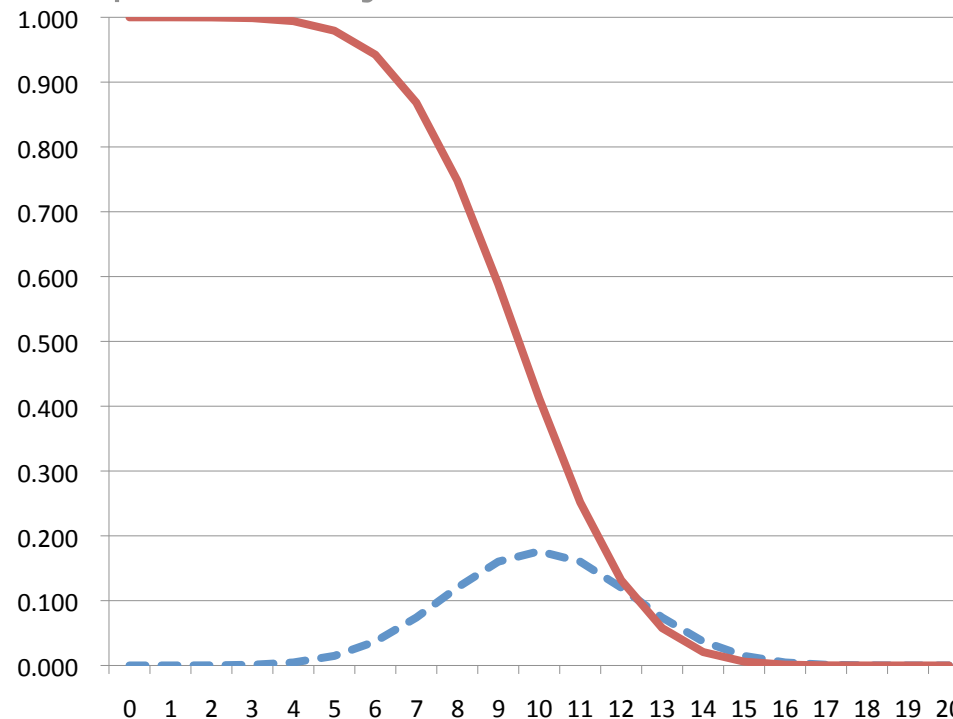
0 – 確率

1 – 累積確率

function type

0 – PDF

1 – CDF



確率

probability

累積確率
の余事象

complement of CDF

心理実験の方法

1. 違いが分かるか / are they different?
2. どう違うのか / different in what aspect?
3. どのくらい違うのか / how much different?

評価軸を作る / attribute elicitation

- 違いの種類と大きさを言い表すための、語彙・尺度が必要 → 評価語・評価軸

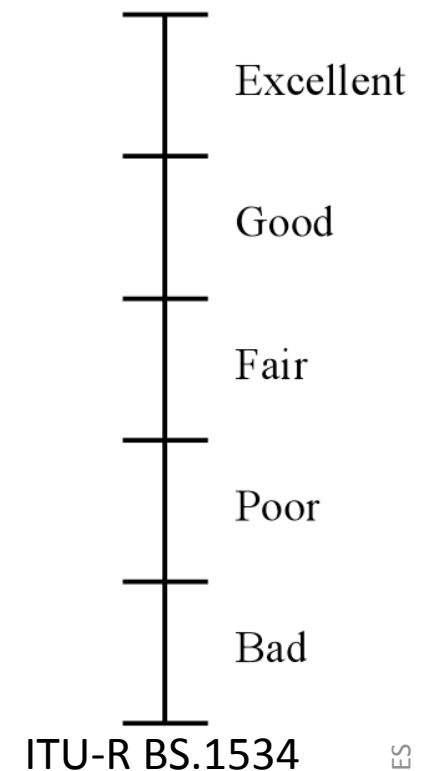
language to express the difference and its size

- 評価語を決めるのは難しく、評価語の選択によって実験の成否も決まってしまう

It's cumbersome to create a new attribute scale, and the success of experiments depend on the scale

- 可能であればITUや他の研究者が定めた評価尺度を利用するほうが失敗が少ない

Better to use standardized scales whenever possible.



尺度作成時に注意すること / to note when creating a scale

- 尺度の説明は明確であること
clear explanation of the scale
 - 刺激のどの部分について判断を求められているか分かる
be able to know what aspects of the sound to listen for
- 尺度が使いやすいこと (特にGUIの場合)
easy to use the scale (esp. when GUI)
- 刺激を区別できるような尺度
have to be able to distinguish between stimuli
 - 刺激の変化を説明できる評価語
descriptors that can tell the difference
 - 十分な分解能 / enough resolution

評価語の抽出法 / attribute elicitation

直接抽出法 direct elicitation

- 感覚とその説明語との間に緊密な関係があると仮定

based on hypothesis that there are close relation between sensation and the description

- 感覚を言語化する方法が必要

Need to verbalize the sensation

- 共通言語抽出法：評定者グループ内で共通の語彙を持つ
common attribute elicitation
- 個別言語抽出法：評定者一人一人が個別に語彙を持つ
individual attribute elicitation

間接抽出法 indirect elicitation

- 感覚とその言語化を分離する方法

to separate perceptual attribute and its verbalization

- 言語化は各評定者の語彙の大きさに依存するため、適切でない場合があるため

verbalization depends on subjects' vocabulary size

- 感覚と説明語の関係が本当に存在するか確かめられないため

no way of knowing whether sensation and the word used for it is actually related

直接抽出法の例 / example

- 三つ組法 (レパートリー・グリッド法を参考に)

Method of Triads (inspired by Repertory Grid Technique)

- 与えられた三つの刺激の中から仲間はずれを見つける

Pick the "odd one" out

- 仲間はずれの理由を書き出す
and describe why it's odd
- 残り二つの共通点を書き出す
then describe commonality
between the rest

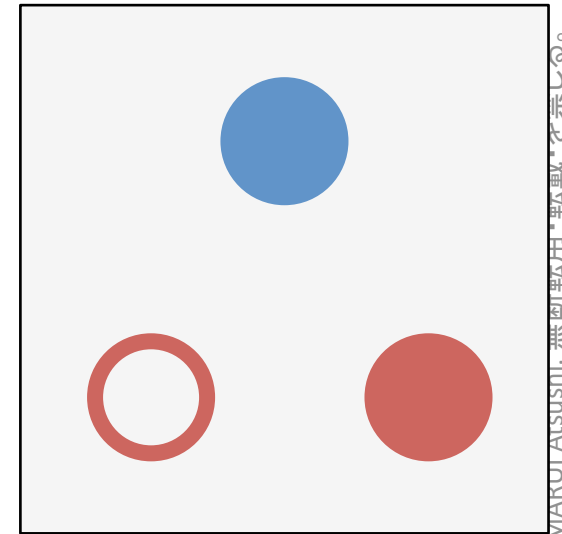
- これをすべての刺激組み合わせで行う

- 10刺激なら ${}_{10}C_3=120$ 組

for all possible triads (120 triads for 10 stimuli).

- 得られた評価語が尺度の両端となることが多い

Obtained words are most likely to be the end point descriptors of a scale.



心理実験の方法

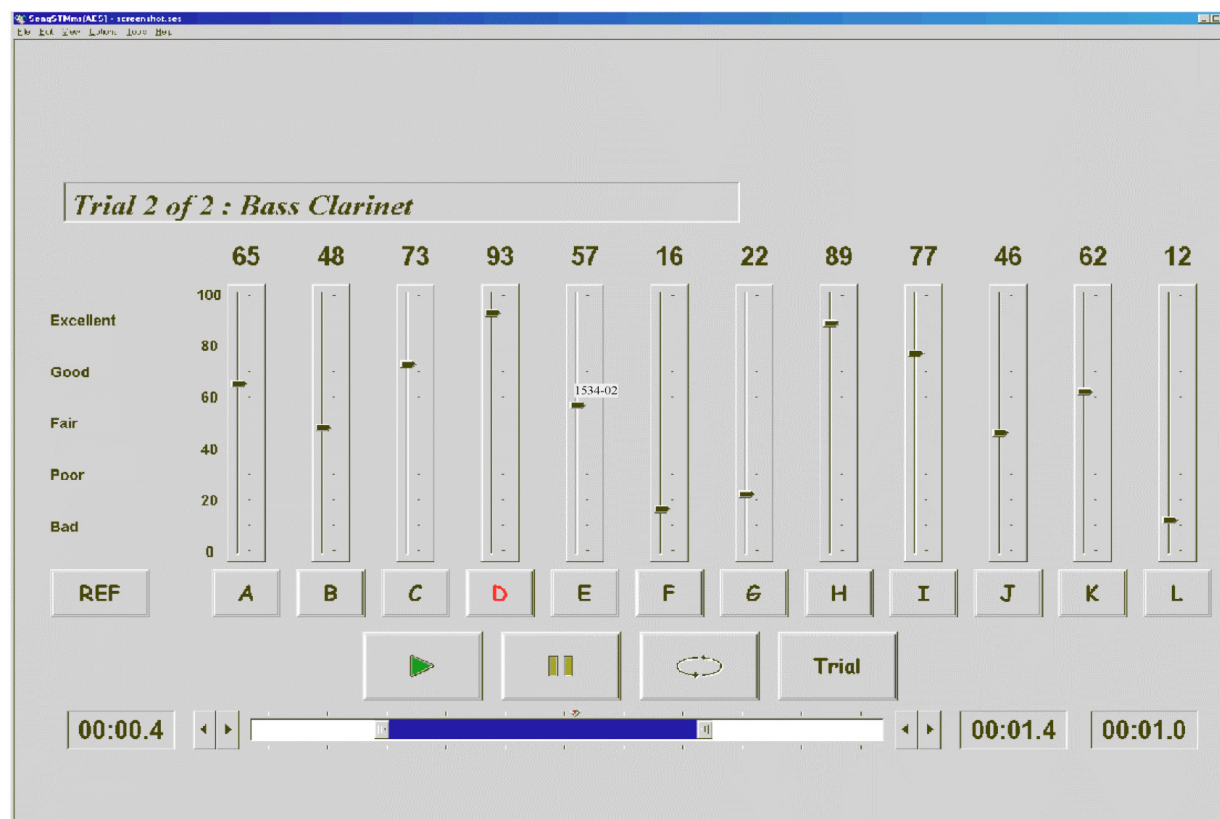
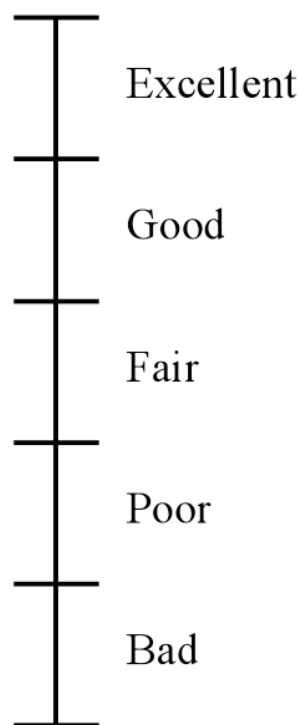
1. 違いが分かるか / are they different?
2. どう違うのか / different in what aspect?
3. どのくらい違うのか / how much different?

例：音声品質聞き比べ / example

- 複数のコーデックで符号化された音ファイルが、原音と比べてどのくらい品質が落ちているか比較

How much degradation was made with a codec?

- ITU-R BS.1534に記載されたMUSHRA法の場合 (下図)
(MULTiple Stimuli with Hidden Reference and Anchors)
below is one way to do an experiment (MUSHRA)



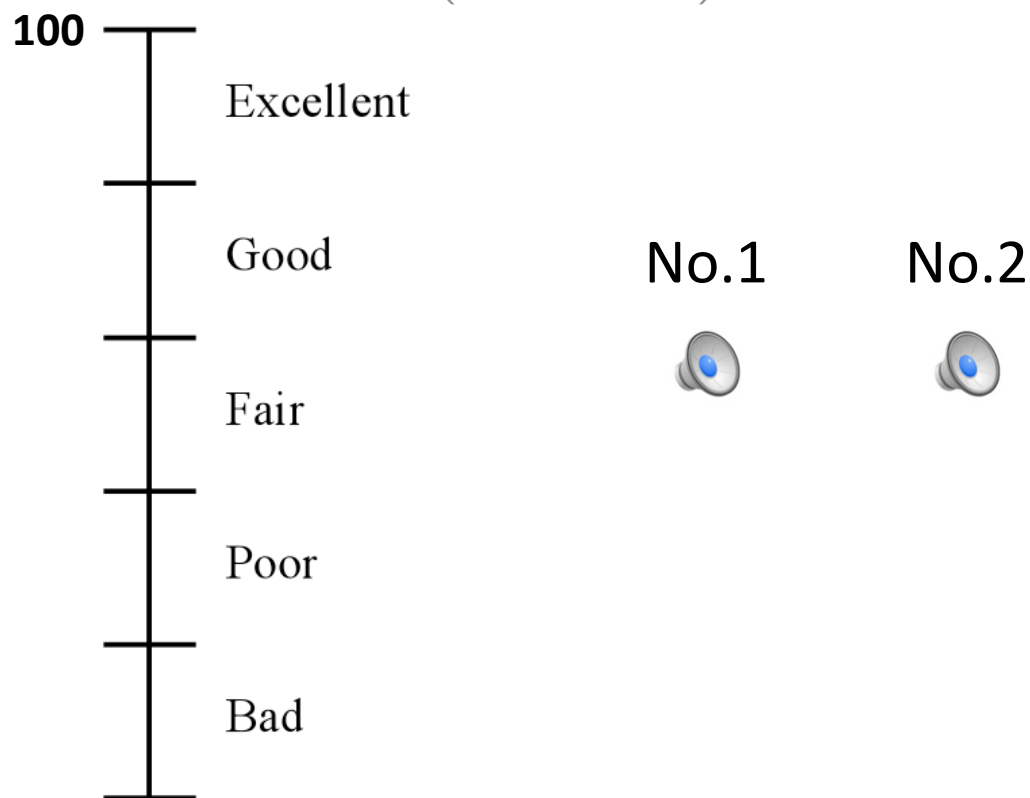
例：音声品質聞き比べ

- コーデック1と2で符号化された音ファイル2つ

Two sounds coded with two codecs

- 原音→符号化音と再生します。どのくらい品質が落ちているか
100点満点で評価して下さい。計2問。

Please rate how much the second sound is degraded compared to the first sound (reference).



評点1	評点2
98	40
92	16
80	60
90	50
75	65
90	70
75	85
54	28
79	75
82	80

平均値の比較

- 二つの平均値の差が、二つの音がどのくらい違うかに関係している

Difference between the two means is related to how different the two stimuli are.

- 差がまぐれによるものではなかったか確認が必要

Need to check if the difference was obtained by chance

- 平均値どうしの差だけでなく、分散の大きさも考慮しないといけない

with both means and variances taken into considerations.

t検定 (t-test) を使う

t検定 / t-test

=TTEST(データ範囲1, データ範囲2, モード, 種類)

- 二つのデータの平均値に統計的な差異があるか調べる

Test if there is a statistical difference between two samples

- モードは必要に応じて選択 (片側：大小関係を調べる / 両側：差異を調べる)

mode should be selected according to the needs

- 分散が等しいかどうかはF検定で調べられる

Use F-test for testing variance equality

モード

- 1 - 片側検定
- 2 - 両側検定

mode

- 1 – one-tail
- 2 – two-tails

種類

- 1 - 対応のある標本
- 2 - 分散の等しい二標本
- 3 - 分散の異なる二標本

type

- 1 – paired samples
- 2 – equal variance
- 3 – unequal variance

信頼区間 / confidence interval

- 信頼区間は母平均を与えられた確率で含むことができる

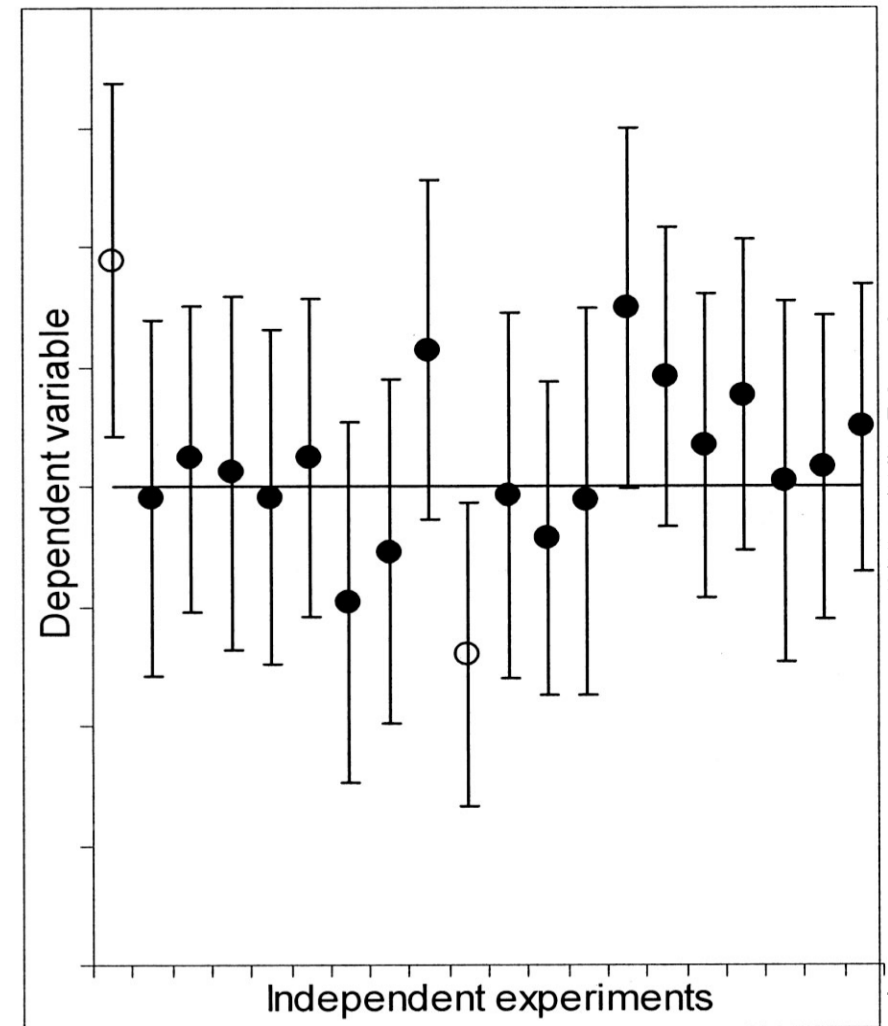
Confidence interval captures the population mean with certain probability (typically set to 95%)

- 確率としては95%がよく使われる
- 「母平均は95%の確率で信頼区間の中にある」とは意味が異なる

This is not the same as “the population mean is within the CI at 95% probability”

- 「同じような実験を100回行ったとき、95回は信頼区間の中に母平均が入ってくる」のほうが正しい

Instead, “if you do similar experiment 100 times, you can include the population mean within the CI for 95 of the times”

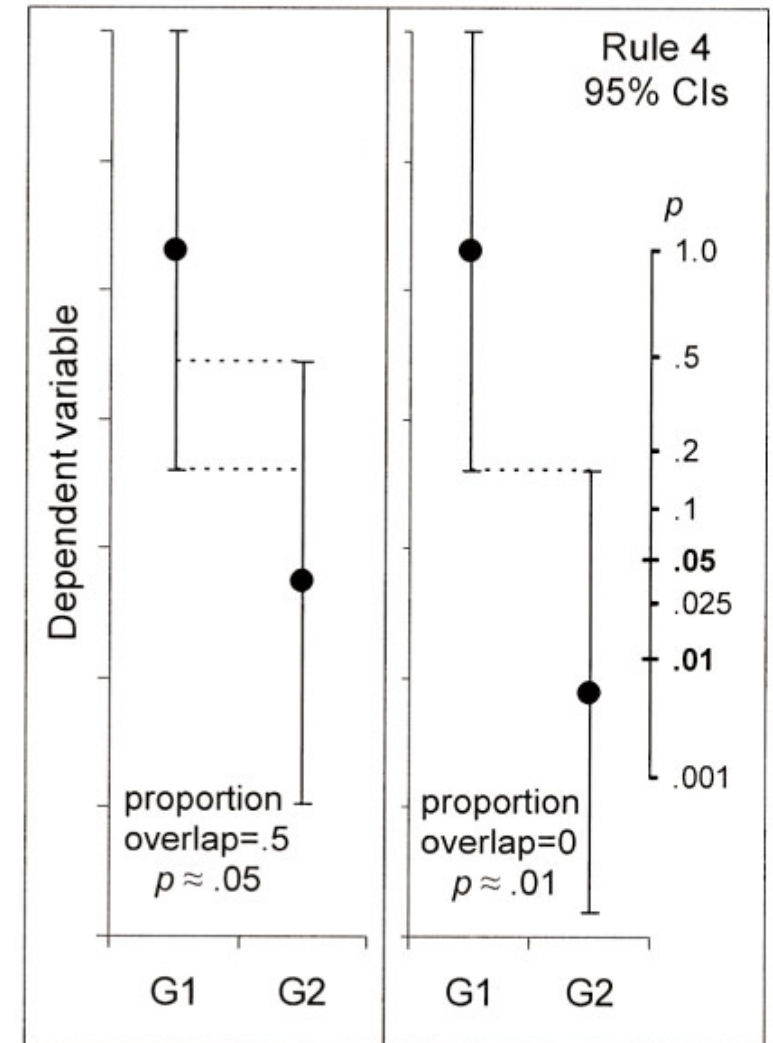


信頼区間 / confidence intervals

- 信頼区間を見て検定する方法
 - Cumming & Finch “Inference by Eye” (2005)
- めやす
 - 95%信頼区間の重なりが半分程度ならおよそ $p=.05$
 - 95%信頼区間が重ならなければおよそ $p=.01$

95% CI proportion overlap=.5 is almost the same as $p=.05$

95% CI proportion overlap=0 is almost the same as $p=.01$

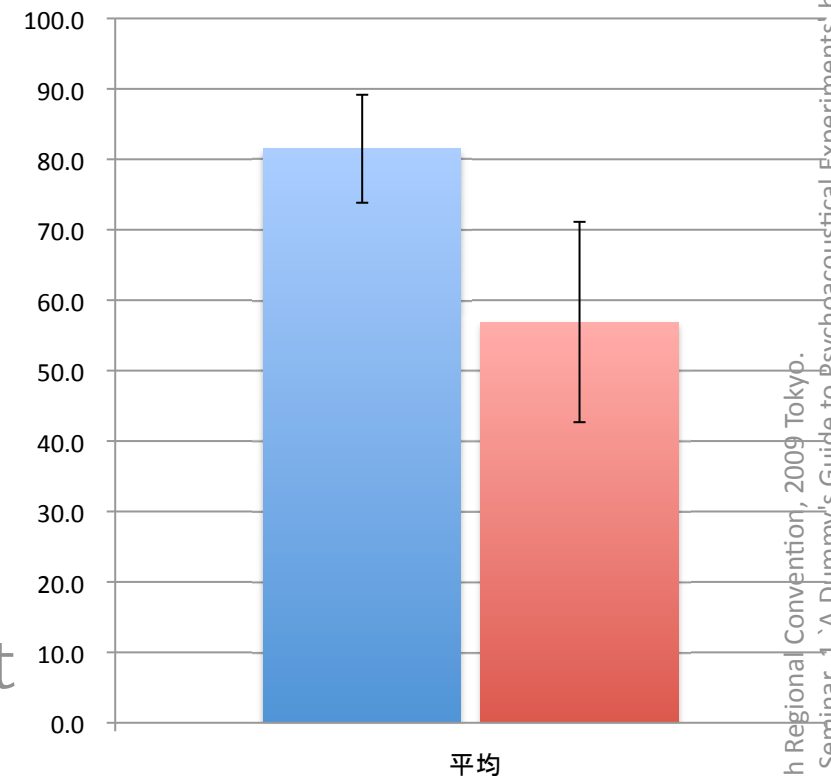


違いがあるかを目視で判断する方法 / inference by eyes

- 音声品質聞き比べの例では、信頼区間が重なっていないので、およそ $p=.01$ で統計的有意差があるのでは、と考えられる

in our example, graph readings tell us that there may be a statistically significant difference with $p=.01$ level

- あくまでも目安なので、きちんと t検定を行う
further t-test is recommended once you get a nice looking result



まとめ

- 不要な実験はやらない
don't do experiments that's unnecessary
 - 参考文献、業界標準、予測モデル、耳 などを利用
much can be learned from literatures, standards, predictive models, and your ears.
- 実験の流れ (一例) / an example of experimental flow
 - 違いが分かるか / are they different?
→ ABX法と二項分布を紹介 / ABX and binomial distribution
 - どう違うのか / different in what aspect?
→ 三組法を紹介 / method of triads
 - どのくらい違うのか / how much different?
→ 評価とt検定 (もしくは 信頼区間)
rating and t-test (and/or confidence interval)
- 音刺激や調査したい内容によって実験手法・分析手法は様々

参考文献 / References

- Søren Bech and Nick Zacharov. *Perceptual Audio Evaluation – Theory, Method and Application*. Wiley. 2006.
- Geoff Cumming and Sue Finch. Inference by Eye: Confidence Intervals and How to Read Pictures of Data. *American Psychologist*, v.60, n.2, pp.170–180. 2005.
- 豊田秀樹 『違いを見抜く統計学』 講談社 1994.
- 田栗正章ら 『やさしい統計学』 講談社 2007.