New Perspectives of p-Value and its Strength of Evidence Measured by Confidence Distribution

Sifan Liu & Minge Xie

Department of Statistics & Biostatistics,
Rutgers University

Joint work with Regina Liu

Conference on Statistical Inference Beyond p-values: Are we all Aligned?
NJ Chapter of ASA
May 25, 2017

Today's talk

- What is a p-value? How/why does it work?
  - A general definition highlighting two *performance characteristics*
- p-value by confidence distribution (CD)
  (setting of nested parametric models)
  - A meaningful interpretation as *evidence in favor of the null* hypothesis, H_0; Thus, for example, $p = .8$ indicates more support of H_0 than $p = .6$ or $.3$

p-value has a long history

- p-value has a long history
  - Early works related to ad hoc p-values date back to the 1770s by Pierre-Simon Laplace.
  - In 1900, Karl Pearson firstly gave a formal definition of p-value in the Pearson's chi-squared test for 2×2 tables.
  - In 1925 and later on, R.A. Fisher popularized p-value and made it an influential statistical inference tool.

- p-values are ubiquitous in practice
  - 4,572,043 p-values in 1,608,736 MEDLINE abstracts and 3,438,299 p-values in 385,393 PubMed Central full-text articles are identified between 1990-2015.
  - From the 151 core clinical journals in 2014, p-values were reported in 33.0% of abstracts, 35.7% of meta-analyses, 38.9% of clinical trials, 54.8% of randomized controlled trials.
  - Ref: Chavalarias et al. (2016).

p-value has a long list of complaints

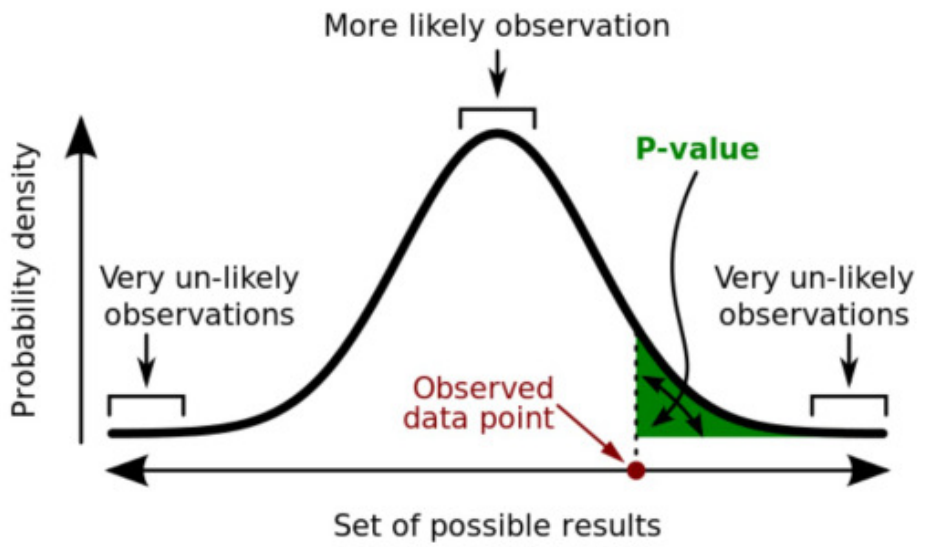
Berger (2003) and references therein -

- p-value is widely misinterpreted either as the probability that the null hypothesis is true, or as an error rate
- p-value provides considerable overstatement of the evidence against H_0 (from the Bayesian prospective)

Wasserstein & Lazar (2016) -

- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.
  - A relatively large p-value does not imply evidence in favor of the null hypothesis.
- Scientific conclusions and business or policy decisions should not be based solely on whether a p-value passes a specific threshold (often 0.05).

What is a p-value?



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Example of a p-value computation. The vertical coordinate is the probability density of each outcome, computed under the null hypothesis. The p-value is the area under the curve past the observed data point.

(Definition tells us how to construct; not directly associated with performance)

Figure 1 : Illustration of p-value from Wikipedia

Current textbook versions of the p-value

Suppose $T(\boldsymbol{X}_n)$ is the test statistic summarizing the information of $\boldsymbol{X}_n$.

1. p-value is an "upper bound" probability,

$$p\text{-value} = \sup_{\theta \in \Theta_0} \mathsf{P}_\theta\{T(\boldsymbol{X}_n) \geq t(\boldsymbol{x}_n)\}, \tag{1}$$

   where $t(\boldsymbol{x}_n)$ is the observed test statistic; cf. e.g., Abell et al. (1999).

2. p-value is the smallest level of significance for which H_0 can be rejected based on $\boldsymbol{x}_n$,

$$p\text{-value} = \inf\{\alpha : t(\boldsymbol{x}_n) \in R_\alpha\}, \tag{2}$$

   where R_α is the rejection region of level $\alpha \in [0, 1]$; cf. e.g., Lehmann & Romano (2005).

◇ *Defined by construction; No direct association with performance*

■ **Question**: Why/how does it work (or not work)?

How/why does it work? – Proof by "Contradiction"

Setup: Need to make a judgment about whether a statement H_0 is true based on observed data $\boldsymbol{X}_n = \boldsymbol{x}_n$

The Logic: Counterfactual [Refs: Fisher (1925, 1935, 1959)]

Assuming that H_0 is true, we intend to design an assessment to evaluate the chance ("how likely") that the observed data $\boldsymbol{x}_n$ is "compatible" with H_0; This value of the assessment is referred to as a *p-value*.
If the p-value is small, then we "consider" there is a conflict (contradiction) and thus in turn indicate H_0 is false.

- p-values intends to indicate how incompatible the data are with a specified statistical model (Wasserstein & Lazar 2016).
- p-value can be viewed as an index of the "strength of evidence" against H_0, with small p indicating an unlikely event and, hence, an unlikely H_0 (cf., e.g., Berger 2003).

However – Need of frequentist interpretation

Unlike the usual proof-by-contradiction in (nonrandom) math problems, there are obstacles due to the random nature in statistics:

- The value of the assessment is small $\not\equiv$ a sure contradiction! (*Resort to*: frequentist interpretation – where 5% is introduced)
- $\Pr(\boldsymbol{X}_n = \boldsymbol{x}_n | H_i)$ is often 0 or small; thus cannot be directly used as a measure of evidence (*Resort to*: "observed plus more extreme than observed")

A frequency (frequentist) argument (if invoking the 5% threshold)

Assume that H_0 is true, we hope that, in 100 tries, at least 95 times we can make the correct decisions in not rejecting H_0.

How to design the assessment (p-value)?

To ensure the *frequency argument* holds, a sufficient condition is –

(I) p-value, as a function of random sample $\boldsymbol{X}_n$, is Uniform[0,1] distributed (or $\overset{sto.}{\geq}$ Uniform[0,1]) under H_0. (cf., Berger & Boos 1994, Liu & Singh 1997, Shafer et al. 2011)

- The probability of mistakenly rejecting a correct statement is less than $100\alpha\%$ (Type-I error $\leq 100\alpha\%$) .

An additional condition that is good to impose (to ensure test power) –

(II) When H_0 is false, the p-value should be getting closer and closer to zero as sample size increases.

- The probability of correctly rejecting a false statement goes to 1 as sample size $n \to \infty$ (power $\to 1$; Type II error $\to 0$).

□ Although not apparent, the conventional (textbook) p-value definitions can be shown to have these two properties!

Formal definition by performance -

- $\boldsymbol{X}_n$: random sample of size n from a distribution indexed by a parameter $\theta \in \Theta$;
- $\mathcal{B}_\Theta$: the Borel algebra of the parameter space Θ;
- $\mathbb{X}^n$: the sample space corresponding to observed sample data $\boldsymbol{x}_n$.

Definition (A) (p-value & limiting p-value)

Suppose a Borel set $\Theta_0 \in \mathcal{B}_\Theta$, and $p(\cdot, \cdot)$ is a mapping:

$$\mathbb{X}^n \times \mathcal{B}_\Theta \mapsto [0, 1].$$

For a given $\boldsymbol{x}_n \in \mathbb{X}^n$, the value of $p(\boldsymbol{x}_n, \Theta_0)$ is called the *p-value* of the statement $H_0 : \theta \in \Theta_0$, if $p(\boldsymbol{X}_n, \Theta_0)$, as a function of the random sample $\boldsymbol{X}_n$, satisfies the following conditions for any $\alpha \in [0, 1]$,

(I) $\mathsf{P}_\theta\{p(\boldsymbol{X}_n, \Theta_0) \leq \alpha\} \leq \alpha$, for all $\theta \in \Theta_0$;

(II) $\mathsf{P}_\theta\{p(\boldsymbol{X}_n, \Theta_0) \leq \alpha\} \to 1$, as $n \to \infty$, for all $\theta \in \Theta / \Theta_0$.

If (I) holds only in limit, then it is called a *limiting p-value*.

Formal definition by performance -

The formal definition is **attractive** -

- General – encompasses all existing definitions of p-value available
  - Including bootstrap p-value (where the calculation is NOT performed under the assumption that H_0 is true!)
- Easy to interpret/understand - it directly follows the logic and highlights on the performance.

More importantly –

- It allows us to broaden the development of p-value concept.
  - How about developing a p-value that measures the strength of evidence in favor of H_0? So we can say $p = .8$ indicates more support of H_0 than $p = .6$ or $.3$?

Commends on current textbook versions

Good news: Have successfully helped solve many real life testing problems

- Can be shown to satisfy the performance described in the formal definition and thus the logic is correct

Bad news: They are defined by construction; Do not provide a "direct" connection to the logic or a clear "evidence-based" interpretation –

- They have caused many confusions and complains in practice!
  - A p-value is often misconstrued as the probability that H_0 is true or an error rate
- Technically,
  - Specific test statistics and/or rejection regions are required;
  - Calculations are under H_0 (even if H_0 is in fact incorrect!);
  - The values do not provide measures of strength in favor of H_0

  Thus, their applicability is limited; e.g., limited to the standard one- or two-sided tests, heavy reliance on $.05$, etc.

Research Question –

Question: Can we find a **new p-value construction** that is simple to implement and interpretation with the following features?

(1) Calculations are NOT performed under the assumption that H_0 is true
   - Can sidestep a specific test statistic

(2) The value is define as a measure of evidence in favor of H_0
   - So we can interpret "large" p-value; i.e., $p = .8$ indicates more support of H_0 than $p = .6$ or $.3$!

(3) The method is applicable for wide range of test problems
   - Specifically, it matches with the conventionally defined p-value for the standard one- or two-sided tests

(4) The p-value can also provide a bridge to the Bayesian posterior probabilities of the null and alternative .

Answer: Yes. We propose to use confidence distribution (CD) to achieve the above goals.

Parameter Estimation: Point, Interval & Distribution Estimators

In statistics, we routinely **estimate unknown parameters** using

- Point estimator
- "Interval estimator": confidence Interval (CI)

A **very simple question** is

— Can we also use a distribution function ("distribution estimator") to estimate an unknown parameter of interest in frequentist inference in the style of a Bayesian posterior?

The **answer is YES** and one such estimator is -

- Confidence distribution (CD)

Example: $X_1, \ldots, X_n$ i.i.d. follows $N(\mu, 1)$

- Point estimate: $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Interval estimate: $(\bar{x}_n - 1.96/\sqrt{n}, \bar{x}_n + 1.96/\sqrt{n})$
- Distribution estimate: $N(\bar{x}_n, \frac{1}{n})$

▶ The idea of the CD approach is to use a sample-dependent distribution (or density) function to estimate the parameter of interest.

Inference using CD: point estimators, confidence intervals, p-values & more

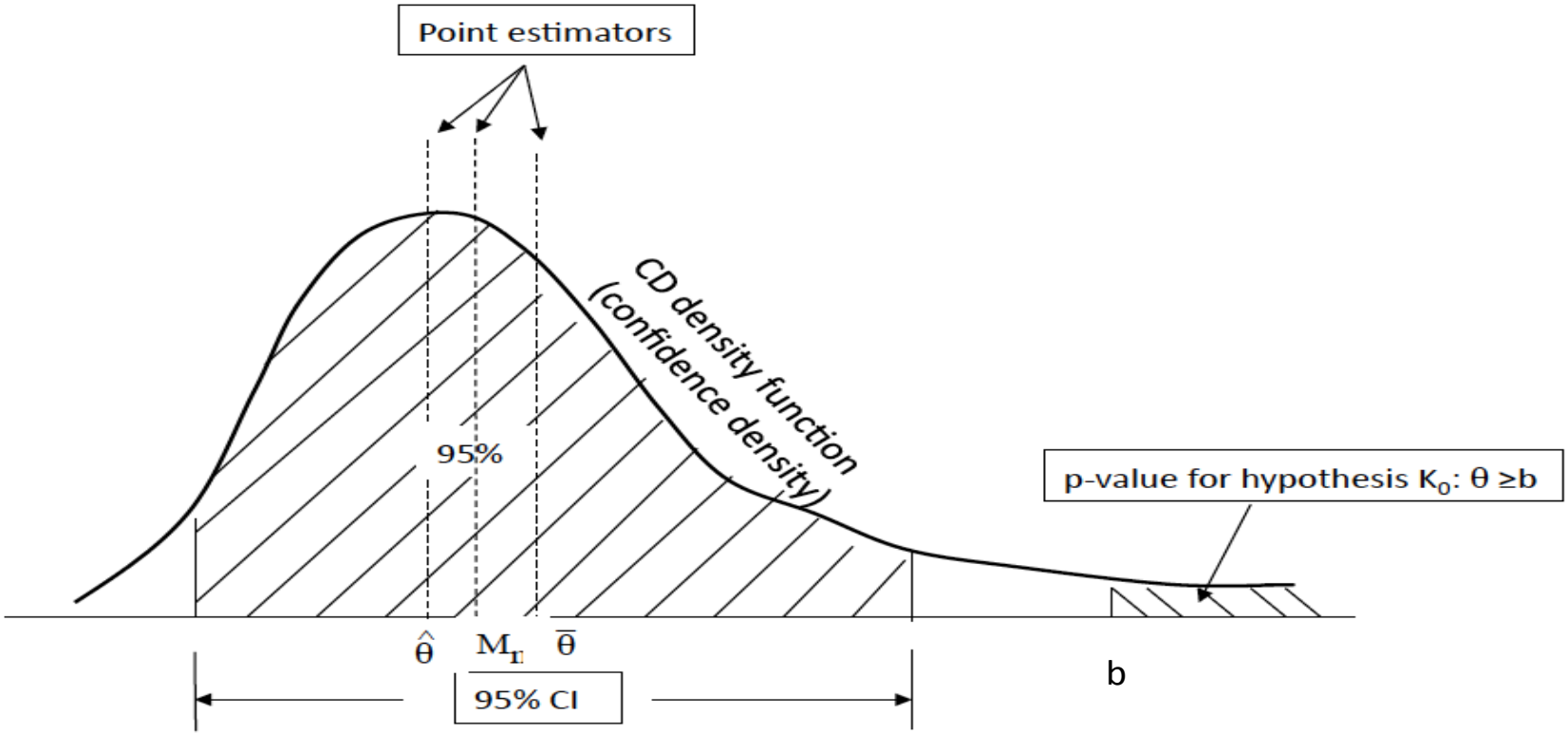


Figure 1. *The plot is a graphical illustration on making inference using a CD, including examples of point estimators (mode $\hat{\theta}$, median M_n and mean $\bar{\theta}$), a level* 95% *confidence interval and a one-sided p-value.*

Example 1: Normal CD function $N(\bar{x}, 1/n) \Longrightarrow$ (i) Point estimator: $\bar{x}$; (ii) level $(1-\alpha)100\%$ CI: $\bar{x} \pm \Phi^{-1}(1-\alpha/2)/\sqrt{n}$; (iii) p-value: $1 - \Phi\big(\sqrt{n}(b-\bar{x})\big)$, etc.

Example: Many ways to obtain the "distribution estimate" $N(\bar{x}_n, \frac{1}{n})$

Method 1: Normalizing likelihood function

- Likelihood function of $x_1, \ldots, x_n$ i.i.d. from $N(\mu, 1)$:

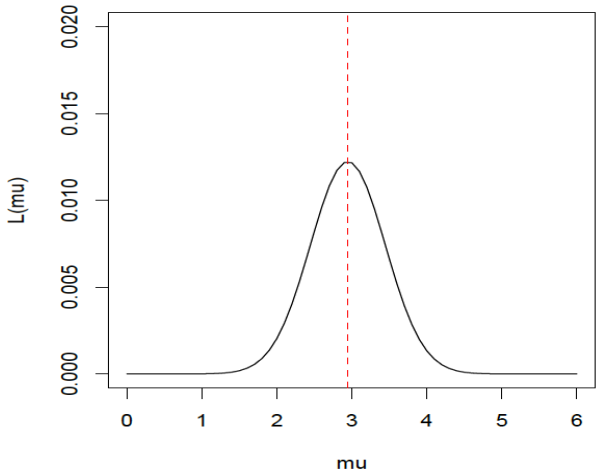
$$L(\mu|data) = \prod f(x_i|\mu) = Ce^{-\frac{1}{2}\sum(x_i-\mu)^2} = Ce^{-\frac{n}{2}(\bar{x}_n-\mu)^2-\frac{1}{2}\sum(x_i-\bar{x}_n)^2}$$

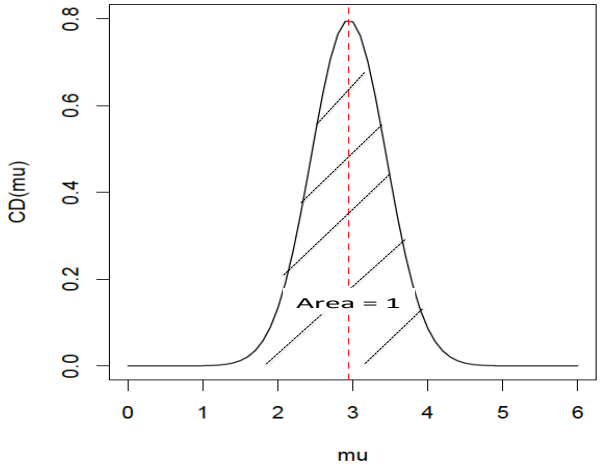
- Normalized with respect to μ

$$\frac{L(\mu|data)}{\int L(\mu|data)d\mu} = \ldots = \frac{1}{\sqrt{2\pi/n}}e^{-\frac{n}{2}(\mu-\bar{x}_n)^2}$$

It is the density of $N(\bar{x}_n, \frac{1}{n})$!

★ Suppose $n = 4$; Observe a sample with mean $\bar{x}_{obs} = 2.945795$:





Example: Many ways to obtain the "distribution estimate" $N(\bar{x}_n, \frac{1}{n})$

Method 2: p-value method

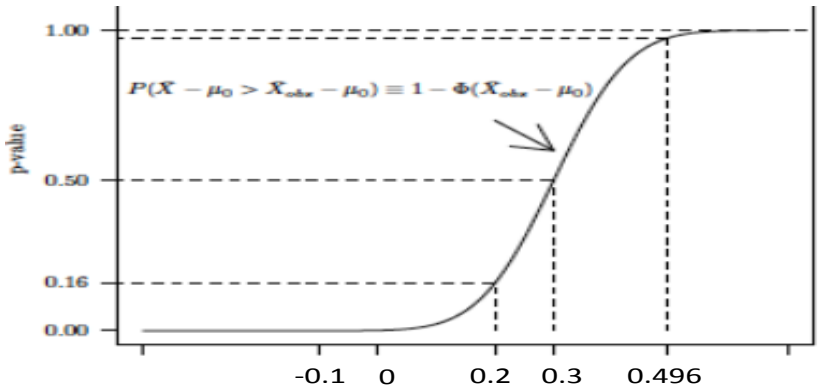
- One-sided test: $K_0 : \mu = \mu_0$ vs $K_a : \mu > \mu_0$

$$p(\mu_0) = P(\bar{X} > \bar{x}_n) = 1 - \Phi(\sqrt{n}\{\bar{x}_n - \mu_0\}) = \Phi(\sqrt{n}\{\mu_0 - \bar{x}_n\}).$$

Varying $\mu_0 \in \Theta$! $\Longrightarrow$ Cumulative distribution function of $N(\bar{x}_n, \frac{1}{n})$!

★ Suppose $n = 100$ and we observe $\bar{x}_n = 0.3$

mu0	infty	-0.1	0	0.1	0.1355	0.2	0.3	0.4645	0.496
P-value	0	.00002	.00135	0.02275	0.05	.15866	0.5	0.95	0.975



Method 3, 4, ... : Bayes (omit), fiducial (omit), ...

Three forms of CD presentations

.

- **Confidence density**: in the form of a density function $\mathrm{h}_n(\theta)$

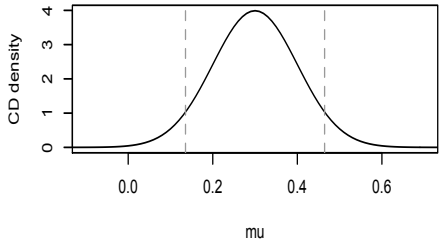
  e.g., $N(\bar{x}_n, \frac{1}{n})$ as $\mathrm{h}_n(\theta) = \frac{1}{\sqrt{2\pi/n}} e^{-\frac{n}{2}(\theta-\bar{x}_n)^2}$.

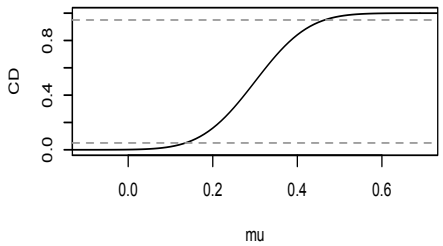
- **Confidence distribution** in the form of a cumulative distribution function $\mathrm{H}_n(\theta)$

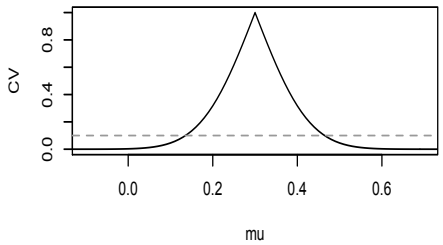
  e.g., $N(\bar{x}_n, \frac{1}{n})$ as $\mathrm{H}_n(\theta) = \Phi\big(\sqrt{n}(\theta - \bar{x}_n)\big)$

- **Confidence curve**:
  $CV_n(\theta) = 2\min\big\{\mathrm{H}_n(\theta), 1 - \mathrm{H}_n(\theta)\big\}$

  e.g., $N(\bar{x}_n, \frac{1}{n})$ as $CV_n(\theta) =$
  $2\min\big\{\Phi\big(\sqrt{n}(\theta - \bar{x}_n)\big), 1 - \Phi\big(\sqrt{n}(\theta - \bar{x}_n)\big)\big\}$







Definition: confidence distribution

A one-sentence version -

- A *confidence distribution* (CD) is a sample-dependent distribution function that can represent confidence intervals (regions) of all levels for a parameter of interest.

Formally,

Definition

A sample-dependent function on the parameter space $(\text{i.e., a function on } \Theta \times \mathcal{X})$ is called a *confidence distribution* (CD) for parameter θ, if:

R1) For each given sample, it is a distribution function on the parameter space;

R2) The function can provide confidence intervals (regions) of all levels for θ.

(Notations: Θ = parameter space; $\mathcal{X}$ = sample space)
Example: $N(\bar{x}_n, \frac{1}{n})$ on $\Theta = (-\infty, \infty)$.

Analogy way (to define point estimators) in point estimation

Point estimation [Point (sample statistic) + Performance]

- Consistent estimator/estimate $\hat{\theta}$

  R1) It is a point (sample statistic) on the parameter space.
  R2) It tends to the true θ_0, as $n \to \infty$.

- Unbiased estimator/estimate $\hat{\theta}$

  R1) It is a point (sample statistic) on the parameter space.
  R2) It is unbiased $E\hat{\theta} = \theta_0$.

Distribution estimation [(Sample-dependent) dist. function + Performance]

- Confidence distribution

  R1) It is a (sample-dependent) dist. function on the parameter space.
  R2) It ensures coverage rates of confidence int'vls/regions of all levels.

  ▶ Performance measurement for CD: frequentist probability coverage

    o Simple and intuitive interpretation & relating to reproducibility

  ▶ Other "distribution estimators"? — perhaps future concepts?

CD — a unifying concept for distributional inference

- Wide range of examples: bootstrap distribution, (normalized) likelihood function, empirical likelihood, p-value functions, fiducial distributions, some informative priors and Bayesian posteriors, among others

Our understanding/interpretation: *Any approach, regardless of being frequentist, fiducial or Bayesian, can potentially be unified under the concept of confidence distributions, as long as it can be used to build confidence intervals of all levels, exactly or asymptotically.*

- May help bridge some gap between different statistical paradigms...

Quotes on confidence distribution

- Efron (1998, *Statist. Sci.*): **Bootstrap distributions** are "distribution estimators" and "**confidence distributions**."
  - "... but here is a safe prediction for the 21st century: ... I believe there is a good chance that ... something like fiducial inference will play an important role ... Maybe Fisher's biggest blunder will become a big hit in the 21st century!"
- Discussion article on CD (Xie & Singh 2013, *Int. Stat. Rev.*)
  - Cox (2013, *Int. Stat. Rev.*): The CD approach is aimed "to provide simple and interpretable summaries of what can reasonably be learned from data (and an assumed model)."
  - Efron (2013, *Int. Stat. Rev.*): The CD development is "a grounding process" to help solve "perhaps the most important unresolved problem in statistical inference" on "the use of Bayes theorem in the absence of prior information."

More examples

Example A: (Normal Mean and Variance) Assume $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$.

- Variance σ^2 is known:

  $\Diamond$ $\mathrm{H}_{\Phi}(\mu) = \Phi\left(\frac{\sqrt{n}(\mu - \bar{X})}{\sigma}\right)$ (i.e., $N(\bar{X}, \sigma/\sqrt{n})$) is a CD for μ

- Variance σ^2 is known:

  $\Diamond$ $\mathrm{H}_{t}(\mu) = F_{t_{n-1}}\left(\frac{\sqrt{n}(\mu - \bar{X})}{s}\right)$ is a CD for μ;

  $\Diamond$ $\mathrm{H}_{\chi^2}(\theta) = 1 - F_{\chi^2_{n-1}}\Big((n-1)s^2/\theta\Big)$ is a CD for σ^2

  (Here, $F_{t_{n-1}}$ and $F_{\chi^2_{n-1}}$ are the cumulative distribution function of t_{n-1} and χ^2_{n-1} distribution, respectively.)

- Asymptotic CD are also available in both cases

More examples

Example B: (Bivariate normal correlation) Let ρ denote the correlation coefficient of a bivariate normal population; r be the sample version.

- Fisher's z

$$z = \frac{1}{2}\log\frac{1+r}{1-r}$$

has the limiting distribution $N\left(\frac{1}{2}\log\frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right)$

$\Longrightarrow$

$$\mathrm{H}_n(\rho) = 1 - \Phi\left(\sqrt{n-3}\left(\frac{1}{2}\log\frac{1+r}{1-r} - \frac{1}{2}\log\frac{1+\rho}{1-\rho}\right)\right)$$

is an asymptotic CD for ρ, when sample size $n \to \infty$.

Many more examples – bootstrap distributions, p-value functions, (normalized) likelihood function, (normalized) empirical likelihood, Bayesian posteriors (often), fiducial distributions ...

- **As long as can be used to create confidence intervals of all levels** – (Parametric & nonpara.; normal & nonnormal; exact & asymptotic ...)

Why confidence distribution (& related inference)?

Powerful all-purpose inference tool: informative, flexible, effective, versatile...

- (CD vs Confidence Interval) CD is more "flexible" and provides a "good summary" (more informative) of sample data and model (Cox 1958, 2013)
- (Unification I) A unifying platform to bridge BFF (Bayesian, Fiducial and Frequestist) inferences
- (Added values in applications) A framework supporting new methodology developments beyond conventional approaches
  - New prediction approaches (predictive distributions)
  - New testing methods (p**-value by CD; C-factor**)
  - New simulation schemes (asymptotic/exact)
  - Fusion learning/combining information from diverse sources (Unification II)

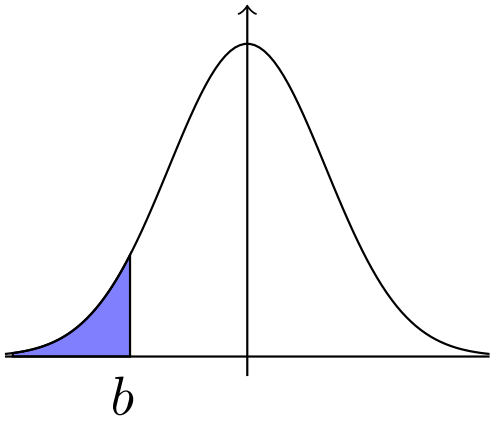
Normal Example: CD and p-value

Sample data: $x_1, \cdots, x_n \overset{ind}{\sim} N(\theta, \sigma^2)$; $\bar{x}_n =$ sam. mean; $s_n^2 =$ sam. variance

- A CD for θ is

$$\mathrm{H}_n(b) = F_{t_{n-1}}\left(\frac{b - \bar{x}_n}{s_n/\sqrt{n}}\right)$$

where $F_{t_{n-1}}$ is the cumulative distribution function of the t_{n-1}-distribution.



- The p-value
for one-sided test: $H_0 : \theta \leq b$ versus $H_A : \theta > b$.

$$p(b) = \sup_{\theta \leq b} \Pr(\bar{X}_n \geq \bar{x}_n) = \ldots = F_{t_{n-1}}\left(\frac{b - \bar{x}_n}{s_n/\sqrt{n}}\right) = \int_{-\infty}^{b} d\mathrm{H}_n(\theta),$$

where $d\mathrm{H}_n$ is the CD density function.

CD direct support and one-sided test

Definition (CD direct support)

Let $E \subseteq \Theta$. The *CD direct support (evidence)* of E is defined as

$$S_n(E) = \int_{\theta \in E} d\mathrm{H}_n(\theta). \tag{3}$$

- If $\mathrm{H}_n(\theta)$ is the Bayesian posterior, then $S_n(E) =$ posterior probability of E

Lemma

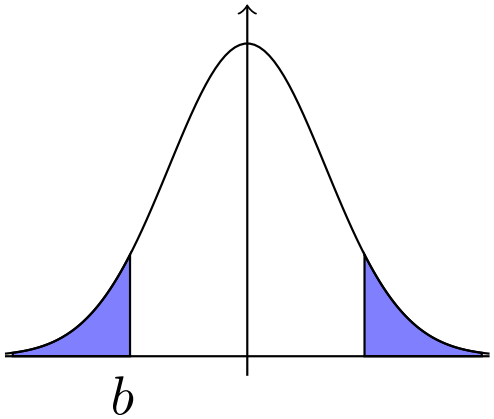
If Θ_0 is of the type $(-\infty, b]$ or $[b, \infty)$, then the CD direct support $S_n(\Theta_0)$ is a p-value for one-sided test $H_0 : \theta \in \Theta_0$ versus $H_A : \theta \notin \Theta_0$.

Normal Example: p-value for two-sided test

Sample data: $x_1, \cdots, x_n \overset{ind}{\sim} N(\theta, \sigma^2)$; $\bar{x}_n =$ sam. mean; $s_n^2 =$ sam. variance

- A CD for θ is

$$\mathrm{H}_n(b) = F_{t_{n-1}}\left(\frac{b - \bar{x}_n}{s_n/\sqrt{n}}\right)$$



- The p-value for two-sided test: $H_0 : \theta = b$ versus $H_A : \theta \neq b$.

$$\begin{aligned} p(b) &= P_{\theta=b}(|\bar{X}_n| \geq |\bar{x}_n|) \\ &= 2\min\left\{F_{t_{n-1}}\left(\frac{b - \bar{x}_n}{s_n/\sqrt{n}}\right), F_{t_{n-1}}\left(\frac{\bar{x}_n - b}{s_n/\sqrt{n}}\right)\right\} \\ &= 2\min\{\mathrm{H}_n(b), 1 - \mathrm{H}_n(b)\}. \end{aligned}$$

CD indirect support and two-sided test

Definition (CD indirect support)

The indirect support (or evidence) of a subset $E \subset \Theta$ is defined as

$$S_n^{IND}(E) = \inf_{\theta_0 \in E} 2\min\{\mathrm{H}_n(\theta_0), 1 - \mathrm{H}_n(\theta_0)\}. \tag{4}$$

- When E is a singleton, say $\{\theta_0\}$, immediately we have

$$S_n^{IND}(\theta_0) \equiv S_n^{IND}(\{\theta_0\}) = 2\min\{\mathrm{H}_n(\theta_0), 1 - \mathrm{H}_n(\theta_0)\}. \tag{5}$$

  which is the value of confidence curve at θ_0. Refs: Birnbaum (1961), Fraser (1991), Xie & Singh (2013).

Lemma

If Θ_0 is a singleton $\{b\}$, then the CD indirect support $S_n^{IND}(\Theta_0)$ is a p-value for two-sided test $H_0 : \theta \in \Theta_0$ versus $H_A : \theta \notin \Theta_0$.

In general, construct a p-value based on CD

Definition

Let

$$p(\boldsymbol{X}_n, \Theta_0) = S_n(\Theta_0) + \inf_{u \in \Theta_0} S_n^{IND}(u). \tag{6}$$

- p$(\boldsymbol{X}_n, \Theta_0)$ is a combination of two parts:

  $=$ direct evidence to Θ_0 + indirect evidence to Θ_0

  $=$ CD probability of Θ_0 + adjustment for evidence against Θ_0^c

  (my direct friends) + (enemy of my enemy is also my friend)

$$= \begin{cases} S_n(\Theta_0), & \text{for a one-sided test;} \\ S_n^{IND}(\Theta_0) & \text{for a two-sided test;} \\ \text{combination of two parts} & \text{for an interval-type null.} \end{cases}$$

Theoretical support

[Regularity Condition]: As $n \to \infty$, for any finite θ_1, θ_2 and positive ϵ, δ,

$$\sup_{\theta \in [\theta_1, \theta_2]} \mathsf{P}_\theta(\max\{\mathrm{H}_n(\theta - \epsilon), 1 - \mathrm{H}_n(\theta + \epsilon)\} > \delta) \to 0, \tag{7}$$

□ $\mathrm{H}_n(\cdot)$ shrinks to 0, uniformly, for θ in a compact set.

Lemma

(i) Let Θ_0 be of the type $(-\infty, b]$ or $[b, \infty)$. Then
$$\sup_{\theta \in \Theta_0} P_\theta\left(p(\boldsymbol{X}_n, \Theta_0) \leq \alpha\right) = \alpha.$$
(ii) Let Θ_0 be a singleton $\{b\}$, $P_{\theta=b}(p(\boldsymbol{X}_n, \Theta_0)) \leq \alpha) = \alpha$.
(iii) Let $\Theta_0 = [a, b]$, $a, b \in \mathcal{R}$. If the regularity condition holds, then
$$\sup_{\theta \in \Theta_0} P_\theta(p(\boldsymbol{X}_n, \Theta_0) \leq \alpha) \to \alpha, \ as\ n \to \infty.$$

Theorem

Under some mild conditions, the mapping $p(\boldsymbol{X}_n, \Theta_0)$ provides a valid p-value (i.e., satisfies the two requirements (I) & (II) in Definition (A))

General comments on p-value by CD

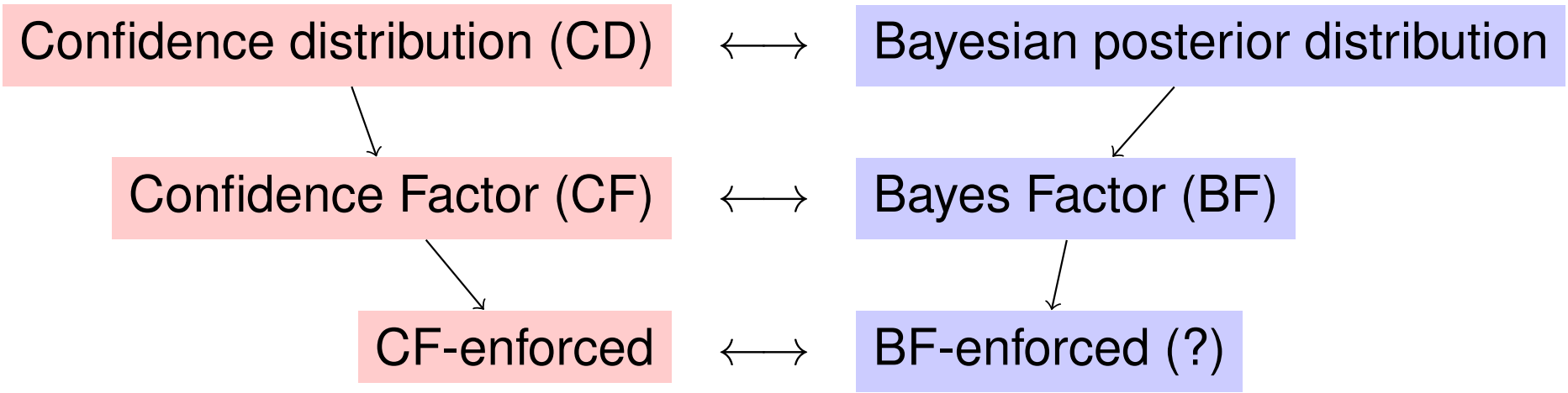
Applicable for a wide range of hypothesis testing problems:

- Some examples of nonstandard tests:
  - Tests with interval-type null hypotheses. Refs: Hodges & Lehmann (1954), Balci & Sargent (1981, 1982), Berger & Delampady (1987), Rousseau (2006);
  - Intersection-union test $H_0 : \theta \in \bigcup_{i=1}^{K} \Theta_{0i} \leftrightarrow H_1 : \theta \in \bigcap_{i=1}^{K} \Theta_{0i}^c$.
    - One important example is the bio-equivalence test: $H_0 : \theta \in (-\infty, \theta_l] \cup [\theta_u, \infty) \leftrightarrow H_1 : \theta \in (\theta_l, \theta_u)$. Refs: Schuirmann (1981, 1987), Anderson & Hauck (1983), Berger & Hsu (1996)
- Results uphold for vector parameter θ!

Ongoing work: confidence factor (CF)

For more complex test problems (e.g. select nonnested models, etc.) –

- We develop a frequentist analog of Bayes factor (BF), called confidence factor (CF)



Real Data Example I: Bio-equivalence test

- Consider a two-period, crossover designed bio-equivalence study provided in Chow & Liu (2008).
  - Compare test (T) and reference (R) formulations of a drug product
  - 24 healthy volunteers equally in two groups (Sequence I & II)

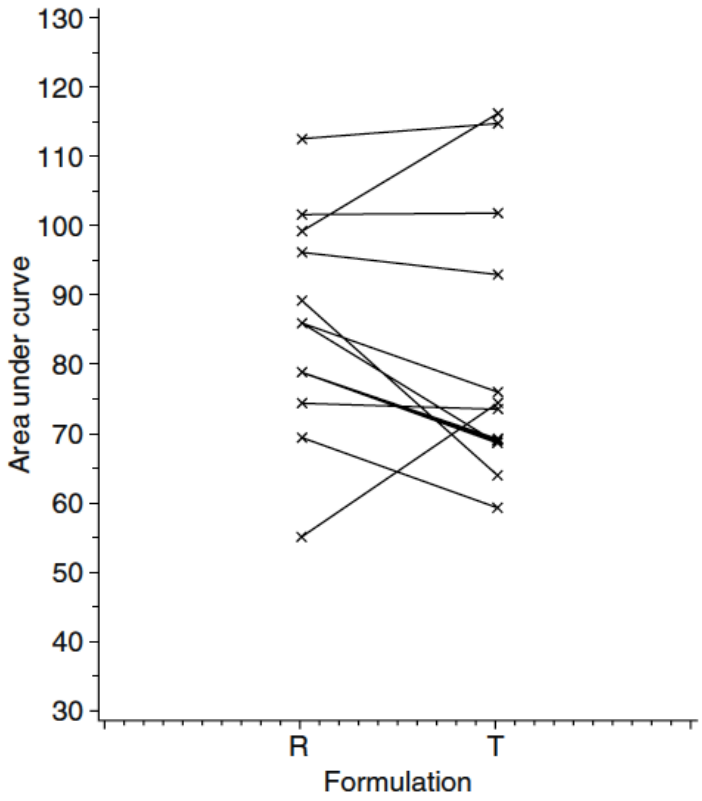
	Period I	Period II
Sequence I	R	T
Sequence II	T	R

  - AUC (area under curve) values from 0 to 32 hours were calculated using the trapezoidal method.
- Let μ_T, μ_R be the population means of AUC from T and R.
- The bio-equivalence test of interest is

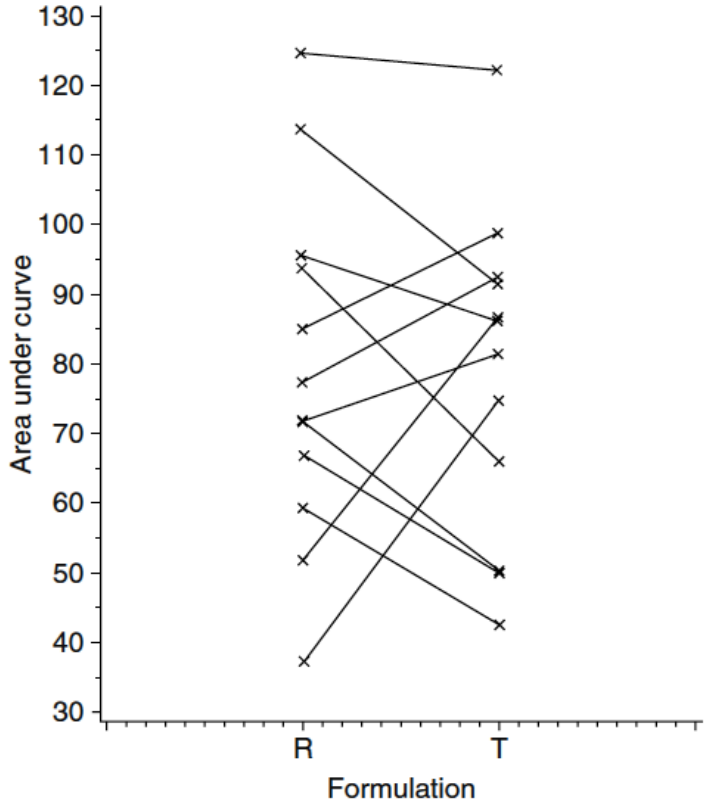
$$H_0 : \theta \in (-\infty, \theta_l] \cup [\theta_u, \infty) \text{ versus } H_1 : \theta \in (\theta_l, \theta_u). \tag{8}$$

where $\theta = \mu_T - \mu_R$, $\theta_l = -16.5$, $\theta_u = 16.5$.

Real Data Example I: Data



Subject profiles sequence = 1.



Subject profiles sequence = 2.

Real Data Example I: p-value approach

- A preliminary analysis of the data shows that there is no significant Carryover Effects or Period Effects [omit].
- A CD of θ is

$$H_{TR}(\theta) = F_{t_{n_1+n_2-1}}\left(\frac{\theta - (\bar{Y}_T - \bar{Y}_R)}{\hat{\sigma}_d\sqrt{1/n_1 + 1/n_2}}\right).$$

  - $\bar{Y}_T$, $\bar{Y}_R$: the sample means for the test and reference formulation;
  - $\hat{\sigma}_d^2$: the pooled sample standard deviation of paired difference;
  - $F_{t_{n_1+n_2-1}}$: CDF of a central student *t* distribution with df. $(n_1 + n_2 - 2)$.

- By (6), the p-value of (8) is $\int_{(-\infty,-\theta_l]\cup[\theta_u,\infty)} dH_{TR}(u)$.

Real Data Example I: Results and Discussions

- Based on the sample data presented, $\bar{Y}_T = 80.272$, $\bar{Y}_R = 82.559$, $\hat{\sigma}_d^2 = 83.623$. The p-value is then 0.000515.
- The bio-equivalence of the test and reference formulations can be claimed based on small p-value. (the FDA bio-equivalence guidelines suggests 0.05 as the significance level.)
- Remarks:
  - The CD-based p-value is easy to implement, especially with different θ_l, θ_u.
  - Alternatively, θ can be chosen as the ratio μ_T/μ_R.
  - Log-transformation is often conducted on AUC values.
  - An alternative way of constructing p-value is by reporting the maximized one-sided p-values based on the two one-sided intervals as the overall p-value (Feng et al. 2006), i.e.,

$$\max\left\{\int_{(-\infty,-16.51]} dH_{TR}(u), \int_{[16.51,\infty)} dH_{TR}(u)\right\} = 0.000479.$$

Real Data Example II: Validation of simulation models

- One of the most important steps in the development of a simulation model is determining whether the simulation model is an accurate representation of the system being studied (Balci & Sargent 1981).
- Specifically, one study is to determine whether the model represents a single server queuing system with two variables:
  - X_1: the average queue length for the first 500 customers;
  - X_2: the average waiting time for the first 500 customers.
- The null hypothesis that model is valid for the acceptable range of accuracy under the set of experimental conditions is specified as

$$H_0 : |\mu_1^d| \leq 0.154, |\mu_2^d| \leq 0.28 \text{ versus } H_1 : |\mu_1^d| > 0.154, |\mu_2^d| > 0.28;$$

  - μ_i^d is the population mean of the differences between the paired observations on X_i from the model and system, $i = 1, 2$.

Real Data Example II: Dataset

- The sample size is 15. An estimate of the variance-covariance matrix of differences between the paired observations on the model and system response variables was given as $\Sigma_U = \begin{pmatrix} 0.2162 & 0.4147 \\ 0.4147 & 0.7959 \end{pmatrix}$.

Difference on X_1	-0.255	0.201	0.008	0.014	-0.146
	0.321	0.097	0.679	0.361	0.269
	0.153	0.329	0.283	0.657	-0.314
Difference on X_2	-0.631	0.372	-0.128	0.035	-0.390
	0.639	0.303	1.240	0.398	0.505
	0.207	0.465	0.438	0.905	-0.458

Table 1 : Selected sample data for validation of simulation model.

Real Data Example II: p-value mapping

- Denote $H_0 : \boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0$, based on bootstrap method and data depth, the p-value mapping can be constructed as

$$p_n(\boldsymbol{\Theta}_0) = P^*(\boldsymbol{\theta}_n^* \in \boldsymbol{\Theta}_0) + P_{G_n^*}(\boldsymbol{\theta}_n^* : D(G_n^*; \boldsymbol{\theta}_n^*) \leq \inf_{\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0} D(G_n^*; \boldsymbol{\theta}_0)).$$

  - $\boldsymbol{\theta}_n^*$ denotes a bootstrap estimate of $\boldsymbol{\theta}_0$. G_n^* denotes the sampling distribution of $\boldsymbol{\theta}_n^*$.
  - $P^*(\boldsymbol{\theta}_n^* \in \boldsymbol{\Theta}_0)$ is the empirical strength probability of $\boldsymbol{\Theta}_0$.
  - Closely related to Hotelling T^2-test, choose *Mahalanobis depth* (Mahalanobis 1936),

$$D(G_n; w) = \left[1 + (w - \mu_U)' \Sigma_U^{-1} (w - \mu_U)\right]^{-1},$$

    where μ_U and Σ_U are the mean vector and variance-covariance matrix.

- The p-value is 0.486, which indicates that the differences between the model and system are acceptable.

Real Data Example II: p-value mapping

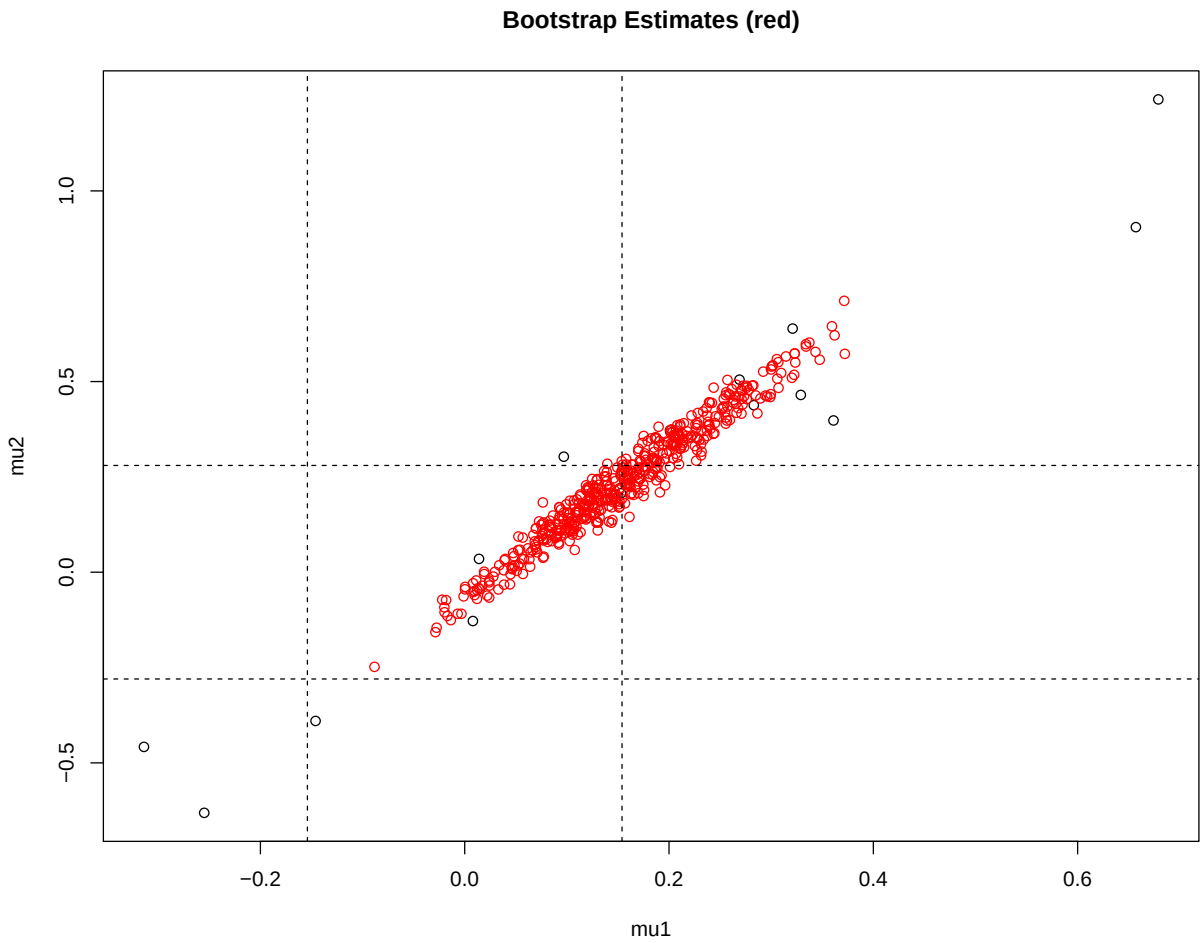


Figure 2 : Red points: bootstrap estimates; Rectangle: the null space.

Summary and discussion

- Proposed a formal p-value definition focusing on performance,
  - Broadens and adds flexibility to the concept of p-value
- Provided a new method to construct a p-value by CD.
  - p-value for $H_0 : \theta \in \Theta_0$ can be interpreted as the CD support of Θ_0.
  - Large p-value indicates large support, therefore, the meaning of non-small p-value is clarified.
  - For Interval-type null hypothesis testing problems (including one-sided and two-sided tests), a unified construction of p-value is proposed.
- More complex cases are also discussed.

Overall: Help address/mitigate some of the concerns/confusions associated p-value

References I

Abell, M. L., Braselton, J. P. & Rafter, J. A. (1999), *Statistics with Mathematica*, Academic Press.

Anderson, S. & Hauck, W. W. (1983), 'A new procedure for testing equivalence in comparative bioavailability and other clinical trials.', *Communications in Statistics—Theory and Methods* **12**, 2663–2692.

Balci, O. & Sargent, R. (1981), 'A methodology for cost-risk analysis in the statistical validation of simulation models', *Communications of the ACM* **24**(4), 190–197.

Balci, O. & Sargent, R. G. (1982), 'Some examples of simulation model validation using hypothesis testing', *Proceedings of the 14th conference on Winter Simulation* **2**, 621–629.

Berger, J. O. (2003), 'Could fisher, jeffreys and neyman have agreed on testing?', *Statistical Science* **18**, 1–32.

Berger, J. O. & Delampady, M. (1987), 'Testing precise hypotheses', *Statistical Science* **2**(3), 317–352.

Berger, R. L. & Boos, D. D. (1994), 'P values maximized over a confidence set for the nuisance parameter', *Journal of the American Satistical Association* **89**(427), 1012–1016.

References II

Berger, R. L. & Hsu, J. C. (1996), 'Bioequivalence trials, intersection-union tests and equivalence confidence sets', *Statistical Science* **11**(4), 283–319.

Birnbaum, A. (1961), 'Confidence curves: An omnibus technique for estimation and testing statistical hypotheses', *Journal of the American Satistical Association* **56**(294), 246–249.

Chavalarias, D., Wallach, J., Li, A. & Ioannidis, J. (2016), 'Evolution of reporting p values in the biomedical literature, 1990-2015', *JAMA* **315**(11), 1141–1148. **URL:** *+ http://dx.doi.org/10.1001/jama.2016.1952*

Chow, S.-C. & Liu, J.-p. (2008), *Design and analysis of bioavailability and bioequivalence studies, third edition*, Chapman and Hall/CRC.

Feng, S., Liang, Q., Kinser, R. D., Newland, K. & Guibaud, R. (2006), 'Testing equivalence between two laboratories or two methods using paired-sample analysis and interval hypothesis testing', *Anal Bioanal Chem* **385**, 975–981.

Fisher, R. (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd.

Fisher, R. (1935), *The design of experiments*, Edinburgh: Oliver and Boyd.

Fisher, R. (1959), *Statistical Methods and Scientific Inference (2nd ed.)*, Edinburgh: Oliver and Boyd.

References III

Fraser, D. (1991), 'Statistical inference: Likelihood to significance', *Journal of the American Satistical Association* **86**, 258–265.

Hodges, J. L. & Lehmann, E. L. (1954), 'Testing the approximate validity of statistical hypotheses', *Journal of the Royal Statistical Society. Series B (Methodological)* **16**(2), 261–268.

Lehmann, E. L. & Romano, J. P. (2005), *Testing Statistical Hypotheses*, Springer-Verlag New York.

Liu, R. Y. & Singh, K. (1997), 'Notions of limiting p-values based on data depth and bootstrap', *Journal of the American Satistical Association* **91**, 266–277.

Mahalanobis, P. C. (1936), 'On the generalized distance in statistics', *Proceedings of the National Academy of India* **12**, 49–55.

Rousseau, J. (2006), 'Approximating interval hypothesis: p-values and bayes factors', *ISBA 8th World Meeting on Bayesian Statistics* .

Schuirmann, D. J. (1981), 'On hypothesis testing to determine if the mean of a normal distribution is continued in a known interval.', *Biometrics* **37**, 617. [abstract]].

Schuirmann, D. J. (1987), 'A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability.', *Journal of Pharmacokinetics and Biopharmaceutics* **15**, 657–680.

References IV

Shafer, G., Shen, A., Vereshchagin, N. & Vovk, V. (2011), 'Test martingales, bayes factors and p-values', *Statistical Science* **26**(1), 84–101.

Wasserstein, R. L. & Lazar, N. A. (2016), 'The ASA's statement on p-values: context, process, and purpose', *The American Statistician* .

Xie, M. & Singh, K. (2013), 'Confidence distribution, the frequentist distribution estimator of a parameter: A review', *International Statitical Review* **81**, 3–39.