Testing One Hypothesis Twice in Observational Studies

Paul R. Rosenbaum

Wharton School, University of Pennsylvania

April 2013

- Structure of the talk

- Structure of the talk
- An entity who was all-knowing and all-wise could reach firmer conclusions from observational data than we can.

- Structure of the talk
- An entity who was all-knowing and all-wise could reach firmer conclusions from observational data than we can.
- With a little statistical theory, we can be somewhat-knowing and somewhat-wise.

- Structure of the talk
- An entity who was all-knowing and all-wise could reach firmer conclusions from observational data than we can.
- With a little statistical theory, we can be somewhat-knowing and somewhat-wise.
- Add a little data, properly analyzed, and we can be almost as effective as that all-knowing, all-wise entity.

Basis for this talk

- Rosenbaum, P. R. (2012), “Testing one hypothesis twice in observational studies,” *Biometrika*, 99, 763-774.
- Rosenbaum, P. R. (2012), “An exact, adaptive test with superior design sensitivity in an observational study of treatments for ovarian cancer,” *AOAS*, 6, 83-105.
- Rosenbaum, P. R. (2011), “A new U-statistic with superior design sensitivity in matched observational studies,” *Biometrics*, 67, 1017-1027.
- Rosenbaum, P. R. (2010), “Design sensitivity and efficiency in observational studies,” *JASA*, 105, 692-702.

Terminology: Some familiar terms

- **Power**: The usual notion: the probability of rejecting the null hypothesis of no effect when there really is an effect.

Terminology: Some familiar terms

- **Power**: The usual notion: the probability of rejecting the null hypothesis of no effect when there really is an effect.
- **Power is**: an aspect of the stochastic process that generated the data and particular methods of analysis.

Terminology: Some familiar terms

- **Power**: The usual notion: the probability of rejecting the null hypothesis of no effect when there really is an effect.
- **Power is**: an aspect of the stochastic process that generated the data and particular methods of analysis.
- **Observational study**: Study of treatment effects when subjects are not randomized to treatment or control.

Terminology: Some familiar terms

- **Power**: The usual notion: the probability of rejecting the null hypothesis of no effect when there really is an effect.
- **Power is**: an aspect of the stochastic process that generated the data and particular methods of analysis.
- **Observational study**: Study of treatment effects when subjects are not randomized to treatment or control.
- **Issue**: Without randomization, treated and control groups may not be comparable. Adjust for observed covariates, perhaps by matching.

Terminology: Some familiar terms

- **Power**: The usual notion: the probability of rejecting the null hypothesis of no effect when there really is an effect.
- **Power is**: an aspect of the stochastic process that generated the data and particular methods of analysis.
- **Observational study**: Study of treatment effects when subjects are not randomized to treatment or control.
- **Issue**: Without randomization, treated and control groups may not be comparable. Adjust for observed covariates, perhaps by matching.
- **Problem**: Adjusting for observed covariates does not typically control unobserved covariates.

Terminology: Some familiar terms

- **Power**: The usual notion: the probability of rejecting the null hypothesis of no effect when there really is an effect.
- **Power is**: an aspect of the stochastic process that generated the data and particular methods of analysis.
- **Observational study**: Study of treatment effects when subjects are not randomized to treatment or control.
- **Issue**: Without randomization, treated and control groups may not be comparable. Adjust for observed covariates, perhaps by matching.
- **Problem**: Adjusting for observed covariates does not typically control unobserved covariates.
- **Sensitivity analysis**: Asks what an unobserved covariate would have to be like to alter the conclusions of a naïve analysis that presumes adjustments for observed covariates suffice. Cornfield et al. (1959).

Terminology: What is Design Sensitivity?

- **Design sensitivity**: Speaking informally, the design sensitivity is the limiting sensitivity to unobserved bias as the sample size increases.

Terminology: What is Design Sensitivity?

- **Design sensitivity**: Speaking informally, the design sensitivity is the limiting sensitivity to unobserved bias as the sample size increases.
- **Design sensitivity is**: (like power and unlike sensitivity analysis) an aspect of the stochastic process that generated the data and particular methods of analysis, evaluated when the sample size is large.

Terminology: What is Design Sensitivity?

- **Design sensitivity**: Speaking informally, the design sensitivity is the limiting sensitivity to unobserved bias as the sample size increases.
- **Design sensitivity is**: (like power and unlike sensitivity analysis) an aspect of the stochastic process that generated the data and particular methods of analysis, evaluated when the sample size is large.
- **Design sensitivity is**: a number, $\widetilde{\Gamma}$, such that, as the sample size increases, the study will eventually be insensitive to biases smaller than $\widetilde{\Gamma}$ and sensitive to biases larger than $\widetilde{\Gamma}$.

Terminology: What is Design Sensitivity?

- **Design sensitivity**: Speaking informally, the design sensitivity is the limiting sensitivity to unobserved bias as the sample size increases.
- **Design sensitivity is**: (like power and unlike sensitivity analysis) an aspect of the stochastic process that generated the data and particular methods of analysis, evaluated when the sample size is large.
- **Design sensitivity is**: a number, $\widetilde{\Gamma}$, such that, as the sample size increases, the study will eventually be insensitive to biases smaller than $\widetilde{\Gamma}$ and sensitive to biases larger than $\widetilde{\Gamma}$.
- **In particular**: in large samples, the limiting power of a sensitivity analysis is determined by the design sensitivity.

Main idea of this talk

- Lacking theoretical guidance, we tend to select statistical methods for use in observational studies based on their power/efficiency in randomized experiments.

Main idea of this talk

- Lacking theoretical guidance, we tend to select statistical methods for use in observational studies based on their power/efficiency in randomized experiments.
- This turns out to be a mistake.

Main idea of this talk

- Lacking theoretical guidance, we tend to select statistical methods for use in observational studies based on their power/efficiency in randomized experiments.
- This turns out to be a mistake.
- A highly efficient method for detecting small treatment effects in randomized experiments need not, and often does not, have the highest power in a sensitivity analysis or the largest design sensitivity.

Main idea of this talk

- Lacking theoretical guidance, we tend to select statistical methods for use in observational studies based on their power/efficiency in randomized experiments.
- This turns out to be a mistake.
- A highly efficient method for detecting small treatment effects in randomized experiments need not, and often does not, have the highest power in a sensitivity analysis or the largest design sensitivity.
- That is, the best procedure assuming that an observational study is effectively a randomized experiment need not be the best procedure under more realistic assumptions

Main idea of this talk

- Lacking theoretical guidance, we tend to select statistical methods for use in observational studies based on their power/efficiency in randomized experiments.
- This turns out to be a mistake.
- A highly efficient method for detecting small treatment effects in randomized experiments need not, and often does not, have the highest power in a sensitivity analysis or the largest design sensitivity.
- That is, the best procedure assuming that an observational study is effectively a randomized experiment need not be the best procedure under more realistic assumptions
- Will present a family of U-statistics for matched pairs that includes Wilcoxon's signed rank statistic, but other members of this family have much higher power in a sensitivity analysis and higher design sensitivity $\widetilde{\Gamma}$.

Main idea of this talk

- Lacking theoretical guidance, we tend to select statistical methods for use in observational studies based on their power/efficiency in randomized experiments.
- This turns out to be a mistake.
- A highly efficient method for detecting small treatment effects in randomized experiments need not, and often does not, have the highest power in a sensitivity analysis or the largest design sensitivity.
- That is, the best procedure assuming that an observational study is effectively a randomized experiment need not be the best procedure under more realistic assumptions
- Will present a family of U-statistics for matched pairs that includes Wilcoxon's signed rank statistic, but other members of this family have much higher power in a sensitivity analysis and higher design sensitivity $\widetilde{\Gamma}$.
- To make full use of this fact, one may have to use multiple tests of one hypothesis, correcting for multiple testing.

Example: Lead and Smoking in NHANES 2008

- $i = 1, \ldots, I = 679$ matched pairs, $j = 1, 2$, one treated, one control.

Example: Lead and Smoking in NHANES 2008

- $i = 1, \ldots, I = 679$ matched pairs, $j = 1, 2$, one treated, one control.
- **Treatment**: Daily smoking of ≥ 10 cigarettes (median $= 20$ cigarettes) every day for the last 30 days ($Z_{ij} = 1$) versus no smoking in the last 30 days and fewer than 100 lifetime cigarettes ($Z_{ij} = 0$).

Example: Lead and Smoking in NHANES 2008

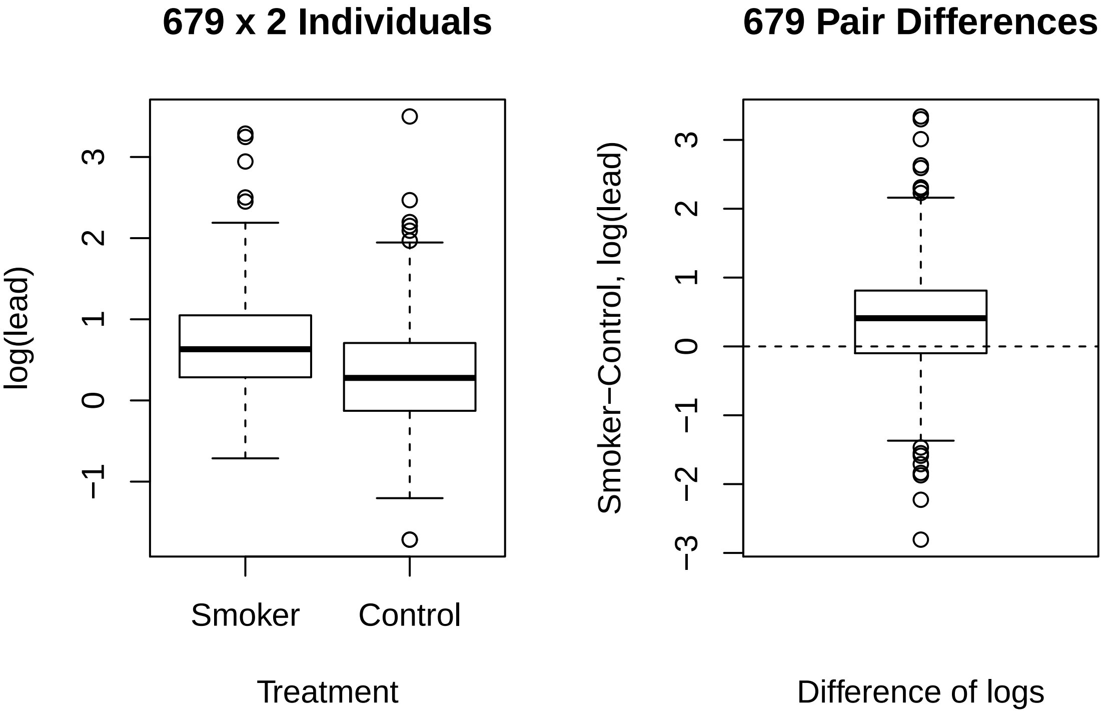
- $i = 1, \ldots, I = 679$ matched pairs, $j = 1, 2$, one treated, one control.
- **Treatment**: Daily smoking of ≥ 10 cigarettes (median $= 20$ cigarettes) every day for the last 30 days ($Z_{ij} = 1$) versus no smoking in the last 30 days and fewer than 100 lifetime cigarettes ($Z_{ij} = 0$).
- **Response**: Blood lead levels in $\mu g / dL$, perhaps transformed, R_{ij}. (Until late in the talk, R_{ij} is the log of the lead level.)

Example: Lead and Smoking in NHANES 2008

- $i = 1, \ldots, I = 679$ matched pairs, $j = 1, 2$, one treated, one control.
- **Treatment**: Daily smoking of ≥ 10 cigarettes (median $= 20$ cigarettes) every day for the last 30 days ($Z_{ij} = 1$) versus no smoking in the last 30 days and fewer than 100 lifetime cigarettes ($Z_{ij} = 0$).
- **Response**: Blood lead levels in $\mu g / dL$, perhaps transformed, R_{ij}. (Until late in the talk, R_{ij} is the log of the lead level.)
- **Matched for**: Gender, age, race, education level,household income level, $\mathbf{x}_{ij}$, $\mathbf{x}_{i1} = \mathbf{x}_{i2}$.

Example: Lead and Smoking in NHANES 2008

- $i = 1, \ldots, I = 679$ matched pairs, $j = 1, 2$, one treated, one control.
- **Treatment**: Daily smoking of ≥ 10 cigarettes (median $= 20$ cigarettes) every day for the last 30 days ($Z_{ij} = 1$) versus no smoking in the last 30 days and fewer than 100 lifetime cigarettes ($Z_{ij} = 0$).
- **Response**: Blood lead levels in $\mu g / dL$, perhaps transformed, R_{ij}. (Until late in the talk, R_{ij} is the log of the lead level.)
- **Matched for**: Gender, age, race, education level,household income level, $\mathbf{x}_{ij}$, $\mathbf{x}_{i1} = \mathbf{x}_{i2}$.
- **Sensitivity to**: an unobserved covariate u_{ij}, possibly with $u_{i1} \neq u_{i2}$.


679 x 2 Individuals
log(lead)
−1
0
1
2
3
Smoker
Control
Treatment
679 Pair Differences
Smoker−Control, log(lead)
−3
−2
−1
0
1
2
3
Difference of logs

Notation

- There are I pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one treated, $Z_{ij} = 1$, the other control, $Z_{ij} = 0$, with $Z_{i1} + Z_{i2} = 1$. $\mathcal{Z}$ is the event $Z_{i1} + Z_{i2} = 1$, $i = 1, \ldots, I$.

Notation

- There are I pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one treated, $Z_{ij} = 1$, the other control, $Z_{ij} = 0$, with $Z_{i1} + Z_{i2} = 1$. $\mathcal{Z}$ is the event $Z_{i1} + Z_{i2} = 1$, $i = 1, \ldots, I$.
- Matched for observed covariates, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$. Possibly differing in term of an unmeasured covariate, $u_{i1} \neq u_{i2}$.

Notation

- There are I pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one treated, $Z_{ij} = 1$, the other control, $Z_{ij} = 0$, with $Z_{i1} + Z_{i2} = 1$. $\mathcal{Z}$ is the event $Z_{i1} + Z_{i2} = 1$, $i = 1, \ldots, I$.
- Matched for observed covariates, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$. Possibly differing in term of an unmeasured covariate, $u_{i1} \neq u_{i2}$.
- Randomized paired experiment, Z_{i1}, $i = 1, \ldots, I$, determined by I independent flips of a coin.

Notation

- There are I pairs, $i = 1, \dots, I$, of two subjects, $j = 1, 2$, one treated, $Z_{ij} = 1$, the other control, $Z_{ij} = 0$, with $Z_{i1} + Z_{i2} = 1$. $\mathcal{Z}$ is the event $Z_{i1} + Z_{i2} = 1$, $i = 1, \dots, I$.
- Matched for observed covariates, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$. Possibly differing in term of an unmeasured covariate, $u_{i1} \neq u_{i2}$.
- Randomized paired experiment, Z_{i1}, $i = 1, \dots, I$, determined by I independent flips of a coin.
- Naïve analysis of an observational study assumes adjustments for $\mathbf{x}$ suffice to remove bias.

Notation

- There are I pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one treated, $Z_{ij} = 1$, the other control, $Z_{ij} = 0$, with $Z_{i1} + Z_{i2} = 1$. $\mathcal{Z}$ is the event $Z_{i1} + Z_{i2} = 1$, $i = 1, \ldots, I$.
- Matched for observed covariates, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$. Possibly differing in term of an unmeasured covariate, $u_{i1} \neq u_{i2}$.
- Randomized paired experiment, Z_{i1}, $i = 1, \ldots, I$, determined by I independent flips of a coin.
- Naïve analysis of an observational study assumes adjustments for $\mathbf{x}$ suffice to remove bias.
- Sensitivity analysis asks: What u would have to be like to alter the conclusions of the naïve analysis?

Causal effects

- Neyman (1923) and Rubin (1974): Each subject ij has two potential responses, r_{Tij} if treated, $Z_{ij} = 1$, or r_{Cij} if control, $Z_{ij} = 0$;

Causal effects

- Neyman (1923) and Rubin (1974): Each subject ij has two potential responses, r_{Tij} if treated, $Z_{ij} = 1$, or r_{Cij} if control, $Z_{ij} = 0$;
- Observed response from ij is $R_{ij} = Z_{ij}\, r_{Tij} + (1 - Z_{ij})\, r_{Cij}$.

Causal effects

- Neyman (1923) and Rubin (1974): Each subject ij has two potential responses, r_{Tij} if treated, $Z_{ij} = 1$, or r_{Cij} if control, $Z_{ij} = 0$;
- Observed response from ij is $R_{ij} = Z_{ij}\, r_{Tij} + (1 - Z_{ij})\, r_{Cij}$.
- Effect of the treatment, $r_{Tij} - r_{Cij}$, on ij is not observed for any subject.

Causal effects

- Neyman (1923) and Rubin (1974): Each subject ij has two potential responses, r_{Tij} if treated, $Z_{ij} = 1$, or r_{Cij} if control, $Z_{ij} = 0$;
- Observed response from ij is $R_{ij} = Z_{ij}\, r_{Tij} + (1 - Z_{ij})\, r_{Cij}$.
- Effect of the treatment, $r_{Tij} - r_{Cij}$, on ij is not observed for any subject.
- Fisher's sharp null hypothesis of no treatment effect asserts $H_0 : r_{Tij} = r_{Cij}$, for $i = 1, \ldots, I$, $j = 1, 2$.

Causal effects

- Neyman (1923) and Rubin (1974): Each subject ij has two potential responses, r_{Tij} if treated, $Z_{ij} = 1$, or r_{Cij} if control, $Z_{ij} = 0$;
- Observed response from ij is $R_{ij} = Z_{ij}\, r_{Tij} + (1 - Z_{ij})\; r_{Cij}$.
- Effect of the treatment, $r_{Tij} - r_{Cij}$, on ij is not observed for any subject.
- Fisher's sharp null hypothesis of no treatment effect asserts $H_0 : r_{Tij} = r_{Cij}$, for $i = 1, \ldots, I$, $j = 1, 2$.
- Write $\mathcal{F} = \{(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, u_{ij})\,,\ i = 1, \ldots, I,\ j = 1, 2\}$.

Causal effects

- Neyman (1923) and Rubin (1974): Each subject ij has two potential responses, r_{Tij} if treated, $Z_{ij} = 1$, or r_{Cij} if control, $Z_{ij} = 0$;
- Observed response from ij is $R_{ij} = Z_{ij}\, r_{Tij} + \left(1 - Z_{ij}\right)\, r_{Cij}$.
- Effect of the treatment, $r_{Tij} - r_{Cij}$, on ij is not observed for any subject.
- Fisher's sharp null hypothesis of no treatment effect asserts $H_0 : r_{Tij} = r_{Cij}$, for $i = 1, \ldots, I$, $j = 1, 2$.
- Write $\mathcal{F} = \left\{\left(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, u_{ij}\right), \, i = 1, \ldots, I, \, j = 1, 2\right\}$.
- H_0 is false if the treatment has an additive effect, $r_{Tij} - r_{Cij} = \tau$ for all ij, $\tau \neq 0$. (Easily replaced by treatment typically has an additive effect, $r_{Tij} - r_{Cij} = \tau + \xi_{ij}$ where the ξ_{ij} are mutually independent, independent of everything else, symmetric about 0.)

Treated-minus-control pair differences

- In pair i, the observed, treated-minus-control difference in responses is $Y_i = (Z_{i1} - Z_{i2})\,(R_{i1} - R_{i2})$.

Treated-minus-control pair differences

- In pair i, the observed, treated-minus-control difference in responses is $Y_i = (Z_{i1} - Z_{i2})\,(R_{i1} - R_{i2})$.
- If the treatment has an additive effect, $r_{Tij} - r_{Cij} = \tau$ for all ij, then Y_i is

$$\begin{aligned} Y_i &= (Z_{i1} - Z_{i2})\,(r_{Ci1} + Z_{i1}\tau - r_{Ci2} - Z_{i2}\tau) \\ &= \tau + \epsilon_i \text{ where } \epsilon_i = (Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2}) \end{aligned}$$

Treated-minus-control pair differences

- In pair i, the observed, treated-minus-control difference in responses is $Y_i = (Z_{i1} - Z_{i2})\,(R_{i1} - R_{i2})$.
- If the treatment has an additive effect, $r_{Tij} - r_{Cij} = \tau$ for all ij, then Y_i is

$$\begin{aligned} Y_i &= (Z_{i1} - Z_{i2})\,(r_{Ci1} + Z_{i1}\tau - r_{Ci2} - Z_{i2}\tau) \\ &= \tau + \epsilon_i \text{ where } \epsilon_i = (Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2}) \end{aligned}$$

- Looking ahead: A sensitivity analysis is an analysis of $Y_1, \ldots, Y_I$. Efficiency, the power of a sensitivity analysis, the design sensitivity refer to a stochastic model that generated the Y_i, such as $Y_i \sim_{iid} N(\tau, 1)$.

General signed rank statistics

- Let $q_i \geq 0$ be some function of the absolute $|Y_i|$'s with the property that $q_i = 0$ if $|Y_i| = 0$.

General signed rank statistics

- Let $q_i \geq 0$ be some function of the absolute $|Y_i|$'s with the property that $q_i = 0$ if $|Y_i| = 0$.
- Let $\operatorname{sgn}(y) = 1$ or 0 for, respectively $y > 0$ or $y \leq 0$.

General signed rank statistics

- Let $q_i \geq 0$ be some function of the absolute $|Y_i|$'s with the property that $q_i = 0$ if $|Y_i| = 0$.
- Let $\text{sgn}(y) = 1$ or 0 for, respectively $y > 0$ or $y \leq 0$.
- A general signed rank statistic is of the form $T = \sum_{i=1}^{I} \text{sgn}(Y_i)\ q_i$

General signed rank statistics

- Let $q_i \geq 0$ be some function of the absolute $|Y_i|$'s with the property that $q_i = 0$ if $|Y_i| = 0$.
- Let $\mathrm{sgn}\left(y\right) = 1$ or 0 for, respectively $y > 0$ or $y \leq 0$.
- A general signed rank statistic is of the form $T = \sum_{i=1}^{I} \mathrm{sgn}\left(Y_i\right)\ q_i$
- The sign test has $q_i = 1$ whenever $|Y_i| > 0$. Wilcoxon's signed rank test has $q_i = \mathrm{rank}\left(|Y_i|\right)$ if $|Y_i| > 0$.

General signed rank statistics

- Let $q_i \geq 0$ be some function of the absolute $|Y_i|$'s with the property that $q_i = 0$ if $|Y_i| = 0$.
- Let $\mathrm{sgn}\,(y) = 1$ or 0 for, respectively $y > 0$ or $y \leq 0$.
- A general signed rank statistic is of the form $T = \sum_{i=1}^{I} \mathrm{sgn}\,(Y_i)\; q_i$
- The sign test has $q_i = 1$ whenever $|Y_i| > 0$. Wilcoxon's signed rank test has $q_i = \mathrm{rank}\,(|Y_i|)$ if $|Y_i| > 0$.
- Randomization creates the null distribution $\Pr\left(T \mid \mathcal{F},\, \mathcal{Z}\right)$ of T under Fisher's H_0 as the distribution of the sum of I independent random variables taking the values q_i or 0 each with probability $\frac{1}{2}$ if $q_i > 0$ or the value 0 with probability 1 if $q_i = 0$. E.g., the binomial distribution for the sign test or the usual reference distribution for Wilcoxon's test.

Sensitivity model

- A sensitivity analysis asks about the magnitude of departure from $\Pr(Z_{ij} = 1 \mid \mathcal{F}, \mathcal{Z}) = \frac{1}{2}$ that would need to be present to alter the qualitative conclusions of a randomization inference.

Sensitivity model

- A sensitivity analysis asks about the magnitude of departure from $\Pr\left(Z_{ij} = 1 \,|\, \mathcal{F}, \mathcal{Z} \right) = \frac{1}{2}$ that would need to be present to alter the qualitative conclusions of a randomization inference.
- A simple model: In the population prior to matching, subjects have independent treatment assignments with unknown probabilities, $\pi_{ij} = \Pr\left(Z_{ij} = 1 \,|\, \mathcal{F} \right)$, such that two subjects, say ij and ij', with the same observed covariates, $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$, may differ in their odds of treatment by at most a factor of $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij} \left(1 - \pi_{ij'}\right)}{\pi_{ij'} \left(1 - \pi_{ij}\right)} \leq \Gamma \quad \text{whenever } \mathbf{x}_{ij} = \mathbf{x}_{ij'};$$

then condition on $Z_{i1} + Z_{i2} = 1$.

Sensitivity model

- A sensitivity analysis asks about the magnitude of departure from $\Pr(Z_{ij}=1 \mid \mathcal{F}, \mathcal{Z}) = \frac{1}{2}$ that would need to be present to alter the qualitative conclusions of a randomization inference.
- A simple model: In the population prior to matching, subjects have independent treatment assignments with unknown probabilities, $\pi_{ij} = \Pr(Z_{ij}=1 \mid \mathcal{F})$, such that two subjects, say ij and ij', with the same observed covariates, $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$, may differ in their odds of treatment by at most a factor of $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}\,(1-\pi_{ij'})}{\pi_{ij'}\,(1-\pi_{ij})} \leq \Gamma \quad \text{whenever } \mathbf{x}_{ij} = \mathbf{x}_{ij'};$$

  then condition on $Z_{i1} + Z_{i2} = 1$.
- For each $\Gamma \geq 1$, obtain a range of possible inference quantities, point estimates, p-values, etc.

Sensitivity analysis for a general signed rank statistic

- Let $\overline{\overline{T}}$ be the sum of I independent random variables taking the value q_i with probability $\Gamma / \left(1+\Gamma\right)$ or 0 with probability $1 / \left(1+\Gamma\right)$. Define $\overline{T}$ similarly with $\Gamma / \left(1+\Gamma\right)$ and $1 / \left(1+\Gamma\right)$ interchanged.

Sensitivity analysis for a general signed rank statistic

- Let $\overline{\overline{T}}$ be the sum of I independent random variables taking the value q_i with probability $\Gamma/\left(1+\Gamma\right)$ or 0 with probability $1/\left(1+\Gamma\right)$. Define $\overline{T}$ similarly with $\Gamma/\left(1+\Gamma\right)$ and $1/\left(1+\Gamma\right)$ interchanged.
- **Bounds**: Under Fisher's H_0 and the sensitivity model with a fixed $\Gamma \geq 1$:

$$\Pr\left(\overline{T} \geq k \,\middle|\, \mathcal{F}, \mathcal{Z}\right) \leq \Pr\left(T \geq k \,\middle|\, \mathcal{F}, \mathcal{Z}\right) \leq \Pr\left(\overline{\overline{T}} \geq k \,\middle|\, \mathcal{F}, \mathcal{Z}\right) \text{ for all } k,$$

with equality for $\Gamma = 1$. Bounds attained for particular π_{ij}.

Sensitivity analysis for a general signed rank statistic

- Let $\overline{\overline{T}}$ be the sum of I independent random variables taking the value q_i with probability $\Gamma/\left(1+\Gamma\right)$ or 0 with probability $1/\left(1+\Gamma\right)$. Define $\overline{T}$ similarly with $\Gamma/\left(1+\Gamma\right)$ and $1/\left(1+\Gamma\right)$ interchanged.
- **Bounds**: Under Fisher's H_0 and the sensitivity model with a fixed $\Gamma \geq 1$:

$$\Pr\left(\overline{T} \geq k \middle| \mathcal{F}, \mathcal{Z}\right) \leq \Pr\left(T \geq k \middle| \mathcal{F}, \mathcal{Z}\right) \leq \Pr\left(\overline{\overline{T}} \geq k \middle| \mathcal{F}, \mathcal{Z}\right) \text{ for all } k,$$

with equality for $\Gamma = 1$. Bounds attained for particular π_{ij}.

- **Approximate bounds**: As $I \rightarrow \infty$,

$$\Pr\left(\overline{\overline{T}} \geq k \middle| \mathcal{F}, \mathcal{Z}\right) \approx 1 - \Phi\left[\frac{k - \left\{\Gamma/\left(1+\Gamma\right)\right\}\sum_{i=1}^{I} q_i}{\sqrt{\left\{\Gamma/\left(1+\Gamma\right)^2\right\}\sum_{i=1}^{I} q_i^2}}\right] \qquad (1)$$

if $\left(\sum_{i=1}^{I} q_i^2\right) / \left(\max_{1\leq i\leq I} q_i^2\right) \rightarrow \infty$. $\left(\Phi\left(\cdot\right)\right.$ is Normal cdf$\left.\right)$

The new U-statistic, described informally

- **Name**: Fix three integers, m, $\underline{m}$, $\overline{m}$ with $1 \leq \underline{m} \leq \overline{m} \leq m < I$. Then $(m, \underline{m}, \overline{m})$ is the name of one U-statistic.

The new U-statistic, described informally

- **Name**: Fix three integers, m, $\underline{m}$, $\overline{m}$ with $1 \leq \underline{m} \leq \overline{m} \leq m < I$. Then $(m, \underline{m}, \overline{m})$ is the name of one U-statistic.
- **Instances**: $(1, 1, 1)$ is the sign test statistic, $(2, 2, 2)$ is (essentially) Wilcoxon's signed rank statistic, and (m, m, m) is a statistic proposed by Stephenson (1981).

The new U-statistic, described informally

- **Name**: Fix three integers, m, $\underline{m}$, $\overline{m}$ with $1 \le \underline{m} \le \overline{m} \le m < I$. Then $(m, \underline{m}, \overline{m})$ is the name of one U-statistic.
- **Instances**: $(1, 1, 1)$ is the sign test statistic, $(2, 2, 2)$ is (essentially) Wilcoxon's signed rank statistic, and (m, m, m) is a statistic proposed by Stephenson (1981).
- **General 1**: Look at every subset of m pairs. Sort the m pair differences Y_i into increasing order by their absolute values, $|Y_i|$.

The new U-statistic, described informally

- **Name**: Fix three integers, m, $\underline{m}$, $\overline{m}$ with $1 \leq \underline{m} \leq \overline{m} \leq m < I$. Then $(m, \underline{m}, \overline{m})$ is the name of one U-statistic.
- **Instances**: $(1,1,1)$ is the sign test statistic, $(2,2,2)$ is (essentially) Wilcoxon's signed rank statistic, and (m,m,m) is a statistic proposed by Stephenson (1981).
- **General 1**: Look at every subset of m pairs. Sort the m pair differences Y_i into increasing order by their absolute values, $|Y_i|$.
- **General 2**: In this order, count the number of positive Y_i among those numbered $\underline{m}, \underline{m}+1, \ldots, \overline{m}$. Average over all $\binom{I}{m}$ subsets.

The new U-statistic, described informally

- **Name**: Fix three integers, m, $\underline{m}$, $\overline{m}$ with $1 \le \underline{m} \le \overline{m} \le m < I$. Then $(m, \underline{m}, \overline{m})$ is the name of one U-statistic.
- **Instances**: $(1, 1, 1)$ is the sign test statistic, $(2, 2, 2)$ is (essentially) Wilcoxon's signed rank statistic, and (m, m, m) is a statistic proposed by Stephenson (1981).
- **General 1**: Look at every subset of m pairs. Sort the m pair differences Y_i into increasing order by their absolute values, $|Y_i|$.
- **General 2**: In this order, count the number of positive Y_i among those numbered $\underline{m}, \underline{m} + 1, \ldots, \overline{m}$. Average over all $\binom{I}{m}$ subsets.
- **One good choice:** $(8, 7, 8)$. Look at 8 pairs. Find the two largest $|Y_i|$'s, and score 0, 1, or 2 depending upon whether neither, one or both Y_i's are positive.

Sensitivity analysis for the NHANES data about blood lead levels

- Compare sign test $(1, 1, 1)$, Wilcoxon test $(2, 2, 2)$, and the new U-statistic with $(m, \underline{m}, \overline{m}) = (8, 7, 8)$ for $I = 679$ smoker-nonsmoker pair differences Y_i in blood lead levels.

Sensitivity analysis for the NHANES data about blood lead levels

- Compare sign test $(1, 1, 1)$, Wilcoxon test $(2, 2, 2)$, and the new U-statistic with $(m, \underline{m}, \overline{m}) = (8, 7, 8)$ for $I = 679$ smoker-nonsmoker pair differences Y_i in blood lead levels.
- Value reported is the upper bound on the one-sided P-value testing the null hypothesis of no effect H_0 when the bias is at most $\Gamma \geq 1$.

Sensitivity analysis for the NHANES data about blood lead levels

- Compare sign test $(1, 1, 1)$, Wilcoxon test $(2, 2, 2)$, and the new U-statistic with $(m, \underline{m}, \overline{m}) = (8, 7, 8)$ for $I = 679$ smoker-nonsmoker pair differences Y_i in blood lead levels.
- Value reported is the upper bound on the one-sided P-value testing the null hypothesis of no effect H_0 when the bias is at most $\Gamma \geq 1$.
-

Γ	1	2	2.5	3	3.5	3.8
Sign test	0.0000	0.0083	0.5961	0.9918	1.0000	1.0000
Wilcoxon	0.0000	0.0000	0.0004	0.0510	0.4224	0.7160
(8,7,8)	0.0000	0.0000	0.0000	0.0009	0.0142	0.0444

Additional sensitivity analyses for the NHANES data about blood lead levels

- Table adds $(m, \underline{m}, \overline{m}) = (5, 4, 5)$, $(20, 14, 20)$ and $(20, 16, 19)$.

Additional sensitivity analyses for the NHANES data about blood lead levels

- Table adds $(m, \underline{m}, \overline{m}) = (5, 4, 5)$, $(20, 14, 20)$ and $(20, 16, 19)$.
- Value reported is the upper bound on the one-sided P-value testing the null hypothesis of no effect H_0 when the bias is at most $\Gamma \geq 1$.

Additional sensitivity analyses for the NHANES data about blood lead levels

- Table adds $(m, \underline{m}, \overline{m}) = (5, 4, 5)$, $(20, 14, 20)$ and $(20, 16, 19)$.
- Value reported is the upper bound on the one-sided P-value testing the null hypothesis of no effect H_0 when the bias is at most $\Gamma \geq 1$.
-

Γ	1	2	2.5	3	3.5	3.8
Wilcoxon	0.0000	0.0000	0.0004	0.0510	0.4224	0.7160
(8,7,8)	0.0000	0.0000	0.0000	0.0009	0.0142	0.0444
(5,4,5)	0.0000	0.0000	0.0000	0.0023	0.0494	0.1530
(20,14,20)	0.0000	0.0000	0.0000	0.0008	0.0147	0.0493
(20,16,19)	0.0000	0.0000	0.0000	0.0009	0.0116	0.0344

The U-statistic is a signed rank statistic

- **Absolute ranks**: Let a_i be the rank of $|Y_i|$, $i = 1, \ldots, I$.

The U-statistic is a signed rank statistic

- **Absolute ranks**: Let a_i be the rank of $|Y_i|$, $i = 1, \ldots, I$.
- **Equivalence**: The Y_i with absolute rank a_i has the ℓth largest $|Y_i|$ in $\binom{a_i-1}{\ell-1}\binom{I-a_i}{m-\ell}$ sets of size m so the statistic $(\overline{m}, \underline{m}, m)$ is:

$$T = \sum_{i=1}^{I} \operatorname{sgn}(Y_i)\ q_i \tag{2}$$

where

$$q_i = \binom{I}{m}^{-1} \sum_{\ell=\underline{m}}^{\overline{m}} \binom{a_i - 1}{\ell - 1} \binom{I - a_i}{m - \ell}. \tag{3}$$

The U-statistic is a signed rank statistic

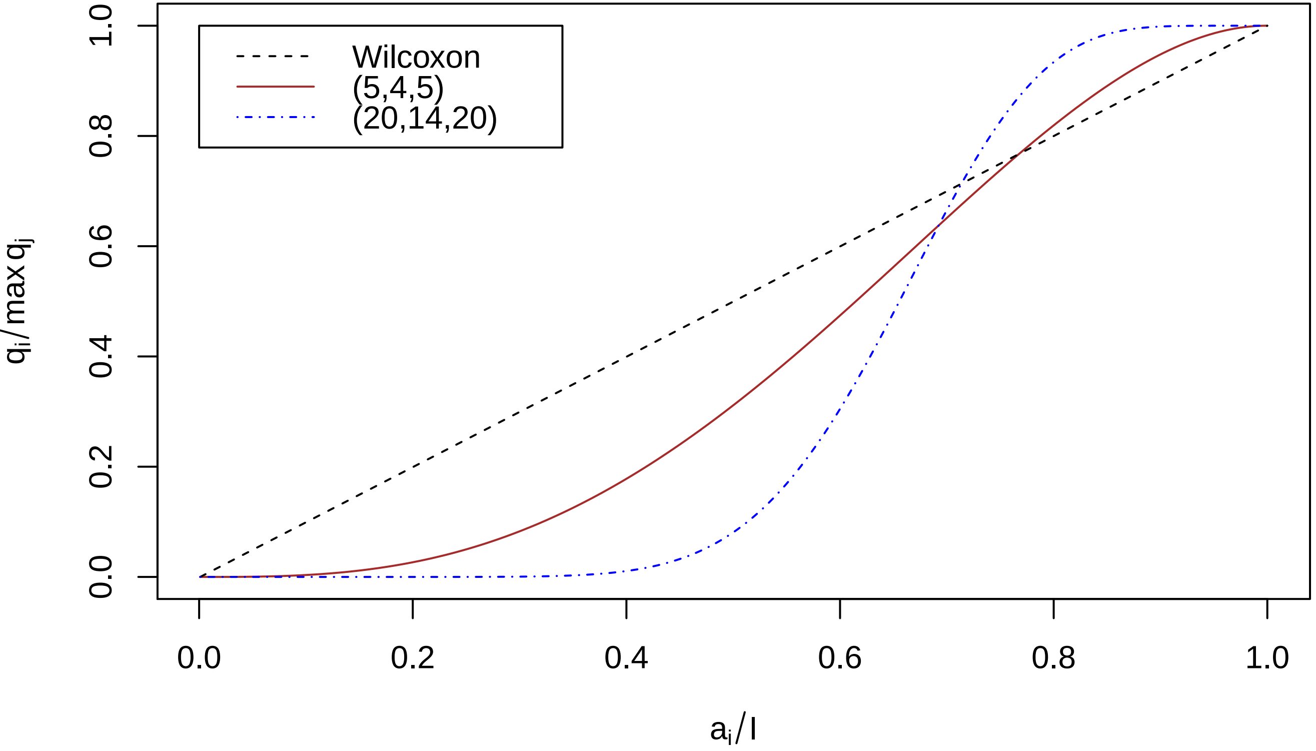
- **Absolute ranks**: Let a_i be the rank of $|Y_i|$, $i = 1, \ldots, I$.
- **Equivalence**: The Y_i with absolute rank a_i has the ℓth largest $|Y_i|$ in $\binom{a_i-1}{\ell-1}\binom{I-a_i}{m-\ell}$ sets of size m so the statistic $(\overline{m}, \underline{m}, m)$ is:

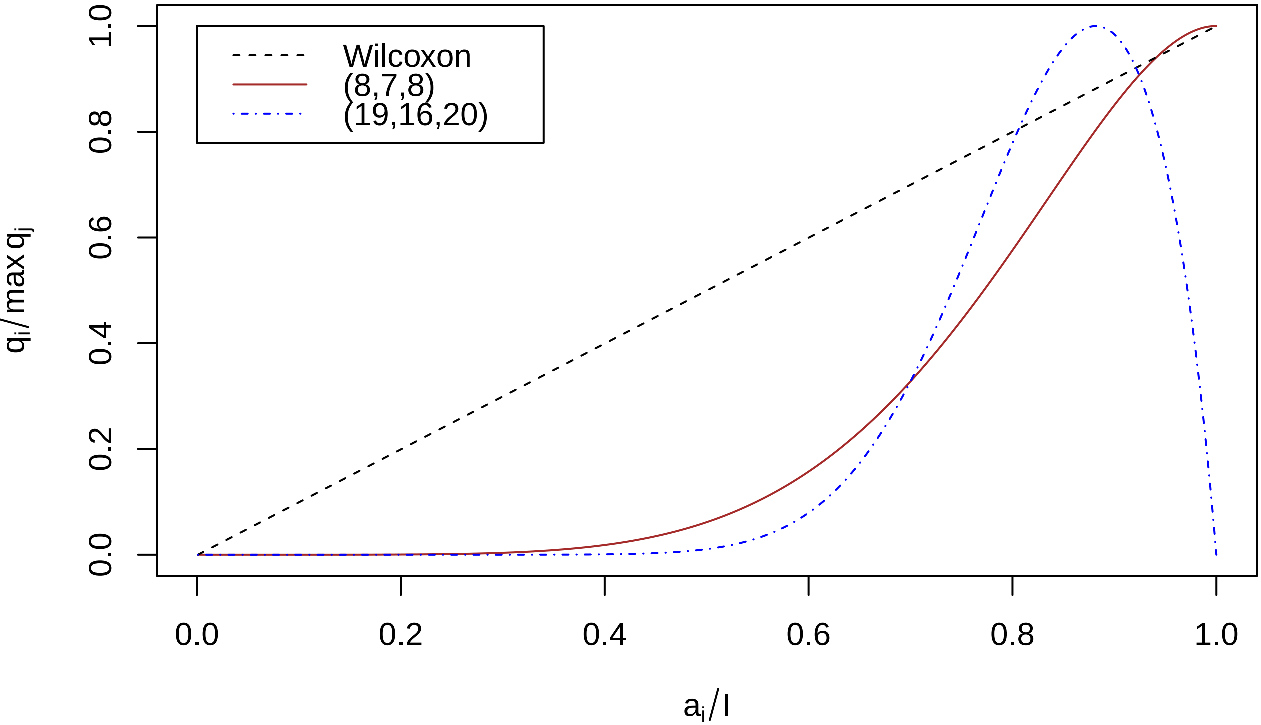
$$T = \sum_{i=1}^{I} \operatorname{sgn}(Y_i)\; q_i \tag{2}$$

where

$$q_i = \binom{I}{m}^{-1} \sum_{\ell=\underline{m}}^{\overline{m}} \binom{a_i - 1}{\ell - 1} \binom{I - a_i}{m - \ell}. \tag{3}$$

- Will plot $q_i / \max q_j$ against a_i / I.


Wilcoxon
(5,4,5)
(20,14,20)
$q_i/\max q_j$
a_i/l
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


Wilcoxon
(8,7,8)
(19,16,20)
$q_i/\max q_j$
a_i/I
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

Power of sensitivity analysis

- If the treatment had an effect and if there was no bias in treatment assignment, $\Pr\left(Z_{ij} \mid \mathcal{F}, \mathcal{Z}\right) = \frac{1}{2}$, then we could not see this in the observed data. The best we can hope to say is that rejection of H_0 at level α is insensitive to small and moderate bias as measured by Γ. The power is the probability that we will be able to say this.

Power of sensitivity analysis

- If the treatment had an effect and if there was no bias in treatment assignment, $\Pr\left(Z_{ij} \mid \mathcal{F}, \mathcal{Z}\right) = \frac{1}{2}$, then we could not see this in the observed data. The best we can hope to say is that rejection of H_0 at level α is insensitive to small and moderate bias as measured by Γ. The power is the probability that we will be able to say this.
- An α-level sensitivity analysis rejects the null hypothesis H_0 of no effect allowing for a bias of $\Gamma \geq 1$ if the upper bound on the P-value is $\leq \alpha$ at this Γ. Conventionally, $\alpha = 0.05$.

Power of sensitivity analysis

- If the treatment had an effect and if there was no bias in treatment assignment, $\Pr\left(Z_{ij} \mid \mathcal{F}, \mathcal{Z}\right) = \frac{1}{2}$, then we could not see this in the observed data. The best we can hope to say is that rejection of H_0 at level α is insensitive to small and moderate bias as measured by Γ. The power is the probability that we will be able to say this.
- An α-level sensitivity analysis rejects the null hypothesis H_0 of no effect allowing for a bias of $\Gamma \geq 1$ if the upper bound on the P-value is $\leq \alpha$ at this Γ. Conventionally, $\alpha = 0.05$.
- **Power is**: the probability that the upper bound on the P-value testing H_0 will be less than or equal to α at this Γ when the Y_i are sampled from some probability model in which there is an effect an no bias, $\Pr\left(T \mid \mathcal{F}, \mathcal{Z}\right) = \frac{1}{2}$, e.g., $Y_i \sim_{iid} N\left(\tau, 1\right)$.

Simulated Power

- **Sampling situation**: $Y_i = \tau + \epsilon_i$ where ϵ_i is standard Normal, standard logistic or t-distributed with 4 degrees of freedom, and no unmeasured bias, $\Pr\left(Z_{ij} = 1 \mid \mathcal{F}, \mathcal{Z}\right) = \frac{1}{2}$.

Simulated Power

- **Sampling situation**: $Y_i = \tau + \epsilon_i$ where ϵ_i is standard Normal, standard logistic or t-distributed with 4 degrees of freedom, and no unmeasured bias, $\Pr\left(Z_{ij} = 1 \,|\, \mathcal{F}, \mathcal{Z} \right) = \frac{1}{2}$.
- **Simulation**: Each situation is sampled 10,000 times, so the standard error of the estimated power is at most $\sqrt{0.5 \times 0.5/10,000} = 0.005$.

Simulated Power

- **Sampling situation**: $Y_i = \tau + \epsilon_i$ where ϵ_i is standard Normal, standard logistic or t-distributed with 4 degrees of freedom, and no unmeasured bias, $\Pr\left(Z_{ij} = 1 \mid \mathcal{F}, \mathcal{Z}\right) = \frac{1}{2}$.
- **Simulation**: Each situation is sampled 10,000 times, so the standard error of the estimated power is at most $\sqrt{0.5 \times 0.5/10,000} = 0.005$.

Table: Power of a one-sided 0.05 level sensitivity analysis with additive effect τ conducted with $\Gamma = 3$ and $I = 250$ pairs. Errors are standard Normal, standard logistic or t-distributed with 4 degrees of freedom. The highest powers in a column are in **bold**.

Errors	Normal	Logistic	t with 4 df
Statistic	$\tau = 1/2$	$\tau = 1$	$\tau = 1$
Wilcoxon	0.08	0.40	0.43
(5,4,5)	0.34	0.67	**0.65**
(8,7,8)	**0.63**	**0.74**	0.57
(20,14,20)	0.53	**0.74**	**0.65**
(20,16,19)	0.52	0.69	0.61

Design sensitivity

- **Definition**: For a given sampling situation with a treatment effect and no unmeasured bias, and for a given test statistic, there is a number $\widetilde{\Gamma}$ such that, as $I \to \infty$, the power of an α-level sensitivity analysis tends to 1 if performed with $\Gamma < \widetilde{\Gamma}$ and to 0 if $\Gamma > \widetilde{\Gamma}$.

Design sensitivity

- **Definition**: For a given sampling situation with a treatment effect and no unmeasured bias, and for a given test statistic, there is a number $\widetilde{\Gamma}$ such that, as $I \to \infty$, the power of an α-level sensitivity analysis tends to 1 if performed with $\Gamma < \widetilde{\Gamma}$ and to 0 if $\Gamma > \widetilde{\Gamma}$.
- **In other words**: In that sampling situation with a treatment effect, eventually (for large enough I) that statistic will be insensitive to all biases smaller than $\widetilde{\Gamma}$ and sensitive to some biases larger than $\widetilde{\Gamma}$.

Design sensitivity

- **Definition**: For a given sampling situation with a treatment effect and no unmeasured bias, and for a given test statistic, there is a number $\widetilde{\Gamma}$ such that, as $I \to \infty$, the power of an α-level sensitivity analysis tends to 1 if performed with $\Gamma < \widetilde{\Gamma}$ and to 0 if $\Gamma > \widetilde{\Gamma}$.
- **In other words**: In that sampling situation with a treatment effect, eventually (for large enough I) that statistic will be insensitive to all biases smaller than $\widetilde{\Gamma}$ and sensitive to some biases larger than $\widetilde{\Gamma}$.
- **Illustration**: For an additive effect of $\tau = 1$ with errors from the t-distribution with 3 degrees of freedom, the Wilcoxon statistic has design sensitivity $\widetilde{\Gamma} = 6.0$ while $(m, \underline{m}, \overline{m}) = (5, 4, 5)$ has design sensitivity $\widetilde{\Gamma} = 6.8$.

Design sensitivity

- **Definition**: For a given sampling situation with a treatment effect and no unmeasured bias, and for a given test statistic, there is a number $\widetilde{\Gamma}$ such that, as $I \rightarrow \infty$, the power of an α-level sensitivity analysis tends to 1 if performed with $\Gamma < \widetilde{\Gamma}$ and to 0 if $\Gamma > \widetilde{\Gamma}$.
- **In other words**: In that sampling situation with a treatment effect, eventually (for large enough I) that statistic will be insensitive to all biases smaller than $\widetilde{\Gamma}$ and sensitive to some biases larger than $\widetilde{\Gamma}$.
- **Illustration**: For an additive effect of $\tau = 1$ with errors from the t-distribution with 3 degrees of freedom, the Wilcoxon statistic has design sensitivity $\widetilde{\Gamma} = 6.0$ while $\left(m, \underline{m}, \overline{m}\right) = \left(5, 4, 5\right)$ has design sensitivity $\widetilde{\Gamma} = 6.8$.
- **Example**: If $I = 100,000$ differences $Y_i = \tau + \epsilon_i$ are sampled from this distribution, the upper bound on the P-value from Wilcoxon's statistic is 0.016 at $\Gamma = 5.8$ and 0.997 at $\Gamma = 6.1$, consistent with $\widetilde{\Gamma} = 6.0$. If $\left(m, \underline{m}, \overline{m}\right) = \left(5, 4, 5\right)$ is used instead, the P-value bound is 0.0028 for $\Gamma = 6.5$ and 0.98 for $\Gamma = 6.9$, consistent with $\widetilde{\Gamma} = 6.8$.

Formula for the design sensitivity of the U-statistic

- **Will assume**: Y_i are *iid* from some distribution $F(\cdot)$ and there is no unobserved bias, $\Pr(Z_{ij} \mid \mathcal{F}, \mathcal{Z}) = \frac{1}{2}$.

Formula for the design sensitivity of the U-statistic

- **Will assume**: Y_i are *iid* from some distribution $F(\cdot)$ and there is no unobserved bias, $\Pr(Z_{ij} \mid \mathcal{F}, \mathcal{Z}) = \frac{1}{2}$.
- **Recall**: $(m, \underline{m}, \overline{m})$ looks at m pair differences Y_i, sorts them into order by $|Y_i|$, and counts the number of positive differences $Y_i > 0$ among those numbered $\underline{m}, \underline{m}+1, \ldots, \overline{m}$, yielding an integer in $\{0, 1, 2, \ldots, \overline{m} - \underline{m} + 1\}$. Let θ be the expectation of this number. It is also the expectation of T.

Formula for the design sensitivity of the U-statistic

- **Will assume**: Y_i are *iid* from some distribution $F(\cdot)$ and there is no unobserved bias, $\Pr(Z_{ij} \mid \mathcal{F}, \mathcal{Z}) = \frac{1}{2}$.
- **Recall**: $(m, \underline{m}, \overline{m})$ looks at m pair differences Y_i, sorts them into order by $|Y_i|$, and counts the number of positive differences $Y_i > 0$ among those numbered $\underline{m}, \underline{m}+1, \ldots, \overline{m}$, yielding an integer in $\{0, 1, 2, \ldots, \overline{m} - \underline{m} + 1\}$. Let θ be the expectation of this number. It is also the expectation of T.
- **Proposition**: Under these assumptions, the design sensitivity of the U-statistic $(m, \underline{m}, \overline{m})$ is:

$$\widetilde{\Gamma} = \frac{\theta}{\overline{m} - \underline{m} + 1 - \theta}$$

Formula for the design sensitivity of the U-statistic

- **Will assume**: Y_i are *iid* from some distribution $F(\cdot)$ and there is no unobserved bias, $\Pr(Z_{ij} \mid \mathcal{F}, \mathcal{Z}) = \frac{1}{2}$.
- **Recall**: $(m, \underline{m}, \overline{m})$ looks at m pair differences Y_i, sorts them into order by $|Y_i|$, and counts the number of positive differences $Y_i > 0$ among those numbered $\underline{m}, \underline{m}+1, \ldots, \overline{m}$, yielding an integer in $\{0, 1, 2, \ldots, \overline{m} - \underline{m} + 1\}$. Let θ be the expectation of this number. It is also the expectation of T.
- **Proposition**: Under these assumptions, the design sensitivity of the U-statistic $(m, \underline{m}, \overline{m})$ is:

$$\widetilde{\Gamma} = \frac{\theta}{\overline{m} - \underline{m} + 1 - \theta}$$

- **Cases**: If $\theta = \overline{m} - \underline{m} + 1$ then $\widetilde{\Gamma} = \infty$. If $\widetilde{\Gamma} < 1$, then the power tends to zero as $I \rightarrow \infty$ for all $\Gamma \geq 1$)

Table of Design Sensitivities

Table: Design sensitivities $\widetilde{\Gamma}$ with additive effect τ. Errors are standard Normal, standard logistic or t-distributed with 3 or 4 degrees of freedom. The largest $\widetilde{\Gamma}$s in a column are in **bold**.

Errors	Normal	Logistic	t with 4 df	t with 3 df
Statistic	$\tau = 1/2$	$\tau = 1$	$\tau = 1$	$\tau = 1$
Wilcoxon	3.2	3.9	6.8	6.0
(5,4,5)	3.9	4.7	8.4	6.8
(8,7,8)	**5.1**	5.5	9.1	6.8
(8,6,7)	3.5	4.5	9.0	7.7
(20,14,20)	4.6	5.3	9.4	7.3
(20,16,19)	4.9	**5.6**	**10.1**	**7.8**

Heuristic Graph I: Where is the evidence that distinguishes effects from unmeasured biases?

- Suppose that the Y_i's are not biased, so each Y_i is telling us about the effects of the treatment. (Of course, we would not know this from the data.)

Heuristic Graph I: Where is the evidence that distinguishes effects from unmeasured biases?

- Suppose that the Y_i's are not biased, so each Y_i is telling us about the effects of the treatment. (Of course, we would not know this from the data.)
- In this case, we would like to say that the results are insensitive to small and moderate biases.

Heuristic Graph I: Where is the evidence that distinguishes effects from unmeasured biases?

- Suppose that the Y_i's are not biased, so each Y_i is telling us about the effects of the treatment. (Of course, we would not know this from the data.)
- In this case, we would like to say that the results are insensitive to small and moderate biases.
- Suppose you could observe an infinite amount of data at any one value of $|Y_i|$, that is, you get to observe $\mathrm{sgn}(Y_i)$.

Heuristic Graph I: Where is the evidence that distinguishes effects from unmeasured biases?

- Suppose that the Y_i's are not biased, so each Y_i is telling us about the effects of the treatment. (Of course, we would not know this from the data.)
- In this case, we would like to say that the results are insensitive to small and moderate biases.
- Suppose you could observe an infinite amount of data at any one value of $|Y_i|$, that is, you get to observe $\mathrm{sgn}\,(Y_i)$.
- What $|Y_i|$ would you pick?

Heuristic Graph II: The abz-function

- Suppose that the Y_i's are not biased, so each Y_i is telling us about the effects of the treatment. (Of course, we would not know this from the data.)

Heuristic Graph II: The abz-function

- Suppose that the Y_i's are not biased, so each Y_i is telling us about the effects of the treatment. (Of course, we would not know this from the data.)
- Suppose that Y_i are *iid* from a continuous distribution $G(\cdot)$ with density $g(\cdot)$.

Heuristic Graph II: The abz-function

- Suppose that the Y_i's are not biased, so each Y_i is telling us about the effects of the treatment. (Of course, we would not know this from the data.)
- Suppose that Y_i are *iid* from a continuous distribution $G(\cdot)$ with density $g(\cdot)$.
- Albers, Bickel and van Zwet (1976) introduced a function $\text{abz}(y)$ defined for $y > 0$, namely

$$\text{abz}(y) = \frac{g(y)}{g(y) + g(-y)} = \Pr\left(Y_i > 0 \,\middle|\, |Y_i| = y \right)$$

Heuristic Graph II: The abz-function

- Suppose that the Y_i's are not biased, so each Y_i is telling us about the effects of the treatment. (Of course, we would not know this from the data.)
- Suppose that Y_i are *iid* from a continuous distribution $G(\cdot)$ with density $g(\cdot)$.
- Albers, Bickel and van Zwet (1976) introduced a function $\mathrm{abz}(y)$ defined for $y > 0$, namely

$$\mathrm{abz}(y) = \frac{g(y)}{g(y) + g(-y)} = \Pr\left(Y_i > 0 \,\middle|\, |Y_i| = y \right)$$

- If $\mathrm{abz}(y) > \Gamma / (1 + \Gamma)$, then at $|Y_i| = y$, positive Y_i occur with a frequency $\mathrm{abz}(y)$ that is too high to be attributed to a bias of magnitude Γ.

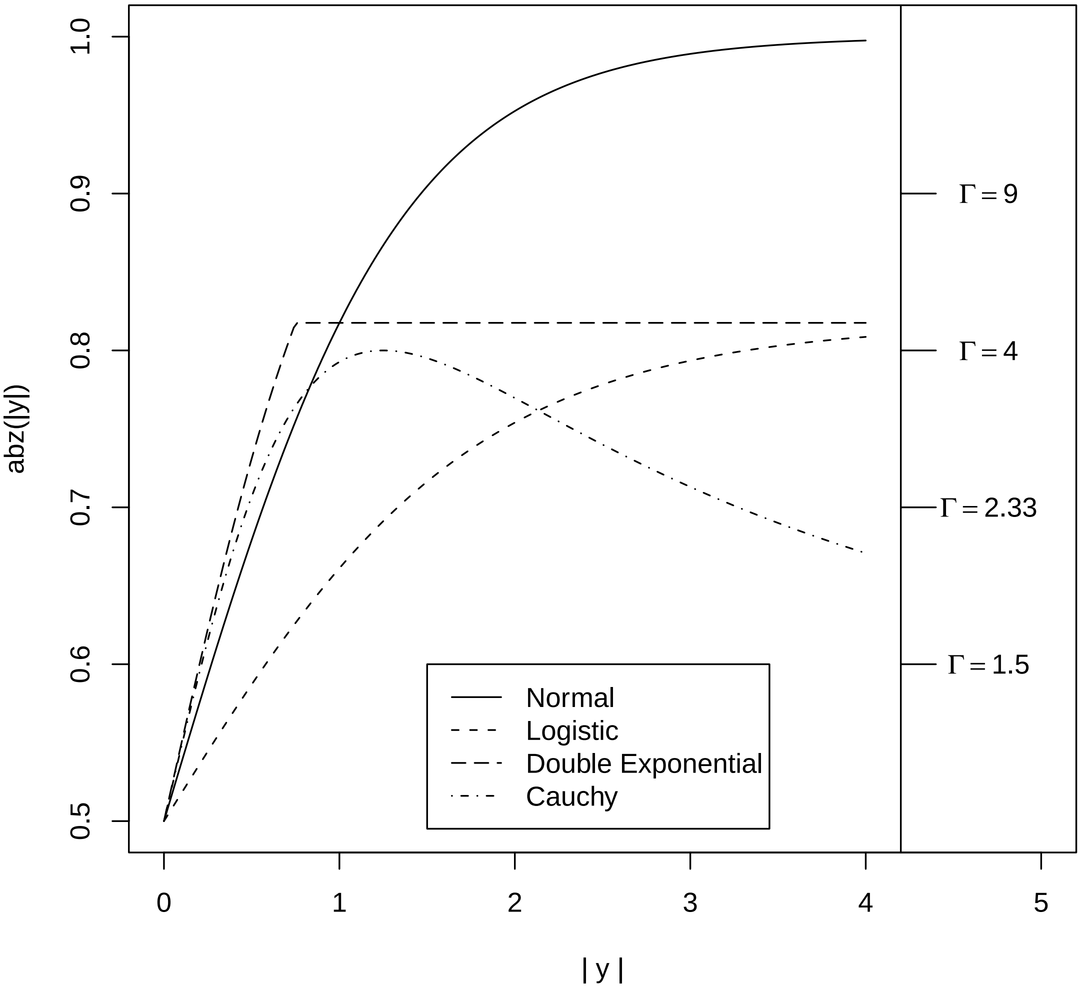


Figure 2: Conditionally given various values of $|Y_i|$, the figure shows the probability of a positive treatment-minus-control difference, $Y_i > 0$, for an additive treatment effect $\tau = ¾$ in the standard forms of four distributions.

Heuristic Graph III

- What do we learn from the heuristic graph?

Heuristic Graph III

- What do we learn from the heuristic graph?
- We actually observe limited data at varied $|Y_i|$.

Heuristic Graph III

- What do we learn from the heuristic graph?
- We actually observe limited data at varied $|Y_i|$.
- Nonetheless, the heuristic graph suggest little weight should be given to small $|Y_i|$.

Heuristic Graph III

- What do we learn from the heuristic graph?
- We actually observe limited data at varied $|Y_i|$.
- Nonetheless, the heuristic graph suggest little weight should be given to small $|Y_i|$.
- What you should do with large $|Y_i|$ depends on the distribution G which you typically do not know.

Stephenson's test: useful when only some people respond to treatment

- **A Lehmann alternative**: Control responses $r_{Cij} \sim F(\cdot)$, treated responses as $r_{Tij} \sim (1-\lambda)\ F(\cdot) + \lambda\ \{F(\cdot)\}^m$, so only a fraction $\lambda \in (0,1)$ respond to treatment.

Stephenson's test: useful when only some people respond to treatment

- **A Lehmann alternative**: Control responses $r_{Cij} \sim F(\cdot)$, treated responses as $r_{Tij} \sim (1-\lambda)\ F(\cdot) + \lambda\ \{F(\cdot)\}^m$, so only a fraction $\lambda \in (0,1)$ respond to treatment.
- **Conover and Salsburg (1988)**: Found the locally most powerful rank test for this problem as $\lambda \to 0$.

Stephenson's test: useful when only some people respond to treatment

- **A Lehmann alternative**: Control responses $r_{Cij} \sim F(\cdot)$, treated responses as $r_{Tij} \sim (1-\lambda)\ F(\cdot) + \lambda\ \{F(\cdot)\}^m$, so only a fraction $\lambda \in (0,1)$ respond to treatment.
- **Conover and Salsburg (1988)**: Found the locally most powerful rank test for this problem as $\lambda \to 0$.
- **Stephenson (1981)**: Based on other considerations, Stephenson had proposed use of ranks that are essentially the same for large I, and have the advantage of permitting a confidence interval for the magnitude of effect; see Rosenbaum (2007).

Stephenson's test: useful when only some people respond to treatment

- **A Lehmann alternative**: Control responses $r_{Cij} \sim F(\cdot)$, treated responses as $r_{Tij} \sim (1-\lambda)\ F(\cdot) + \lambda\ \{F(\cdot)\}^m$, so only a fraction $\lambda \in (0,1)$ respond to treatment.
- **Conover and Salsburg (1988)**: Found the locally most powerful rank test for this problem as $\lambda \to 0$.
- **Stephenson (1981)**: Based on other considerations, Stephenson had proposed use of ranks that are essentially the same for large I, and have the advantage of permitting a confidence interval for the magnitude of effect; see Rosenbaum (2007).
- **The U-statistic**: is Stephenson's statistic for $(m, \underline{m}, \overline{m}) = (m, m, m)$. That is, look at the sign of Y_i for the one pair of m with the largest $|Y_i|$.

Testing one hypothesis twice

- **How should one select** $(m, \underline{m}, \overline{m})$? Have seen that the sign test $(1, 1, 1)$ and Wilcoxon's test $(2, 2, 2)$ are poor choices for $\Gamma > 1$. Some good choices are $(m, \underline{m}, \overline{m}) = (8, 7, 8)$ and $(20, 14, 20)$ for general use, and $(20, 16, 19)$ for thicker tails with larger samples I.

Testing one hypothesis twice

- **How should one select** $(m, \underline{m}, \overline{m})$? Have seen that the sign test $(1, 1, 1)$ and Wilcoxon's test $(2, 2, 2)$ are poor choices for $\Gamma > 1$. Some good choices are $(m, \underline{m}, \overline{m}) = (8, 7, 8)$ and $(20, 14, 20)$ for general use, and $(20, 16, 19)$ for thicker tails with larger samples I.
- **Testing one hypothesis twice**: Use more than one test statistic and correct for multiple testing.

Testing one hypothesis twice

- **How should one select** $(m, \underline{m}, \overline{m})$? Have seen that the sign test $(1, 1, 1)$ and Wilcoxon's test $(2, 2, 2)$ are poor choices for $\Gamma > 1$. Some good choices are $(m, \underline{m}, \overline{m}) = (8, 7, 8)$ and $(20, 14, 20)$ for general use, and $(20, 16, 19)$ for thicker tails with larger samples I.
- **Testing one hypothesis twice**: Use more than one test statistic and correct for multiple testing.
- **Bonferroni**: Obviously, one could perform two tests (i.e., two sensitivity analyses at Γ) of the same null hypothesis of no treatment effect H_0, rejecting H_0 if the smaller of the two (upper bounds on) P-values is at most $\alpha = 0.025$. This would control the chance of falsely rejecting H_0 at $\alpha = 0.05$ in the presence of a bias of at most Γ. This is ok, but we can do much better.

Two ways to work with the joint sensitivity distribution

- **Better approach**: Use the joint null sensitivity distribution of two test statistics, allowing for the high positive correlation between two tests of one H_0 based on the same data.

Two ways to work with the joint sensitivity distribution

- **Better approach**: Use the joint null sensitivity distribution of two test statistics, allowing for the high positive correlation between two tests of one H_0 based on the same data.
- **Exact joint null sensitivity distribution**: For simple statistics, the exact joint sensitivity distribution is available for each $\Gamma \geq 1$. See Rosenbaum (2012 AOAS) for discussion and Small (2012) for an implementation in `R`.

Two ways to work with the joint sensitivity distribution

- **Better approach**: Use the joint null sensitivity distribution of two test statistics, allowing for the high positive correlation between two tests of one H_0 based on the same data.
- **Exact joint null sensitivity distribution**: For simple statistics, the exact joint sensitivity distribution is available for each $\Gamma \geq 1$. See Rosenbaum (2012 AOAS) for discussion and Small (2012) for an implementation in `R`.
- **Large sample approximation null sensitivity distribution**: For many statistics, a large sample multivariate Normal approximation to the joint sensitivity distribution is available for each $\Gamma \geq 1$. (Rosenbaum 2012 Biometrika).

Choice of test statistic affects reported sensitivity to bias

Table: Five tests of no effect, using Wilcoxon's test on lead levels, (8,7,8) and (8,6,7) on lead levels and on logs of lead levels. Tabled are upper bound on the one-sided *P*-value testing no treatment effect for the given value of Γ.

	Wilcoxon	U-statistic		U-statistic on logs	
Γ		(8,7,8)	(8,6,7)	(8,7,8)	(8,6,7)
1	0.000	0.000	0.000	0.000	0.000
2.5	0.016	0.026	0.000	0.000	0.000
2.8	0.147	0.119	0.015	0.000	0.001
3	–	–	0.050	0.001	0.004
3.4	–	–	–	0.009	0.041
3.6	–	–	–	0.022	0.095

Testing one hypothesis four times, correcting for multiple testing

Table: Testing one hypothesis four times, correcting for multiple testing. The combined test uses both U-statistics on both lead levels and logs of lead levels. Tabled are upper bound on the one-sided *P*-value testing no treatment effect for the given value of Γ.

	Testing 4-times	U-statistic		U-statistic on logs	
Γ		(8,7,8)	(8,6,7)	(8,7,8)	(8,6,7)
1	0.000	0.000	0.000	0.000	0.000
2.5	0.000	0.026	0.000	0.000	0.000
2.8	0.000	0.119	0.015	0.000	0.001
3	0.003	–	0.050	0.001	0.004
3.4	0.022	–	–	0.009	0.041
3.6	0.049	–	–	0.022	0.095

Two test statistics and their respective bounds

- Suppose there are two tests of H_0 using the same Y_i but different scores, $T = \sum_{i=1}^{I} \operatorname{sgn}(Y_i)\ q_i$ and $T' = \sum_{i=1}^{I} \operatorname{sgn}(Y_i)\ q_i^{'}$, where $q_i \geq 0$ and $q_i^{'} \geq 0$.

Two test statistics and their respective bounds

- Suppose there are two tests of H_0 using the same Y_i but different scores, $T = \sum_{i=1}^{I} \mathrm{sgn}\left(Y_i\right) q_i$ and $T' = \sum_{i=1}^{I} \mathrm{sgn}\left(Y_i\right) q_i^{'}$, where $q_i \geq 0$ and $q_i^{'} \geq 0$.
- It is important here that T and T' both receive a nonnegative contribution whenever $\mathrm{sgn}\left(Y_i\right) = 1$ or $Y_i \geq 0$.

Two test statistics and their respective bounds

- Suppose there are two tests of H_0 using the same Y_i but different scores, $T = \sum_{i=1}^{I} \mathrm{sgn}\left(Y_i\right)\, q_i$ and $T' = \sum_{i=1}^{I} \mathrm{sgn}\left(Y_i\right)\, q_i^{'}$, where $q_i \geq 0$ and $q_i^{'} \geq 0$.
- It is important here that T and T' both receive a nonnegative contribution whenever $\mathrm{sgn}\left(Y_i\right) = 1$ or $Y_i \geq 0$.
- In the sensitivity analysis, there are now two upper bound random variables, $\overline{\overline{T}}$ and $\overline{\overline{T}}'$, which are each the sum of I independent random variables, both taking the value 0 with probability $1/\left(1+\Gamma\right)$ or else the values q_i and $q_i^{'}$ with probability $\Gamma/\left(1+\Gamma\right)$.

Two test statistics and their respective bounds

- Suppose there are two tests of H_0 using the same Y_i but different scores, $T = \sum_{i=1}^{I} \operatorname{sgn}(Y_i)\, q_i$ and $T' = \sum_{i=1}^{I} \operatorname{sgn}(Y_i)\, q_i^{'}$, where $q_i \geq 0$ and $q_i^{'} \geq 0$.
- It is important here that T and T' both receive a nonnegative contribution whenever $\operatorname{sgn}(Y_i) = 1$ or $Y_i \geq 0$.
- In the sensitivity analysis, there are now two upper bound random variables, $\overline{\overline{T}}$ and $\overline{\overline{T}}'$, which are each the sum of I independent random variables, both taking the value 0 with probability $1/(1+\Gamma)$ or else the values q_i and $q_i^{'}$ with probability $\Gamma/(1+\Gamma)$.
- Under mild conditions on the scores, q_i and $q_i^{'}$, as $I \rightarrow \infty$, the joint distribution of $\overline{\overline{T}}$ and $\overline{\overline{T}}'$ tends to a bivariate Normal distribution.

The maximum of two standardized deviates

- To repeat, there are two tests of H_0 using the same Y_i but different scores, $T = \sum_{i=1}^{I} \operatorname{sgn}(Y_i)\ q_i$ and $T' = \sum_{i=1}^{I} \operatorname{sgn}(Y_i)\ q_i^{'}$, where $q_i \geq 0$ and $q_i^{'} \geq 0$.

The maximum of two standardized deviates

- To repeat, there are two tests of H_0 using the same Y_i but different scores, $T = \sum_{i=1}^{I} \operatorname{sgn}(Y_i)\ q_i$ and $T' = \sum_{i=1}^{I} \operatorname{sgn}(Y_i)\ q_i^{'}$, where $q_i \geq 0$ and $q_i^{'} \geq 0$.
- Let μ_Γ and $\mu_\Gamma^{'}$ be the expectations and ω_Γ and $\omega_\Gamma^{'}$ be the standard deviations of $\overline{\overline{T}}$ and $\overline{\overline{T}}'$.

The maximum of two standardized deviates

- To repeat, there are two tests of H_0 using the same Y_i but different scores, $T = \sum_{i=1}^{I} \operatorname{sgn}(Y_i)\ q_i$ and $T' = \sum_{i=1}^{I} \operatorname{sgn}(Y_i)\ q_i^{'}$, where $q_i \geq 0$ and $q_i^{'} \geq 0$.
- Let μ_Γ and $\mu_\Gamma^{'}$ be the expectations and ω_Γ and $\omega_\Gamma^{'}$ be the standard deviations of $\overline{\overline{T}}$ and $\overline{\overline{T}}'$.
- The test statistic will be

$$\max\left(\frac{T-\mu_\Gamma}{\omega_\Gamma}, \frac{T'-\mu_\Gamma^{/}}{\omega_\Gamma^{'}}\right).$$

The maximum of two standardized deviates

- To repeat, there are two tests of H_0 using the same Y_i but different scores, $T = \sum_{i=1}^{I} \operatorname{sgn}(Y_i)\, q_i$ and $T' = \sum_{i=1}^{I} \operatorname{sgn}(Y_i)\, q_i^{'}$, where $q_i \geq 0$ and $q_i^{'} \geq 0$.
- Let μ_Γ and $\mu_\Gamma^{'}$ be the expectations and ω_Γ and $\omega_\Gamma^{'}$ be the standard deviations of $\overline{\overline{T}}$ and $\overline{\overline{T}}'$.
- The test statistic will be

$$\max\left(\frac{T-\mu_\Gamma}{\omega_\Gamma}, \frac{T'-\mu_\Gamma^{/}}{\omega_\Gamma^{'}}\right).$$

- So we need a bound on the distribution of this quantity.

The respective bounds provide the joint bound

- The bounding statistics $\left(\overline{\overline{T}},\ \overline{\overline{T}}'\right)$ are jointly stochastically larger than $\left(T,\ T'\right)$, so

$$\Pr\left\{\max\left(\frac{\overline{\overline{T}}-\mu_\Gamma}{\omega_\Gamma}, \frac{\overline{\overline{T}}'-\mu_\Gamma^{\prime}}{\omega_\Gamma^{\prime}}\right)\geq k \,\middle|\, \mathcal{F},\mathcal{Z}\right\} \tag{4}$$

$$\geq \Pr\left\{\max\left(\frac{T-\mu_\Gamma}{\omega_\Gamma}, \frac{T'-\mu_\Gamma^{\prime}}{\omega_\Gamma^{\prime}}\right)\geq k \,\middle|\, \mathcal{F},\mathcal{Z}\right\}$$

The respective bounds provide the joint bound

- The bounding statistics $\left(\overline{\overline{T}},\ \overline{\overline{T}}'\right)$ are jointly stochastically larger than $(T,\ T')$, so

$$\Pr\left\{\max\left(\frac{\overline{\overline{T}}-\mu_\Gamma}{\omega_\Gamma}, \frac{\overline{\overline{T}}'-\mu_\Gamma^{'}}{\omega_\Gamma^{'}}\right) \geq k \,\middle|\, \mathcal{F}, \mathcal{Z}\right\} \tag{4}$$

$$\geq \Pr\left\{\max\left(\frac{T-\mu_\Gamma}{\omega_\Gamma}, \frac{T'-\mu_\Gamma^{'}}{\omega_\Gamma^{'}}\right) \geq k \,\middle|\, \mathcal{F}, \mathcal{Z}\right\}$$

- For all $\Gamma \geq 1$, the correlation between $\overline{\overline{T}}$ and $\overline{\overline{T}}'$ is the same, not dependent on Γ, namely $\rho = \sum_{i=1}^{I} q_i\, q_i^{'} / \sqrt{\sum_{i=1}^{I} q_i^2 \sum_{i=1}^{I} q_i^{'2}}$.

The respective bounds provide the joint bound

- The bounding statistics $\left(\overline{\overline{T}}, \overline{\overline{T}}'\right)$ are jointly stochastically larger than (T, T'), so

$$\Pr\left\{\max\left(\frac{\overline{\overline{T}}-\mu_\Gamma}{\omega_\Gamma}, \frac{\overline{\overline{T}}'-\mu_\Gamma'}{\omega_\Gamma'}\right) \geq k \,\middle|\, \mathcal{F}, \mathcal{Z}\right\} \tag{4}$$

$$\geq \Pr\left\{\max\left(\frac{T-\mu_\Gamma}{\omega_\Gamma}, \frac{T'-\mu_\Gamma'}{\omega_\Gamma'}\right) \geq k \,\middle|\, \mathcal{F}, \mathcal{Z}\right\}$$

- For all $\Gamma \geq 1$, the correlation between $\overline{\overline{T}}$ and $\overline{\overline{T}}'$ is the same, not dependent on Γ, namely $\rho = \sum_{i=1}^{I} q_i\, q_i' / \sqrt{\sum_{i=1}^{I} q_i^2 \sum_{i=1}^{I} q_i'^2}$.
- Consider a bivariate Normal distribution with expectations 0, variances 1, and correlation ρ. Let $1 - \Upsilon_\rho(k)$ be the probability that both coordinates of this distribution are less than k. (In R, calculate $\Upsilon_\rho(k)$ using the `mvtnorm` package.) Then as $I \to \infty$ for given Γ, the left side of (4) tends to $\Upsilon_\rho(k)$.

Design sensitivity of the joint procedure

Lemma

If T has design sensitivity $\widetilde{\Gamma}$ and T' has design sensitivity $\widetilde{\Gamma}'$, then

$$\max\left(\frac{T-\mu_{\Gamma}}{\omega_{\Gamma}}, \frac{T'-\mu_{\Gamma}^{\prime}}{\omega_{\Gamma}^{\prime}}\right)$$

has design sensitivity $\max\left(\widetilde{\Gamma}, \widetilde{\Gamma}'\right)$.

- This is consistent with what we saw in the example. The corrected multiple test was almost as insensitive to unmeasured bias as the best of four individual procedures.

Proof of the lemma

Lemma

If T has design sensitivity $\widetilde{\Gamma}$ and T' has design sensitivity $\widetilde{\Gamma}'$, then testing twice has design sensitivity $\max\left(\widetilde{\Gamma}, \widetilde{\Gamma}'\right)$.

Proof.

If $\widetilde{\Gamma} \geq \widetilde{\Gamma}'$, then the power of the test based on T is tending to 1 for any nonzero level in a sensitivity analysis with $\Gamma < \widetilde{\Gamma}$, so for sufficiently large I, with arbitrarily high probability, the deviate $\left(T - \mu_{\Gamma}\right)/\omega_{\Gamma}$ will be greater than k such that $\Upsilon_{\rho}\left(k\right) = \alpha$, so the multiple test procedure will reject H_0. Analogously, for $\Gamma > \widetilde{\Gamma}$, the power based on T and T' is tending to 0. So the design sensitivity is $\widetilde{\Gamma} = \max\left(\widetilde{\Gamma}, \widetilde{\Gamma}'\right)$. The proof for $\widetilde{\Gamma} \leq \widetilde{\Gamma}'$ is parallel. □

Remarks

- The same logic works for more than 2 test statistics. The table above worked with 4 test statistics.

Remarks

- The same logic works for more than 2 test statistics. The table above worked with 4 test statistics.
- The paper considered 12 test statistics (different tests, different scores, using lead levels or logs of lead levels, weighting or not by the amount smoked). Correction for all 12 tests is almost as insensitive as using the best test.

Remarks

- The same logic works for more than 2 test statistics. The table above worked with 4 test statistics.
- The paper considered 12 test statistics (different tests, different scores, using lead levels or logs of lead levels, weighting or not by the amount smoked). Correction for all 12 tests is almost as insensitive as using the best test.
- The median of the pairwise correlations among the 12 upper bounds was 0.82. With such high correlations, the correction using the joint distribution is much less severe than is the Bonferroni inequality.

The Bonferroni inequality is very conservative

- Bonferroni is quite conservative when the correlation ρ is high. A property of the multivariate Normal distribution.

The Bonferroni inequality is very conservative

- Bonferroni is quite conservative when the correlation ρ is high. A property of the multivariate Normal distribution.
- The (multivariate) quantity $\Upsilon_{\rho}(k)$ determines the true size of a procedure that rejects when the maximum standardized deviate is at least k. When the true size is 0.05, what does Bonferroni report as the nominal level?

The Bonferroni inequality is very conservative

- Bonferroni is quite conservative when the correlation ρ is high. A property of the multivariate Normal distribution.
- The (multivariate) quantity $\Upsilon_\rho(k)$ determines the true size of a procedure that rejects when the maximum standardized deviate is at least k. When the true size is 0.05, what does Bonferroni report as the nominal level?

Table: Nominal or reported level using the Bonferroni inequality to correct for multiple testing when the true size is 0.05 with an L-dimensional Normal random variable with equal correlations ρ.

L	Bonferroni's Nominal Level		
	$\rho = 0$	$\rho = 0.8$	$\rho = 0.9$
2	0.051	0.065	0.072
4	0.051	0.086	0.108
6	0.051	0.103	0.137
10	0.051	0.131	0.189

Sample splitting: an alternative to testing twice

- **Sample splitting**: Split the sample into 10% and 90%. Make decisions using the 10%, then discard it. Do one analysis of the 90%.

Sample splitting: an alternative to testing twice

- **Sample splitting**: Split the sample into 10% and 90%. Make decisions using the 10%, then discard it. Do one analysis of the 90%.
- **Example**: Form 679 matched pairs. Sample 68 pairs. Plan the study using 68 pairs. Do a planned analysis of $679 - 68 = 611$ pairs.

Sample splitting: an alternative to testing twice

- **Sample splitting**: Split the sample into 10% and 90%. Make decisions using the 10%, then discard it. Do one analysis of the 90%.
- **Example**: Form 679 matched pairs. Sample 68 pairs. Plan the study using 68 pairs. Do a planned analysis of $679 - 68 = 611$ pairs.
- **Meta-Proposition**: For finitely many well-defined choices that can be decided with data, sample splitting attains the best design sensitivity. (As $I \to \infty$, $I/10$ pairs make choices correctly, and the replacement of I pairs by $9I/10$ pairs is inconsequential for design sensitivity.)

Sample splitting: an alternative to testing twice

- **Sample splitting**: Split the sample into 10% and 90%. Make decisions using the 10%, then discard it. Do one analysis of the 90%.
- **Example**: Form 679 matched pairs. Sample 68 pairs. Plan the study using 68 pairs. Do a planned analysis of $679 - 68 = 611$ pairs.
- **Meta-Proposition**: For finitely many well-defined choices that can be decided with data, sample splitting attains the best design sensitivity. (As $I \rightarrow \infty$, $I/10$ pairs make choices correctly, and the replacement of I pairs by $9I/10$ pairs is inconsequential for design sensitivity.)
- **Reflection in light of evidence**: Splitting permits thoughtful planning of ill-defined choices in the presence of data.

Sample splitting: an alternative to testing twice

- **Sample splitting**: Split the sample into 10% and 90%. Make decisions using the 10%, then discard it. Do one analysis of the 90%.
- **Example**: Form 679 matched pairs. Sample 68 pairs. Plan the study using 68 pairs. Do a planned analysis of $679 - 68 = 611$ pairs.
- **Meta-Proposition**: For finitely many well-defined choices that can be decided with data, sample splitting attains the best design sensitivity. (As $I \rightarrow \infty$, $I/10$ pairs make choices correctly, and the replacement of I pairs by $9I/10$ pairs is inconsequential for design sensitivity.)
- **Reflection in light of evidence**: Splitting permits thoughtful planning of ill-defined choices in the presence of data.
- **Reference**: Heller, Small and Rosenbaum (JASA, 2009).

Sample splitting, continued

- **Selecting one of several outcomes**: In a sensitivity analysis, $\Gamma > 1$, with $K = 2$, 4, 8, or 16 possible outcomes, a 10/90 split of $I = 1000$ pairs outperforms use of the Bonferroni inequality (although both attain the best design sensitivity). (That is, the sensitivity analysis has higher power).

Sample splitting, continued

- **Selecting one of several outcomes**: In a sensitivity analysis, $\Gamma > 1$, with $K = 2, 4, 8,$ or 16 possible outcomes, a 10/90 split of $I = 1000$ pairs outperforms use of the Bonferroni inequality (although both attain the best design sensitivity). (That is, the sensitivity analysis has higher power).
- **Weighted combination of several outcomes**: In a sensitivity analysis, $\Gamma > 1$, with $K = 8$ outcomes, a 10/90 split of 1000 pairs to determine a weighted combination of outcomes outperformed (i) use of the Bonferroni inequality (except when only one outcome was affected, and then the difference was small), (ii) a fixed weighting (expect when the fixed weighting of $K = 8$ outcomes coincided with the optimal weighting).

Sample splitting, continued

- **Selecting one of several outcomes**: In a sensitivity analysis, $\Gamma > 1$, with $K = 2$, 4, 8, or 16 possible outcomes, a 10/90 split of $I = 1000$ pairs outperforms use of the Bonferroni inequality (although both attain the best design sensitivity). (That is, the sensitivity analysis has higher power).
- **Weighted combination of several outcomes**: In a sensitivity analysis, $\Gamma > 1$, with $K = 8$ outcomes, a 10/90 split of 1000 pairs to determine a weighted combination of outcomes outperformed (i) use of the Bonferroni inequality (except when only one outcome was affected, and then the difference was small), (ii) a fixed weighting (expect when the fixed weighting of $K = 8$ outcomes coincided with the optimal weighting).
- **Reference**: Heller, Small and Rosenbaum (JASA, 2009).

Summary

- **Design sensitivity** $\widetilde{\Gamma}$: The power of a sensitivity analysis performed at Γ will tend to 1 if $\Gamma < \widetilde{\Gamma}$ and to 0 if $\Gamma > \widetilde{\Gamma}$.

Summary

- **Design sensitivity** $\widetilde{\Gamma}$: The power of a sensitivity analysis performed at Γ will tend to 1 if $\Gamma < \widetilde{\Gamma}$ and to 0 if $\Gamma > \widetilde{\Gamma}$.
- **Choice of test statistic**: In a given sampling situation, the design sensitivity $\widetilde{\Gamma}$ will be different for different test statistics.

Summary

- **Design sensitivity** $\widetilde{\Gamma}$: The power of a sensitivity analysis performed at Γ will tend to 1 if $\Gamma < \widetilde{\Gamma}$ and to 0 if $\Gamma > \widetilde{\Gamma}$.
- **Choice of test statistic**: In a given sampling situation, the design sensitivity $\widetilde{\Gamma}$ will be different for different test statistics.
- **Wilcoxon's signed rank statistic**: has poor design sensitivity if $Y_i = \tau + \epsilon_i$ with ϵ_i Normal, logistic, or t on 3 or 4 degrees of freedom.

Summary

- **Design sensitivity** $\widetilde{\Gamma}$: The power of a sensitivity analysis performed at Γ will tend to 1 if $\Gamma < \widetilde{\Gamma}$ and to 0 if $\Gamma > \widetilde{\Gamma}$.
- **Choice of test statistic**: In a given sampling situation, the design sensitivity $\widetilde{\Gamma}$ will be different for different test statistics.
- **Wilcoxon's signed rank statistic**: has poor design sensitivity if $Y_i = \tau + \epsilon_i$ with ϵ_i Normal, logistic, or t on 3 or 4 degrees of freedom.
- **In terms of** $\widetilde{\Gamma}$: several choices of $(m, \underline{m}, \overline{m})$ increase $\widetilde{\Gamma}$ relative to Wilcoxon's statistic for all of these sampling situations.

Summary

- **Design sensitivity** $\widetilde{\Gamma}$: The power of a sensitivity analysis performed at Γ will tend to 1 if $\Gamma < \widetilde{\Gamma}$ and to 0 if $\Gamma > \widetilde{\Gamma}$.
- **Choice of test statistic**: In a given sampling situation, the design sensitivity $\widetilde{\Gamma}$ will be different for different test statistics.
- **Wilcoxon's signed rank statistic**: has poor design sensitivity if $Y_i = \tau + \epsilon_i$ with ϵ_i Normal, logistic, or t on 3 or 4 degrees of freedom.
- **In terms of** $\widetilde{\Gamma}$: several choices of $(m, \underline{m}, \overline{m})$ increase $\widetilde{\Gamma}$ relative to Wilcoxon's statistic for all of these sampling situations.
- **Testing twice**: In exchange for a small correction for multiple testing, one obtains the design sensitivity of the best of several tests.

Typically additive effects are similar to additive effects

- Treatment typically has an additive effect, $r_{Tij} - r_{Cij} = \tau + \xi_{ij}$ where the ξ_{ij} are mutually independent, independent of everything else, symmetric about 0.

Typically additive effects are similar to additive effects

- Treatment typically has an additive effect, $r_{Tij} - r_{Cij} = \tau + \xi_{ij}$ where the ξ_{ij} are mutually independent, independent of everything else, symmetric about 0.
- If the treatment typically has an additive effect, $r_{Tij} - r_{Cij} = \tau + \xi_{ij}$, then

$$\begin{aligned}
Y_i &= (Z_{i1} - Z_{i2}) \left(r_{Ci1} + Z_{i1}\tau + Z_{i1}\xi_{i1} - r_{Ci2} - Z_{i2}\tau \right. \\
&= \tau + \epsilon_i^{'} \text{ where } \epsilon_i^{'} = \epsilon_i + \xi_i^{'} \\
\text{where, as before, } \epsilon_i &= (Z_{i1} - Z_{i2}) \left(r_{Ci1} - r_{Ci2} \right), \\
\text{and now } \xi_i^{'} &= \left(Z_{i1}\xi_{i1} - Z_{i2}\xi_{i2} \right).
\end{aligned}$$

Typically additive effects are similar to additive effects

- Treatment typically has an additive effect, $r_{Tij} - r_{Cij} = \tau + \xi_{ij}$ where the ξ_{ij} are mutually independent, independent of everything else, symmetric about 0.
- If the treatment typically has an additive effect, $r_{Tij} - r_{Cij} = \tau + \xi_{ij}$, then

$$\begin{aligned} Y_i &= (Z_{i1} - Z_{i2})\,(r_{Ci1} + Z_{i1}\tau + Z_{i1}\xi_{i1} - r_{Ci2} - Z_{i2} \\ &= \tau + \epsilon_i^{'} \text{ where } \epsilon_i^{'} = \epsilon_i + \xi_i^{'} \\ \text{where, as before, } \epsilon_i &= (Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2})\,, \\ \text{and now } \xi_i^{'} &= (Z_{i1}\xi_{i1} - Z_{i2}\xi_{i2})\,. \end{aligned}$$

- Because ξ_{ij} is independent of everything else and symmetric about 0, $\xi_i^{'} = (Z_{i1}\xi_{i1} - Z_{i2}\xi_{i2})$ has the same distribution as ξ_{ij}, is symmetric about 0, and is independent of the Z_{ij}.

Typically additive effects are similar to additive effects

- Treatment typically has an additive effect, $r_{Tij} - r_{Cij} = \tau + \xi_{ij}$ where the ξ_{ij} are mutually independent, independent of everything else, symmetric about 0.
- If the treatment typically has an additive effect, $r_{Tij} - r_{Cij} = \tau + \xi_{ij}$, then

$$\begin{aligned} Y_i &= (Z_{i1} - Z_{i2}) \left(r_{Ci1} + Z_{i1}\tau + Z_{i1}\xi_{i1} - r_{Ci2} - Z_{i2}\tau \right. \\ &= \tau + \epsilon_i^{'} \text{ where } \epsilon_i^{'} = \epsilon_i + \xi_i^{'} \\ \text{where, as before, } \epsilon_i &= (Z_{i1} - Z_{i2}) \left(r_{Ci1} - r_{Ci2} \right), \\ \text{and now } \xi_i^{'} &= \left(Z_{i1}\xi_{i1} - Z_{i2}\xi_{i2} \right). \end{aligned}$$

- Because ξ_{ij} is independent of everything else and symmetric about 0, $\xi_i^{'} = (Z_{i1}\xi_{i1} - Z_{i2}\xi_{i2})$ has the same distribution as ξ_{ij}, is symmetric about 0, and is independent of the Z_{ij}.
- If $H_{\tau_0} : \tau = \tau_0$ were true in a randomized experiment, then $Y_i - \tau_0 = \epsilon_i^{'}$ would be independent of Z_{ij} and symmetric about 0, and this is the basis for inference about the (typical) effect τ.

Typically additive effects are similar to additive effects

- Treatment typically has an additive effect, $r_{Tij} - r_{Cij} = \tau + \xi_{ij}$ where the ξ_{ij} are mutually independent, independent of everything else, symmetric about 0.
- If the treatment typically has an additive effect, $r_{Tij} - r_{Cij} = \tau + \xi_{ij}$, then

$$\begin{aligned}
Y_i &= (Z_{i1} - Z_{i2})\,(r_{Ci1} + Z_{i1}\tau + Z_{i1}\xi_{i1} - r_{Ci2} - Z_{i2}\tau \\
&= \tau + \epsilon_i^{'} \text{ where } \epsilon_i^{'} = \epsilon_i + \xi_i^{'} \\
\text{where, as before, } \epsilon_i &= (Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2})\,, \\
\text{and now } \xi_i^{'} &= (Z_{i1}\xi_{i1} - Z_{i2}\xi_{i2})\,.
\end{aligned}$$

- Because ξ_{ij} is independent of everything else and symmetric about 0, $\xi_i^{'} = (Z_{i1}\xi_{i1} - Z_{i2}\xi_{i2})$ has the same distribution as ξ_{ij}, is symmetric about 0, and is independent of the Z_{ij}.
- If $H_{\tau_0} : \tau = \tau_0$ were true in a randomized experiment, then $Y_i - \tau_0 = \epsilon_i^{'}$ would be independent of Z_{ij} and symmetric about 0, and this is the basis for inference about the (typical) effect τ.

The new U-statistic

- Fix three integers, m, $\underline{m}$, $\overline{m}$ with $1 \leq \underline{m} \leq \overline{m} \leq m < I$. Let $\mathcal{K}$ be the set containing the $\binom{I}{m}$ sequences $\mathcal{I} = \langle i_1, \ldots, i_m \rangle$ of m distinct integers $1 \leq i_1 < \cdots < i_m \leq I$, and write $\mathbf{Y}_{\mathcal{I}} = \langle Y_{i_1}, \ldots, Y_{i_m} \rangle$.

The new U-statistic

- Fix three integers, m, $\underline{m}$, $\overline{m}$ with $1 \leq \underline{m} \leq \overline{m} \leq m < I$. Let $\mathcal{K}$ be the set containing the $\binom{I}{m}$ sequences $\mathcal{I} = \langle i_1, \ldots, i_m \rangle$ of m distinct integers $1 \leq i_1 < \cdots < i_m \leq I$, and write $\mathbf{Y}_{\mathcal{I}} = \langle Y_{i_1}, \ldots, Y_{i_m} \rangle$.
- A U-statistic (Hoeffding 1948) has the form

$$T = \binom{I}{m}^{-1} \sum_{\mathcal{I} \in \mathcal{K}} h\left(\mathbf{Y}_{\mathcal{I}}\right)$$

where $h(\cdot)$ is a symmetric function.

The new U-statistic

- Fix three integers, m, $\underline{m}$, $\overline{m}$ with $1 \leq \underline{m} \leq \overline{m} \leq m < I$. Let $\mathcal{K}$ be the set containing the $\binom{I}{m}$ sequences $\mathcal{I} = \langle i_1, \ldots, i_m \rangle$ of m distinct integers $1 \leq i_1 < \cdots < i_m \leq I$, and write $\mathbf{Y}_{\mathcal{I}} = \langle Y_{i_1}, \ldots, Y_{i_m} \rangle$.
- A U-statistic (Hoeffding 1948) has the form

$$T = \binom{I}{m}^{-1} \sum_{\mathcal{I} \in \mathcal{K}} h\left(\mathbf{Y}_{\mathcal{I}}\right)$$

where $h(\cdot)$ is a symmetric function.

- For $\mathcal{I} = \langle i_1, \ldots, i_m \rangle \in \mathcal{K}$, sort $Y_{i_1}, \ldots, Y_{i_m}$ to $Y_{[\mathcal{I},1]}, \ldots, Y_{[\mathcal{I},m]}$ to be increasing in absolute value, $0 < \left|Y_{[\mathcal{I},1]}\right| < \cdots < \left|Y_{[\mathcal{I},m]}\right|$.

The new U-statistic

- Fix three integers, m, $\underline{m}$, $\overline{m}$ with $1 \leq \underline{m} \leq \overline{m} \leq m < I$. Let $\mathcal{K}$ be the set containing the $\binom{I}{m}$ sequences $\mathcal{I} = \langle i_1, \ldots, i_m \rangle$ of m distinct integers $1 \leq i_1 < \cdots < i_m \leq I$, and write $\mathbf{Y}_{\mathcal{I}} = \langle Y_{i_1}, \ldots, Y_{i_m} \rangle$.
- A U-statistic (Hoeffding 1948) has the form

$$T = \binom{I}{m}^{-1} \sum_{\mathcal{I} \in \mathcal{K}} h\left(\mathbf{Y}_{\mathcal{I}}\right)$$

  where $h(\cdot)$ is a symmetric function.
- For $\mathcal{I} = \langle i_1, \ldots, i_m \rangle \in \mathcal{K}$, sort $Y_{i_1}, \ldots, Y_{i_m}$ to $Y_{[\mathcal{I},1]}, \ldots, Y_{[\mathcal{I},m]}$ to be increasing in absolute value, $0 < \left| Y_{[\mathcal{I},1]} \right| < \cdots < \left| Y_{[\mathcal{I},m]} \right|$.
- In the new u-statistic, $h\left(\mathbf{Y}_{\mathcal{I}}\right)$ is the number of positive differences among $Y_{[\mathcal{I},\underline{m}]}, \ldots, Y_{[\mathcal{I},\overline{m}]}$, so $h\left(\mathbf{Y}_{\mathcal{I}}\right)$ is an integer in $\left\{0, 1, \ldots, \overline{m} - \underline{m} + 1\right\}$.

Familiar instances of the new U-statistic

- To repeat: $0 < \left|Y_{[\mathcal{I},1]}\right| < \cdots < \left|Y_{[\mathcal{I},m]}\right|$, $h\left(\mathbf{Y}_{\mathcal{I}}\right)$ is the number of positive differences among $Y_{[\mathcal{I},\underline{m}]}, \ldots, Y_{[\mathcal{I},\overline{m}]}$, $T = \binom{I}{m}^{-1} \sum_{\mathcal{I}\in\mathcal{K}} h\left(\mathbf{Y}_{\mathcal{I}}\right)$

Familiar instances of the new U-statistic

- To repeat: $0 < \left|Y_{[\mathcal{I},1]}\right| < \cdots < \left|Y_{[\mathcal{I},m]}\right|$, $h\left(\mathbf{Y}_{\mathcal{I}}\right)$ is the number of positive differences among $Y_{[\mathcal{I},\underline{m}]}, \ldots, Y_{[\mathcal{I},\overline{m}]}$, $T = \binom{I}{m}^{-1} \sum_{\mathcal{I}\in\mathcal{K}} h\left(\mathbf{Y}_{\mathcal{I}}\right)$
- **Sign test**: if $m = \overline{m} = \underline{m} = 1$, then $h\left(\mathbf{Y}_{\mathcal{I}}\right) = \operatorname{sgn}\left(Y_{i_1}\right) = \operatorname{sgn}\left(Y_{[\mathcal{I},1]}\right)$ and T is the sign statistic.

Familiar instances of the new U-statistic

- To repeat: $0 < \left|Y_{[\mathcal{I},1]}\right| < \cdots < \left|Y_{[\mathcal{I},m]}\right|$, $h(\mathbf{Y}_{\mathcal{I}})$ is the number of positive differences among $Y_{[\mathcal{I},\underline{m}]}, \ldots, Y_{[\mathcal{I},\overline{m}]}$,
  $T = \binom{I}{m}^{-1} \sum_{\mathcal{I} \in \mathcal{K}} h(\mathbf{Y}_{\mathcal{I}})$
- **Sign test**: if $m = \overline{m} = \underline{m} = 1$, then $h(\mathbf{Y}_{\mathcal{I}}) = \mathrm{sgn}(Y_{i_1}) = \mathrm{sgn}\left(Y_{[\mathcal{I},1]}\right)$ and T is the sign statistic.
- **Wilcoxon's signed rank**: If $m = \overline{m} = \underline{m} = 2$, then $h(\mathbf{Y}_{\mathcal{I}}) = \mathrm{sgn}\left(Y_{[\mathcal{I},2]}\right)$, and T is the u-statistic that closely approximates Wilcoxon's signed rank statistic (Lehmann 1975, p. 337).

Familiar instances of the new U-statistic

- To repeat: $0 < \left|Y_{[\mathcal{I},1]}\right| < \cdots < \left|Y_{[\mathcal{I},m]}\right|$, $h\left(\mathbf{Y}_{\mathcal{I}}\right)$ is the number of positive differences among $Y_{[\mathcal{I},\underline{m}]}, \dots, Y_{[\mathcal{I},\overline{m}]}$, $T = \binom{I}{m}^{-1} \sum_{\mathcal{I}\in\mathcal{K}} h\left(\mathbf{Y}_{\mathcal{I}}\right)$
- **Sign test**: if $m = \overline{m} = \underline{m} = 1$, then $h\left(\mathbf{Y}_{\mathcal{I}}\right) = \mathrm{sgn}\left(Y_{i_1}\right) = \mathrm{sgn}\left(Y_{[\mathcal{I},1]}\right)$ and T is the sign statistic.
- **Wilcoxon's signed rank**: If $m = \overline{m} = \underline{m} = 2$, then $h\left(\mathbf{Y}_{\mathcal{I}}\right) = \mathrm{sgn}\left(Y_{[\mathcal{I},2]}\right)$, and T is the u-statistic that closely approximates Wilcoxon's signed rank statistic (Lehmann 1975, p. 337).
- **Stephenson's statistic**: If $m = \overline{m} = \underline{m} \geq 1$, then $h\left(\mathbf{Y}_{\mathcal{I}}\right) = \mathrm{sgn}\left(Y_{[\mathcal{I},m]}\right)$ and T is Stephenson's (1981) statistic. Excellent power when only a subset of treated subjects respond to treatment; see Conover and Salsburg (1988) and Rosenbaum (2007; 2010a, §16).