ExpEcon Methods: Fay & Proschan (2010) Perspectives for Hypothesis Testing

ECON 8877

P.J. Healy

First version thanks to Sungmin Park

Updated 2023-11-17

Two Popular Statistical Tests

Two samples: $Y^0 = (Y_1^0, \ldots, Y_n^0)$ and $Y^1 = (Y_1^1, \ldots, Y_m^1)$. Which is "bigger"?
Given data $X = (Y^0, Y^1)$ and significance level α, reject H if...

- **Student's *t*-test**: reject if

$$\left| \frac{\hat{\mu}_1 - \hat{\mu}_0}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}} \right| > t_{n-2}^{-1}(1-\alpha/2),$$

where $\hat{\sigma}^2$ is the pooled sample variance, $n = n_1 + n_0$, and $t_d(\cdot)$ is the CDF of Student t distribution with degree of freedom d.

  - Parametric. Test of means? Assumes normality?

- **Wilcoxon/Mann-Whitney rank-sum test**: reject if

$$\sum_{i=1}^{n} \sum_{j=1}^{m} S(Y_i^0, Y_j^1) < U_{n_0,n_1}^{-1}(1-\alpha/2) \quad \text{where } S(x, y) = \begin{cases} 1, & \text{if x > y,} \\ \frac{1}{2}, & \text{if x = y,} \\ 0, & \text{if x < y.} \end{cases}$$

  - Non-parametric. Test of medians??? Assumes what???

Fay and Proschan [2010]

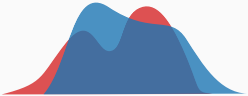
Question

When is it appropriate to use Wilcoxon-Mann-Whitney (WMW) test or *t*-test to compare two samples?

- When is it valid & consistent? When is it optimal?

Answer

They are appropriate for different pairs of null and alternative hypotheses ("perspectives")



Illustration

Illustration: 9th Grade Math Ability of Boys & Girls

Abbildung 1: Histograms of math ability

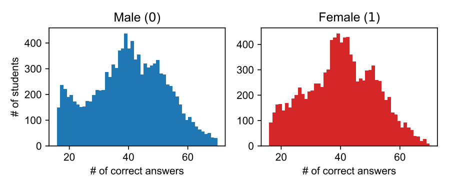


Tabelle 1: Summary statistics of math ability

Statistic		Sample (j)	
		Male (0)	Female (1)
Obs.	n_j	10,887	10,557
Mean	$\hat{\mu}_j$	40.17	40.20
Median		40.44	40.36
Variance	$\hat{\sigma}_j^2$	152.00	134.74

Source: High School Longitudinal Study (HSLS) of 2009

- Assuming each obs is independent, should we use t-test? WMW test? To test what?
- Fay and Proschan (2010) say that the answer depends on your perspective(s).
- A perspective is a pair of null (H) and alternative (K) hypotheses.

One perspective you know from Stats 101

Perspective (Shift in normal distribution)
Let Y denote a random variable. The ***shift-in-normal perspective*** *is*

$$H : \mathbb{E}_F(Y) = \mathbb{E}_G(Y) \quad \textit{versus} \quad K : \mathbb{E}_F(Y) \neq \mathbb{E}_G(Y),$$

where F and G are two normal distributions with the same variance.

- **Student's t-test** (decision rule): Given data X and significance level α, reject H if

$$\left| \frac{\hat{\mu}_1 - \hat{\mu}_0}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}} \right| > t^{-1}_{n-2}(1-\alpha/2),$$

where $\hat{\sigma}^2$ is the pooled sample variance, $n = n_1 + n_0$, and $t_d(\cdot)$ is the CDF of Student t distribution with degree of freedom d.

- Under the above, Student's t-test is not only valid (α works as intended) but also uniformly most powerful (UMP) unbiased. It's also asymptotically most powerful (AMP).

A relaxed perspective, also from Stats 101

Perspective (Behrens-Fisher)
The ***Behrens-Fisher perspective*** *is*

$$H : \mathbb{E}_F(Y) = \mathbb{E}_G(Y) \quad \textit{versus} \quad K : \mathbb{E}_F(Y) \neq \mathbb{E}_G(Y),$$

where F and G are two normal distributions with possibly different variances.

- Under this relaxed perspective, Student's t-test is no longer valid.
- Instead, **Welch's t-test** is asymptotically valid and asymptotically most powerful:

$$\left| \frac{\hat{\mu}_1 - \hat{\mu}_0}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}}} \right| > t_{d_W}^{-1}(1-\alpha/2), \qquad \text{where } d_W = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}\right)^2}{\frac{(\hat{\sigma}_1^2/n_1)^2}{n_1-1} + \frac{(\hat{\sigma}_0^2/n_0)^2}{n_0-1}}$$

$\Rightarrow$ Each statistical test can have multiple valid perspectives. The authors call this idea the Multiple perspective decision rules (MPDR) framework

Even more relaxed perspective

Perspective (Distributions equal or not)

$$H : F = G \quad versus \quad K : F \neq G,$$

where F and G are any two distributions.

- Under this perspective, the t-tests are asymptotically valid and the WMW test is valid. But neither are consistent! (power approaches 1 as $n \to \infty$)
- The WMW test (or Mann-Whitney U test or Wilcoxon rank-sum test) is to reject if
$$\sum_{i=1}^{n} \sum_{j=1}^{m} S(Y_i^0, Y_j^1) < U_{n_0,n_1}^{-1}(1-\alpha/2) \quad \text{where } S(x,y) = \begin{cases} 1, & \text{if x > y,} \\ \frac{1}{2}, & \text{if x = y,} \\ 0, & \text{if x < y.} \end{cases}$$
- Neither t-tests nor WMW test reject the null hypothesis for the 9th-graders' data

Philosophy behind the MPDR framework

- The Multiple perspective decision rules (MPDR) framework has practical value because it suits the nature of scientific theories.
- A scientific theory is often a vague idea or a qualitative result that can be described by more than one statistical model.
  - In biological sciences, for example, the Physicians' Health Study (PHS) aims to test a theory that says prolonged low-dose aspirin decreases cardiovascular mortality.
  - Researchers testing this theory assume a particular statistical model to formulate the null hypothesis, but that model is just one way of representing the data's randomness.
- So we should consider the set of possible statistical assumptions behind a scientific theory to assess which statistical tests (decision rules) are the most useful.

Framework

Terminology

Data	$X \in \mathcal{X}$, where $\mathcal{X}$ is the sample space. Write X_n to denote number of observations n
"Probability model"	A distribution $P \in \mathcal{P}$ on $\mathcal{X}$, where $\mathcal{P} = \{P_\theta \| \theta \in \Theta\}$ with a given parameter space Θ
Null hypothesis	$H = \{P_\theta \| \theta \in \Theta_H\}$
Alternative hypothesis	$K = \{P_\theta \| \theta \in \Theta_K\}$ (Θ_H and Θ_K are disjoint subsets of Θ)
"Assumption"	$A = (\mathcal{X}, H, K)$
Decision rule (test)	$\delta(X, \alpha) \in \{0\,(\text{not reject}), 1\,(\text{reject})\}$, for all data $X \in \mathcal{X}$ and significance level $\alpha \in (0, 0.5)$

Terminology about decision rule (test) δ

"Power" $Pow[\delta(X_n, \alpha); \theta] = \Pr[\delta(X_n, \alpha) = 1; \theta]$ (Probability of rejecting)

"Size" $\alpha_n^* = \sup_{\theta \in \Theta_H} Pow[\delta(X_n, \alpha); \theta]$. (Max. prob. of rejecting given null)

Validity A test δ is valid if $\alpha_n^* \leq \alpha$ for all n.
A test δ is uniformly asymptotically valid (UAV) if $\limsup_{n \to \infty} \alpha_n^* \leq \alpha$.
A test δ is pointwise asymptotically valid (PAV) if, for all $\theta \in \Theta_H$,

$$\limsup_{n \to \infty} Pow[\delta(X_n, \alpha); \theta] \leq \alpha.$$

p-value $p(X) = \inf\{\alpha' : \delta(X, \alpha') = 1\}$ (the strictest α' that rejects)

Terminology about optimal decision rules

Bias A test δ is unbiased if, for all $\theta \in \Theta_K$, power $\geq$ size.

Consistency A test δ is consistent if, for all $\theta \in \Theta_K$, the power approaches 1 as $n \to \infty$.

Optimality A test δ is uniformly most powerful (UMP) if, $\forall \delta'$ and $\forall \theta \in \Theta_K$,

$$Pow[\delta(X, \alpha); \theta] \geq Pow[\delta'(X, \alpha); \theta].$$

A test is UMP unbiased if it is UMP among all unbiased tests.

A test is asymptotically most powerful (AMP) if, as θ_n approaches θ_0,

$$\limsup_{n \to \infty} Pow[\delta(X_n, \alpha); \theta_n] - Pow[\delta'(X_n, \alpha); \theta_n] \geq 0$$

as $\theta_n \in \Theta_K$ approaches $\theta_0 \in \Theta_H$.

Perspectives

Perspective 1

Perspective (Difference in means; same null distribution)

$$H = \{F, G : F = G\}$$
$$K = \{F, G : \mathbb{E}_F(Y) \neq \mathbb{E}_G(Y)\}$$

- Weird ("focusing") perspective because it leaves out many pairs of distributions
- Still, the alternative hypotheses K is a pretty large set
- The WMW test is valid but inconsistent
- The paper doesn't mention how the t-tests fare, but they are likely inconsistent, too.
- So, don't take this perspective.

Perspective 2

Perspective (Stochastic ordering)
Let Ψ_C denote the set of continuous distributions. Write $F <_{st} G$ if G has first-order stochastic dominance over F (i.e. $F(y) \geq G(y)$ for all y and $F(y) > G(y)$ for some y).

$$H = \{F, G : F = G; F \in \Psi_C\}$$
$$K = \{F, G : F <_{st} G \text{ or } G <_{st} F; F, G \in \Psi_C\}$$

- Under this perspective, the WMW test is valid and consistent (Mann and Whitney, 1947). It's also unbiased (Lehmann, 1951)
- The t-tests (both Student's and Welch's) are asymptotically valid and consistent
- So, both the WMW test and t-tests work under this perspective!

Perspective 3

Perspective (Mann-Whitney Functional)
Let $Y_F \sim F$ and $Y_G \sim G$. Define the Mann-Whitney functional ϕ as

$$\phi(F, G) = \Pr[Y_F > Y_G] + \frac{1}{2}\Pr[Y_F = Y_G]$$

The Mann-Whitney functional perspective is

$$H = \{F, G : F = G; F \in \Psi_C\},$$

$$K = \{F, G : \phi(F, G) \neq \frac{1}{2}; F, G \in \Psi_C\}.$$

- A natural perspective by construction. Especially appropriate for ordinal data
- The WMW test is valid and consistent, whereas the t-tests are inconsistent
- So don't use t-tests under this perspective. Use the WMW test

Perspective 4

Perspective (Distribution equal or not)

$$H = \{F, G : F = G\}$$
$$K = \{F, G : F \neq G\}$$

- The WMW test is valid but inconsistent. The t-tests are asymptotically valid but iconsistent.
- If you take this perspective, find a different test like Kolmogorov-Smirnov

Perspectives 5–8: Shifts & scale in distributions

Let Ψ_L, Ψ_C, and Ψ_{LG} denote the sets of logistic, continuous, and log-gamma distributions.
Let Ψ_{D_k} denote the set of discrete distributions with sample space $\{1, 2, \ldots, k\}$

Perspective (Shift in logistic distribution)

$$H = \{F, G : F = G; F \in \Psi_L\}$$

$$K = \{F, G : G(y) = F(y + \Delta); \Delta \neq 0; F \in \Psi_L\}$$

Perspective (Shift in continuous distribution)

$$H = \{F, G : F = G; F \in \Psi_C\}$$

$$K = \{F, G : G(y) = F(y + \Delta); \Delta \neq 0; F \in \Psi_C\}$$

Perspective (Shift in log-gamma distribution)

$$H = \{F, G : F = G; F \in \Psi_{LG}\}$$

$$K = \{F, G : G(y) = F(y + \Delta); \Delta \neq 0; F \in \Psi_{LG}\}$$

Perspective (Proportional odds)

$$H = \{F, G : F = G; F \in \Psi_{D_k}\}$$

$$K = \{F, G : \frac{F(y)}{1-F(y)} = \frac{G(y)}{1-G(y)}\Delta; \Delta \neq 1; F \in \Psi_{D_k}\}$$

- The WMW test is valid and consistent. The t-tests are asymptotically valid and consistent.

Perspective 11: Differences in means assuming normality with same variance

Perspective (Shift in normal distribution)

$$H = \{F, G : F = G; F \in \Psi_N\}$$
$$K = \{F, G : G(y) = F(y + \Delta); \Delta \neq 0; F \in \Psi_N\}$$

where Ψ_N is the set of normal distributions.

- The first perspective you've seen at the beginning.
- The WMW test and the Student's t-test are valid and consistent. The Student's t-test is optimal, because it is UMP unbiased and asymptotically most powerful. The Welch's t-test is asymptotically valid and consistent.

Perspective 14: Differences in means assuming normality with different variance

Perspective (Behrens-Fisher: Diference in normal means, different variances)

$$H = \{F, G : \mathbb{E}_F(Y) = \mathbb{E}_G(Y); F, G \in \Psi_N\}$$

$$K = \{F, G : \mathbb{E}_F(Y) \neq \mathbb{E}_G(Y); F, G \in \Psi_N\}$$

where Ψ_N is the set of normal distributions.

- Both the WMW test and the Student's t-test are invalid and inconsistent
- Welch's t-test is uniformly asymptotically valid and consistent
- So, use Welch's t-test if you take this perspective... but better ones exist:

Perspectives 12–13: Differences in means without assuming normality

Perspective (Finite variances)

$H = \{F, G : F = G; F \in \Psi_{fv}\}$

$K = \{F, G : \mathbb{E}_F(Y) \neq \mathbb{E}_G(Y); F, G \in \Psi_{fv}\}$

where Ψ_{fv} is the set of distributions with finite variances.

- The WMW test is valid but inconsistent
- The t-tests are pointwise asymptotically valid and consistent

Perspective (Finite 4th moments)

$H = \{F, G : F = G; F \in \Psi_{B_\epsilon}\}$

$K = \{F, G : \mathbb{E}_F(Y) \neq \mathbb{E}_G(Y); F, G \in \Psi_{B_\epsilon}\}$

where Ψ_{B_ϵ} is the set of distributions with $Var(Y) \geq \epsilon > 0$ and $\mathbb{E}(Y^4) \leq B < \infty$.

- The WMW test is valid but inconsistent
- The t-tests are uniformly asymptotically valid and consistent

$\Rightarrow$ t-tests are clearly preferable in large samples

Perspective 15: Seemingly natural but invalid perspective

Perspective (Difference in means; any distributions)

$$H = \{F, G : \mathbb{E}_F(Y) = \mathbb{E}_G(Y)\}$$
$$K = \{F, G : \mathbb{E}_F(Y) \neq \mathbb{E}_G(Y)\}$$

- There exists no valid decision rule with some power greater than its significant level
- If you take this loose perspective, nothing works!
- Your perspective needs more structure

If you want to see the full picture...

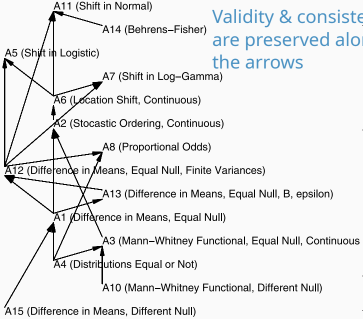


Fig 1. *Relationship between assumptions. $A_i \leftarrow A_j$ denotes that $A_i \sqsubset A_j$ (i.e., A_i are more restrictive assumptions than A_j).*

Validity & consistency are preserved along the arrows

Valid & consistent

Asymptotically valid & consistent

Table 1
Validity and Consistency of Two Sample MPDRs

Decision Rules

WMW =Wilcoxon-Mann-Whitney (exact)
NBF_a =Nonparametric Behrens-Fisher (asymptotic)
NBF_p =Nonparametric Behrens-Fisher (permutation)
t =t-test (pooled variance)
t_W =Welch's t-test (Satterthwaite's df)
t_H =Hsu's t-test (df=min(n_i)-1)
t_p =permutation t-test
t_{BFp} =permutation test on Behrens-Fisher statistic

	Perspective	WMW	NBF_a	NBF_p	t	t_W	t_H	t_p	t_{BFp}
11.	Normal Shift	yy	uy	yy	yy	uy	yy	yy	yy
14.	Behrens-Fisher	n-	ay	ay	n-	uy	yy	n-	ay
5.	Shift in Logistic	yy	uy	yy	ay	ay	ay	yy	yy
7.	Shift in Log-Gamma	yy	uy	yy	ay	ay	ay	yy	yy
6*.	Location Shift, fv	yy	uy	yy	ay	ay	ay	yy	yy
2*.	Stochastic Ordering, SN, fv	yy	uy	yy	ay	ay	ay	yy	yy
8.	Proportional Odds, SN	yy	uy	yy	ay	ay	ay	yy	yy
12.	Diff in Means, SN,fv	yn	un	yn	py	py	py	yy	yy
13.	Diff in Means, SN, $B\epsilon$	yn	un	yn	uy	uy	uy	yy	yy
3*.	Mann-Whitney Func., SN, fv	yy	uy	yy	an	an	an	yn	yn
4*.	Distributions Equal or Not, fv	yn	un	yn	an	an	an	yn	yn
15*.	Diff in Means, DN, fv	n-	n-	n-	n-	n-	n-	n-	n-
10*.	Mann-Whitney Func., DN, fv	n-	ay	ay	n-	n-	n-	n-	n-
9.	Randomization Model	y-	- -	y-	- -	- -	- -	y-	y-

Perspective numbers with * have the additional assumption that $F, G \in \Psi_{fv}$ in both H and K.
SN=Same Null Distns., DN=Different Null Distns., fv=Finite Var.,
$B\epsilon = \left\{E(Y^4) \leq B \text{ and } Var(Y) \geq \epsilon\right\}$

Each hypothesis test is represented by 2 sets of symbols representing the 2 properties: (i) validity, and (ii) (pointwise) consistency, where each character answers the question, This test has this property: y=yes, n=no, and − = not applicable.
For validity we also have the symbols: u=UAV, a = PAV, p =PNUAV.

Discussion

Takeaways

So... WMW test or t-test?

- It's important to identify your perspective first! Be *precise*!
- *t*-test is usually only asymptotically valid...
- In the math ability example, maybe use Welch's t-test since $n, m \geq 10,000$
- But depending on the application, the WMW test may be more appropriate
  - For example, if the variable is ordinal. Also, the authors argue that the WMW test is often more powerful than the t-tests in small samples
- In any case, the decision should not depend on whether the data look normally distributed or not, because there are valid perspectives without the normality assumption
- But, stay tuned for the permutation test!

Literatur

Michael P. Fay and Michael A. Proschan. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, 4:1–39, 2010.

The End!