

ExpEcon Methods: Intro to Hypothesis Testing and Fay & Proschan's (2010) "Perspectives"

ECON 8877

P.J. Healy

First version thanks to Sungmin Park

Updated 2026-01-28

A Primer/Reminder on Hypothesis Tests

- Assumed model/DGP (“population”): $F(X; \theta)$, $\theta \in \Theta$
- Sample (r.v.): X . Often $X = (X_1, \dots, X_n)$. Realization: x
- Sample statistic: $W(X)$
 - Ex: 2 groups. $X_i = (Y_i^0, Y_i^1)$, $W(X) = \sum_i Y_i^0/n - \sum_i Y_i^1/n$
- Hypotheses: $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta_1$
 - Here: Assume $\Theta_1 = \Theta \setminus \Theta_0$
- Do not reject H_0 if $W(x) \in W_0 \subseteq \mathbb{R}$
- Reject if $W(x) \in W_0^C$. Let $R = \{x : W(x) \in W_0^C\}$
 - Decision rule: $\delta(x; W_0^C) = 1$ if $W(x) \in W_0^C$
 - Decision rule: $\delta(x; W_0^C) = 0$ if $W(x) \notin W_0^C$
- Rejection region (in sample space): reject H_0 if $x \in R$
 - $\text{Pr}(\text{Type I Error at } \theta \in \Theta_0) = P_\theta(X \in R)$
 - $\text{Pr}(\text{Type II Error at } \theta \in \Theta_0^C) = P_\theta(X \notin R) = 1 - P_\theta(X \in R)$
- Power function: $\beta(\theta) = P_\theta(X \in R)$
 - Ideal: $\beta(\theta) = 0 \forall \theta \in \Theta_0$, $\beta(\theta) = 1 \forall \theta \in \Theta_0^C$

Primer: One-Sided Test

Example: $X = (X_1, \dots, X_n) \in \mathbb{R}^n$, iid $N(\theta, \sigma^2)$, σ^2 known.

- Question: $H_0 : \theta \leq \theta_0$ for some θ_0 (given by research question)
- Test statistic: $W(X) = \frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}}$ (average “z-score”)
- Reject if $W(X) > c$ for some c (that you choose)
 - Why? Equivalent to LRT $\sup_{\theta \in \Theta_0} L(\theta|x) / \sup_{\theta} L(\theta|x) < \lambda$
 - Parametric distribution $\Rightarrow L(\theta|x)$ known
- How to pick c ? Want $\beta(\theta) \leq 0.05$ for all $\theta \leq \theta_0$. So

$$\begin{aligned}\beta(\theta) &= P_{\theta} \left(\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c \right) \\ &= P_{\theta} \left(\frac{\bar{X} - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} \right) \\ &= P_{\theta} \left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} \right) = 1 - \Phi \left(c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} \right)\end{aligned}$$

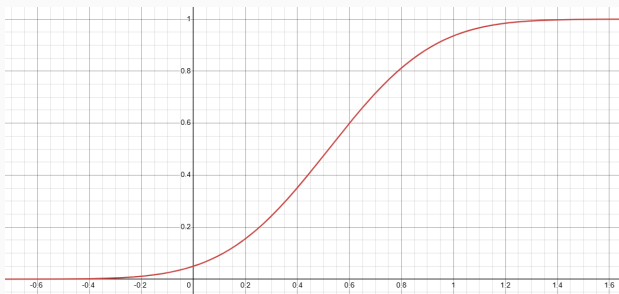
- Increasing in $\theta \Rightarrow$ Want $\beta(\theta_0) = 1 - \Phi(c) = 0.05 \Rightarrow$ Set $c = 1.645$

Primer: One-Sided Test

Example: $X = (X_1, \dots, X_n) \in \mathbb{R}^n$, iid $N(\theta, \sigma^2)$, σ^2 known.

Test: Reject if $W(X) = \frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c = 1.645$

Graph: $\beta(\theta)$ for $\theta_0 = 0$, $n = 10$



Maximum type-I error (among $\theta \leq \theta_0 = 0$) is 0.05 (by design)

Size of test: $\alpha := \sup_{\theta \in \Theta_0} \beta(\theta)$

"I want 80% power." OK but... at which θ ???

Primer: p -Values

So, what's a p -value?

- In general, just another statistic $p(X)$
- But it's an alternative (equivalent) way to run the same test
- But most commonly, rejection rule is $R = \{x : p(x) < \alpha\}$ where

$$p(x) = \sup_{\theta \in \Theta_0} P_{\theta}(W(X) \geq W(x))$$

- “Under H_0 , what the probability of a more-extreme $W(X)$?”
- Reject iff $p(x) < \alpha$
 - Decision rule: $\delta(x; \alpha) = 1$ if $p(x) < \alpha$, $\delta(x; \alpha) = 0$ if $p(x) \geq \alpha$
 - Previously you chose c , reject if $W(X) > c$
 - Now you choose α , reject if $p(X) < \alpha$
- This will generate a valid test for any α
- One-sided test: reject if $W(X) > w^{**}$
- Two-sided test: reject if $W(x) \notin (w^*, w^{**})$

The Example

Again, $X_i \sim N(\theta, \sigma^2)$, $H_0 : \theta \leq \theta_0 (= 0)$.

$$\begin{aligned} p(x) &= \sup_{\theta \in \Theta_0} P_{\theta}(W(X) \geq W(x)) \\ &= \sup_{\theta \in \Theta_0} P_{\theta} \left(\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} \geq \frac{\bar{x} - \theta_0}{\sigma/\sqrt{n}} \right) \\ &= P_{\theta_0} \left(\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} \geq \frac{\bar{x} - \theta_0}{\sigma/\sqrt{n}} \right) \\ &= P_{\theta_0} (Z \geq W(x)) \end{aligned}$$

- Usual rule: reject H_0 iff $p(x) < 0.05$
- Note: $p(x) < 0.05$ iff $W(x) > 1.645$
- So, reject if $W(x) > 1.645$. Same rule as before!!! $c = 1.645$
- Has same meaning regardless of $W(\cdot)$ and W_0
- p -value gives a useful measure of “how close you were” to rejecting/not rejecting

How To Simulate It

- In the normal example, $\beta(\theta)$ has analytic solution
- In general might not exist/too hard to solve
- We can simulate it! Steps:
 1. Set a grid of θ values
 2. Choose a sample size n and number of “runs” R
 3. At each “true” θ generate R iid samples of size n
 - r^{th} sample is $x_r^\theta = (x_{1,r}^\theta, \dots, x_{n,r}^\theta)$ where $x_{i,r}^\theta \stackrel{iid}{\sim} F(\cdot, \theta)$
 4. For each x_r^θ determine $\delta(x_r^\theta; \alpha)$
 5. Estimated power function: $\hat{\beta}(\theta) = \sum_{r=1}^R \frac{1}{R} \delta(x_r^\theta; \alpha)$
 6. Redo this for various $n, \alpha, W(x)$, whatever...
Plot them all and check:
 - 6.1 Is $\hat{\beta}(\theta) \leq 0.05$ when $\theta \in \Theta_0$?
 - 6.2 Which has the greatest $\hat{\beta}(\theta)$ when $\theta \notin \Theta_0$?
- Very useful for picking your actual sample size!

Example MATLAB Code

```
1 %% Set parameters
2 tnot = 0; % the cutoff theta_0
3 tgrid = [-2.5 -1 -0.5 -0.1 0 0.1 0.5 1 1.5 2 2.5 5 10]; %
   true means (theta)
4 n = 100;
5 c = 1.645;
6 sig = 10; %true std deviation
7 runs = 5000; %how many samples to generate for each t and n
```

Part 1: Setting Parameters

Example MATLAB Code

```
1 %% Now run the simulation
2 for ti=1:length(tgrid)
3     t = tgrid(ti); % current "true" theta
4     fdist = makedist('Normal','mu',t,'sigma',sig); %true
        dist'n
5     for r = 1:runs
6         xr = random(fdist,n,1); %rth sample, an nx1 vector
7         W(ti,r) = (mean(xr)-tnot)/(sig/sqrt(n)); %sample
            statistic
8         rej(ti,r) = W(ti,r) > c; %reject or not
9         pval(ti,r) = 1-normcdf(W(ti,r),tnot,1); %our p-value
10    end
11    rejavg(ti) = mean(rej(ti,:));
12    pvalavg(ti) = mean(pval(ti,:));
13 end
```

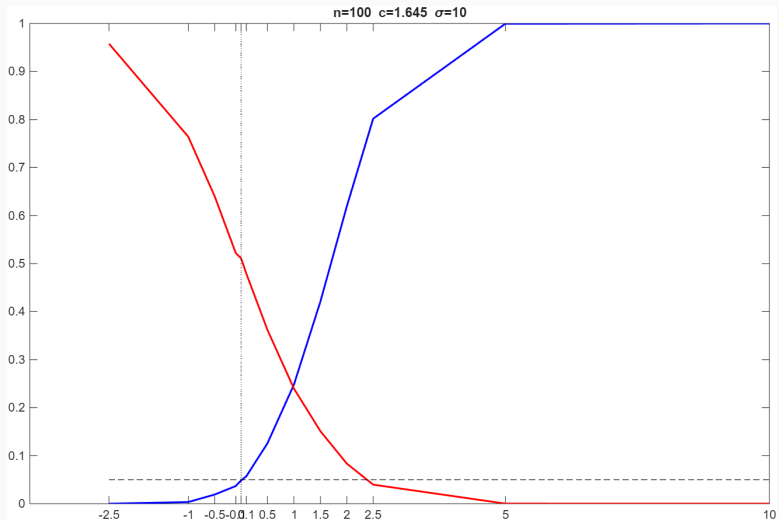
Part 2: Running the Simulation

Example MATLAB Code

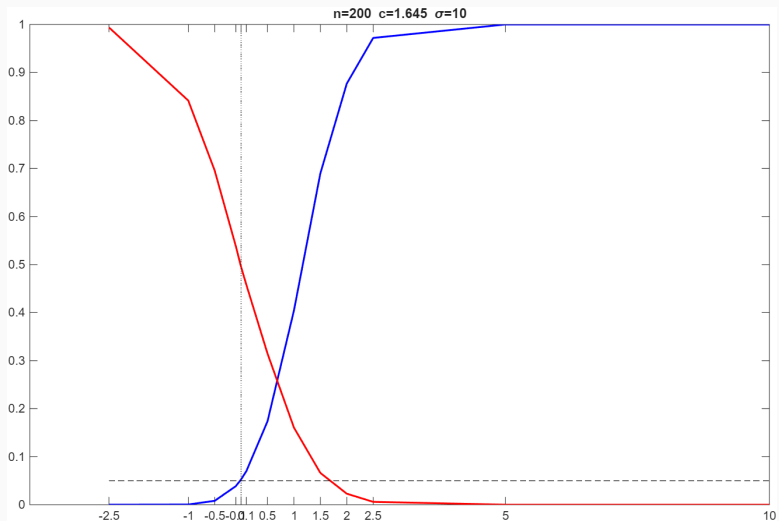
```
1 %% Now plot the output
2 figure; %open new figure
3 plot(tgrid,rejavg,"LineWidth",1.5,"Color","blue"); %plot
    rejavg (power)
4 title("n="+n+" c="+c+" \sigma="+sig);
5 xticks(tgrid);
6 ylim([0 1]);
7 hold on; %allows multiple plots
8 plot(tgrid,pvalavg,"LineWidth",1.5,"Color","red"); %plot
    pvalavg
9 plot(tgrid,0.05*ones(length(tgrid)),"LineStyle","--","Color
    ","k"); %plot y=0.05 as black dashed line
10 plot([tnot tnot],[0 1],"LineStyle",":","Color","k"); %plot x
    =0
11 hold off;
```

Part 3: Plotting Output

Example MATLAB Simulation



Example MATLAB Simulation



Primer: Two-Sided t -Test

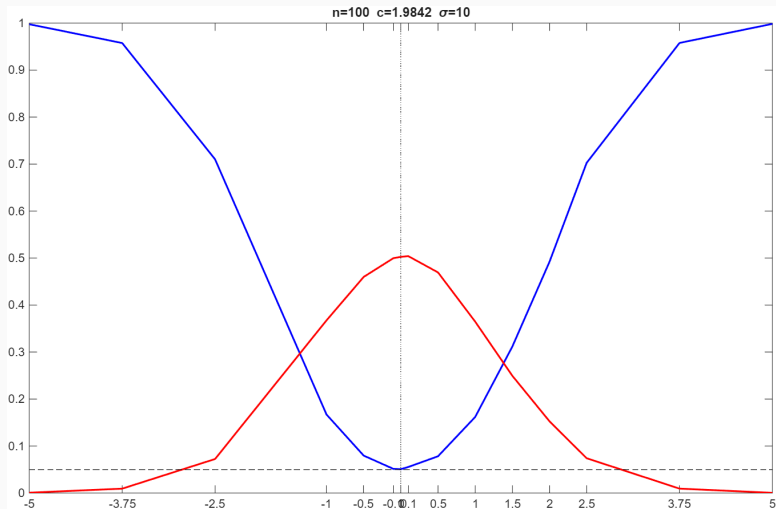
Example: $X = (X_1, \dots, X_n) \in \mathbb{R}^n$, iid $N(\theta, \sigma^2)$, σ^2 **NOT** known.

- Question: $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$
- Test statistic: $W(X) = \frac{\bar{X} - \theta_0}{\hat{\sigma}/\sqrt{n}}$ where $\hat{\sigma} = \sqrt{\sum_i \frac{1}{n-1} (X_i - \bar{X})^2}$
- Reject if $W(X) \notin [-c, c]$ for some c . How to pick c ?
- Want $\beta(\theta) \leq 0.05$ for all $\theta \in \Theta_0 = \{\theta_0\}$. So

$$\begin{aligned}\beta(\theta_0) &= 2 \cdot P_{\theta_0} \left(\frac{\bar{X} - \theta_0}{\hat{\sigma}/\sqrt{n}} > c \right) \\ &= 2 \cdot P_{\theta_0} \left(\frac{\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}}}{\sqrt{\frac{\sum_i (X_i - \bar{X})^2}{\sigma^2} \frac{1}{n-1}}} > c \right) \\ &= 2 \cdot P_{\theta_0} \left(\frac{Z}{\sqrt{\chi^2/(n-1)}} > c \right) = 2 \cdot (1 - T_{n-1}(c))\end{aligned}$$

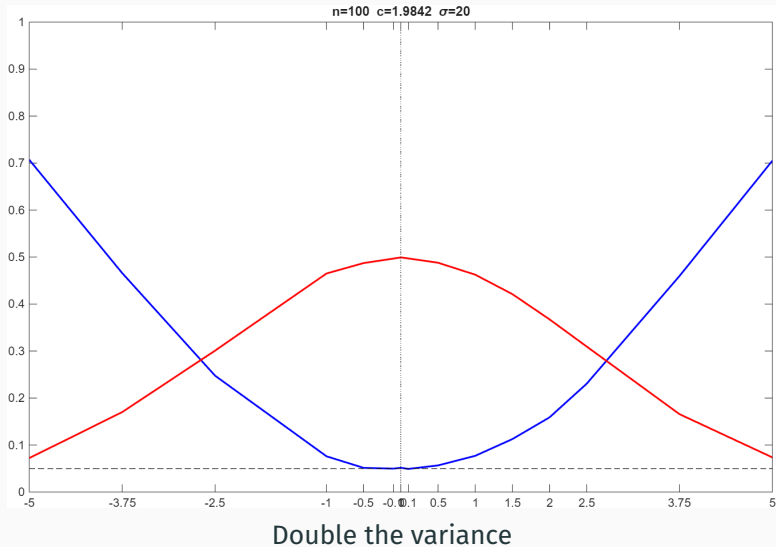
- Want $\beta(\theta_0) = 0.05 \Rightarrow$ Set $c = T_{n-1}^{-1}(0.975)$. Note: depends on n
- Does a closed-form solution exist for $\beta(\theta)$? Let's just simulate!!

Primer: Two-Sided Test



x-axis: True mean. Blue: Power function. Red: Average p -value.

Primer: Two-Sided Test



Comparing tests

- Suppose you have a class of tests \mathcal{C} for a fixed H_0
- Fix n . Test w/ power function $\beta \in \mathcal{C}$ is **uniformly most powerful (UMP)** if $\forall \beta' \in \mathcal{C}, \forall \theta \in \Theta_0^c, \beta(\theta) \geq \beta(\theta')$
 - May not exist
- Test with β is asymptotically (uniformly) most powerful if it becomes UMP as $n \rightarrow \infty$
- A test is **valid** for size α if $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$
- It is **asymptotically valid** if it's valid as $n \rightarrow \infty$

Now We're Ready

OK, now onto Fay & Proschan (2010)...

Two Popular Statistical Tests

Two samples: $Y^0 = (Y_1^0, \dots, Y_n^0)$ and $Y^1 = (Y_1^1, \dots, Y_m^1)$

Assume $Y_i^0 \stackrel{iid}{\sim} F$ and $Y_j^1 \stackrel{iid}{\sim} G$. Is G “bigger” than F ?

In what sense? Mean? Median? FOSD? What does a given test measure?

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

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

where $\hat{\sigma}^2$ is the pooled sample variance 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???

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

Figure 1: Histograms of math ability

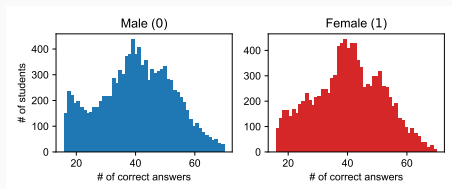


Table 1: Summary statistics of math ability

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 $H : \mathbb{E}_F(Y) = \mathbb{E}_G(Y)$ versus $K : \mathbb{E}_F(Y) \neq \mathbb{E}_G(Y)$,
where F and G are two **normal** distributions with the **same variance**. (Difference must be in means.)

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

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

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 .

- 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

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

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

- Under this relaxed perspective, Student's t -test is no longer valid because it pools the variances.
- **Welch's t -test** uses separate variance estimates, thus is **asymptotically valid** and **asymptotically most powerful**:

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

⇒ 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)

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

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 $n \rightarrow \infty$)

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

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

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 $n \rightarrow \infty$.

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

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

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 ,

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

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

Perspectives

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 (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 (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 inconsistent.
- 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 $\{1, 2, \dots, k\}$

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)

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

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

- WMW: valid and consistent. t-tests: asymp. 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: Difference 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 $\text{Var}(Y) \geq \epsilon > 0$ and $\mathbb{E}(Y^4) \leq B < \infty$.

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

⇒ t-tests are clearly preferable in large samples

Perspective 15: Seemingly natural but invalid perspective

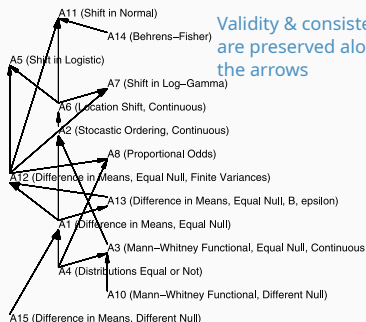
Perspective (Difference in means; any distributions)

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

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

- 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...



Validity & consistency
are preserved along
the arrows

TABLE 1
Validity and Consistency of Two Sample MPDRs

Perspective	Decision Rules									
	WMW	NBF _a	NBF _p	t	t _W	t _H	t _p	t _{BF}	t _{BF}	t _{BF}
11. Normal Shift	yy	yy	yy	yy	yy	yy	yy	yy	yy	yy
14. Behrens-Fisher	n-	ay	ay	n-	yy	yy	n-	ay	yy	yy
5. Shift in Logistic	yy	yy	yy	yy	yy	yy	yy	yy	yy	yy
7. Shift in Log-Gamma	yy	yy	yy	yy	yy	yy	yy	yy	yy	yy
6*. Location Shift, fv	yy	yy	yy	yy	yy	yy	yy	yy	yy	yy
2*. Stochastic Ordering, SN, fv	yy	yy	yy	yy	yy	yy	yy	yy	yy	yy
8. Proportional Odds, SN	yy	yy	yy	yy	yy	yy	yy	yy	yy	yy
12. Diff in Means, SN, fv	yn	un	yn	py	py	py	yy	yy	yy	yy
13. Diff in Means, SN, Be	yn	un	yn	yy	yy	yy	yy	yy	yy	yy
3*. Mann-Whitney Func., SN, fv	yy	yy	yy	an	an	an	yn	yn	yn	yn
4*. Distributions Equal or Not, fv	yn	un	yn	an	an	an	yn	yn	yn	yn
15*. Diff in Means, DN, fv	n-	n-	n-	n-	n-	n-	n-	n-	n-	n-
10*. Mann-Whitney Func., DN, fv	n-	ay	ay	n-	n-	n-	n-	n-	n-	n-
9. Randomization Model	y-	-	y-	-	-	-	y-	y-	y-	y-

Perspective numbers with * have the additional assumption that $F, G \in \Psi_{f_0}$ in both H and K .
 SN=Same Null Dists., DN=Different Null Dists., fv=Finite Var.,
 $Be = \{E(Y^4) \leq B \text{ and } Var(Y) \geq c\}$

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.

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).

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!

References

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!

