Economics 361
Hypothesis Testing and Tennis

Jun Ishii *
Department of Economics
Amherst College

Fall 2023

This handout discusses the hypothesis testing undertaken by the authors, Mark Walker and John Wooders, in their December 2001 *American Economic Review* article, "Minimax Play at Wimbledon." Their basic contention is that while experimental results may fail to find evidence for mixed strategy equilibrium, games played among experienced players may provide evidence consistent with a mixed strategy equilibrium.

Overview of Mixed Strategy Equilibrium

The authors consider the following 2×2 game characterizing strategies associated with tennis serves

		Receiver L	Receiver R
Server	L	Π_{LL}	Π_{LR}
	R	Π_{RL}	Π_{RR}

where Π_{XY} refers to the probability of the Server winning the point when the Server plays X and the Receiver Y. e.g. Π_{LR} is the probability that the Server wins when the Server serves **L**eft and the Receiver anticipates **R**ight.

A single point is at stake for each serve. So Π_{XY} also indicates the expected payoff to the Server when Server plays X and the Receiver Y. Similarly, $1 - \Pi_{XY}$ is the probability that the Receiver wins and corresponds to the expected payoff to the Receiver when Server plays X and Receiver Y.

Let θ_s be the probability with which the server plays L and θ_r the probability with which the receiver plays L. In order for $\{\theta_s, \theta_r\}$ to make up a proper mixed strategy equilibrium, each player must be indifferent to the actual pure-game strategy (L or R) they end up playing.

*Office: Converse Hall 315 Phone: (413) 542-2901 E-mail: jishii@amherst.edu

For the Receiver

$$\begin{aligned}
\text{Expected Payoff for Receiver from Playing L} &= \theta_s\,(1-\Pi_{LL}) + (1-\theta_s)\,(1-\Pi_{RL}) \\
\text{Expected Payoff for Receiver from Playing R} &= \theta_s\,(1-\Pi_{LR}) + (1-\theta_s)\,(1-\Pi_{RR}) \\
\text{Expected Payoff for Receiver from Playing L} &= \text{Expected Payoff for Receiver from Playing R}
\end{aligned}$$

This implies that $\theta_s = \frac{\Pi_{RL}-\Pi_{RR}}{(\Pi_{RL}-\Pi_{RR})-(\Pi_{LL}-\Pi_{LR})}$ for a mixed strategy equilibrium.

Similarly, $\theta_r = \frac{\Pi_{LR}-\Pi_{RR}}{(\Pi_{LR}-\Pi_{RR})-(\Pi_{LL}-\Pi_{RL})}$ for a mixed strategy equilibrium.

Consider the numerical example given in Figure 1

		Receiver L	Receiver R
Server	L	0.58	0.79
	R	0.73	0.49

Based on the earlier calculation, we know that in the mixed strategy equilibrium, the server will play L with the probability of $\frac{8}{15} \approx 0.53$ and the receiver will play L with the probability of $\frac{2}{3} \approx 0.67$.

In Table 1, the authors provide the requisite data from 10 championship matches. They note that the payoff matrix for point games in the championship match may differ depending on [1] who is serving and [2] which side of the court they are serving. Consequently, they break down the championship game further into 4 different types of point games. Each of the 4 types is considered a separate experiment.

For each point game, they provide the number of serves made to each side and the number of points won by serves to a given side. If the players achieve a mixed strategy equilibrium, then the win rates should be the same whether the serve is actually to the left or to the right. Otherwise, the player would prefer not to play a mixed strategy and, instead, commit to whichever of L or R that provides the higher win rate.

Pearson Chi-squared Test

This is a test statistic used to determine whether the observed frequency of an outcome in a random sample is consistent with the expected frequency from an assumed distribution. In general, the test statistic is of the following form:

$$\text{TS} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency of outcome i, E_i is the expected frequency of outcome i, and k is the total number of possible outcomes.

So for a sample of realizations for a random variable with one of two outcomes ("**Bernoulli trial**"), the test statistic is

$$\text{TS} = \frac{(O_1 - E_1)^2}{E_1} + \frac{((N - O_1) - (N - E_1))^2}{E_2}$$

Note that $O_2 = N - O_1$ and $E_2 = N - E_1$ when $k = 2$.

Let p be the probability that outcome 1 is realized. So O_1 is a **Binomial** random variable with mean Np and variance $Np(1-p)$. Therefore, $E_1 = Np$ and $E_2 = N(1-p)$.

$$\begin{aligned} TS &= \frac{(O_1 - Np)^2}{Np} + \frac{((N - O_1) - (N(1-p)))^2}{N(1-p)} \\ &= \left(\frac{O_1 - Np}{\sqrt{Np(1-p)}} \right)^2 \end{aligned}$$

As long as we have a random sample, we know from the **Central Limit Theorem** that

$$\frac{O_1 - Np}{\sqrt{Np(1-p)}} \overset{a}{\sim} N(0,1)$$

Moreover, we know that the square of a statistic distributed standard normal is **Chi-squared** with one degree of freedom. Therefore

$$\text{TS} = \left(\frac{O_1 - Np}{\sqrt{Np(1-p)}} \right)^2 \overset{a}{\sim} \chi_1^2$$

So under the null hypothesis (H_o) that each of the N observed random event is independently and identically distributed Bernoulli with probability p of getting outcome 1, the Pearson test statistic is distributed χ_1^2 asymptotically (convergence in distribution as $N \to +\infty$).

In "Minimax Play at Wimbledon," it is not the frequency of one set of outcomes but, rather, the frequency of two sets of outcomes that are **jointly** being tested. One set of frequency is the outcome associated with serving L and the other set of frequency is the outcome associated with serving R. They should both be governed by the same Bernoulli distribution.

$$H_o : \begin{cases} \text{Prob(winning with serve L)} = p \\ \text{and} \\ \text{Prob(winning with serve R)} = p \end{cases}$$

$$H_a : \begin{cases} \text{Prob(winning with serve L)} \neq p \\ \text{or} \\ \text{Prob(winning with serve R)} \neq p \end{cases}$$

The Pearson test statistic can be amended to allow for these joint hypothesis test by summing together the two associated, individual test statistic. If the outcome of a point associated with serving left is not only independent of other points associated with serving left but *also* other points associated with serving right, then the test statistic for testing whether the win rate for

serving left is p is independent of the test statistic for testing whether the win rate for serving right is p. This means that the sum of the test statistic is the sum of two independent χ^2_1 random variables, yielding a random variable that is distributed χ^2_2.

$$\text{TS'} \quad = \quad \left(\frac{O_1^L - N^L p}{\sqrt{N^L p(1-p)}} \right)^2 + \left(\frac{O_1^R - N^R p}{\sqrt{N^R p(1-p)}} \right)^2 \quad \overset{a}{\sim} \quad \chi^2_2$$

where $\{O_1^L, O_1^R\}$ are the observed frequency of points won serving L and R respectively and $\{N^L, N^R\}$ are the number of total serves L and R respectively.

Note that the above test statistics require the knowledge of p, the "true" probability of a serve winning. In the case where p is replaced with a consistent estimate, the test statistic loses a degree of freedom. So the above test statistic would be distributed χ^2_1. A consistent estimate of the p (under the H_o) is simply the average win rate

$$\hat{p} \quad = \quad \frac{O_1^L + O_1^R}{N^L + N^R}$$

Thus,

$$\begin{aligned}
\hat{\text{TS'}} \quad &= \quad \left(\frac{O_1^L - N^L \hat{p}}{\sqrt{N^L \hat{p}(1-\hat{p})}} \right)^2 + \left(\frac{O_1^R - N^R \hat{p}}{\sqrt{N^R \hat{p}(1-\hat{p})}} \right)^2 \quad \overset{a}{\sim} \quad \chi^2_1 \\
\hat{\text{TS'}} \quad &= \quad \frac{(O_1^L - \hat{E}_1^L)^2}{\hat{E}_1^L} + \frac{((N^L - O_1^L) - (N^L - \hat{E}_1^L))^2}{N^L - \hat{E}_1^L} \\
&\quad + \frac{(O_1^R - \hat{E}_1^R)^2}{\hat{E}_1^R} + \frac{((N^R - O_1^R) - (N^R - \hat{E}_1^R))^2}{N^R - \hat{E}_1^R} \\
&\overset{a}{\sim} \quad \chi^2_1
\end{aligned}$$

where $\hat{E}_1^L = N^L \hat{p}$ and $\hat{E}_1^R = N^R \hat{p}$

The appropriate critical regions for the hypothesis test can be found on the χ^2_1 table, available in the "back" in many standard statistics / econometrics textbooks. Remember: choose the critical region (value) such that the test statistic rejects the H_o when the H_o is true α (significance level) percent of the time.

Power Function

The power function (PF) is simply defined as

$$\begin{aligned} PF(\theta) &= \text{Prob}(\ TS \in \text{Critical Region} \mid \theta) \\ &= \text{Prob}(\ \text{Reject } H_o \mid \theta) \\ &\quad \text{where } H_o : \theta = \theta_o \text{ and } H_a : \theta \neq \theta_o \end{aligned}$$

In words, the power function is the probability of rejecting the null hypothesis, given some test statistic and chosen critical region, as a function of the "true" parameter values governing the data generating process. It is the function used to evaluate a test statistic under the Neyman-Pearson hypothesis testing framework.

The power function evaluated at the null hypothesis ($\theta = \theta_o$) yields the significance level (α) associated with that test statistic and critical region

$$PF(\theta_o) = \text{Prob}(\ \text{Reject } H_o \mid \theta = \theta_o) = \alpha$$

This is because the critical region is chosen such that the probability of rejecting H_o is equal to α when H_o is true ($\theta = \theta_o$). (This is what is meant by "choosing the significance level" of a test)

So, the power function, evaluated at $\theta = \theta_0$, yields the probability of a type I error associated with the hypothesis testing. The power function, evaluated at other non-null values ($\theta \neq \theta_o$), provides 1-probability(type II error) associated with that θ. Ideally, the hypothesis test should have a small value (close to 0) of the power function at $\theta = \theta_0$ and a large value (close to 1) at all other θ.

REMEMBER: In a Neyman-Pearson framework, for a chosen, **fixed** probability of type I error (α), you try to find a hypothesis test that allows you to minimize the probability of type II error.

The "p-value" listed in some tables (e.g. Table 1 in the "Minimax Play at Wimbledon" article) is a concept related to the power function. It is the lowest significance level (α) for which a test based on the calculated test statistic will reject the null hypothesis. From a practical point of view, it tells you that the test statistic will reject the null hypothesis for any test that involves a significance level higher than the p-value. Some researchers prefer to report the p-value rather than the result from a specific α significance hypothesis test. It is a meta-analysis tool.

The simplest way to calculate the power function associated with a hypothesis test is to calculate the **empirical** power function associated with the test, using simulated data.

Consider the following tri-variate, linear regression example:

$$\begin{aligned} Y_i &= \underbrace{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}}_{=X_i\beta} + \epsilon_i \\ Y_i \mid X &\overset{i.i.d.}{\sim} N(X_i\beta, \sigma^2) \end{aligned}$$

Suppose you have data $\{Y_i, X_{1i}, X_{2i}\}_{i=1}^N$ and you estimated the above linear regression equation using OLS on the data. This gives you

$$\hat{\beta}_{\text{OLS}} = b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \sigma^2 (X'X)^{-1} \right)$$

$$\text{where } X = \begin{pmatrix} 1 & X_{11} & X_{21} \\ \vdots & \vdots & \vdots \\ 1 & X_{1N} & X_{2N} \end{pmatrix}$$

For simplicity, let us denote

$$(X'X)^{-1} = \begin{pmatrix} \Sigma_{00} & \Sigma_{01} & \Sigma_{02} \\ \Sigma_{10} & \Sigma_{11} & \Sigma_{12} \\ \Sigma_{20} & \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Suppose you want to test whether variable X_1 has any impact on Y

$$\begin{aligned} H_o: &\quad \beta_1 = 0 \\ H_a: &\quad \beta_1 \neq 0 \end{aligned}$$

If σ^s is known, then a Z-test statistic could be used

$$\text{Z-test} = \frac{b_1 - 0}{\sqrt{\sigma^2 \Sigma_{11}}} \sim N(0,1) \text{ under } H_o$$

But if σ^2 must be estimated using $s^2 = \frac{e'e}{N-3}$, then a t-test statistic might be used instead

$$\text{t-test} = \frac{b_1 - 0}{\sqrt{\sigma^2 \Sigma_{11}}} \sim t_{N-3} \text{ under } H_o$$

Note that the t-test can be used even if $Y \mid X$ is not distributed Normal (you still need *i.i.d.* but can be *i.i.d.* some other distribution). In this case, you do not know the distribution of the t-test under H_o in **small samples** but you do **asymptotically** (as $N \to +\infty$).

For now, let us consider the t-test. Further, suppose $N = 103$. So the critical value should be drawn from the t-distribution table for $103 - 3 = 100$ degrees of freedom. For a two-sided hypothesis test of 5% significance level ($\alpha = 0.05$), the critical region is $\{\text{t-test} < -1.984 \text{ or t-test} > +1.984\}$

So the power function for this $\alpha = 0.05$ hypothesis test is

$$\begin{aligned} PF(\beta, \sigma^2) &= \text{Prob}(\text{ t-test} < -1.984 \text{ or t-test} > +1.984 \mid \beta, \sigma^2 \text{ }) \\ &= 1 - \text{Prob}(\text{ } -1.984 \leq \text{t-test} \leq +1.984 \mid \beta, \sigma^2 \text{ }) \end{aligned}$$

Analytically evaluating a power function may be tricky. So one can evaluate it numerically (or "empirically") using simulated data. Suppose we want to know how the power function varies **fixing** $\{\beta_0 = b_0, \beta_2 = b_2, \sigma^2 = s^2\}$ but allowing β_1 to vary.

Step 1: Choose the range of β_1 values over which to evaluate the Power Function

Make sure that the range includes both the null hypothesis value ($\beta_1 = 0$) and the estimated value ($\beta_1 = b_1$). Choose M evenly dispersed values that span this range. The smoothness with which you can graph your Power Function will depend on the value of M: the more values you consider from this range, the smoother the Power Function.

Step 2: For each of the β_1 values from Step 1 (M in total), create J simulated samples

You will want to keep the same X values as your data. but you will want to draw new Y values associated with the X. Take the case of $\beta_1 = 0.5$. For each observation i (from 1 to N), draw

$$Y_i(j) \sim N(\ b_0 + 0.5 \times X_{1i} + b_2 \times X_{2i}\ ,\ s^2\ )$$

The "j" indexes the simulation run: $j = 1$ to J. For some statistical programs, they may only allow you to draw from the Standard Normal distribution. If you can draw from the standard normal distribution, you can "draw" $Y_i(s)$ by doing the following transformation

$$\begin{aligned} \tilde{Y}_i(j) &\sim N(0,1) \\ Y_i(j) &= \sqrt{s^2}\ \times\ \tilde{Y}_i(j) + (\ b_0 + 0.5 \times X_{1i} + b_2 \times X_{2i}\ ) \end{aligned}$$

In the end, you will have J samples, each with N observations, for each of the M values of β_1 being considered

$$\begin{array}{lll} Y_1(1)\cdots & Y_i(1) & \cdots Y_N(1) \\ & \vdots & \\ Y_1(j)\cdots & Y_i(j) & \cdots Y_N(j) \qquad \text{per each of the } \beta_1 \text{ values} \\ & \vdots & \\ Y_1(J)\cdots & Y_i(J) & \cdots Y_N(J) \end{array}$$

Step 3: For each simulated sample, conduct the hypothesis test

- Run OLS on the simulated sample
- Use b_1 and $s^2 A_{11}$ from OLS to calculate the t-test

  Note: A_{11} should be the same for all simulations as you are using the same X
- See whether the t-test falls within the critical region (< -1.984 or $> +1.984$)

  If it does, the test rejects H_o

  If it does not, the test fails to reject H_o
- Store the outcome as $I_j(\beta_1)$

$$I_{ij}(\beta_1) = \begin{cases} 1 & \text{if } H_o \text{ rejected for simulation } j \text{ and } \beta_1 \text{ value} \\ 0 & \text{if } H_o \text{ not rejected for simulation } j \text{ and } \beta_1 \text{ value} \end{cases}$$

Step 4: Calculate the empirical power function as follows

$$EPF(\beta_1) \quad = \quad \frac{\sum_{j=1}^{J} I_j(\beta_1)}{J}$$

Note: This gives you the value of the empirical power function where β_0 is fixed to be b_0, β_2 is fixed to be b_2, σ^2 is fixed to be s^2 but for M different β_1 values.

Step 5: Graph $EPF(\beta_1)$ across the M values of β_1

Steps 1-5 above can be amended to consider cases where ...

- ... a different hypothesis is being tested
- ... $\{\beta_0, \beta_2, \sigma^2\}$ are either not fixed or fixed to different values

Empirical power functions are not difficult to calculate but does necessitate some modest programming to automate the process.