# Topic 3: Tests in Parametric Models

## Hypothesis Testing By Likelihood Methods

- Let $H_0$ denote a null hypothesis to be tested. Typically, we may represent $H_0$ as a specified family $\mathcal{F}_0$ of distributions for the data.
- For any test procedure $T$, we shall denote by $T_n$ the version based on a sample of size $n$.
- The function
$$\beta_n(T, F) = P_F(T_n \text{ rejects } H_0),$$
defined for distribution function $F$, is called the *power function* of $T_n$ (or of $T$).
  - For $F \in \mathcal{F}_0$, $\beta_n(T, F)$ represents the probability of a Type I error.
  - The quantity
$$\alpha_n(T, F) = \sup_{F \in \mathcal{F}_0} \beta_n(T, F)$$
is called the *size* of the test.
  - For $F \notin \mathcal{F}_0$, the quantity $1 - \beta_n(T, F)$ represents the probability of a Type II error.
- Usually, attention is confined to *consistent* tests: for fixed $F \notin \mathcal{F}_0$, $\beta_n(T, F) \to 0$ as $n \to \infty$.
- Also, usually attention is confined to *unbiased* tests: for $F \notin \mathcal{F}_0$, $\beta_n(T, F) \geq \alpha_n(T, \mathcal{F}_0)$.

A general way to compare two such test procedures is through their power functions. In this regard we shall use the concept of *asymptotic relative efficiency* (ARE).

- For two test procedures $T_A$ and $T_B$, suppose that a performance criterion is tightened in such a way that the respective sample sizes $n_A$ and $n_B$ for $T_A$ and $T_B$ to perform "equivalently" tend to $\infty$ but have ratio $n_A/n_B$ tending to some limit. Then the limit represents the ARE of procedure $T_B$ relative to procedure $T_A$ and is denoted by $e(T_B, T_A)$.
- The earliest approach to ARE was introduced by Pitman (1949). In this approach, two tests sequences $T = \{T_n\}$ and $U = \{U_n\}$ are compared as the Type I and Type II error probabilities tend to positive limits $\alpha$ and $1 - \beta$, respectively.

- In order that $\alpha_n \to \alpha > 0$ and simultaneously $1 - \beta_n \to 1 - \beta > 0$, it is necessary to consider $\beta_n(\cdot)$ evaluated at an alternative $F^{(n)}$ converging at a suitable rate to the null hypothesis $\mathcal{F}_0$.
- In justification of this approach, we might argue that large sample sizes would be relevant in practice only if the alternative of interest were close to the null hypothesis and thus hard to distinguish with only a small sample.

To demonstrate the above point, we consider the following example.
**Example 3.11** Let $X_1, \ldots, X_n$ be iid with $X_1 \sim N(\mu, 1)$.

- Test $H_0 : \mu = 0$ versus $H_1 : \mu = \mu_0 > 0$.
- Construct a test with $\alpha = 0.05$ and $\beta = 0.2005$.
- Reject $H_0$ if $\sqrt{n}\bar{X}_n > 1.645$.
- Note that

$$\beta = P(\sqrt{n}\bar{X}_n \leq 1.645 | \mu = \mu_0) = \Phi(1.645 - \sqrt{n}\mu_0).$$

- If $n \to \infty$ and $\mu_0$ is a fixed positive constant, $\beta \to 0$.
- To ensure $\beta = 0.2005$, it requires that

$$1.645 - \sqrt{n}\mu_0 = -0.84$$

  or $\mu_0 = 2.485n^{-1/2}$.
- Do you notice that $\mu_0$ will change with $n$ which is no longer a fixed alternative?

## Test Statistics for A Simple Null Hypothesis

Although the theory of the following three tests are of most value for composite null hypotheses, it is convenient to begin with simple null hypothesis. Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0 \in R^s$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}^0$.

## Likelihood Ratio Test

- A *likelihood ratio* statistic,

$$\Lambda_n = \frac{L(\boldsymbol{\theta}^0; \mathbf{x})}{\sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \mathbf{x})}$$

  was introduced by Neyman and Pearson (1928).
- $\Lambda_n$ takes values in the interval $[0, 1]$ and $H_0$ is to be rejected for sufficiently small values of $\Lambda_n$.

- The rationale behind LR tests is that when $H_0$ is true, $\Lambda_n$ tends to be close to 1, whereas when $H_1$ is true, $\Lambda_n$ tends to be close to 0,
- The test may be carried out in terms of the statistic

$$\lambda_n = -2 \log \Lambda_n.$$

- For finite $n$, the null distribution of $\lambda_n$ will generally depend on $n$ and on the form of pdf of $X$.
- LR tests are closely related to MLE's.
- Denote MLE by $\hat{\boldsymbol{\theta}}$. For asymptotic analysis, expanding $\lambda_n$ at $\hat{\boldsymbol{\theta}}$ in a Taylor series, we get

$$\begin{aligned} \lambda_n &= -2\left\{-\sum_{i=1}^{n} \log f(X_i, \hat{\boldsymbol{\theta}}) + \sum_{i=1}^{n} \log f(X_i, \boldsymbol{\theta}^0)\right\} \\ &= 2\left\{\frac{1}{2}(\boldsymbol{\theta}^0 - \hat{\boldsymbol{\theta}})^T \left(-\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log f(x;\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right)(\boldsymbol{\theta}^0 - \hat{\boldsymbol{\theta}})\right\}, \end{aligned}$$

  where $\hat{\boldsymbol{\theta}}$ lies between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^0$.
- Since $\boldsymbol{\theta}^*$ is consistent,

$$\lambda_n = n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)^T \left(-\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta_j \partial\theta_k} L(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) + o_P(1).$$

  By the asymptotic normality of $\hat{\boldsymbol{\theta}}$ and

$$-n^{-1}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta_j \partial\theta_k} L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} \xrightarrow{P} \mathbf{I}(\boldsymbol{\theta}^0),$$

  $\lambda_n$ has, under $H_0$, a limiting chi-squared distribution on $s$ degrees of freedom.

**Example 3.12** Consider the testing problem $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ based on iid $X_1, \ldots, X_n$ from the uniform distribution $U(0, \theta)$.

- $L(\theta_0; \mathbf{x}) = \theta_0^{-n} 1_{\{x_{(n)} < \theta_0\}}$
- $\hat{\theta} = x_{(n)}$ (MLE) and $\sup_{\theta\in\Theta} L(\theta; \mathbf{x}) = x_{(n)}^{-n} 1_{\{x_{(n)} < \theta\}}$
- We have

$$\Lambda_n = \begin{cases} (X_{(n)}/\theta^0)^n & X_{(n)} \leq \theta_0 \\ 0 & X_{(n)} > \theta_0 \end{cases}$$

- Reject $H_0$ if $X_{(n)} > \theta_0$ or $X_{(n)}/\theta_0 < c^{1/n}$.
- What is the asymptotic distribution of $\lambda_n$?

- What is $P(n\log(X_{(n)}/\theta^0) \le c)$ where $c < 0$? It is not a $\chi^2$ distribution. (Why???)

**Example 3.13** Consider the testing problem $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 \neq \sigma_0^2$ based on iid $X_1, \ldots, X_n$ from the normal distribution $N(\mu_0, \sigma^2)$.

- $L(\theta^0; \mathbf{x}) = (2\pi\sigma_0^2)^{-n/2} \exp\left[-\sum_i (x_i - \mu_0)^2/2\sigma_0^2\right]$
- $\hat{\sigma}^2 = n^{-1}\sum_i (x_i - \mu_0)^2$ (MLE) and

$$\sup_{\theta\in\Theta} L(\theta; \mathbf{x}) = (2\pi\hat{\sigma}^2)^{-n/2}\exp(-n/2).$$

- We have

$$\Lambda_n = \left(\frac{\hat{\sigma}^2}{\sigma_0^2}\right)^{n/2} \exp\left(\frac{n}{2} - \frac{\sum_i (x_i - \mu_0)^2}{2\sigma_0^2}\right)$$

  or under $H_0$

$$\lambda_n = -n\left\{\ln\left(\frac{1}{n}\sum_{i=1}^n Z_i^2\right) - \left[1 - \left(\frac{1}{n}\sum_{i=1}^n Z_i^2\right)\right]\right\},$$

  where $Z_1, \ldots, Z_n$ are iid $N(0, 1)$.

- Fact: Using CLT, we have

$$\frac{n^{-1}\sum_{i=1}^n Z_i^2 - 1}{\sqrt{2/n}} \xrightarrow{d} N(0, 1)$$

  or

$$\frac{n}{2}\left(\frac{1}{n}\sum_{i=1}^n Z_i^2 - 1\right)^2 \xrightarrow{d} \chi_1^2.$$

- Note that $\ln u \approx -(1-u) - (1-u)^2/2$ when $u$ is near 1 and $n^{-1}\sum_{i=1}^n Z_i^2 \to 1$ in probability by LLN.

- A common question to be asked in Taylor's series approximation is that how many terms we should consider. In this example, it refers to the use of approximation $\ln u \approx -(1-u)$ as a contrast to the second order approximation we use. If we do use the first order approximation, we will end up the difficulty of finding $\lim_n a_n b_n$ when $\lim_n a_n = \infty$ and $\lim_n b_n = 0$.

- We conclude that $\lambda_n$ has a limiting chi-squared distribution with 1 degree of freedom.

## The Wald Test

- Let $\hat{\boldsymbol{\theta}}_n$ denote a consistent, asymptotically normal, and asymptotically efficient sequence of solutions of the likelihood equations.

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\theta}))$$

  as $n \to \infty$.

- Because $\mathbf{I}(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$, we have

$$\mathbf{I}(\hat{\boldsymbol{\theta}}_n) \xrightarrow{P} \mathbf{I}(\boldsymbol{\theta})$$

as $n \to \infty$.

- Replace the matrix $\left(-\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta_j \partial\theta_k} L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}\right)$ by $\mathbf{I}(\hat{\boldsymbol{\theta}}_n)$ in large sample approximation of $\lambda_n$, we get a second statistic,

$$W_n = n(\hat{\boldsymbol{\theta}}_n - \theta^0)^T \mathbf{I}(\hat{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0),$$

which was introduced by Wald (1943).

- By Slutsky's theorem, $W_n$ converges in distribution to $\chi^2_s$.

- For the construction of confidence region, one generates $\{\boldsymbol{\theta}^0 : W_n \leq \chi^2_{s,\alpha}\}$ which is an ellipsoid in $R^s$.

- As a remark, for the construction of confidence region based on $\lambda_n$, one generates $\{\boldsymbol{\theta}^0 : \lambda_n \leq \chi^2_{s,\alpha}\}$ which is not necessary an ellipsoid in $R^s$.

## The Rao Score Tests

- Both the Wald and likelihood ratio tests requires evaluation of $\hat{\boldsymbol{\theta}}_n$. Now we consider a test for which this is not necessary.

- Denote the likelihood score vector

$$q(\mathbf{x}; \boldsymbol{\theta}) = (q_1(\mathbf{x}; \boldsymbol{\theta}), \ldots, q_s(\mathbf{x}; \boldsymbol{\theta}))^T$$

where

$$q_j(\mathbf{x}; \boldsymbol{\theta}) = \frac{\partial}{\partial\theta_j} \log f(\mathbf{x}; \boldsymbol{\theta}).$$

- Write $Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} q(X_i; \boldsymbol{\theta})$. By the central limit theorem,

$$n^{-1/2} Q(\boldsymbol{\theta}^0) \xrightarrow{d} N(0, \mathbf{I}(\boldsymbol{\theta}^0)).$$

- A third statistic,

$$V_n = [n^{-1/2} Q(\boldsymbol{\theta}^0)]^T \mathbf{I}^{-1}(\boldsymbol{\theta}^0)[n^{-1/2} Q(\boldsymbol{\theta}^0)] = n^{-1} Q(\boldsymbol{\theta}^0)^T \mathbf{I}^{-1}(\boldsymbol{\theta}^0) Q(\boldsymbol{\theta}^0),$$

was introduced by Rao (1948).
Again, it has a limiting $\chi^2_s$ distribution.

**Example 3.14** Consider a sample $X_1, \ldots, X_n$ from the logistic distribution with density

$$f_\theta(x) = \frac{e^{x-\theta}}{(1 + e^{x-\theta})^2}.$$

- $q(x;\theta) = -1 + 2e^{x-\theta}/(1+e^{x-\theta})$ and

$$Q(\theta^0) = -n + 2\sum_{i=1}^{n} \frac{e^{x_i-\theta^0}}{1+e^{x_i-\theta^0}}.$$

- $I(\theta) = 1/3$ for all $\theta$.
- The Rao scores test therefore rejects $H_0$ with test statistic

$$\sqrt{\frac{3}{n}} \sum_{i=1}^{n} \frac{e^{x_i-\theta^0}-1}{1+e^{x_i-\theta^0}}.$$

- In this case, the MLE does not have an explicit expression and therefore the Wald and likelihood ratio tests are less convenient.

The three test statistics we discuss are asymptotically equivalent under $H_0$. However, they do differ in computation and ease of interpretation.

- All three statistics have the same limit chi-squared distribution with degree of freedom $s$ under the null hypothesis. The limiting distribution can be found by the following lemma.

  **Lemma 1** *Under regularity conditions,*
  *(i)* $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0) \xrightarrow{d} N(0, \mathbf{I}^{-1}(\boldsymbol{\theta}^0))$*;*
  *(ii)* $n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0)^T \mathbf{I}(\boldsymbol{\theta}^0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0) \xrightarrow{d} \chi^2_s$*;*
  *(iii)* $n^{-1}Q(\boldsymbol{\theta}^0)^T \mathbf{I}^{-1}(\boldsymbol{\theta}^0) Q(\boldsymbol{\theta}^0) \xrightarrow{d} \chi^2_s$*;*
  *(iv)* $\lambda_n - n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0)^T \mathbf{I}(\boldsymbol{\theta}^0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0) \xrightarrow{P} 0.$

- Both the likelihood ratio test and the Wald test require calculating an efficient estimator $\hat{\boldsymbol{\theta}}_n$, while the Rao test does not and is therefore the most convenient from the computational point of view.

- The Wald test, being based on the studentized difference

$$\mathbf{I}^{1/2}(\hat{\boldsymbol{\theta}}_n)[\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^T]$$

  is more easily interpretable and has the advantage immediately yields confidence regions for $\boldsymbol{\theta}$.

- The Wald test has the drawback, not shared by the other two, that it is only asymptotically but not exactly invariant under reparametrization.
  For simplicity, consider $s = 1$ and $\eta = g(\theta)$. Here we assume that $g$ is differentiable and strictly increasing. The Wald statistic for testing $\eta = \eta^0 (= g(\theta^0))$ is

$$[g(\hat{\theta}_n) - g(\theta_0)]\sqrt{nI(\hat{\eta}_n)} = \sqrt{nI(\hat{\theta}_n)}(\hat{\theta}_n - \theta_0)\frac{g(\hat{\theta}_n) - g(\theta_0)}{\hat{\theta}_n - \theta_0} \cdot \frac{1}{g^{(1)}(\hat{\theta}_n)}.$$

  The product of the second and third factor tends to 1 as $\hat{\theta}_n \to \theta_0$ but typically will differ from 1 for finite $n$.

**Example 3.15** Consider a sequence of $n$ independent trials, with $s$ possible outcomes for each trials.

- Let $\theta_j$ denote the probability of occurrence of the $j$th outcome in any given trial.
- Let $N_j$ denote the number of occurrences of the $j$th outcome in the series of $n$ trials.
- The MLE of $\theta_j$'s are $N_j/n$.
- The three test statistics $\lambda_n$, $W_n$ and $V_n$ for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0$ against $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}^0$ are easily seen to be

$$\begin{aligned} \lambda_n &= 2\sum_{j=1}^{s} N_j \log(\frac{N_j}{n\theta_j^0}), \\ W_n &= \sum_{j=1}^{s} \frac{(N_j - n\theta_j^0)^2}{N_j}, \\ V_n &= \sum_{j=1}^{s} \frac{(N_j - n\theta_j^0)^2}{n\theta_j^0}. \end{aligned}$$

- Both $W_n$ and $V_n$ are referred to as chi-squared goodness of fit statistics; the latter often called the Pearson chi-squared distribution. The large sample properties was first derived by Pearson (1900).
  Pearson's chi-square statistic is easily remembered as

$$\chi^2 = sum\frac{(Observed - Expected)^2}{Expected}.$$

- Let us now consider the behavior of $\lambda_n$, $W_n$ and $V_n$ under "local" alternatives, that is, for a sequence $\{\boldsymbol{\theta}_n\}$ of the form

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + n^{-1/2}\Delta,$$

  where $\Delta = (\Delta_1, \ldots, \Delta_s)^T$.
- Suppose that the convergences expressed in the above lemma may be established uniformly in $\Theta$ for $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}^0$.
- It then would follow that

$$\begin{aligned} n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) &= n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_n) + \Delta \xrightarrow{d} N(\Delta, \mathbf{I}^{-1}(\boldsymbol{\theta}^0)), \\ n^{-1/2}Q(\boldsymbol{\theta}^0) &= n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_n)\mathbf{I}(\boldsymbol{\theta}^0) + o_{P_{\boldsymbol{\theta}_n}}(1) \xrightarrow{d} N(\mathbf{I}(\boldsymbol{\theta}^0)\Delta, \mathbf{I}(\boldsymbol{\theta}^0)), \end{aligned}$$

  and

$$\lambda_n - W_n \xrightarrow{P_{\boldsymbol{\theta}_n}} 0,$$

- It then follow that the statistics $\lambda_n$, $W_n$ and $V_n$ each converge in distribution to $\chi^2_s(\Delta^T \mathbf{I}(\boldsymbol{\theta}^0)\Delta)$.
- Therefore, under appropriate regularity conditions, the statistics $\lambda_n$, $W_n$ and $V_n$ are asymptotically *equivalent* in distribution, both under the null hypothesis and under local alternatives converging sufficiently fast.
- However, at fixed alternatives these equivalences are not anticipated to hold.

**Example 3.16** (Testing a Genetic Theory)

- In experiments on pea breading, Mendel observed the different kinds of seeds obtained by crosses from peas with round yellow seeds and peas with wrinkled green seeds.
- Possible types of progeny were: (1) round yellow; (2) wrinkled yellow; (3) round green; and (4) wrinkled green.
- Assume the seeds are produced independently. We can think of each seed as being the outcome of a multinomial trial with possible outcomes numbered $1$, $2$, $3$, $4$ as above and associated probabilities of occurrence $\theta_1$, $\theta_2$, $\theta_3$, $\theta_4$.
- Mendel's theory predicted that $\theta_1 = 9/16$, $\theta_2 = \theta_3 = 3/16$, $\theta_4 = 1/16$.
- Data: $n = 556$, $n_1 = 315$, $n_2 = 101$, $n_3 = 108$, $n_4 = 32$.
- Pearson's chi-square statistic is

$$\frac{(315 - 556 \times 9/16)^2}{312.75} + \frac{(3.25)^2}{104.25} + \frac{(3.75)^2}{104.25} + \frac{(2.75)^2}{34.75} = 0.47,$$

  which has a $p$ value of $0.9$ when referred to a $\chi^2_3$ table.
  There is insufficient evidence to reject Mendel's hypothesis. (Why don'twe state that *we accept Mendel's hypothesis?)*

# Topic 5: Tests in Nonparametric Models

## Sign, permutation, and rank tests

- In a nonparametric problem, a UMP, UMPU, or UMPI test usually does not exist.
- Nonparametric tests are derived using some intuitively appealing ideas. They are commonly referred to as *distribution-free* tests, since almost No assumption is imposed on the population under consideration.
- Sign test:
  - Let $X_1, \ldots, X_n$ be iid random variables from $F$, $u$ be a fixed constant, and $p = F(u)$.

- Consider the problem of testing $H_0 : p \leq p_0$ versus $H_1 : p > p_0$, or testing $H_0 : p = p_0$ versus $H_1 : p \neq p_0$, where $p_0$ is a fixed constant in $(0, 1)$.
- Let $\triangle_I = 1_{X_i - u \leq 0}$, $i = 1, \ldots, n$. Then $\triangle_1, \ldots, \triangle_n$ are iid binary random variables with $p = P(\triangle_i = 1)$.
- For testing $H_0 : p \leq p_0$ versus $H_1 : p > p_0$, it follows from Neymann-Pearson lemma and monotone likelihood ratio that the test

$$T^*(Y) = \begin{cases} 1 & Y > m \\ \gamma & Y = m \\ 0 & Y < m \end{cases}$$

  is UMP among tests based on $\triangle_i$'s, where $Y = \sum_{i=1}^n \triangle_i$.
- For testing $H_0 : p = p_0$ versus $H_1 : p \neq p_0$, the test

$$T^*(Y) = \begin{cases} 1 & Y < c_1 \text{ or } Y > c_2 \\ \gamma_i & Y = c_i, i = 1, 2, \\ 0 & c_1 < Y < c_2 \end{cases}$$

  is UMPU among tests based on $\triangle_i$'s.
- Since $Y$ is equal to the number of nonnegative signs of $(u - X_i)$'s, tests based on $T^*$ are called sign tests.
- One can easily extend the sign tests to the case where $p = P(X_1 \in B)$.

- Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ (matched pairs) be iid random variables from $F$. By using $\triangle_i = X_i - Y_i - u$, one can obtain sign tests for hypotheses concerning $P(X_1 - Y_1 \leq u)$.

- Permutation tests:
  - Let $X_{i1}, \ldots, X_{in_i}$, $i = 1, 2$, be two independent samples iid from $F_i$, $i = 1, 2$, respectively. Here $F_i$'s are cdf's on $R$.
  - Think of two-sample problem in parametric setting (normal). Such type of problems arise from the comparison of two treatments.
  - Remove the parametric assumption and assume that $F_i$'s are in t he non-parametric family $\mathcal{F}$ containing all continuous cdf's on $R$.
  - Consider the problem of testing

$$H_0 : F_1 = F_2 \quad \text{versus} \quad H_1 : F_1 \neq F_2.$$

  - Let $\mathbf{X} = (X_{ij}, j = 1, \ldots, n_i, i = 1, 2)$, $n = n_1 + n_2$, and $\alpha$ be a given significance level. A test $T(\mathbf{X})$ satisfying

$$\frac{1}{n!} \sum_{\mathbf{z} \in \pi(\mathbf{x})} T(\mathbf{z}) = \alpha$$

is called a permutation test, where $\pi(x)$ is the set of $n!$ points obtained from $\mathbf{x} \in R^n$ by permuting the components of $\mathbf{x}$.

- For rank tests, we only consider Wilcoxon rank-sum test.

## Rank Tests for Comparing Two Treatments

- For comparing a new treatment or procedure with the standard method, $N$ subjects (patients, students, etc.) are divided at random into a group of $n$ who will receive a new treatment and a control group of $m$ who will be treated by the standard method.

- At the termination of the study, the subjects are ranked either directly or according to some response that measures the success of the treatment such as a test score in an educational or pyschological investigation.

- The hypothesis $H_0$ of no treatment effect is rejected, and the superiority of the new treatment acknowledged, if the ranking the $n$ treated subjects rank sufficiently high. (Here it is assumed that the success of the treatment is indicated by an increased response; if instead the aim is to decrease the response, $H_0$ is rejected when the $n$ treated subjects rank sufficiently low.)

- Let the ranks of the treated subjects be denoted by $S_1, \ldots, S_n$, where we shall assume that they are numbered in increasing order. Denote the sum of the treatment ranks $W_S = S_1 + \cdots + S_n$.

- The hypothesis $H_0$ is then rejected and the treatment judged to be effective when $W_S$ is sufficiently large, say, when $W_S \geq c$. Here the constant $c$ is determined by the equation

$$P_{H_0}(W_S \geq c) = \alpha.$$

- The test defined above is known as the *Wilcoxon rank-sum test*.

- Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent, the $X$'s identically distributed with distribution $F$ and the $Y$'s identically distributed with distribution $G$. Here the $Y$'s are responses to a treatment.

- Then $H_0 : F = G$ and $H_a : Y$ is stochastically larger than $X$, i.e., $G(t) \leq F(t)$ for all $t$ but $G \neq F$.

- Let the ranks of the $X'$s be denoted by $R_1, \ldots, R_m$. If we substitute $R'$s for $X$'s and $S$'s for $Y$'s in the two-sample t-test statistic, we obtain

$$\left(\frac{nm}{N}\right)^{1/2} \frac{\frac{1}{n}\sum_{i=1}^{n} S_i - \frac{1}{m}\sum_{j=1}^{m} R_j}{(N-2)^{-1}\left[\{\sum_{i=1}^{n}(S_i - \frac{N+1}{2})^2 + \sum_{j=1}^{m}(R_j - \frac{N+1}{2})^2\right]\}^{1/2}}.$$

- This statistic is equivalent to the Wilcoxon statistic $W_S$, the sum of the ranks of the treatment group.
  - Write $W_{XY}$ as the number of pairs $(X_i, Y_j)$ with $X_i < Y_j$.
  - It can be shown that
$$W_S - \frac{1}{2}n(n+1) = W_{XY}.$$
  - $W_{XY}$ is usually known as the Mann-Whitney statistic.
  - Let $\phi(X_i, Y_j) = 1$ if $X_i < Y_j$, and $0$ otherwise. Then
$$W_{XY} = \sum_{i=1}^{m}\sum_{j=1}^{n}\phi(X_i, Y_j) \tag{1}$$
  - We shall prove that $W_{XY}$ is asymptotically normal as $m$ and $n$ tend to infinity.
- The method of proof consists in replacing the variable $W_{XY}$ by a sum of independent random variables, which is asymptotically equivalent to $W_{XY}$ and to which the central limit theorem can then be applied.
- It is natural for this purpose to try a sum of the form
$$S = \sum_{i=1}^{m} a_i(X_i) + \sum_{j=1}^{n} b_j(Y_j) \tag{2}$$
  but how should one choose the functions $a_i$ and $b_j$?
- The following "projection mathod" introduced in a different context by Hajek (1961), produces the $a_i$ and $b_j$ most likely to succeed in the sense of minimizing $E(W_{XY} - S)^2$.
- This approach is due to Hoeffding (1948), and is applicable to a large class of statistics, the so-called U-statistics.
  - Note that
$$\theta(F, G) = \int F dG = P(X \leq Y).$$
  - An unbiased estimator of $\theta(F, G)$ is
$$U = \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n} I(X_i \leq Y_j),$$
    which is the $W_{XY}$.
  - A statistic can be written in the form is called a U-statistics.
  - Note that the popularity of this projection method is due to Hajek (1968), who gives the following result.

**Lemma 2** *(Hoeffding) Let $Z_1, \ldots, Z_n$ be independent random variables and $S = S(Z_1, \ldots, Z_n)$ any statistic satisfying $E(S^2) < \infty$. Then the random variable*

$$S^* = \sum_{i=1}^{n} E(S|Z_i) - (n-1)E(S)$$

*satisfies $E(S^*) = E(S)$ and*

$$E(S - S^*)^2 = Var(S) - Var(S^*).$$

**Remarks:**

1. The random variables $S^*$ is called the *projection* of $S$ on $Z_1, \ldots, Z_n$.
2. Note that it is conveniently a sum of *independent and identically distributed* random variables.
3. In cases that $E(S - S^*)^2 \to 0$ at a suitable rate as $n \to \infty$, the asymptotic normality of $S$ may be established by applying classical theory to $S^*$.

**Proof of Hoeffding's Lemma.**

- Without loss of generality, we can assume that $E(S) = 0$.
- Consider the problem of finding the sum

$$T = \sum_{i=1}^{n} k_i(Z_i) \tag{3}$$

  for which $E(S - T)^2$ is as small as possible; the minimizing $T$ may be considered the "projection" of $S$ onto the linear space formed by the functions $T$.

- Let

$$r_i(z_i) = E(S|Z_i = z_i) \tag{4}$$

  be the conditional expectation of $S$ given $Z_i = z_i$, and let

$$S^* = \sum_{i=1}^{n} r_i(Z_i). \tag{5}$$

  That $S^*$ is the desired minimizing function is an immediate consequence of the following identity, which holds for all statistics $T$ and $S$ with mean zero and satisfying (3) for which the required expectation exist:

$$E(S - T)^2 = E(S - S^*)^2 + E(S^* - T)^2. \tag{6}$$

- To prove the above identity, write

$$E(S - T)^2 = E[(S - S^*) + (S^* - T)]^2.$$

– Squaring the right-hand side proves (6) if it can be shown that

$$E[(S - S^*)(S^* - T)] = 0. \tag{7}$$

– Since the left-hand side of (7) is the sum of the expectations of

$$[r_i(Z_i) - k_i(Z_i)](S - S^*) \tag{8}$$

it is enough to show that the expectation of (8) given $Z_i$ is zero for all $i$.

– We shall prove this by showing that the conditional expectation of (8) given $Z_i$ is zero.

– In the conditional expectation of this product, the first factor can be taken out of the expectation sign since it depends only on $Z_i$, so that it is finally only necessary to show that the conditional expectation of $S - S^*$ given $Z_i$ is zero.

– Now

$$E[(S - S^*)|Z_i] = E\{S - r_i(Z_i) - \sum_{j \neq i} r_j(Z_j)|Z_i\}.$$

– From the definition of $r_i(Z_i)$, it is seen that the conditional expectation of $S - r_i(Z_i)$ given $Z_i$ is zero.

– On the other hand, since $Z_i$ and $Z_j$ are independent, the conditional expectation of $r_j(Z_j)$ given $Z_i$ is equal to the unconditional expectation of $r_j(Z_j)$, which by the definition of $r_j$ is equal to $E(S)$ and hence equal to zero.

– This completes the proof of (7) and therefore of (6).

- A useful special case of (6) is obtained by putting $T = 0$, which gives after arrangement

$$E(S - S^*)^2 = E(S^2) - E(S^{*2}) = Var(S) - Var(S^*). \tag{9}$$

- Before we apply Hoeffeding lemma to the $W_{XY}$-statistic (1), we will calculate the expectation and variance of $W_{XY}$.

- Set $\theta = (F, G)$,

$$E_\theta[\phi(X, Y)] = P_\theta[X < Y]$$

and we obtain

$$E_\theta(W_{XY}) = mnp \tag{10}$$

where $p = P_\theta[X < Y]$.

- Similarly, we have

$$Var_\theta(W_{XY}) = nmp(1-p) + nm(n-1)(q_1 - p^2) + nm(m-1)(q_2 - p^2) \tag{11}$$

where $q_1 = P_\theta[X_1 < \min(Y_1, Y_2)]$ and $q_2 = P_\theta[Y_1 > \max(X_1, X_2)]$.

- Note that under $H_0$, if $F$ is continuous, $p = 1/2$ while $q_1 = q_2 = 1/3$, since, among three independent identically distributed variables, each one is equally likely to be the minimum or the maximum.
- We then have $E_\theta(W_{XY}) = mn/2$ and $Var_\theta(W_{XY}) = mn(N+1)/12$ under $H_0$.
- Put
$$\psi(x,y) = \phi(x,y) - p. \tag{12}$$
Note that
$$E[\psi(X_\alpha, Y_\beta)|X_i = x] = \begin{cases} E\psi(x, Y_\beta) & \text{if } \alpha = i \\ 0 & \text{if } \alpha \neq i \end{cases}$$
and
$$E[\psi(X_\alpha, Y_\beta)|Y_j = y] = \begin{cases} E\psi(X_\alpha, y) & \text{if } \beta = j \\ 0 & \text{if } \beta \neq j \end{cases}$$
- Put $\psi_{10}(x) = E_Y\psi(x,Y)$ and $\psi_{01}(y) = E_X\psi(X,y)$.
- The projection of $W_{XY} - mnp$ by Hoeffeding Lemma is $n\sum_{i=1}^m \psi_{10}(X_i) + m\sum_{j=1}^n \psi_{01}(Y_j)$. Consider
$$U = \sqrt{m}\left[\frac{1}{m}\sum_{i=1}^m \psi_{10}(X_i) + \frac{1}{n}\sum_{j=1}^n \psi_{01}(Y_j)\right]$$
and $S = \sqrt{m}[(mn)^{-1}W_{XY} - p]$.
- Note that
$$\begin{aligned} Var(S) &\to q_1 - p^2 + \frac{m}{n}(q_2 - p^2), \\ Var(U) &= Var(\psi_{10}(X)) + \frac{m}{n}Var(\psi_{01}(Y)), \\ E(S-U)^2 &= Var(S) - Var(U). \end{aligned}$$
Observe that for $j \neq k$, $Var(\psi_{10}(X)) = q_1 - p^2$ and $Var(\psi_{01}(Y) = q_2 - p^2$. (i.e. $E\psi(x_1, Y_j)\psi(x_1, Y_k) = [\psi_{10}(x_1)]^2$ and
$$E_X[\psi_{10}(X)]^2 = E\psi(X, Y_j)\psi(X, Y_k) = Cov(\psi(X, Y_j), \psi(X, Y_k)).$$
We then conclude that $E(S-U)^2 \to 0$.

**Theorem 1** *Suppose that $F$ and $G$ are continuous and that $0 < P_\theta[X < Y] < 1$. Then*
$$\frac{S - E_\theta(S)}{\sqrt{Var_\theta(S)}} \xrightarrow{d} N(0,1) \text{ as } \min(n,m) \to \infty.$$

**Remark.** Reject $H_0$ when

$$\frac{W_{XY} - \frac{1}{2}nm}{\sqrt{\frac{1}{12}nm(N+1)}} \geq z(1-\alpha).$$

Pitman efficiency of the Wilcoxon rank-sum test to the two-sample t-test

We turn now to the comparison of the performance of the Wilcoxon and two-sample t tests. At first sight it would appear that a good reason for using the Wilcoxon is that it has a guaranteed probability of type I error and a good reason against using the Wilcoxon is its inefficient use of the data.

- We assume that the $X$'s and $Y$'s have the same variance $\sigma^2$ and means $\mu_1$ and $\mu_2$.
- Although the t test does not have a guaranteed probability of type I error, if $n$ and $m$ are moderately large, $H_0$ is true, and $F$ has a finite second moment, then the probability of type I error of the t test is fairly close to that specified by the normal model.
- Recall that the two-sample t statistic is given by

$$T = \sqrt{\frac{nm}{N}}\frac{\bar{Y} - \bar{X}}{s_2} \tag{13}$$

where

$$s_2 = \frac{\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{j=1}^{n}(Y_j - \bar{Y})^2}{N-2}. \tag{14}$$

- We start by obtaining an approximation to the critical value and power of the t test. Note that $s_2^2 \xrightarrow{P} \sigma^2$ as $\min(n,m) \to \infty$. It follows from Slutsky's theorem and central limit theorem that when $\mu_1 = \mu_2$, $T$ converges in law to a $N(0,1)$ random variable as $\min(n,m) \to \infty$.
- Then the t test that rejects $H_0$ when $T \geq t_{N-2}(1-\alpha)$ has approximately level $\alpha$ regardless of the shape of $F$ and $G$ and $z(1-\alpha)$ is an approximate critical value as we claimed above.
- If $\mu_1 \neq \mu_2$, let $\delta = (\mu_2 - \mu_1)/\sigma$. Then, arguing as above, if $\sqrt{nm/N}\delta$ stays bounded $T - \sqrt{nm/N}\delta$ has approximately a $N(0,1)$ random distribution for all $F$ and $G$ with $\sigma^2 < \infty$. We then can approximate the probability $P_\theta(T \geq t_{N-2}(1-\alpha))$ by

$$\beta_T = P_\theta[T \geq z(1-\alpha)] = 1 - \Phi(z(1-\alpha) - \sqrt{nm/N}\delta) = \Phi(z(\alpha) + \sqrt{nm/N}\delta).$$

- For Wilcoxon test,

$$\begin{aligned}\beta_N &= P_\theta\left[W_{XY} \geq \frac{1}{2}nm + z(1-\alpha)\sqrt{\frac{1}{12}nm(N+1)}\right] \\ &= P_\theta\left[\frac{W_{XY} - E_\theta(W_{XY})}{\sqrt{var_\theta(W_{XY})}} \geq \frac{nm(\frac{1}{2}-p) + z(1-\alpha)\sqrt{\frac{1}{12}nm(N+1)}}{\sqrt{var_\theta(W_{XY})}}\right] \\ &\approx \Phi\left(\frac{nm(\frac{1}{2}-p) + z(1-\alpha)\sqrt{\frac{1}{12}nm(N+1)}}{\sqrt{var_\theta(W_{XY})}}\right).\end{aligned}$$

- Consider the case that $X \sim N(\mu_1, \sigma^2)$, $Y \sim N(\mu_2, \sigma^2)$, $n = m$ and $\alpha = 0.05$. Note that $\delta = (\mu_2 - \mu_1)/\sigma = 0.5$.

- Suppose we want to have $\beta = 0.9$.
  For t-test, solve

$$-1.645 + \sqrt{\frac{\frac{N}{2}\frac{N}{2}}{N}}0.5 = 1.282$$

  and get $N = 16 \cdot (2.927)^2 \approx 140$.
  For Wilcoxon test:

$$p = P_\theta(X < Y) = \Phi\left(\frac{\mu_2 - \mu_1}{\sqrt{2}}\sigma\right), \qquad q_1 = P\left(Z_1 < \frac{\Delta}{\sqrt{2}}, Z_2 < \frac{\Delta}{\sqrt{2}}\right),$$
$$q_2 = P\left(Z_1 < \frac{\Delta}{\sqrt{2}}, Z_3 < \frac{\Delta}{\sqrt{2}}\right),$$

  where $Z_1 = [X_1 - Y_1 - (\mu_1 - \mu_2)]/\sqrt{2}\sigma$, $Z_2 = [X_1 - Y_2 - (\mu_1 - \mu_2)]/\sqrt{2}\sigma$, $Z_3 = [X_2 - Y_1 - (\mu_1 - \mu_2)]/\sqrt{2}\sigma$.

- Note that $(Z_1, Z_2) \sim N(0, 0, 1, 1, 1/2)$, $(Z_1, Z_3) \sim N(0, 0, 1, 1, 1/2)$. When $\Delta = 0.5$, $p = 0.638$, $q_1 = q_2 = 0.483$, we have $\beta_W \approx \Phi(-1.729 + 0.355\sqrt{N/2}) = 0.9$. Hence, $N \approx 144$.