5 Introduction to the Theory of Order Statistics and Rank Statistics

- This section will contain a summary of important definitions and theorems that will be useful for understanding the theory of order and rank statistics. In particular, results will be presented for *linear rank statistics*.

- Many nonparametric tests are based on test statistics that are linear rank statistics.

  - For one sample: The Wilcoxon-Signed Rank Test is based on a linear rank statistic.
  - For two samples: The Mann-Whitney-Wilcoxon Test, the Median Test, the Ansari-Bradley Test, and the Siegel-Tukey Test are based on linear rank statistics.

- Most of the information in this section can be found in Randles and Wolfe (1979).

5.1 Order Statistics

- Let $X_1, X_2, \ldots, X_n$ be a random sample of continuous random variables having cdf $F(x)$ and pdf $f(x)$.

- Let $X_{(i)}$ be the i^{th} smallest random variable ($i = 1, 2, \ldots, n$).

- $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are referred to as the **order statistics** for $X_1, X_2, \ldots, X_n$. By definition, $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$.

Theorem 5.1: Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ be the order statistics for a random sample from a distribution with cdf $F(x)$ and pdf $f(x)$. The joint density for the order statistics is

$$\begin{aligned} g(x_{(1)}, x_{(2)}, \ldots, x_{(n)}) &= n! \prod_{i=1}^{n} f(x_{(i)}) \quad \text{for } -\infty < x_{(1)} < x_{(2)} < \cdots < x_{(n)} < \infty \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{16}$$

Theorem 5.2: The marginal density for the j^{th} order statistic $X_{(j)}$ ($j = 1, 2, \ldots, n$) is

$$g_j(t) = \frac{n!}{(j-1)!(n-j)!} [F(t)]^{j-1} [1 - F(t)]^{n-j} f(t) \qquad -\infty < t < \infty.$$

- For random variable X with cdf $F(x)$, the **inverse distribution** $F^{-1}(\cdot)$ is defined as

$$F^{-1}(y) = \inf\{x : F(x) \geq y\} \qquad 0 < y < 1.$$

- If $F(x)$ is strictly increasing between 0 and 1, then there is only one x such that $F(x) = y$. In this case, $F^{-1}(y) = x$.

Theorem 5.3 (Probability Integral Transformation): Let X be a continuous random variable with distribution function $F(x)$. The random variable $Y = F(X)$ is uniformly distributed on $(0, 1)$.

- Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ be the order statistics for a random sample from a continuous distribution. Application of Theorem 5.3, implies that $F(X_{(1)}) < F(X_{(2)}) < \cdots < F(X_{(n)})$ are distributed as the order statistics from a uniform distribution on $(0, 1)$.

- Let $V_j = F(X_{(j)}$ for $j = 1, 2, \ldots, n$. Then, by Theorem 5.2, the marginal density for each V_j has the form

$$g_j(t) \;=\; \frac{n!}{(j-1)!(n-j)!}\, t^{j-1}\,[1-t]^{n-j} \qquad\qquad -\infty < t < \infty$$

  because $F(t) = t$ and $f(t) = 1$ for a uniform distribution on $(0, 1)$.

- Thus, V_j has a beta distribution with parameters $\alpha = j$ and $\beta = n - j + 1$. Therefore, the moments of V_j are

$$E(V_j^r) \;=\; \frac{n!\,\Gamma(r+j)}{(j-1)!\,\Gamma(n+r+1)}$$

  where $\Gamma(k) = (k-1)!$.

- Thus, when V_j is the j^{th} order statistic from a uniform distribution,

$$E(V_j) \;=\; \frac{j}{n+1} \qquad\qquad Var(V_j) \;=\; \frac{j(n-j+1)}{(n+1)^2(n+2)}$$

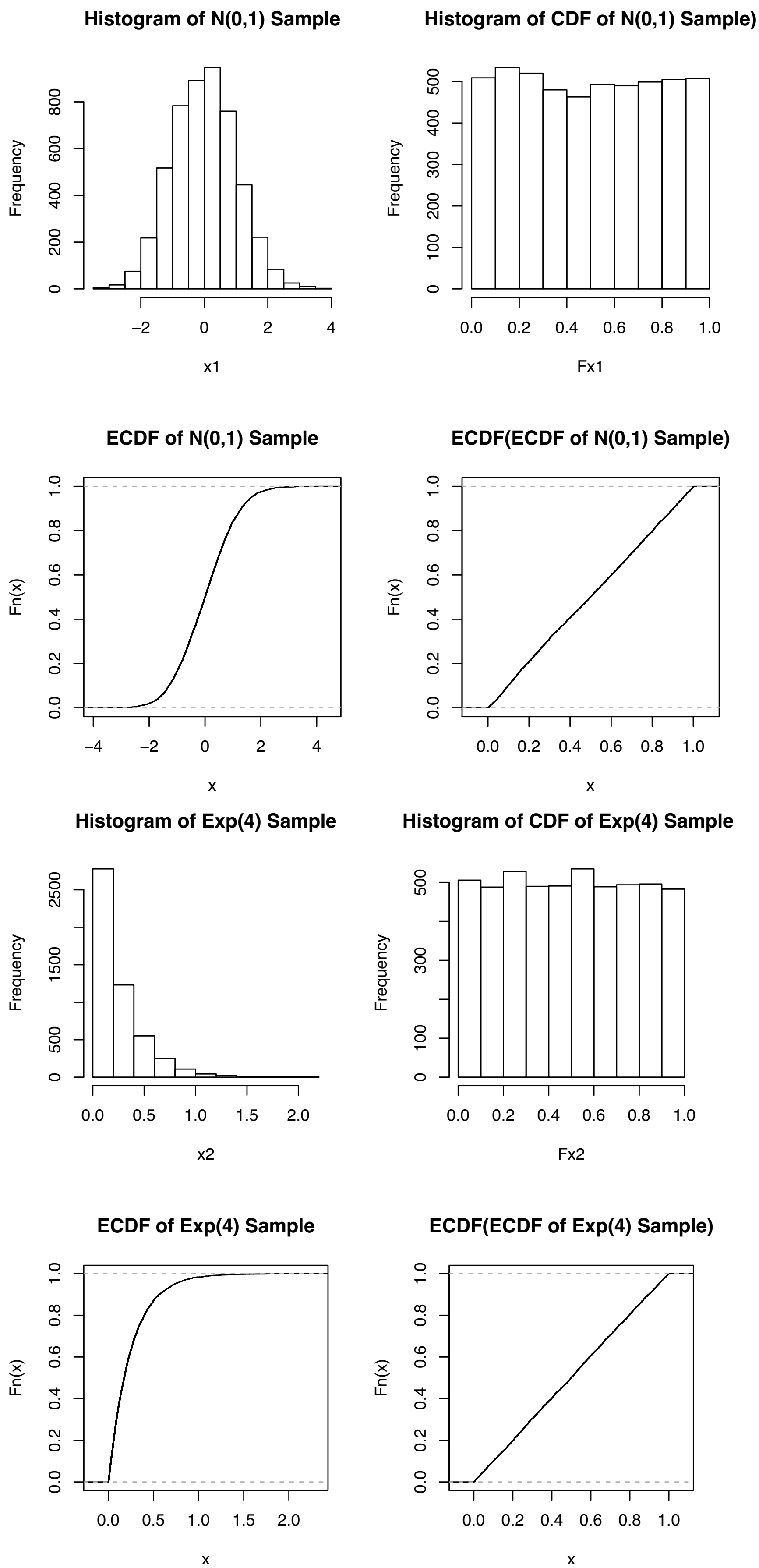
Simulation to Demonstrate Theorem 5.3 (Probability Integral Transformation)

Case 1: $N(0, 1)$ Distribution

1. Generate a random sample $(x_1, x_2, \ldots, x_{5000})$ of 5000 values from a normal $N(0, 1)$ distribution.
2. Determine the 5000 empirical cdf $\widehat{F}(x_i)$ values.
3. Plot the histograms and empirical cdf of the original $N(0, 1)$ sample. Note how they represent a sample from a standard normal distribution.
4. Plot the histograms and empirical cdf of the $\widehat{F}(x_i)$ values. Note the histograms and empirical cdf of the $\widehat{F}(x_i)$ values represent a sample from a uniform $U(0, 1)$ distribution (as supported by Theorem 5.3).

Case 2: $Exp(4)$ Distribution

1. Generate a random sample $(x_1, x_2, \ldots, x_{5000})$ of 5000 values from an exponential $Exp(4)$ distribution.
2. Determine the 5000 empirical cdf $\widehat{F}(x_i)$ values.
3. Plot the histograms and empirical cdf of the original $Exp(4)$ sample. Note how they represent a sample from an exponential $Exp(4)$ distribution.
4. Plot the histograms and empirical cdf of the $\widehat{F}(x_i)$ values. Note the histograms and empirical cdf of the $\widehat{F}(x_i)$ values represent a sample from a uniform $U(0, 1)$ distribution (as supported by Theorem 5.3).


Histogram of N(0,1) Sample
Frequency
0
200
400
600
800
−2
0
2
4
x1
Histogram of CDF of N(0,1) Sample)
Frequency
0
100
200
300
400
500
0.0
0.2
0.4
0.6
0.8
1.0
Fx1
ECDF of N(0,1) Sample
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
−4
−2
0
2
4
x
ECDF(ECDF of N(0,1) Sample)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
x
Histogram of Exp(4) Sample
Frequency
0
500
1500
2500
0.0
0.5
1.0
1.5
2.0
x2
Histogram of CDF of Exp(4) Sample
Frequency
0
100
300
500
0.0
0.2
0.4
0.6
0.8
1.0
Fx2
ECDF of Exp(4) Sample
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
x
ECDF(ECDF of Exp(4) Sample)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
x

R Code for Simulation of Theorem 5.3 (Probability Integral Transformation)

```
n = 5000  # size of random sample

# CASE 1: Random Samples from N(0,1) Distribution
x1  <- rnorm(n,0,1)
x1[1:10]            # view first 10 values
Fx1 <- pnorm(x1)
Fx1[1:10]

windows()
par(mfrow=c(2,2))
hist(x1,main="Histogram of N(0,1) Sample")
hist(Fx1,main="Histogram of CDF of N(0,1) Sample)")
plot(ecdf(x1),main="ECDF of N(0,1) Sample")
plot(ecdf(Fx1),main="ECDF(ECDF of N(0,1) Sample)")

# CASE 2: Random Samples from Exponential(4) Distribution
x2  <- rexp(n,4)
x2[1:10]            # view first 10 values
Fx2 <- pexp(x2,4)
Fx2[1:10]

windows()
par(mfrow=c(2,2))
hist(x2,main="Histogram of Exp(4) Sample")
hist(Fx2,main="Histogram of CDF of Exp(4) Sample)")
plot(ecdf(x2),main="ECDF of Exp(4) Sample")
plot(ecdf(Fx2),main="ECDF(ECDF of Exp(4) Sample)")
```

5.2 Equal-in-Distribution Results

- Two random variables S and T are **equal in distribution** if S and T have the same cdf.

  To denote equal in distribution, we write $S \overset{d}{=} T$.

Theorem 5.4 A random variable X has a distribution that is symmetric about some number μ if and only if $(X - \mu) \overset{d}{=} (\mu - X)$.

Theorem 5.5 Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables. Let $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ denote any permutation of the integers $(1, 2, \ldots, n)$. Then $(X_1, X_2, \ldots, X_n) \overset{d}{=} (X_{\alpha_1}, X_{\alpha_2}, \ldots, X_{\alpha_n})$.

- A set of random variables $X_1, X_2, \ldots, X_n$ is **exchangeable** if for every permutation $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ of the integers $1, 2, \ldots, n$,

$$(X_1, X_2, \ldots, X_n) \overset{d}{=} (X_{\alpha_1}, X_{\alpha_2}, \ldots, X_{\alpha_n}).$$

- If $X_1, X_2, \ldots, X_n$ are i.i.d random variables, then the set $X_1, X_2, \ldots, X_n$ is exchangeable.
- The statistic $t(\cdot)$ is
  1. a **translation** statistic if $t(x_1 + k, x_2 + k, \ldots, x_n + k) = t(x_1, x_2, \ldots, x_n) + k$
  2. a **translation-invariant** statistic if $t(x_1 + k, x_2 + k, \ldots, x_n + k) = t(x_1, x_2, \ldots, x_n)$

  for every k and $x_1, x_2, \ldots, x_n$.

5.3 Ranking Statistics

- Let $Z_1, Z_2, \ldots, Z_n$ be a random sample from a continuous distribution with cdf $F(z)$, and let $Z_{(1)} < Z_{(2)} < \cdots < Z_{(n)}$ be the corresponding order statistics.
- Z_i has **rank** R_i among $Z_1, Z_2, \ldots, Z_n$ if $Z_i = Z_{(R_i)}$ assuming the R_i^{th} order statistic is uniquely defined.
- By "uniquely defined" we are assuming that ties are not possible. That is, $Z_{(i)} \neq Z_{(j)}$ for all $i \neq j$.
- Let $\mathcal{R} = \{\mathbf{r} : \mathbf{r} \text{ is a permutation of the integers } (1, 2, \ldots, n)\}$. That is, $\mathcal{R}$ is the set of all permutations of the integers $(1, 2, \ldots, n)$.

Theorem 5.6 Let $\mathbf{R} = (R_1, R_2, \ldots, R_n)$ be the vector of ranks where R_i is the rank of Z_i among $Z_1, Z_2, \ldots, Z_n$. Then $\mathbf{R}$ is uniformly distributed over $\mathcal{R}$. That is, $P(\mathbf{R} = \mathbf{r}) = 1/n!$ for each permutation $\mathbf{r}$.

Theorem 5.7 Let $Z_1, Z_2, \ldots, Z_n$ be a random sample from a continuous distribution, and let $\mathbf{R}$ be the corresponding vector of ranks where R_i is the rank of Z_i for $i = 1, 2, \ldots, n$. Then

$$\begin{aligned} P[R_i = r] &= 1/n \qquad \text{for } r = 1, 2, \ldots, n \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

and, for $i \neq j$,

$$\begin{aligned} P[R_i = r\, ,\ R_j = s] &= \frac{1}{n(n-1)} \qquad \text{for } r \neq s,\ r, s = 1, 2, \ldots, n \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

Corollary 5.8 Let $\mathbf{R}$ be the vector of ranks corresponding to a random sample from a continuous distribution. Then

$$E[R_i] = \frac{n+1}{2} \qquad \text{and} \qquad Var[R_i] = \frac{(n+1)(n-1)}{12} \text{ for } i = 1, 2, \ldots, n$$

$$Cov[R_i, R_j] = \frac{-(n+1)}{12} \text{ for } i \neq j.$$

- Let $V_1, V_2, \ldots, V_n$ be random variables with joint distribution function D, where D is a member of some collection $\mathcal{A}$ of possible joint distributions. Let $T(V_1, V_2, \ldots, V_n)$ be a statistic based on $V_1, V_2, \ldots, V_n$.
- The statistic T is **distribution-free over** $\mathcal{A}$ if the distribution of T is the same for every joint distribution in $\mathcal{A}$.

Corollary 5.9 Let $Z_1, Z_2, \ldots, Z_n$ be a random sample from a continuous distribution, and let $\mathbf{R}$ be the corresponding vector of ranks. If $V(\mathbf{R})$ is a statistic based only on $\mathbf{R}$, then $V(\mathbf{R})$ is distribution-free over the class $\mathcal{A}$ of joint distributions of n i.i.d. continuous random variables.

- A statistic (such as $V(\mathbf{R})$) that is a function of $Z_1, Z_2, \ldots, Z_n$ only through the rank vector $\mathbf{R}$ is called a **rank statistic**.

Example of a distribution-free statistic: Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be independent random samples from continuous distributions with cdfs $F(x)$ and $G(x) = F(x - \Delta)$, respectively ($-\infty < \Delta < \infty$). That is, Δ is a **shift parameter**.

– Combine the X and Y samples. Let R_i ($i = 1, 2, \ldots, n$) and Q_j ($j = 1, 2, \ldots, m$) be the ranks of the n X-values and the m Y-values in the combined sample. Thus, R_i and Q_j take on values $1, 2, \ldots, (m + n)$.

– Thus, the rank vector $\mathbf{R} = (R_1, R_2, \ldots, R_n, Q_1, Q_2, \ldots, Q_m)$ is simply a permutation of the integers $(1, 2, \ldots, (m + n))$ which satisfy the constraint

$$\sum_{i=1}^{n} R_i + \sum_{j=1}^{m} Q_j = \sum_{k=1}^{m+n} k = \frac{(m+n)(m+n+1)}{2}.$$

– To construct a test for $H_0 : \Delta = 0$ vs $H_1 : \Delta > 0$ based on the ranks in rank vector $\mathbf{R}$, we compare the X-ranks $(R_1, R_2, \ldots, R_n)$ to the Y-ranks $(Q_1, Q_2, \ldots, Q_m)$.

– If we know the X-ranks $(R_1, R_2, \ldots, R_n)$, then we also know the Y-ranks. Thus, it will be sufficient to consider a statistic based only on the X-ranks, say $W(R_1, R_2, \ldots, R_n)$.

– The test statistic proposed by Wilcoxon is $W = \sum_{i=1}^{n} R_i$. That is, W is the sum of the X-ranks. W is known as a **ranksum statistic**.

– Note that the statistic W is a function of the data only through the rank vector $\mathbf{R} = (R_1, R_2, \ldots, R_n, Q_1, Q_2, \ldots, Q_m)$. That is, once we have $\mathbf{R}$, we no longer need $(X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m)$ to calculate W.

– If $H_0 : \Delta = 0$ is true, then the data $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$ are i.i.d. continuous random variables. Applying Corollary 5.9, the rank statistic W is distribution-free over the class $\mathcal{A}$ of all continuous distributions. That is, for any continuous cdf $F \in \mathcal{A}$, the distribution of W does not depend on the choice of F.

Theorem 5.10: Let W be the rank sum statistic when $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ are independent random samples from $F(x)$ and $G(y) = F(y - \Delta)$, respectively. If $H_0 : \Delta = 0$ is true, then the discrete distribution of W is given by

$$\begin{aligned} P_0[W = w] &= \frac{t_{m,n}(w)}{\binom{m+n}{n}} \quad \text{for } w = \frac{n(n+1)}{2}, \; \frac{n(n+1)}{2} + 1, \; \ldots, \; \frac{n(2m+n+1)}{2} \\ &= 0 \quad \text{otherwise} \end{aligned}$$

where $t_{m,n}(w)$ is the number of subsets of n integers selected without replacement from $(1, 2, \ldots, (m + n))$ such that their sum $= w$.

- Thus, to calculate $P_0[W = w]$ for a given m and n, we need to (i) generate all $\binom{m+n}{n}$ possible assignments of $(m + n)$ ranks to the X and Y observations, (ii) calculate W for each assignment, and (iii) count the number of cases where $W = w$.

- For example consider the case with $n = 2$ and $m = 4$. There are $\binom{6}{2} = 15$. Thus, there will be two X-ranks (R_1, R_2) from the six possible ranks $(1, 2, 3, 4, 5, 6)$. $W = R_1 + R_2$ is then calculated for all possible assignments of the 6 ranks.

- The following table shows the 15 assignments of the 6 ranks and the corresponding W statistic values.

X-ranks R_1, R_2	Y-ranks Q_1, Q_2, Q_3, Q_4	$W = R_1 + R_2$	X-ranks R_1, R_2	Y-ranks Q_1, Q_2, Q_3, Q_4	$W = R_1 + R_2$
5,6	1,2,3,4	11	2,4	1,3,5,6	6
4,6	1,2,3,5	10	2,3	1,4,5,6	5
4,5	1,2,3,6	9	1,6	2,3,4,5	7
3,6	1,2,4,5	9	1,5	2,3,4,6	6
3,5	1,2,4,6	8	1,4	2,3,5,6	5
3,4	1,2,5,6	7	1,3	2,4,5,6	4
2,6	1,3,4,5	8	1,2	3,4,5,6	3
2,5	1,3,4,6	7			

For each of the 15 unordered assignments of ranks within samples, there are $4! \times 2! = 48$ ordered assignments yielding the same W value. Thus, overall there are $6! = 720 = (15)(48)$ ordered assignments of the 6 ranks.

- The distribution of W is

w	3	4	5	6	7	8	9	10	11
$P_0[W = w]$	1/15	1/15	2/15	2/15	3/15	2/15	2/15	1/15	1/15

- Suppose that $W = 9$. Then for the test of $H_0 : \Delta = 0$ vs $H_1 : \Delta > 0$:

$$\begin{aligned} p - \text{value} &= \text{the probability of getting a test statistic } W \text{ that is at least 9} \\ &= 2/15 \ + \ 1/15 \ + \ 1/15 \ = \ 4/15 \ \approx \ .27. \end{aligned}$$

Note that $w \in \{3, 4, \ldots, 11\} = \left\{ \frac{n(n+1)}{2}, \ \frac{n(n+1)}{2} + 1, \ \ldots, \ \frac{n(2m+n+1)}{2} \right\}$ as stated in Theorem 5.10.

Theorem 5.11 Let $W = \sum_{j=1}^{n}$ be the ranksum statistic. If $H_0 : \Delta = 0$ is true (i.e. $F = G$), then the distribution of W is symmetric about the value $\mu = n(m+n+1)/2$ and

$$E_0[W] \ = \ \mu \qquad Var[W] \ = \ \frac{mn(m+n+1)}{12}.$$

5.3.1 Statistics Based on Counting and Ranking

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous distribution that is symmetric about value μ.
- Let $Z_1, Z_2, \ldots, Z_n = (X_1 - \mu, X_2 - \mu, \ldots, X_n - \mu)$. Then $Z_1, Z_2, \ldots, Z_n$ is a random sample that is symmetric about 0.
- Define $\Psi_i = \Psi(Z_i)$ to be an indicator variable where

$$\Psi(t) = 1 \ \text{ if } \ t > 0 \qquad \text{and} \qquad \Psi(t) = 0 \ \text{ if } \ t \leq 0$$

Lemma 5.12 Let Z be a random variable that is symmetrically distributed about 0. Then the random variables $|Z|$ and $\Psi = \Psi(Z)$ are stochastically independent. That is,

$$P(\Psi = 1, |Z| \leq t) \;=\; P(\Psi = 1)P(|Z| \leq t) \quad \text{and} \quad P(\Psi = 0, |Z| \leq t) \;=\; P(\Psi = 0)P(|Z| \leq t).$$

- For random variables $Z_1, Z_2, \ldots, Z_n$, the **absolute rank** of Z_i, denoted R_i^+, is the rank of $|Z_i|$ among $|Z_1|, |Z_2|, \ldots, |Z_n|$.
- The **signed rank** of Z_i is $\Psi_i R_i^+$. Thus, (i) $\Psi_i = |Z_i|$ if $Z_i > 0$ and (ii) $\Psi_i = 0$ if $Z_i \leq 0$.
- A **signed rank statistic** is a statistic that is a function of $\Psi_1 R_1^+,\ \Psi_2 R_2^+,\ \ldots,\ \Psi_n R_r^+$.
- The following theorem establishes properties of the joint distribution of $\mathbf{\Psi} = (\Psi_1, \Psi_2, \ldots, \Psi_n)$ and $\mathbf{R}^+ = (R_1^+, R_2^+, \ldots, R_n^+)$.

Theorem 5.13 Let $Z_1, Z_2, \ldots, Z_n$ be a random sample from a continuous distribution that is symmetric about 0. Then $\Psi_1, \Psi_2, \ldots, \Psi_n, \mathbf{R}^+$ are mutually independent. Moreover, each Ψ_i is a Bernoulli random variable with $p = 1/2$, and $\mathbf{R}^+$ is uniformly distributed over $\mathcal{R}$ (the set of all permutations of the integers $(1, 2, \ldots, n)$).

Proof of Theorem 5.13

- $Z_1, Z_2, \ldots, Z_n$ are are independent because they are a random sample. Lemma 5.12 implies that $\Psi_1, |Z_1|, \Psi_2, |Z_2|, \ldots, \Psi_n, |Z_n|$ are $2n$ mutually independent random variables.
- Each Ψ_i is a Bernoulli random variable with parameter $p = P[Z_i > 0] = 1/2$ because Z_i is continuous and symmetrically distributed about 0.
- The $\mathbf{R}^+$ is independent of $\Psi_1, \Psi_2, \ldots, \Psi_n$ because it is a function only of $|Z_1|, |Z_2|, \ldots, |Z_n|$. That is, $\mathbf{R}^+$ does not depend on any Ψ_i.
- Because $\mathbf{R}^+$ is a rank vector of n i.i.d. continuous random variables, application of Theorem 5.6 shows that $\mathbf{R}^+$ is uniformly distributed over $\mathcal{R}$ (the set of permutations of the integers $(1, 2, \ldots, n)$.

Let $\mathcal{A}_0$ be the set of joint distributions of n i.i.d. continuous random variables that are symmetrically distributed about 0.

Corollary 5.14 Let $S(\mathbf{\Psi}, \mathbf{R}^+)$ be a statistic that depends on $Z_1, Z_2, \ldots, Z_n$ only through $\mathbf{\Psi} = \Psi_1, \Psi_2, \ldots, \Psi_n$ and $\mathbf{R}^+ = (R_1^+, R_2^+, \ldots, R_n^+)$. Then the statistic $S(\cdot)$ is distribution-free over $\mathcal{A}_0$.

Proof of Corollary 5.14 This result follows from Theorem 5.13 because $\mathbf{\Psi}$ and $\mathbf{R}^+$ have the same joint distribution for every joint distribution $F_0(Z_1, Z_2, \ldots, Z_n) \in \mathcal{A}_0$. That is, the joint distribution of $\mathbf{\Psi}$ and $\mathbf{R}^+$ does not depend on the choice of $F_0(Z_1, Z_2, \ldots, Z_n) \in \mathcal{A}_0$.

- We will often be interested in functions of $\mathbf{\Psi}$ and $\mathbf{R}^+$ that are symmetric functions of the signed ranks $\Psi_1 R_1^+, \Psi_2 R_2^+, \ldots, \Psi_n R_n^+$. If this is the case, then the following theorem can help establish the distribution of such a statistic.

Theorem 5.15 Let $Z_1, Z_2, \ldots, Z_n$ be a random sample from a continuous distribution that is symmetric about 0. Let Q be the number of positive Zs. For $Q = q$, let $S_1 < S_2 < \cdots < S_q$ denote the ordered absolute ranks of those Zs that are positive (i.e., $S_1 < S_2 < \cdots < S_q$ are the positive signed ranks in numerical order). Then

$$\begin{aligned} P[Q = q, S_1 = s_1, S_2 = s_2, \ldots, S_q = s_q] &= (1/2)^n \text{ for } q = 0, 1, \ldots, n \text{ and each of} \\ & \text{the } q - \text{tuples } (s_1, s_2, \ldots, s_q) \text{ such that} \\ & s_i \text{ is an integer and } 1 \leq s_1 < s_2 < \cdots < s_q \leq n \\ &= 0 \quad \text{otherwise} \end{aligned}$$

- Recall: Suppose $X_1, X_2, \ldots, X_n$ be a random sample from a continuous distribution that is symmetric about μ. Then $Z_1, Z_2, \ldots, Z_n = (X_1 - \mu, X_2 - \mu, \ldots, X_n - \mu)$ is a random sample that is symmetric about 0.
- Thus, all of the preceding results also apply to the $(X_i - \mu)$ random variables. That is, we can generalize the results to $\mathcal{A}_\mu$ = the class of continuous distributions that are symmetric about μ for any $-\infty < \mu < \infty$.

Example:

- Suppose we have a random sample $X_1, X_2, \ldots, X_n$ from a distribution in $\mathcal{A}_\mu$.
- The Wilcoxon signed rank statistic W^+ is defined as

$$W^+ = \sum_{i=1}^{n} \Psi_i R_i^+.$$

  That is, W^+ is the sum of the signed ranks.
- To test $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$, we would reject H_0 if W^+ is "too large". That is, we would reject H_0 if the p-value is small (e.g., p-value $< .05$). So how do we calculate the p-value?

Corollary 5.16 Let W^+ be the Wilcoxon signed rank statistic for testing $H_0 : \theta = \theta_0$. For a random sample of size n, the distribution of W^+ assuming H_0 is true is

$$\begin{aligned} P_0[W^+ = k] &= \frac{c_n(k)}{2^n} \qquad \text{for } k = 0, 1, \ldots, \frac{n(n+1)}{2} \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

where $c_n(k)$ = the number of subsets of integers $\{1, 2, \ldots, n\}$ for which W^+ is equal to k.

- Suppose $n = 4$. The following table list the 2^4 combinations of signed ranks and the corresponding W^+ values.

Subset of $\{1,2,3,4\}$	W^+	Subset of $\{1,2,3,4\}$	W^+
$\emptyset$	0	{2,3 }	5
{1}	1	{2,4}	6
{2}	2	{3,4}	7
{3}	3	{1,2,3}	6
{4}	4	{1,2,4}	7
{1,2}	3	{1,3,4}	8
{1,3}	4	{2,3,4}	9
{1,4}	5	{1,2,3,4}	10

Thus, the distribution of W^+ is

k	0	1	2	3	4	5	6	7	8	9	10
$P[W^+ = k]$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$

– Suppose the data are $(X_1, X_2, X_3, X_4) = (24.6, 25.1, 25.6, 25.7)$, and we want to test $H_0 : \mu = 25$ vs $H_1 : \mu > 25$.

– Next calculate the deviations from $\mu_0 = 25$. That is, $(Z_1, Z_2, Z_3, Z_4) = (-.4, .1, .6, .7)$.

  and the vector of absolute values is $(|Z_1|, |Z_2|, |Z_3|, |Z_4|) = (.4, .1, .6, .7)$.

– The absolute rank vector $\mathbf{R}^+ = (R_1^+, R_2^+, R_3^+, R_4^+) = (2, 1, 3, 4)$.

– $\Psi_i = 1$ if $Z_i > 0$ (or equivalently, if $X_i > 25$)), and is 0 otherwise. Thus, $(\Psi_1, \Psi_2, \Psi_3, \Psi_4) = (0, 1, 1, 1)$.

– Therefore the signed rank statistic $W^+ = \sum_{i=1}^{n} \Psi_i R_i^+$ is

$$W^+ \;=\; (0)(2) + (1)(1) + (1)(3) + (1)(4) \;=\; 8.$$

- The p-value is the probability of getting a W^+ value that is <u>*at least*</u> 8.

  Therefore, the p-value $= P[W^+ = 8, 9, \text{ or } 10] = (1 + 1 + 1)/16 = 3/16 = .1875$.

Theorem 5.17 The distribution of the Wilcoxon signed rank statistic W^+ is symmetric about its mean $\mu_{W^+} = [n(n+1)/4]$ <u>if $H_0 : \mu = \mu_0$ is true</u>.

5.4 Linear Rank Statistics

- Earlier we studied the ranksum statistic $W = \sum_{i=1}^{n} R_i$ where R_i is the rank of X_i among a combined sample $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$.
- If $H_0 : \Delta = 0$ is true, then the random variables $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$ are i.i.d, and by Corollary 5.9, W is distribution-free over the class of continuous distributions $\mathcal{A}$.
- The test statistic W has two important properties:
  1. W maintains the desired α-level over a very broad class of distributions ($\mathcal{A}$).
  2. The power of W is excellent for detecting a shift for many distributions, especially for a medium-tailed distribution (such as the normal or logistic).
- We now consider a <u>general class</u> of rank statistics (which includes W).
- Let $\mathbf{R} = (R_1, R_2, \ldots, R_N)$ be a vector of ranks. Let $a(1), a(2), \ldots, a(N)$ and $c(1), c(2), \ldots, c(N)$ be two sets of n constants. A statistic of the form

$$S \;=\; \sum_{i=1}^{N} c(i)\, a(R_i)$$

  is called a **linear rank statistic**. The constants $a(1), a(2), \ldots, a(n)$ are called the **scores**, and $c(1), c(2), \ldots, c(n)$ are called the **regression constants**.
- The choice of $c(1), c(2), \ldots, c(n)$ will depend on the specific testing problem of interest.

Case I:

- In two-sample problems $\mathbf{R}$ is the rank vector of $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$. In general, let $R_1, R_2, \ldots, R_n$ be the ranks of $X_1, X_2, \ldots, X_n$ and $R_{n+1}, R_{n+2}, \ldots, R_{m+n}$ be the ranks of $Y_1, Y_2, \ldots, Y_m$. If

$$\begin{aligned} c(i) &= 1 \qquad \text{for } i = 1, 2, \ldots, n \\ &= 0 \qquad \text{for } i = n+1, n+2, \ldots, m+n \end{aligned} \tag{17}$$

  then $S = \sum_{i=1}^{m+n} c(i)\, a(R_i) = \sum_{i=1}^{n} a(R_i)$ which is the sum of the scores associated with the ranks of $X_1, X_2, \ldots, X_n$.

- The constants $c(i)$ in (17) are called **two-sample regression constants**.

Case II:

- For Case I, if we also let $a(i) = i$ for $i = 1, 2, \ldots, m+n$, then $S = \sum_{i=1}^{n} R_i$ which is the ranksum statistic W. The scores $a(i) = i$ are called the **Wilcoxon scores**.

Case III:

- It is clear that a different choice of $a(1), a(2), \ldots, a(N)$ scores for the two-sample problem will yield a test statistic with different properties.

- Let $\widehat{M}$ = the median of the combined sample $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$, and define

$$\begin{aligned} a(i) &= 0 \qquad \text{if } i \leq \frac{m+n+1}{2} \\ &= 1 \qquad \text{if } i > \frac{m+n+1}{2} \end{aligned} \tag{18}$$

  Consider S with these $a(i)$ scores and the two-sample regression constants in Case I:

$$\begin{aligned} S &= \sum_{i=1}^{n} a(R_i) \\ &= \text{the number of } X_i \text{ values larger than the sample median } \widehat{M} \end{aligned}$$

- This S is the linear rank statistic for the **two-sample median test**, and the scores in (18) are called the **median scores**.

5.4.1 Linear Rank Statistics under H_0

- In this section, general properties of linear rank statistics will be studied under the **null hypothesis** where "null hypothesis" refers to any set of assumptions that will result in the rank vector $\mathbf{R}$ being uniformly distributed over $\mathcal{R}$ (the set of permutations of the integers $1, 2, \ldots, N$).

- In future sections, we will study the null hypothesis for specific testing problems.

Lemma 5.18 Let $a(1), a(2), \ldots, a(N)$ be a set of N constants. Then, if $\mathbf{R}$ is uniformly distributed over permutation set $\mathcal{R}$,

$$\begin{aligned} E[a(R_i)] &= \frac{1}{N}\sum_{i=1}^{N} a(i) \;=\; \overline{a} \qquad \text{for } i = 1, 2, \ldots, N \\ Var[a(R_i)] &= \frac{1}{N}\sum_{k=1}^{N} (a(i) - \overline{a})^2 \\ Cov[a(R_i), a(R_j)] &= \frac{-1}{N(N-1)}\sum_{k=1}^{N} (a(i) - \overline{a})^2 \;=\; \frac{1}{N-1} Var[a(R_i)] \quad \text{for } i \neq j \end{aligned}$$

- The proof of Lemma 5.18 involves using Theorem 5.7 and the definitions of $E(\cdot)$, $Var(\cdot)$, and $Cov(\cdot, \cdot)$.
- Lemma 5.18 is used to establish the mean and variance of a linear rank statistic under the null hypothesis.

Theorem 5.19 Let S be a linear rank statistic with regression constants $c(1), c(2), \ldots, c(N)$ and scores $a(1), a(2), \ldots, a(N)$. If $\mathbf{R}$ is uniformly distributed over $\mathcal{R}$, then

$$E[S] \;=\; N\overline{ca} \qquad \text{and}$$

$$Var[S] \;=\; \frac{1}{N-1}\left[\sum_{i=1}^{N} (c(i) - \overline{c})^2\right]\left[\sum_{k=1}^{N} (a(k) - \overline{a})^2\right]$$

where $\overline{a} = (1/N)\sum_{i=1}^{N} a(i)$ and $\overline{c} = (1/N)\sum_{i=1}^{N} c(i)$.

5.5 Asymptotic Normality of Rank Statistics (Supplemental)

- The regression constants $c(1), c(2), \ldots, c(N)$ are determined by the problem of interest. Thus, we will only place a weak restriction on these constants.
- The restriction essentially requires that asymptotically no individual c_i value is much larger than the other constants. Specifically, the restriction is

$$\frac{\sum_{i=1}^{N} (c(i) - \overline{c})^2}{\max_{1 \leq i \leq n} (c(i) - \overline{c})^2} \;\rightarrow\; \infty \;\text{ as }\; N \rightarrow \infty \tag{19}$$

  where $(1/N)\sum_{i=1}^{N} c_i$.

  This is known as **Noether's condition**.

- Let ϕ be a real-valued function defined on $(0, 1)$ that (i) does not depend on N, (ii) can be written as the difference $\phi = \phi_i - \phi_2$ of two non-decreasing functions, and (iii) satisfies

$$0 \;<\; \int_0^1 \left[\phi(u) - \overline{\phi}\right]^2 du \;<\; \infty \quad \text{with } \overline{\phi} = \int_0^1 \phi(u)du.$$

  A function $\phi(\cdot)$ with these properties is called a **square integrable score function**.

- For a square integrable function, $\int_0^1 \left[\phi(u) - \overline{\phi}\right]^2 du = \int_0^1 \phi^2(u)du - [\bar{(\phi)}]^2$.
- Let ϕ be a square integrable score function and $a(1), a(2), \ldots, a(N)$ be scores that satisfy any of the following three conditions:

(A1) $a(i) = \phi\left(\frac{i}{N+1}\right)$.

(A2) $a(i) = N \int_{(i-i)/N}^{i/N} \phi(u)du$ for $i = 1, 2, \ldots, N$.

(A3) $a(i) = E[\phi(U_{(i)})]$ where $U_{(i)}$ is the i^{th} order statistic from a random sample of size N from a uniform $(0,1)$ distribution.

Let $S = \sum_{i=1}^{N} c(i)\, a(R_i)$.

Let $S^+ = \sum_{i=1}^{N} c(i)\, \Psi(i)\, a(R_i)$.

Theorem 5.20 (Asymptotic Normality of Linear Rank Statistics): Under H_0 for a linear rank statistic S, and assuming Noether's condition and condition A1, A2 or A3, then

$$\frac{S - E(S)}{\sqrt{Var(S)}} \xrightarrow{d} N(0,1) \text{ as } N \to \infty$$

Theorem 5.21 (Asymptotic Normality of Signed Rank Statistics): Under H_0 for a linear rank statistic S^+, and assuming Noether's condition and condition A1, A2 or A3, then

$$\frac{S^+ - E(S^+)}{\sqrt{Var(S^+)}} \xrightarrow{d} N(0,1) \text{ as } N \to \infty$$

- The linear rank statistics and signed rank statistics discussed in this course all all have asymptotic $N(0,1)$ distributions after standardizing.