MATH 524 - NONPARAMETRIC STATISTICS

GREG TAM

Contents

1. September 5th, 2012

Classically: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$
$\mathcal{F}_\theta = \{f_\theta, \theta \in \Theta\}$
We will focus more on hypothesis testing in this course. Advanced topics include estimation of F and f. We may compare $\widehat{F}$ to $F_{\widehat{\theta}}$ or $\widehat{f}$ to $f_{\widehat{\theta}}$.

Kolmogorov-Smirnov statistic: $\sup_{x \in \mathbb{R}} \left| F_{\widehat{\theta}}(x) - \widehat{F}(x) \right|$

Cramér von Mises statsitic: $\displaystyle\int_{-\infty}^{\infty} (F_{\widehat{\theta}}(x) - \widehat{F}(x))^2 \, \mathrm{d}x$

1.1. **Notion of rank.** Given a random sample $X_1, \ldots, X_N$ from an arbitrary distribution F, let

$$R_i = \text{rank of } X_i$$

1.2. **Working hypothesis.** We generally assume that the data come from a continuous distribution such that there are no ties in rank almost surely. The advantage of using ranks is that it is robust. The assumptions under which the procedures are carried out are minimal. A disadvantage is that there is a loss of information. The mean of the ranks is constant.

1.3. **Distribution of $(R_1, \ldots, R_N)$.** The distribution of the random vector $(R_1, \ldots R_N)$ is uniformly distributed.

1.4. **Invariance with respect to F.** If F is continuous and strictly increasing, then

$$\begin{aligned} R_i \leq R_j &\Leftrightarrow X_i \leq X_j \\ &\Leftrightarrow F(X_i) \leq F(X_j) \end{aligned}$$

Proof.

$$\int_{-\infty}^{\infty} F(x)f(x)\,\mathrm{d}x = \int_0^1 u\,\mathrm{d}u = \frac{1}{2}$$

by changing $u = F(x)$, $\mathrm{d}u = f(x)\,\mathrm{d}x$ □

Theorem 1. *Let X be a random variable with CDF F, F continuous. Then the distribution of $F(X)$ is Uniform on $(0,1)$.*

Proof.

$$F^{-1}(u) = \inf\{x : F(X) \geq u\}$$

$\forall u, x$, it holds that $F^{-1}(u) \leq x \Leftrightarrow F(x) \geq u$.

- If F is continuous, then $F(F^{-1}(u)) = u$
- If F is continuous, then $F(X)$ is continuous, i.e. $\forall x \in \mathbb{R}$, $P(F(X) = u) = 0$

$P(F(X) = u) = 0$ is the same as $P(X = F^{-1}(u))$ or $P(X$ lies in the flat part of F).

$$P(\underbrace{F(X)}_{\in[0,1]} \leq u) = \begin{cases} 0 & u < 0 \\ 1 & u \geq 1 \end{cases}$$

If $u \in [0,1)$,

$$\begin{aligned} P(F(X) \leq u) &= 1 - P(F(X) > u) \\ &= 1 - P(F(X) \geq u) \text{ by continuity of } F \\ &= 1 - P(X \geq F^{-1}(u) \\ &= P(X \leq F^{-1}(u)) \text{ by continuity of } F \\ &= F(F^{-1}(u)) = u \end{aligned}$$

□

Remark 1. *Continuity of F is crucial*

$X \sim \text{Bernoulli}(p)$
$P(X = 0) = 1 - p$
$P(X = 1) = p$

$$F(X) = \begin{cases} F(0) = 1 - p & \text{with prob. } 1 - p \\ F(1) = 1 & \text{with prob. } p \end{cases}$$

2. September 10th, 2012

$$X \prec Y \Leftrightarrow G(y) \leq F(y)\ \forall\, y$$

where X has distribution F and Y has distribution G.

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

Our hypothesis is of the form

$$H_0 : F = G \qquad H_a : F \neq G \text{ or } F \prec G$$

but this really means that the distributions have the same location parameter, not that they are exactly the same.

$$W_{XY} = W_S - \frac{n(n+1)}{2}$$

Under H_0, any assignment of ranks is equally likely. Using this fact we can tabulate the distribution of W_{XY} and make inference.

3. September 12th, 2012

$$\mathrm{Var}(W_{XY}) = \sum_{(i,j)=(i',j')} \mathrm{Var}(H_{ij}) + \sum_{(i,j)\neq(i',j')} \mathrm{Cov}(H_{ij}, H_{i'j'})$$

$$\begin{aligned}\mathrm{Var}(H_{ij}) &= \frac{1}{2}\left(1-\frac{1}{2}\right)\\ &= \frac{1}{4}\end{aligned}$$

a) $i \neq i'$, $j \neq j'$

$$\begin{aligned}\mathbb{E}\left[H_{ij}H_{i'j'}\right] &= P(X_i < Y_j, X_{i'} < Y_{j'})\\ &= \frac{1}{4}\end{aligned}$$

b) $i = i'$, $j \neq j'$

$$\begin{aligned}\mathbb{E}\left[H_{ij}H_{i'j'}\right] &= P(X_i < Y_j, X_i, Y_{j'})\\ &= P(X_i < \min(Y_j, Y_{j'}))\end{aligned}$$

$$\mathbb{E}\left[\overline{Y}_S\right] = \mu = \frac{1}{N}\sum_{i=1}^{N} i = \frac{N+1}{2}$$

$$\mathrm{Var}(\overline{Y}_s) = \underbrace{\frac{N-n}{N-1}/n}_{\text{finite population factor}} \times \frac{1}{N}\sum_{i=1}^{N}(Y_i-\mu)^2$$

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}(Y_i-\mu)^2\\ &= \frac{1}{N}\left\{\sum_{i=1}^{N} Y_i^2 - N\mu^2\right\}\\ &= \frac{1}{N}\left\{\frac{N(N+1)(2N+1)}{6} - N\left(\frac{N+1}{2}\right)^2\right\}\\ &= \frac{N+1}{N}\left\{\frac{N^2-N}{12}\right\}\\ &= \frac{(N-1)(N+1)}{12}\\ \mathrm{Var}(\overline{Y}_s) &= \frac{N-n}{N-1}/n\frac{(N-1)(N+1)}{12}\\ &= \frac{1}{n}\frac{(N-n)(N+1)}{12}\end{aligned}$$

Treatment: 78, 64, 75, 45, 82 $\sim F$
Control: 110, 70, 53, 51 $\sim G$
We must decide now whether the treatment has any effect without assuming any distribution.

$$H_0 : F = G \qquad H_1 : F \prec G \ (\forall\, x : F(x) \geq G(x))$$

4.1. **Dealing with Ties.**

$$\underbrace{17\ 17\ 17}_{d_1=3}\ \underbrace{19}_{d_2=1}\ \underbrace{21}_{d_3=1}$$

4.1.1. *General Result on Sampling.* $(y_1, \ldots, y_N)$ = Population
Draw a sample without replacement at random.
sample: $\{y_i : i \in S\}$ $n = |S| \leq N$
$T = \sum_{i \in S} y_i$
We wish to compute $\mathbb{E}[T]$ and $\mathrm{Var}(T)$.
$\mathbb{E}[T] = \sum_{i \in S} \mathbb{E}[y_i] = n \cdot \overline{y}$ The y_i's are identically distributed (not independent) and

$$\mathbb{E}[y_i] = \sum_{j=1}^{N} y_j \frac{1}{N} = \overline{y}$$

$$\begin{aligned}\mathrm{Var}(T) &= \sum_{i \in S} \mathrm{Var}(y_i) + \sum_{i,j \in S, i \neq j} \mathrm{Cov}(y_i, y_j) \\ &= n \cdot \underbrace{\mathrm{Var}(y_i)}_{\frac{1}{N}\sum_{j=1}^{N}(y_j - \overline{y})^2 = \tau^2} + n(n-1)\underbrace{\mathrm{Cov}(y_i, y_j)}_{=\lambda}\end{aligned}$$

To get the covariance, let $n = N$.

$$\mathrm{Var}(T) = 0 = N\tau^2 + N(N-1)\lambda \Rightarrow \lambda = \frac{\tau^2}{N-1}$$

For general n,

$$\mathrm{Var}(T) = n\tau^2 - \frac{n(n-1)\tau^2}{N-1} = n\tau^2\left(1 - \frac{n-1}{N-1}\right) = n\tau^2\frac{N-n}{N-1}$$

$$\begin{aligned}\underbrace{W_S^*}_{=T} &= n\left(\frac{d_1 R_1 + d_2 R_2 + \ldots + d_l R_l}{N}\right) \\ &= \frac{n}{N}\left(\sum_{i=1}^{d_1} i + \sum_{i=1}^{d_2}(d_1 + i) + \ldots \sum_{i=1}^{d_l}(d_1 + \ldots + d_{l-1} + i)\right) \\ &= \frac{n}{N}\frac{N(N+1)}{2} \\ &= \frac{n(N+1)}{2}\end{aligned}$$

Since

$$R_1 = \frac{1}{d_1}\sum_{i=1}^{d_1} i = \frac{d_1+1)}{2}$$

$$R_2 = \frac{1}{d_2}\sum_{i=1}^{d_2}(d_1+i) = \frac{d_2 d_1}{d_2} + \frac{d_2(d_2+1)}{2d_2}$$

$$R_3 = \frac{1}{d_3}\sum_{i=1}^{d_3}(d_1+d_2+i) = d_1+d_2+\frac{d_3+1}{2}$$

$$\vdots$$

$$R_l = d_1+d_2+\cdots+d_{l-1}+\frac{d_l+1}{2}$$

Note that for any $a_1,\ldots,a_k \in \mathbb{R}$, $\overline{a} = \frac{1}{k}\sum a_i$,

$$\sum_{i=1}^{k}(a_i-\overline{a})^2 = \sum_{i-1}^{k} a_i^2 - k\overline{a}^2 \Leftrightarrow k\overline{a}^2 = \sum_{i-1}^{k} a_i^2 - \sum_{i=1}^{k}(a_i-\overline{a})^2$$

Now set $a_i = i$, $\overline{a} = (k+1)/2$

$$\begin{aligned} i = 1^k\left(i-\frac{k+1}{2}\right)^2 &= \sum_{i=1}^{k} i^2 - k\left(\frac{k+1}{2}\right)^2 \\ &= \frac{k(k+1)(2k+1)}{6} - i = 1^k\left(i-\frac{k+1}{2}\right)^2 \\ &= \frac{k(k^2-1)}{12} \\ &= \frac{k^3-k}{12} \end{aligned}$$

Hence

$$k\overline{a}^2 - \sum a_i^2 - \frac{k^3-k}{12}$$

$$\sum_{i=1}^{d_1} R_i^2 = d_1 R_1^2 = d_1\left(\frac{d_1+1}{2}\right)^2 = \sum_{i=1}^{d_1} i^2 - \frac{d_1^3-d}{12}$$

$$\sum_{i=d_1+1}^{d_2} R_i^2 = d_2 R_2^2 = d_2\underbrace{\left(d_1+\frac{d_2+1}{2}\right)^2}$$

$$\overline{a} = \frac{1}{d_2}\sum_{i=1}^{d_2}(d_1+i)$$

$$= \sum_{i=1}^{d_2}(d_1+i)^2 - \frac{d_2^3-d_2}{12}$$

$$\sum_{i=1}^{d_l} R_l^2 = d_l \underbrace{R_l^2}_{} = \sum i = 1^{d_l}(d_1 + \ldots + d_{l-1} + i)^2 - \frac{d_l^3 - d_l}{12}$$
$$\frac{1}{d_l} = \sum_{i=1}^{d_l}(d_1 + \ldots + c_{l-1} + i)$$

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\left(R_i^* - \frac{N+1}{2}\right)^2 &= \frac{1}{N}\sum (R_i^*)^2 - \left(\frac{N+1}{2}\right)^2 \\
&= \frac{1}{N}\left(\sum_{i=1}^{d_1} i^2 - \frac{d_1^3 - d_1}{12} + \sum_{i=1}^{d_2}(d_1 + i)^2 - \frac{d_2^3 - d_2}{12} + \right. \\
&\left. \ldots + \sum_{i=1}^{d_l}(d_1 + \ldots + d_{l-1} + i)^2 - \frac{d_l^3 - d_l}{12}\right) \\
&= \frac{1}{N}\left(\sum_{i=1}^{N} i^2\right) - \frac{1}{N}\sum_{j=1}^{l}\frac{d_j^3 - d_j}{12} - \frac{(N+1)^2}{4} \\
&= \frac{N^2 - 1}{12} - \frac{1}{N}\sum_{j=1}^{l}\frac{d_j^3 - d_j}{12} \\
&= \tau^2
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(W_S^*) &= n\frac{(N-n)(N^2-1)}{(N-1)12} - \frac{n(N-n)}{(N-1)N}\sum\frac{d_j^3 - d_j}{12} \\
&= \frac{nm(N+1)}{12} - \frac{nm}{N(N-1)12}\sum\frac{d_j^3 - d_j}{12}
\end{aligned}$$

$\max_i d_i/N$ cannot be too close to 1 or the variance is equal to 0.

$$H_0 : F = G \qquad H_1 : F \succ G$$

$$\begin{aligned}
P(W_S^* \geq 1720) &= 1 - P(W_S^* < 1720) \\
&= 1 - P(W_S^* \leq 1719)
\end{aligned}$$

5. September 19th, 2012

Definition 1 (Omnibus Test)**.**

$$\begin{cases} H_0 : F = G \\ H_1 : F \neq G \end{cases}$$

Idea:
Estimate F by F_m and G by G_n. We check whether the "distance" between F_m and G_n is large.

$$\|F_m - G_n\| = \begin{cases} \sup_t |F_m(t) - G_n(t)| \\ \int (F_m(t) - G_n(t))^2 \,\mathrm{d}t \end{cases}$$

$$F_m(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}(X_i \leq x)$$

$$Z_i = \mathbb{1}(X_i \leq x) \sim \text{Bernoulli}(F(x))$$
$$P(Z_i = 1) = P(X_i \leq x) = F(x)$$

so

$$\begin{aligned} F_m(x) &= \frac{1}{m}\sum_{i=1}^{m} Z_i \\ &\to F(x) \text{ almost surely by SLLN} \\ &= \mathbb{E}[Z] \end{aligned}$$

Stronger result(Glivenko-Cantelli Theorem):

$$\sup_{x\in\mathbb{R}} |F_m(x) - F(x)| \to^{a.s.} 0$$

- $H_1 : F \neq G$

$$D_{m,n} = \sup_{t\in\mathbb{R}} |F_m(t) - G_n(t)|$$

- $H_1 : H \prec G$ $(F(t) \geq G(t))$

$$D^+_{m,n} = \sup_{t\in\mathbb{R}} (F_m(t) - G_n(t))$$

- $H_1 : H \succ G$

$$D^-_{m,n} = \sup_{t\in\mathbb{R}} (G_n(t) - F_m(t))$$

Given $(X_1, \ldots, X_m)$ and $(Y_1, \ldots, Y_n)$, we look a the pooled sample, $(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = (X^*_1, \ldots, X^*_{n+m})$ and rank them $(R^*_1, \ldots, R^*_{m+n}) = (R_1, \ldots, R_m, S_1, \ldots, S_n)$

$$\begin{aligned} &\sup_x F_m(x) - G_n(x) \\ &= \max_{1\leq i\leq n+m} F_m(X^*_i) - G_n(X^*_i) \\ &= \max_{1\leq i\leq n+m} \left(\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}(X_j \leq X^*_i) - \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}(Y_j \leq X^*_i) \right) \\ &= \frac{n+m}{nm} \max_{1\leq i\leq n+m} \left(\frac{n}{n+m}\sum_{j=1}^{m} \mathbb{1}(X_j \leq X^*_i) - \frac{m}{n+m}\sum_{j=1}^{n} \mathbb{1}(Y_j \leq X^*_i) \right) \\ &= \frac{n+m}{nm} \max_{1\leq i\leq n+m} \left(\frac{n}{n+m}\sum_{j=1}^{m} \mathbb{1}(X_j \leq X^*_i) - \frac{m}{n+m}\sum_{j=1}^{n} \mathbb{1}(Y_j \leq X^*_i) + \right. \\ &\left. \frac{n}{n+m}\sum_{j=1}^{m} \mathbb{1}(Y_j \leq X^*_i) - \frac{n}{n+m}\sum_{j=1}^{m} \mathbb{1}(Y_j \leq X^*_i) \right) \\ &= \frac{n+m}{nm} \max_{1\leq i\leq n+m} \left(\frac{n}{n+m}\sum_{j=1}^{n+m} \mathbb{1}(X^*_j \leq X^*_i) - \sum_{j=1}^{n} \mathbb{1}(Y_j \leq X^*_i) \left[\frac{m}{n+m} + \frac{n}{m+n} \right] \right) \\ &= \frac{n+m}{nm} \max_{1\leq k\leq n+m} \left(\frac{n}{n+m}k - \sum_{j=1}^{n} \mathbb{1}(S_j \leq k) \right) \end{aligned}$$

For the asymptotic result given by Gnedenko and Koklyuk, we require that

$$\frac{n}{n+m} \to \eta \in (0,1)$$

since we require it to be a two sample problem.

6. September 24th, 2012

$$G(x) = \mathbb{P}\left(X + \Delta \leq x\right) = \mathbb{P}\left(X \leq x - \Delta\right) = F(x - \Delta)$$

We have

$$G(x) - F(x - \Delta) \leq F(x)$$

since F is increasing and G is stochastically bigger than F.

The power, $\pi(\Delta)$, is equal to

$$\mathbb{P}_{F,G}\left(\text{Test rejects } H_0\right)$$

or equivalently in our case,

$$\mathbb{P}_{F,\Delta}\left(\text{Test rejects } H_0\right)$$

$$\pi(0) = \alpha \approx \mathbb{P}_{F,\Delta}(W_s > c)$$

Since the Wilcoxon test is discrete, we cannot find a c such that the power is exactly α. If m and n are small, then $\pi(0) < \alpha$. If $\min(m,n) \to \infty$, we can approximate the statistic by the normal distribution and $\pi(0) \approx \alpha$.

We hope that $\pi(\Delta)$ is as close to 1 as possible.
Is it true that $0 < \Delta_1 < \Delta_2 \Rightarrow \pi(\Delta_1) \leq \pi(\Delta_2)$?

Proof.

$$Y_1, \ldots, Y_n \sim F(x - \Delta_1)$$
$$Y_1 - \Delta_1, \ldots, Y_n - \Delta_1 \sim F$$
$$Y_1 + \Delta_2 - \Delta_1, \ldots, Y_n + \Delta_2 - \Delta_1 \sim F(x - \Delta_2)$$
$$\mathbb{P}\left(Y_i + \Delta_2 - \Delta_1 \leq x\right) = \mathbb{P}\left(Y_i - \Delta_1 \leq x - \Delta_2\right) = F(x - \Delta_2)$$

We let $V_i = Y_i + \Delta_2 - \Delta_1$.

$$W_{XY} = W_S - \frac{n(n+1)}{2} = \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbb{1}(X_i < Y_j)$$

$$W_{XV} = \sum_{i=1}^{m}\sum_{j=1}^{n} \underbrace{\mathbb{1}(X_i < V_j)}_{\mathbb{1}(X_i < Y_j + \underbrace{\Delta_2 - \Delta_1}_{>0})}$$

$$W_{XY} \leq W_{XV} \Rightarrow W_{XY} > c + \frac{n(n+1)}{2} \Rightarrow W_{XV} > c + \frac{n(n+1)}{2}$$

Our treatment group is:
$Y_1, \ldots, Y_n \sim F(x - \Delta_1)$
$V_1^*, \ldots, V_n^* \sim F(x - \Delta_2)$
However, it is not necessarily true that $V_i^* = Y_i + \Delta_2 - \Delta_1$. Hence

$$\{W_{XY} > c\} \subset \{W_{XV} > c\} \Rightarrow \mathbb{P}\left(W_{XY} > c\right) \leq \mathbb{P}\left(W_{XV} > c\right)$$

$$\Rightarrow \pi(\Delta_1) \leq \pi(\Delta_2)$$

□

From symmetry of the normal distribution, we have

$$1 - \Phi(z) = \Phi(-z)$$

Then

$$\begin{aligned}
\pi(F, G) &= \Phi\left(\frac{\mathbb{E}\left[W_{XY}\right] - c}{\sqrt{\mathrm{Var}(W_{XY})}}\right) \\
&= \Phi\left(\frac{mnp_1 - z_\alpha\sqrt{\frac{mn(N+1)}{2}} - \frac{mn}{2}}{\sqrt{\mathrm{Var}(W_{XY})}}\right) \\
p_1 &= \mathbb{P}\left(X < Y\right) \qquad \left(= \frac{1}{2} \text{ under} H_0\right) \\
&= \int_{-\infty}^{\infty} F(y)g(y)\,\mathrm{d}y \\
F^*(t) &= \mathbb{P}\left(X - X' \leq t\right),\ X \perp X',\ X \sim X' \sim F \\
&= \int_{-\infty}^{\infty} F(t + x')f(x')\,\mathrm{d}x' \\
f^*(t) &= \int_{-\infty}^{\infty} f(t + x')f(x')\,\mathrm{d}x' \\
p_1 &= \mathbb{P}\left(X < X' + \Delta\right) \\
&= \mathbb{P}\left(X - X' < \Delta\right) \\
&= F^*(\Delta) \\
F^*(\Delta) &\approx F^*(0) + \frac{f^*(0)}{1!}\Delta + \dots \\
p_1 &\approx \frac{1}{2} + \frac{f^*(0)}{1!}\Delta
\end{aligned}$$

Now we have

$$\frac{mn(\frac{1}{2} + f^*(0)\Delta - \frac{1}{2}) - z_\alpha\sqrt{\frac{mn(N+1)}{12}}}{\sqrt{\frac{mn(N+1)}{12}}} = \sqrt{\frac{12}{N+1}}f^*(0)\Delta\sqrt{mn} - z_\alpha$$

7. September 26th, 2012

$$f^*(0) = \int_{-\infty}^{\infty} f(t)^2\,\mathrm{d}t$$

$$F \sim \mathcal{N}(0, \sigma^2) \qquad G \sim \mathcal{N}(\Delta, \sigma^2)$$

f^* is the density of $X - X'$, $X', X \sim F$, and $X' \perp X$. We have $X - X' \sim \mathcal{N}(0, 2\sigma^2)$.

$$f^*(0) = \frac{1}{2\sqrt{\pi}\sigma}$$

$$\pi(\Delta) \approx \Phi\left(\sqrt{\frac{12mn}{m+n+1}}\Delta f^*(0) - z_\alpha\right)$$

$$\Phi(z_\alpha) = 1 - \alpha$$

Example 1. $\Delta \approx 2$, $\sigma^2 \approx 2$

$$\pi(\Delta) = 1 - \beta$$

To make it easier, we assume $m = n$, which implies

$$\sqrt{\frac{12mn}{n+m+1}} = \sqrt{\frac{12n^2}{2n+1}} \approx \sqrt{6n}$$

$$\begin{aligned}\pi(\Delta) = 1 - \beta &= \Phi(z_\beta) \\ &\approx \Phi\left(\sqrt{6n}\frac{\Delta}{2\sqrt{\pi}\sigma} - z_\alpha\right) \\ \Leftrightarrow \sqrt{6n}\frac{1}{2\sqrt{\pi}}\frac{\Delta}{\sigma} - z_\alpha &= z_\beta\end{aligned}$$

so we have

$$n = \frac{(z_\beta + z_\alpha)^2 4\pi\sigma^2}{6\Delta^2}$$

For a large number of repetitions ($N = 1000$) do:

Step 1. Sample $X_1, \ldots, X_m$ from $\mathcal{N}(0, \sigma^2)$ and sample $Y_1, \ldots, Y_n$ from $\mathcal{N}(\Delta, \sigma^2)$ independently.

Step 2. Wilcoxon Test (X, Y). return the p-value. Store all p-values.

End. We will end up with $N = 1000$ p-values.

How often did the test reject? We look at the number of p-values which are less than α. The probability of the rejection is thus

$$\frac{\text{rej}}{N} \approx \mathbb{P}\left(\text{test rejecting } H_0 \text{ under our scenario}\right)$$

$$\begin{aligned}\mathbb{P}\left(\text{rejection}\right) &= \mathbb{P}\left(W_S > c\right) \\ &= 12\mathbb{P}\left(X_1 < X_2 < Y_1 < Y_2 < Y_3\right)\end{aligned}$$

$$X \sim F \qquad Y \sim G \qquad \Rightarrow \frac{X}{a} \sim G$$

$\log X$, $\log Y = \log X - \log a$

$$\begin{aligned}\mathbb{P}\left(\log X \leq x\right) &= \mathbb{P}\left(X \leq e^x\right) \\ &= 1 - \exp(-e^x) \text{ Gumbel distribution}\end{aligned}$$

Here, $\Delta = -\log a$.

Suppose that $\text{Var}(F) < \infty$ $(H_0 : F = G)$

To compute the critical level, we need the distribution of the t-statistic under H_0.

$$\frac{\overline{Y} - \overline{X}}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

Central Limit Theorem:

$$\sqrt{n}\frac{\overline{Y} - \mathbb{E}\left[Y\right]}{\sqrt{\text{Var}(F)}} \overset{n\to\infty}{\rightsquigarrow} \mathcal{N}(0, 1)$$

$$\sqrt{m}\frac{\overline{X}-\mathbb{E}[X]}{\sqrt{\mathrm{Var}(F)}} \overset{n\to\infty}{\rightsquigarrow} \mathcal{N}(0,1)$$

$$\frac{S}{\sigma} \to^P 1$$

$$\begin{aligned}\frac{\overline{Y}-\overline{X}}{S}\sqrt{\frac{mn}{m+n}} &= \frac{\sigma}{S}\frac{\overline{Y}-\overline{X}\pm\mu}{\sigma}\sqrt{\frac{mn}{m+n}} \\ &= \frac{\sigma}{S}\left(\frac{(\overline{Y}-\mu)\sqrt{n}\sqrt{\frac{m}{n+m}}}{\sigma} - \frac{(\overline{X}-\mu)\sqrt{m}\sqrt{\frac{n}{n+m}}}{\sigma}\right)\end{aligned}$$

The $\frac{\sigma}{S}$ term converges to 1 and the rest converges to

$$\mathcal{N}(0,1)\lambda + \mathcal{N}(0,1)(1-\lambda) \sim \mathcal{N}(0,1)$$

where $\lambda = \sqrt{\frac{m}{n+m}}$ as $(m,n)\to\infty$ and $\frac{m}{n+m}\to\lambda\in(0,1)$.

8. October 1st, 2012

Wilcoxon test: for Δ small,

$$\text{Power} = \pi(\Delta) \approx \Phi\left(\sqrt{\frac{12mn}{m+n}}\underbrace{f^*(0)}_{\int_{-\infty}^{\infty} f(x)^2\,\mathrm{d}x}\Delta - z_\alpha\right)$$

T-test:

$$\begin{aligned}\pi(\Delta) &\approx \Phi\left(\Delta\frac{1}{\sigma\sqrt{1/m+1/n}} - z_\alpha\right) \\ &= \Phi\left(\sqrt{\frac{mn}{m+n}}\frac{1}{\sigma}\Delta - z_\alpha\right)\end{aligned}$$

8.1. **Comparing Wilcoxon to t-test.** How do we compare the Wilcoxon and t-test for the shift model when under H_0, the hypothesized distribution is F.
Idea 1: look at the power curve π in a neighbourhood of $\Delta = 0$.

$$\Delta_n = \frac{h}{\sqrt{n}}$$

Wilcoxon: (Test 1)

$$\pi(\Delta_n) = \Phi\left(\sqrt{6}f^*(0)h - z_\alpha\right)$$

T-test: (Test 2)

$$\pi(\Delta_n) = \Phi\left(\frac{1}{\sqrt{2}\sigma}h - z_\alpha\right)$$

These curves are asymptotic powers. Test 1 is better than test 2 if the slope of the asymptotic power curve at 0 is bigger for test 1 than for test 2.
Test 1:

$$\Phi'(-z_\alpha)\sqrt{6}f^*(0)$$

Test 2:

$$\Phi'(-z_\alpha)\frac{1}{\sqrt{2}\sigma}$$

So this is true if

$$\Phi'(-z_\alpha)\sqrt{6}f^*(0) > \Phi'(-z_\alpha)\frac{1}{\sqrt{2}\sigma}$$

or equivalently if

$$\sqrt{12}f^*(0)\sigma > 1$$

Idea 2:
Test 1: Compute the sample size to reach power β at level α.

$$n_1 = \frac{(z_\alpha + z_\beta)^2}{\Delta^2 6\{f^*(0)\}^2}$$

Test 2:

$$n_2 = \frac{(z_\alpha + z_\beta)^2\sigma^2 2}{\Delta^2}$$

We prefer the test that has a smaller sample size so we look at the relative efficiency.

$$\begin{aligned}\frac{n_2}{n_1} &= \frac{\frac{(z_\alpha+z_\beta)^2\sigma^2 2}{\Delta^2}}{\frac{(z_\alpha+z_\beta)^2}{\Delta^2 6\{f^*(0)\}^2}} \\ &= 12\sigma^2\{f^*(0)\}^2\end{aligned}$$

Test 1 is better than test 2 if $12\sigma^2\{f^*(0)\}^2 > 1$. The ratio

$$\frac{n_2}{n_1} \underset{n\to\infty}{\longrightarrow} \text{ is called Pitman's asymptotic relative efficiency}$$

8.2. **Optimal Tests.**

$$W_S = \sum_{i=1}^{n} S_i$$

$$\underbrace{X_1, \ldots, X_m}_{\text{controls}}, \underbrace{Y_1, \ldots, Y_n}_{\text{treatments}}$$

We have $R_1^*, \ldots R_N^*$ which are the ranks of the pooled sample.

$$W_S = \sum_{i=m+1}^{N} R_i^* = \sum_{i=1}^{N} C_{N,i} a_{N,R_i^*}$$

The $C_{N,i}$ are called coefficients and the typical choice of these coefficients in the two-sample problem is:

$$C_{N,i} = \begin{cases} 0 & \text{if } i \in \{1, \ldots, m\} \\ 1 & \text{if } i \in \{m+1, \ldots, N\} \end{cases}$$

Wilcoxon test: $a_{N,R_i^*} = R_i^*$
Question: How can we compute the scores as to achieve maximal power?

General Result: Assume that $X_1, \ldots, X_n$ have density f and $Y_1, \ldots, Y_n$ have density f_θ. Then under smoothness conditions on f, the optimal score function

$$a_{N,k} = \mathbb{E}_{\theta_0}\left(-\frac{\frac{\partial f}{\partial\theta}\big|_{\theta_0}}{f_{\theta_0}} \circ F_{\theta_0}^{-1}(\mathcal{U}_{(k)})\right)$$

where

$$\mathcal{U}_1, \ldots \mathcal{U}_n \text{ are from } \mathcal{U}(0,1)$$

and

$$\mathcal{U}_{(1)} \leq \mathcal{U}_{(2)} \leq \ldots \leq \mathcal{U}_{(N)}$$

If $X_1, \ldots, X_n$ are from F_{θ_0}, then

$$X_{(1)} \leq \ldots \leq X_{(N)} \stackrel{d}{=} F_{\theta_0}^{-1}(\mathcal{U}_{(1)}) \leq \ldots \leq F_{\theta_0}^{-1}(\mathcal{U}_{(N)})$$

Example 2. *f_{θ_0} is the density of $\mathcal{N}(0,1)$. f_θ is the density of $\mathcal{N}(\Delta,1)$.*

$$f_\theta = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-\Delta)^2}{2}\right\}$$

$$\left.\frac{\partial f_\Delta}{\partial \Delta}\right|_{\Delta=0} = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}(-x)$$

$$\frac{\left.\frac{\partial f_\Delta}{\partial \Delta}\right|_{\Delta=0}}{f_0} = x$$

So our optimal scores are thus

$$\mathbb{E}\left[\Phi^{-1}(\mathcal{U}_{(k)})\right] = \mathbb{E}\left[X_{(k)}\right]$$

Suppose that the exact scores are

$$A_{N,k} = \mathbb{E}\left[\phi(\mathcal{U}_{(k)})\right]$$

Then if ϕ is well behaved, then the test using the approximative scores:

$$A^*_{N,k} = \phi(\mathbb{E}\left[\mathcal{U}_{(k)}\right]) = \phi\left(\frac{k}{N+1}\right)$$

achieves the same asymptotic power (i.e. the same efficiency)

Example 3. *Suppose we had the shift model*

$$X_1, \ldots, X_m \text{ with density } f(x)$$

$$Y_1, \ldots, Y_n \text{ with density } f(x_\Delta)$$

$$\frac{\partial f}{\partial \Delta} = -f'(x-\Delta)$$

$$\left.\frac{\partial f}{\partial \Delta}\right|_{\Delta=0} = -f'(x)$$

Optimal exact scores: $a_{N,i} = \mathbb{E}\left[-\frac{f'}{f}(F^{-1}(\mathcal{U}_{(k)})\right]$
Approximative scores: $a_{N,i} = -\frac{f'}{f}(F^{-1}\left(\frac{k}{N+1}\right))$
Now let $F(X) = \frac{1}{1+e^{-x}}$, $x \in \mathbb{R}$ *(logistic distribution)*

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2},\ x \in \mathbb{R}$$

Optimal scores in the logistic shift model

$$f'(x) = \frac{(-e^{-x})(1+e^{-x}-2e^{-x})}{(1+e^{-x})^3}$$

$$= \frac{(-e^{-x})(1-e^{-x})}{(1+e^{-x})^3}$$

$$F^{-1}(u) = \ln\left(\frac{u}{1-u}\right)$$

$$
\begin{aligned}
-\frac{f'(x)}{f(x)} &= \frac{e^{-x}(1-e^{-x})(1+e^{-x})^2}{(1+e^{-x})^3 e^{-x}} \\
&= \frac{1-e^{-x}}{1+e^{-x}} \\
-\frac{f'}{f}(F^{-1}(u)) &= \frac{1-\frac{1-u}{u}}{1+\frac{1-u}{u}} \\
&= \frac{u-1+u}{u+1-u} \\
&= 2u-1
\end{aligned}
$$

exact: $\mathbb{E}\left[2\mathcal{U}_{(k)}-1\right] = 2\frac{k}{N+1}-1$ *so we have that our optimal test*

$$\sum_{i=1}^{n} 2\frac{S_i}{N+1} - 1 \propto W_S$$

Example 4. *$X_1,\dots,X_m$ have density f and $Y_1,\dots,Y_n$ have density $\theta f(1-F)^\theta$*
$1-(1-F)^\theta$ are called the Lehmann alternatives

$$\frac{f_\theta}{1-F_\theta} = \theta\frac{f}{1-F}$$

$$\phi(u) = 1+\ln(1-u)$$

(in the book there is a minus sign instead of a plus but Johanna said she did the calculation multiple times..) In the exact case this is equal to

$$\mathbb{E}\big[-\underbrace{\ln(1-\mathcal{U}_{(k)})}_{X_{(k)}}\big]$$

and in the approximate case we have

$$-\ln\left(1-\frac{k}{N+1}\right)$$

9. October 3rd, 2012

9.1. **Lehmann Alternatives.**

$$H_1 : G = 1-(1-F)^\Delta$$

then

$$g = \Delta(1-F)^{\Delta-1}f$$

Hazard rate: $\dfrac{\mathbb{P}(X=t)}{\mathbb{P}(X>t)} = \dfrac{f(t)}{1-F(t)} = \lambda(t)$

$$\frac{\Delta(1-F(t))^{\Delta-1}f(t)}{(1-F(t))^\Delta} = \Delta\frac{f(t)}{1-F(t)}$$

$$
\begin{aligned}
H_0 &: \lambda = \mu \\
H_1 &: \mu = \Delta\cdot\lambda,\ \Delta > 0
\end{aligned}
$$

9.2. **Estimation of Treatment Effect.** Assume F is continuous.

$$H_0 : F = G$$
$$H_1 : G = F(x - \Delta)$$

Suppose we had $X \sim F$, $Y \sim G$. Then

$$Y \stackrel{d}{=} X + \Delta$$

We have that $Y \stackrel{d}{=} X + \Delta$. Then

$$\mathbb{E}[Y - X] = \mathbb{E}[Y] - \mathbb{E}[X] = \Delta$$

or also

$$Y - X \stackrel{d}{=} X' + \Delta - X = (X' - X) + \Delta$$

where $X' - X$ is symmetric around 0, this is all symmetric around Δ. We also have

$$\text{med}(Y - X) = \Delta$$

Example 5. *Suppose we have* $X_1, \dots X_n \sim \mathcal{N}(\mu, 1)$

$$\overline{X} \sim \mathcal{N}\left(0, \frac{1}{n}\right) \qquad \overline{X} \sim \mathcal{N} - \mu\left(0, \frac{1}{n}\right)$$

and there is no dependency on μ.

$(X_1, \dots, X_m) = X$
$(Y_1, \dots, Y_n) = Y$

$$\widehat{\Delta}(X, Y) = \text{med}(Y_j - X_i),\ i \in \{1, \dots, m\}, j \in \{1, \dots, n\}$$

so

$$\widehat{\Delta}(aX + b, aY + b) = a\widehat{\Delta}(X, Y)$$

$$\begin{aligned}\widehat{\Delta} \text{ is symmetric around } \Delta &\Leftrightarrow \widehat{\Delta} - \Delta \text{ is symmetric around } 0 \\ &\Leftrightarrow \widehat{\Delta} \text{ is symmetric around } 0 \text{ in the case } F = G\end{aligned}$$

Example 6. *Given* Y_1, X_1, X_2*, we have our possible differences are* $Y_1 - X_1$ *and* $Y_1 - X_2$.

$$\begin{aligned}\widehat{\Delta} &= \frac{Y_1 - X_1 + Y_1 - X_2}{2} \\ &= Y_1 + \frac{X_1 + X_2}{2} \stackrel{d}{\neq} \frac{X_1 + X_2}{2} - Y_1 \\ &\qquad Y_1 - X_1 \stackrel{d}{=} X_1 - Y_1\end{aligned}$$

If F is continuous, it can be shown that $\widehat{\Delta}$ is also continuous, i.e. $\mathbb{P}\left(\widehat{\Delta} = x\right) = 0$ $\forall\, x \in \mathbb{R}$. If mn is odd, then

$$\mathbb{P}\left(\widehat{\Delta} < \Delta\right) = \mathbb{P}\left(\widehat{\Delta} > \Delta\right) = \frac{1}{2}$$

In the case where mn is even,

$$D_{(k)} < \widehat{\Delta} = \frac{D_{(k)} + D_{(k+1)}}{2} < D_{(k+1)}$$
$$\mathbb{P}\left(\widehat{\Delta} \leq 0\right) \leq \mathbb{P}\left(D_{(k)} \leq 0\right)$$

$$\mathbb{P}\left(\widehat{\Delta} \leq 0\right) \geq \mathbb{P}\left(D_{(k+1)} \leq 0\right)$$

10. October 10th, 2012

10.1. **Consistency of $\widehat{\Delta}$.**

$$\widehat{\Delta}_{mn} \xrightarrow{\mathbb{P}} \Delta$$

$$\forall\, \varepsilon > 0,\ \mathbb{P}\left(|\widehat{\Delta} - \Delta| > \varepsilon\right) \to 0 \text{ as } mn \to \infty \Leftrightarrow \mathbb{P}\left(|\widehat{\Delta} - \Delta| \leq \varepsilon\right) \to 1$$

If mn is odd, then it is equal to $2k+1$, $\widehat{\Delta} = D_{(k)}$

$$\begin{aligned}
\mathbb{P}_\Delta(|\widehat{\Delta} - \Delta| \leq a) &= \mathbb{P}_\Delta(\Delta - a \leq \widehat{\Delta} \leq \Delta + a) \\
&= \mathbb{P}_\Delta(\widehat{\Delta} - \Delta \leq a) - \mathbb{P}_\Delta(\widehat{\Delta} - \Delta \geq -a) \\
&= \mathbb{P}_0(\widehat{\Delta} \leq a) - \mathbb{P}_0(\widehat{\Delta} \geq -a) \\
&= \mathbb{P}_0(D_{(k)} \leq a) - 1 + \mathbb{P}_0(D_{(k)} \leq -a) \\
&= \mathbb{P}_0(W_{X,Y-a} \leq k+1) - 1 + \mathbb{P}_0(W_{X,Y+a} \leq k+1) \\
&= \mathbb{P}_0\left(\frac{W_{X,Y-a} - p_1 mn}{\sqrt{\mathrm{Var}(W_{X,Y-a})}} \leq \frac{\frac{mn}{2} + \frac{1}{2} - p_1 mn}{\sqrt{\mathrm{Var}(W_{X,Y-a})}}\right) - 1 \\
&\quad + \mathbb{P}_0\left(\frac{W_{X,Y+a} - p_1^* mn}{\sqrt{\mathrm{Var}(W_{X,Y+a})}} \leq \frac{\frac{mn}{2} + \frac{1}{2} - p_1^* mn}{\sqrt{\mathrm{Var}(W_{X,Y+a})}}\right) \\
&= \Phi\left(\frac{mn(\frac{1}{2} - p_1)}{\sqrt{\mathrm{Var}(W_{X,Y-a})}}\right) - 1 + \Phi\left(\frac{mn(\frac{1}{2} - p_1)}{\sqrt{\mathrm{Var}(W_{X,Y+a})}}\right)
\end{aligned}$$

where

$$\begin{aligned}
p_1 &= \mathbb{P}\left(X < Y - a\right) \\
p_1^* &= \mathbb{P}\left(X < Y + a\right)
\end{aligned}$$

10.2. **Paired Z-test versus Classical Z-test.** We observe that the power of the paired test is bigger than the power of the classical test iff $\varrho > 0$.

Suppose we have

	$i=1$	$i=2$	$i=3$
C/T	5.2	4.8	4.1
T/C	5.4	4.2	4.0

where we do not know which one corresponds to the control and which one corresponds to the treatment. We have that

$$H_0 : \text{Treatment and control allocation}$$

$$\mathbb{P}\left(\text{1st is control}\right) = \mathbb{P}\left(\text{2nd is control}\right) = \frac{1}{2}$$

Our S_N is binomially distributed.

$$\frac{\sqrt{N}\frac{1}{N}S_n - \frac{1}{2}}{\sqrt{\frac{1}{4}}} \approx \mathcal{N}(0,1)$$

$$\frac{2}{\sqrt{N}}S_n - \sqrt{N} \approx Z$$

so

$$S_N \approx Z\frac{\sqrt{N}}{2} + \frac{N}{2} \approx \mathcal{N}\left(\frac{N}{2}, \frac{N}{4}\right)$$

11. October 15th, 2012

Suppose we have

Controls	X_i	5.2	4.7	5.8
Treatment	Y_i	5.1	5.0	5.2
	$Y_i - X_i$	-0.1	0.3	-0.6

$$S_n = 2 \qquad \mathbb{P}(\underbrace{S_N}_{\text{Bin}\left(3, \frac{1}{2}\right)} \geq 2)$$

$\lvert Y_i - X_i \rvert$	0.1	0.3	0.6
Ranks of $\lvert Y_i - X_i \rvert$	-1	2	-3

$$V_r = 1 + 3 = 4 \qquad \mathbb{P}\left(V_r \geq 4\right)$$

The random part of this is the sign ($+$ or $-$). To calculate the p-value, we use this fact. Our possibilities are

	Pr	V_S
$---$	1/8	0
$--+$	1/8	3
$-+-$	1/8	2
$+--$	1/8	1
$-++$	1/8	5
$+-+$	1/8	4
$++-$	1/8	3
$+++$	1/8	6

So

$$V_S = \begin{cases} 0 & \text{with prob. } \frac{1}{8} \\ 1 & \text{with prob. } \frac{1}{8} \\ 2 & \text{with prob. } \frac{1}{8} \\ 3 & \text{with prob. } \frac{2}{8} \\ 4 & \text{with prob. } \frac{1}{8} \\ 5 & \text{with prob. } \frac{1}{8} \\ 6 & \text{with prob. } \frac{1}{8} \end{cases}$$

and we have

$$\mathbb{P}\left(V_S \leq 2\right) = \mathbb{P}\left(V_S \in \{0, 1, 2\}\right) = \frac{3}{8}$$

11.1. **Moments of V_S^*.** Denote the mid-ranks of $|D_1|, \ldots, |D_n|$ as $a_1, \ldots, a_N$.

$$\begin{aligned} \mathbb{E}\left[V_S^*\right] &= \mathbb{E}\left[\sum_{\{i: X_i \neq Y_i\}} a_i I_i\right] \qquad I_i \sim \text{Ber}(1/2) \\ &= \sum_{\{i: X_i \neq Y_i\}} a_i \frac{1}{2} \end{aligned}$$

$$= \sum_{i=1}^{N} a_i \frac{1}{2} - \frac{1}{2}\left(\frac{d_0+1}{2}\right) d_0$$
$$= \frac{1}{2}\frac{N(N+1)}{2} - \frac{1}{2}\left(\frac{d_0+1}{2}\right) d_0$$

We must note that we have $\mathbb{P}(I_i = 1) \neq \mathbb{P}(Y_i > X_i)$ but actually is equal to $\mathbb{P}(Y_i > X_i | X_i \neq Y_i)$

$$\begin{aligned}
\operatorname{Var}(V_S^*) &= \sum_{\substack{i=1 \\ X_i \neq Y_i}}^{N} a_i^2 \frac{1}{4} \\
&= \frac{1}{4} \underbrace{\sum_{i=1}^{N} a_i^2}_{} - \frac{(d_0+1)^2}{4} d_0 \frac{1}{4} \\
&\frac{N(N+1)(2N+1)}{6} - \sum_{j=0}^{l} \frac{d_j(d_j^2-1)}{12} \\
&= \frac{N(N+1)(2N+1)}{24} \\
&\quad - \frac{d_0(d_0+1)(2d_0+1)}{24} \\
&\quad - \frac{\sum_{i=1}^{l} d_i(d_i+1)(d_i-1)}{48}
\end{aligned}$$

$$V_S^* = 17.5$$
$$\mathbb{E}[V_S^*] = \frac{7 \cdot 8}{4} - \frac{3 \cdot 4}{4} = 14 - 3 = 11$$
$$\operatorname{Var}(V_S^*) = \frac{7 \cdot 8 \cdot 15}{24} - \frac{3 \cdot 4 \cdot 7}{24} - \frac{2 \cdot 3 + 2 \cdot 3}{48} = 31.25$$

We usually do not use the continuity correction for approximation of p-values since the values of the statistics are not necessarily integers.

12. October 22nd, 2012

12.1. **Multiple Data Blocks.** Suppose we would like to compare whether live lectures or online lectures. Let's say the grades are tabulate as follows:

Live (G)	72	75	83	95	100
Online (F)	68	69	74	82	93

We could test $H_0 : G = F$ versus $H_1 : G \succ F$, in which case we could have

$$W_{S_1} = 3 + 5 + 7 + 9 + 10 = 34$$

where

$$\mathbb{P}(W_S \geq 34) = 0.11$$

Suppose that the data were from year 1 and we look at the data again for year 2.

Live (G_2)	54	59	60	70
Online (F_2)	47	51	52	56

However, the distribution of F may not be the same as F_2. Now if we rank the data for year 2, we get

$$W_{S_2} = 4 + 6 + 7 + 8 = 25$$

We take

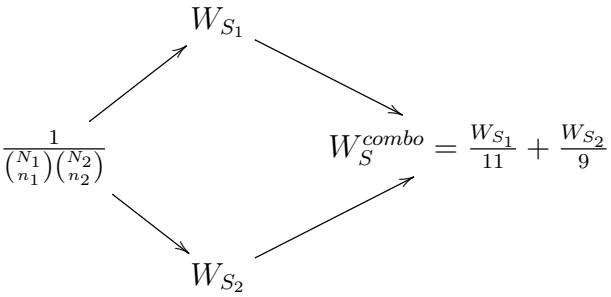
$$W_S^{combo} = \sum_{k=1}^{b} c_k W_{S_k}$$

where

$$c_k = \frac{1}{N_k + 1}$$

This is our optimal choice. So we have

$$W_S^{combo} = \frac{34}{11} + \frac{25}{9} = 5.869$$



$$\mathbb{E}\left[W_S^{combo}\right] = \frac{5+4}{2} = 4.5$$
$$\mathrm{Var}(W_S^{combo}) = \frac{5 \cdot 5}{12 \cdot 11} + \frac{4 \cdot 4}{12 \cdot 9} = 0.3375$$

Using this knowledge, we approximate

$$\begin{aligned}\mathbb{P}\left(W_S^{combo} \geq 5.869\right) &= \mathbb{P}\left(\frac{W_S^{combo} - 4.5}{\sqrt{0.3375}} \geq \frac{5.869 - 4.5}{\sqrt{0.3375}}\right) \\ &\approx 1 - \Phi(2.3558) \\ &= 0.00924\end{aligned}$$

12.2. **Power of the Sign Test.** Suppose we have

$$(X_1, Y_1), \ldots, (X_N, Y_N) \quad iid$$

and we want to test H_0 that there is no treatment effect against H_1 that the treatment increases the response. However here, X_i and Y_i are not necessarily independent. We would like to see that under the null, $\mathbb{P}(Y - X > 0) = 1/2$ and under the alternative $\mathbb{P}(Y - X > 0) > 1/2$.

Looking at $\mathbb{P}(Y - X > 0) = 1/2$ is not good enough however. We may also consider if the distribution of $Y - X$ is symmetric around 0 ($Y - X \stackrel{d}{=} X - Y$). But even this isn't good enough!

Caveat for exchangeability:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \stackrel{d}{=} \begin{pmatrix} Y \\ X \end{pmatrix} \Rightarrow X \stackrel{d}{=} Y \ (F = G)$$

but the converse is not true.

Under the shift model, if $Y_i - \Delta \stackrel{d}{=} X_i$, then

$$\begin{pmatrix} X \\ Y - \Delta \end{pmatrix}$$

is exchangeable.

Exchangeability implies $Y - X$ is symmetric around 0 which implies $P(Y - X > 0) = 1/2$. However, the converses are certainly not true. But even in reality, we are actually using the fact that

$$\mathbb{P}(Y_i - X_i > 0) = \frac{1}{2} \quad \forall\, i$$

13. October 24th, 2012

13.1. **Sign Test.** Denote

$$Z = Y - X$$

Under the null, we assume $Z \sim L$ where L is a cdf.

Sign Test:

- (X_i, Y_i) for $i = 1, \ldots, n$ are iid.
  H_0: $(X, Y) \stackrel{d}{=} (Y, X)$
  $H_1 : (X, Y - \Delta),\ \Delta > 0$ is exchangeable.
  This is the most restrictive hypothesis.
- $Z_1, \ldots, Z_n$ iid and
  $H_0 : L$ is symmetric about 0
  $H_1 :$ L is slanted towards positive values.
  or also $H_1^* : Y - \Delta - X$ is symmetric about 0 with $\Delta > 0$.
- $Z_1, \ldots, Z_n$ are independent.
  $H_0 : \mathbb{P}(Z_i > 0) = \frac{1}{2}\ \forall i = 1, \ldots, N$
  $H_1 : \mathbb{P}(Z_i > 0) = p > \frac{1}{2}\ \forall i = 1, \ldots, N.$

We have our statistic

$$S_N = \sum_{i=1}^{N} \mathbb{1}(Z_i > 0) \sim \text{Bin}(N, p)$$

and so

$$\frac{S_N - Np}{\sqrt{Np(1-p)}} \sim \mathcal{N}(0, 1)$$

Hence we use

$$\begin{aligned} \mathbb{P}_0(S_N > c) &= \mathbb{P}\left(\frac{S_N - \frac{N}{2}}{\sqrt{\frac{N}{4}}} > \frac{c - \frac{N}{2}}{\sqrt{\frac{N}{4}}} \right) \\ &\approx 1 - \Phi\left(\frac{c - \frac{N}{2}}{\sqrt{\frac{N}{4}}} \right) \\ &\approx \alpha \end{aligned}$$

and we choose c such that

$$\frac{c - \frac{N}{2}}{\sqrt{\frac{N}{4}}} = z_\alpha$$

Now if we want our power to be π_0, that is we want

$$\pi(p) = \pi_0$$

then we must choose

$$\frac{N(\frac{1}{2} - p) + z_\alpha \sqrt{\frac{N}{4}}}{\sqrt{Np(1-p)}} = z_{\pi_0} = -z_{1-\pi_0}$$

$$DIAGRAMS$$

And hence we solve N such that

$$\frac{\sqrt{N}(\frac{1}{2} - p) + \frac{z_\alpha}{2}}{\sqrt{p(1-p)}} \leq z_{\pi_0}$$

which results in

$$\begin{aligned}
\sqrt{N} &\geq \frac{\frac{z_\alpha}{2} - z_{\pi_0}\sqrt{p(1-p)}}{p - \frac{1}{2}} \\
&\geq \frac{\frac{z_\alpha}{2} + z_{1-\pi_0}\sqrt{p(1-p)}}{p - \frac{1}{2}} \\
N &\geq \frac{\left(\frac{z_\alpha}{2} + z_{1-\pi_0}\sqrt{p(1-p)}\right)^2}{(p - \frac{1}{2})^2}
\end{aligned}$$

13.2. **Sign Wilcoxon Test.** Suppose that L is symmetric about 0. Then

$$\underbrace{\mathbb{1}(Z > 0)}_{\text{Bernoulli}(1/2)}$$

is independent of $|Z|$. To prove this we would like to show that

$$\mathbb{P}\left(\mathbb{1}(Z > 0) = 1,\ |Z| \leq z\right) = \frac{1}{2}\mathbb{P}\left(|Z| \leq z\right)$$

and

$$\mathbb{P}\left(\mathbb{1}(Z > 0) = 0,\ |Z| \leq z\right) = \frac{1}{2}\mathbb{P}\left(|Z| \leq z\right)$$

$\forall z \in \mathbb{R}$.

Proof.

$$\begin{aligned}
\mathbb{P}\left(Z < 0, |Z| \leq z\right) &= \mathbb{P}\left(0 < Z < z\right) \\
&= \frac{1}{2}\mathbb{P}\left(|Z| \leq z\right) \qquad \text{by symmetry}
\end{aligned}$$

□

We have that

$$\begin{aligned}
(\mathbb{1}(Z_1 > 0), \ldots, \mathbb{1}(Z_N > 0)) &\perp\!\!\!\perp (|Z_1|, \ldots, |Z_N|) \\
&\perp\!\!\!\perp \text{the ranks of } |Z_1|, \ldots, |Z_N|
\end{aligned}$$

so the probability that we have any combination of $+$ and $-$ signs given the ranks is $\left(\frac{1}{2}\right)^N$.

$$\mathbb{P}\left((+, -, +, \ldots, -) | S_1 = s_1, \ldots, S_N = s_n\right) = \left(\frac{1}{2}\right)^N$$

13.3. **Power of Wilcoxon's Signed-rank Test.** We could like to compute

$$\pi(\Delta) = \mathbb{P}_\Delta(V_S \geq c)$$

where

$$Y_i - \Delta - X_i \sim L$$

The Mann-Whitney statistic was given as

$$W_{XY} = \sum_{i=1}^{m}\sum_{j=1}^{n} \underbrace{\mathbb{1}(Y_j > X_i)}_{\mathbb{1}(Y_j - X_i > 0)}$$

so similarly, we have

$$V_S = \sum_{j=1}^{N}\sum_{i=1}^{j} \underbrace{\mathbb{1}(Z_j + Z_i > 0)}_{\mathbb{1}(Y_j - X_j + Y_i - X_i > 0)}$$

We note that

$$\sum_{i=1}^{j} \mathbb{1}(Z_j + Z_i > 0) = \begin{cases} 0 & \text{if } Z_j < 0 \\ j & \text{if } Z_j > 0 \end{cases}$$

If we want to compute the expectation, then we have

$$\mathbb{E}[V_S] = \sum_{j=1}^{N}\sum_{i=1}^{j} \mathbb{E}[\mathbb{1}(Z_i + Z_j > 0)]$$

- if $i \neq j$

$$\mathbb{E}[\mathbb{1}(Z_i + Z_j > 0)] = \mathbb{P}(Z + Z' > 0) = q$$

- if $i = j$

$$\mathbb{E}[\mathbb{1}(Z_i + Z_i > 0)] = \mathbb{P}(Z_i > 0) = p$$

$$\begin{aligned} \mathbb{E}[V_S] &= Np + \binom{N}{2} q \\ &= Np + \frac{N(N-1)}{2} q \end{aligned}$$

To compute the variance, we write it as

$$V_S = \underbrace{\sum_{i<j} \mathbb{1}(Z_i + Z_j > 0)}_{V_S^*} + \underbrace{\sum_{i=1}^{N} \mathbb{1}(Z_i > 0)}_{S_N}$$

and then we have

$$\mathrm{Var}(V_S) = \mathrm{Var}(V_S^*) + \underbrace{\mathrm{Var}(S_N)}_{Np(1-p)} + 2\mathrm{Cov}(V_S^*, S_N)$$

Since

$$\mathrm{Var}(V_S^*) = \underbrace{\sum_{i<j} \mathrm{Var}(\mathbb{1}(Z_i + Z_j > 0))}_{\frac{N(N-1)}{2} q(1-q)} + \underbrace{\sum_{i<j}\sum_{k<l} \mathrm{Cov}(\mathbb{1}(Z_i + Z_j > 0), \mathbb{1}(Z_k + Z_l > 0))}_{A}$$

Now we need to look at A. We have to consider if any of the indices are different. If they are all different, then the covariance is 0. If there are three of them, then we have either $i = k$, $i = l$, $j = k$, or $j = l$. Hence

$$\mathbb{P}\left(Z_k + Z_j > 0, Z_k + Z_l < 0\right) - \mathbb{P}\left(Z_k + Z_j > 0\right)\mathbb{P}\left(Z_k + Z_l < 0\right) = r - q^2$$

So we have

$$\begin{aligned} A &= (r-q)^2 \left\{ \binom{N}{2}\binom{N}{2} - \binom{N}{2} - \binom{N}{2}\binom{N-2}{2} \right\} \\ &= (r - q^2)N(N-1)(N-2) \end{aligned}$$

14. October 29th, 2012

Looking at last class notes, we had

$$p = \mathbb{P}\left(Z > 0\right)$$

where $Z = Y - X$ which is not necessarily symmetric about zero.

$$q = \mathbb{P}\left(Z + Z' > 0\right)$$

where $Z \perp\!\!\!\perp Z'$ and $Z \overset{d}{=} Z'$.

$$r = \mathbb{P}\left(Z + Z' > 0, Z + Z'' > 0\right)$$

where $Z \overset{d}{=} Z' \overset{d}{=} Z''$ and $Z \perp\!\!\!\perp Z' \perp\!\!\!\perp Z''$.

Now if H_0 holds, i.e. $Y - X$ has a continuous distribution which is symmetric about zero. We have

$$p = \frac{1}{2}$$

since Z is symmetric.

$$q = \frac{1}{2}$$

also since $Z \overset{d}{=} -Z$ meaning $Z + Z' \overset{d}{=} -Z - Z'$, so $Z + Z'$ is also symmetric about zero.

$$r = \frac{1}{3}$$

We get this by doing the following calculation:

$$\begin{aligned} &\int_{-\infty}^{\infty} \mathbb{P}\left(Z + Z' > 0, Z + Z'' > 0 | Z = z\right) \mathrm{d}L(z) \\ &= \int_{-\infty}^{\infty} \mathbb{P}\left(Z' > -z, Z'' > -z\right) \mathrm{d}L(z) \\ &= \int_{-\infty}^{\infty} \mathbb{P}\left(-Z' \leq z, -Z'' \leq -z\right) \mathrm{d}L(z) \\ &= \int_{-\infty}^{\infty} \{L(z)\}^2 \, \mathrm{d}L(z) \\ &= \mathbb{E}\left[\{L(z)\}^2\right] \\ &= \mathbb{E}\left[\mathcal{U}^2\right] \\ &= \int_0^1 u^2 \, \mathrm{d}u \\ &= \frac{1}{3} \end{aligned}$$

Then we have that under H_0,

$$\begin{aligned}
\mathbb{E}[V_S] &= \frac{1}{4}N(N-1) + \frac{N}{2} = \frac{N(N+1)}{4} \\
\mathrm{Var}(V_S) &= N(N-1)(N-2)\frac{1}{12} + \frac{N}{4} + \frac{1}{2}N(N-1)\frac{3}{4} \\
&= \frac{N(2N^2 - 6N + 4 + 9N - 9 + 6)}{24} \\
&= \frac{N(2N^2 + 3N + 1)}{24} \\
&= \frac{N(2N+1)(N+1)}{24}
\end{aligned}$$

Recall:

$$\mathbb{P}\left(\underbrace{\frac{V_S - \mathbb{E}_0(V_S)}{\sqrt{\mathrm{Var}_0(V_S)}}}_{\sim\mathcal{N}(0,1)} > \frac{c - \mathbb{E}_0(V_S)}{\sqrt{\mathrm{Var}_0(V_S)}} \right) \approx 1 - \Phi\left(\frac{c - \mathbb{E}_0(V_S)}{\sqrt{\mathrm{Var}_0(V_S)}} \right)$$

so then

$$\frac{c - \mathbb{E}_0(V_S)}{\sqrt{\mathrm{Var}_0(V_S)}} \approx z_\alpha$$

so we choose c such that

$$c = z_\alpha \sqrt{\mathrm{Var}_0(V_S)} + \mathbb{E}_0(V_S)$$

So if we have H_1: $Y - \Delta - X$ is symmetric about zero (and continuous) then

$$\pi(\Delta) = \mathbb{P}_\Delta(V_S > c)$$

and we want

$$\pi(0) \approx \alpha$$

Again, we should expect this function to be increasing with respect to Δ.

We estimate this curve with the normal

$$\begin{aligned}
\mathbb{P}\left(\underbrace{\frac{V_S - \mathbb{E}_\Delta(V_S)}{\sqrt{\mathrm{Var}_\Delta(V_S)}}}_{\sim\mathcal{N}(0,1)} > \frac{c - \mathbb{E}_\Delta(V_S)}{\sqrt{\mathrm{Var}_\Delta(V_S)}} \right) &\approx 1 - \Phi\left(\frac{c - \mathbb{E}_\Delta(V_S)}{\sqrt{\mathrm{Var}_\Delta(V_S)}} \right) \\
&= \Phi\left(\frac{\mathbb{E}_\Delta(V_S) - z_\alpha\sqrt{\mathrm{Var}_0(V_S)} - \mathbb{E}_0(V_S)}{\sqrt{\mathrm{Var}_\Delta(V_S)}} \right)
\end{aligned}$$

If we have $(X, Y) \sim \mathcal{N}_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma \right)$, then

$$Y - X \sim \mathcal{N}(\underbrace{\Delta}_{=\mu_2 - \mu_1}, \overbrace{\tau^2}^{=\sigma_{11}+\sigma_{22}-2\sigma_{12}})$$

Under H_0: $Y - X \sim \mathcal{N}(0, \tau^2)$

Shift alternative H_1: $Y - X \sim \mathcal{N}(\Delta, \tau^2)$, $\Delta > 0$. We have that

$$p = \mathbb{P}_\Delta(Z > 0)$$

$$= \mathbb{P}_\Delta\left(\underbrace{\frac{Z-\Delta}{\tau}}_{\sim\mathcal{N}(0,1)} > -\frac{\Delta}{\tau}\right)$$
$$= 1 - \Phi\left(\frac{\Delta}{\tau}\right)$$
$$= \Phi\left(\frac{\Delta}{\tau}\right)$$

We also have

$$Z + Z' \sim \mathcal{N}(2\Delta, 2\tau^2)$$

so

$$q = \mathbb{P}\left(Z + Z' > 0\right)$$
$$= \Phi\left(\frac{2\Delta}{\sqrt{2}\tau}\right)$$
$$= \Phi\left(\sqrt{2}\frac{\Delta}{\tau}\right)$$

Now if $S = Z + Z'$ and $T = Z + Z''$, then

$$\begin{aligned}\mathrm{Cov}(S,T) &= \mathrm{Cov}(Z+Z', Z+Z'')\\ &= \mathrm{Var}(Z) + \mathrm{Cov}(Z,Z'') + \mathrm{Cov}(Z',Z) + \mathrm{Cov}(Z',Z'')\\ &= \mathrm{Var}(Z)\end{aligned}$$

since all the other covariances are zero by independence.

$$\begin{pmatrix} Z+Z' \\ Z+Z'' \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} Z \\ Z' \\ Z'' \end{pmatrix}}$$
$$\sim \mathcal{N}_3\left(\underbrace{\begin{pmatrix}\Delta\\\Delta\\\Delta\end{pmatrix}}_{\mu}, \underbrace{\begin{pmatrix}\tau^2 & 0 & 0\\ 0 & \tau^2 & 0\\ 0 & 0 & \tau^2\end{pmatrix}}_{\Sigma}\right)$$

and we have

$$A \cdot \mu = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}\begin{pmatrix}\Delta\\\Delta\\\Delta\end{pmatrix} = \begin{pmatrix}2\Delta\\2\Delta\end{pmatrix}$$

and

$$r = \mathbb{P}\left(Z + Z' > 0, Z + Z'' > 0\right)$$
$$= \mathbb{P}\left(S > 0, T > 0\right)$$
$$= ?$$

Let us consider the power function. If $\Delta \approx 0$, then $\mathrm{Var}_\Delta(V_S) \approx \mathrm{Var}_0(V_S)$. So then

$$\pi(\Delta) \approx \Phi\left(\frac{\mathbb{E}_\Delta(V_S) - \mathbb{E}_0(V_S)}{\sqrt{\mathrm{Var}_0(V_S)}} - z_\alpha\right)$$

We use the fact that

$$\mathbb{E}_\Delta(V_S) = \frac{N(N-1)}{2}q + Np$$

so

$$\mathbb{E}_\Delta(V_S) - \mathbb{E}_0(V_S) = \frac{N(N-1)}{2}\left\{q - \frac{1}{2}\right\} + N\left\{p - \frac{1}{2}\right\}$$

We know that

$$q = \mathbb{P}_\Delta(Z + Z' > 0)$$

Let us denote the CDF of $Z + Z'$ by L^*. Then we have

$$\begin{aligned} q &= \mathbb{P}_\Delta(Z + Z' > 0) \\ &= \mathbb{P}_\Delta(Z - \Delta + Z' - \Delta > -2\Delta) \\ &= 1 - L^*(-2\Delta) \\ &= L^*(2\Delta) \end{aligned}$$

where the last line is true as L^* is symmetric about zero. Now we note that

$$\begin{aligned} q - \frac{1}{2} &= L^*(2\Delta) - L^*(0) \\ &\approx l^*(0)2\Delta \end{aligned}$$

where $l^*(0) = (L^*)'(0)$.

Similarly, we have

$$\begin{aligned} p - \frac{1}{2} &= \mathbb{P}_\Delta(Z > 0) - \frac{1}{2} \\ &= \mathbb{P}_\Delta(Z - \Delta > -\Delta) - \frac{1}{2} \\ &= 1 - L(-\Delta) - \frac{1}{2} \\ &= L(\Delta) - L(0) \\ &\approx l(0)\Delta \end{aligned}$$

Now when we plug these in, we get

$$\begin{aligned} \mathbb{E}_\Delta(V_S) - \mathbb{E}_0(V_S) &= \frac{N(N-1)}{2}\left\{q - \frac{1}{2}\right\} + N\left\{p - \frac{1}{2}\right\} \\ &\approx \Delta\left\{\frac{N(N-1)}{2}2l^*(0) + Nl(0)\right\} \\ &= \Delta\left\{N(N-1)l^*(0) + Nl(0)\right\} \end{aligned}$$

Now suppose we want to choose an N such that $\pi(\Delta) = 1 - \beta$. We wish to choose N such that

$$\frac{N(N-1)l^*(0) + Nl(0)}{\sqrt{\frac{N(N+1)(2N+1)}{24}}}\Delta - z_\alpha \geq z_\beta$$

This can be approximated to roughly

$$\frac{N^2 l^*(0) + Nl(0)}{\sqrt{\frac{2N^3}{24}}} \geq \frac{z_\beta + z_\alpha}{\Delta} \Leftrightarrow \sqrt{12}\sqrt{N}l^*(0) + \frac{l(0)}{\sqrt{N}} \geq \frac{z_\beta + z_\alpha}{\Delta}$$

For N large, the $\frac{l(0)}{\sqrt{N}}$ is not significant, so we ignore it. So we choose N such that

$$\sqrt{12}\sqrt{N}l^*(0) \geq \frac{z_\beta + z_\alpha}{\Delta}$$

which is the same as

$$N \geq \frac{(z_\alpha + z_\beta)^2}{\Delta^2 12\{l^*(0)\}^2}$$

This is almost like the previous case, but instead of $f^*(0)$ being the derivative of F^* where F^* is the cdf of $X - X'$, we have $l^*(0)$ as the derivative of L^*, which is the cdf of $Z + Z' \stackrel{d}{=} Z - Z'$.

$$Z + Z' \sim \mathcal{N}(0, 2\tau^2)$$

so

$$l^*(t) = \frac{1}{\sqrt{2\pi}}\frac{1}{\tau\sqrt{2}}e^{-t^2/2}$$
$$l^*(0) = \frac{1}{\sqrt{2\pi}\sqrt{2}\tau}$$

If we want to compute the asymptotic power, we use the fact that

$$\frac{\sqrt{N}\bar{Z}}{S} \sim \mathcal{N}(0,1)$$

and under H_0, $Z \sim \mathcal{N}(\Delta, \tau^2)$.
Paired t-test: $\pi(\Delta) = \Phi\left(\frac{\Delta\sqrt{N}}{\tau} - z_\alpha\right)$
Note that

$$p - \frac{1}{2} = L(\Delta) - L(0) \approx l(0)\Delta$$

For the sign test, for $\Delta \approx 0$,

$$\pi(\Delta) \approx 1 - \Phi\{z_\alpha - 2\sqrt{N}\Delta l(0)\}$$

15. October 31st, 2012

15.1. **ARE of Wilcoxon Sign Test versus Paired *t*-test.**

***t*-test:**

$$N_t \geq \frac{(z_\alpha + z_\beta)^2\tau^2}{\Delta^2}$$

Wilcoxon test:

$$N_{V_S} \geq \frac{(z_\alpha + z_\beta)^2}{\Delta^2 \cdot 12\{l^*(0)\}^2}$$

where $l^*(0)$ is the density of L^*, with CDF $Z + Z'$.

$$e_{V_S,t} = \frac{\frac{(z_\alpha+z_\beta)^2\tau^2}{\Delta^2}}{\frac{(z_\alpha+z_\beta)^2}{\Delta^2\cdot 12\{l^*(0)\}^2}} = 12\tau^2\{l^*(0)\}^2$$

If $L \sim \mathcal{N}(0, \tau^2)$, then we know that L^* is a CDF of mean 0 and variance $2\tau^2$.

$$l^*(x) = \frac{1}{\sqrt{2\pi}}\frac{1}{\tau\sqrt{2}}e^{-\frac{x^2}{4\tau^2}}$$

In the case when $L \sim \mathcal{N}(0, \tau^2)$

$$e_{V_S,t} = 12\tau^2 \frac{1}{\tau^2 4\pi} = \frac{3}{\pi}$$

15.2. **Comparison between the three tests.**
Sign test:

$$\pi(\Delta) = \Phi\left(\frac{\sqrt{N}(p - \frac{1}{2}) - z_{\alpha/2}}{\sqrt{p(1-p)}}\right)$$

where

$$\begin{aligned} p &= \mathbb{P}(Z > 0) \\ &= \mathbb{P}(Z - \Delta > -\Delta) \\ &= 1 - L(-\Delta) \\ &= L(\Delta) \end{aligned}$$

When $\Delta \approx 0$, we have

$$\sqrt{p(1-p)} \approx \frac{1}{2}$$

and

$$p - \frac{1}{2} = L(\Delta) - L(0) \approx \Delta \cdot l(0)$$

This means that

$$\pi(\Delta) \approx \Phi\left(\frac{\sqrt{N} l(0) \Delta}{\frac{1}{2} - z_\alpha}\right)$$

and so we want

$$2\sqrt{N_s} l(0)\Delta - z_\alpha \geq z_\beta$$

which means we want a sample size of

$$N_s \geq \frac{(z_\alpha + z_\beta)^2}{4\Delta^2 \{l(0)\}^2}$$

15.3. **ARE of Sign Test versus Wilcoxon Test.**

$$E_{S,V_S} = \frac{\frac{(z_\alpha + z_\beta)^2}{\Delta^2 12 \{l^*(0)\}^2}}{\frac{(z_\alpha + z_\beta)^2}{\Delta^2 4 \{l(0)\}^2}} = \frac{\{l(0)\}^2}{\{3l^*(0)\}^2}$$

Note also that

$$l^*(0) = \int_{-\infty}^{\infty} \{l(t)\}^2 \, \mathrm{d}t$$

Now if we have $L \sim \mathcal{N}(0, \tau^2)$, then

$$\begin{aligned} l^*(0) &= \frac{1}{2\tau\sqrt{\pi}} \\ l(0) &= \frac{1}{\tau\sqrt{2\pi}} \end{aligned}$$

which gives us

$$e_{S,t} = \frac{2}{\pi} \qquad e_{S,V_S} = \frac{2}{3}$$

15.4. **Estimation of the Treatment Effect.** We have that our efficiencies between the three tests are given by

$$e_{V_S,t} = 12\tau^2\{l^*(0)\}^2$$
$$e_{S,t} = 4\tau^2\{l(0)\}^2$$
$$e_{S,V_S} = \frac{\{l(0)\}^2}{3\{l^*(0)\}^2}$$

Now if $Z = Y - X \sim L_\Delta$ and

$$H_0 : L \text{ symmetric about } 0$$
$$H_1 : L_\Delta = L(\bullet - \Delta) \text{ symmetric about } 0$$

Now suppose that $\mathbb{E}[Z] < \infty$. Then under H_1, $\mathbb{E}[Z] = \Delta$, $\text{med}(Z) = \Delta$, and $\text{med}\left(\frac{Z+Z'}{2}\right) = \Delta$.

If

$$\Delta^*(Z_1, \dots Z_n)$$

is symmetric about Δ if and only if

$$\Delta^*(Z_1, \dots, Z_n) - \Delta$$

is symmetric about 0.

$$\Delta^*(\underbrace{Z_1 - \Delta}_{\sim L}, \dots, \underbrace{Z_n - \Delta}_{\sim L}) \stackrel{d}{=} \Delta^*(\Delta - Z_1, \dots, \Delta - Z_n) = -\Delta^*(Z_1 - \Delta, \dots, Z_n - \Delta)$$

since we have that $\forall i$, $Z_i - \Delta \stackrel{d}{=} \Delta - Z_i$ as L is symmetric about 0.

When will be choose $\bar{\Delta}$ over $\widehat{\Delta}$? We have this when

$$\tau^2 \leq \frac{1}{12\{l^*(0)\}^2}$$

which occurs exactly when

$$12\tau^2\{l^*(0)\}^2 \leq 1$$

which is when the t-test is preferable to the Wilcoxon test!

Similarly, $\bar{\Delta}$ is better than $\tilde{\Delta}$ if

$$\tau^2 \leq \frac{1}{4\{l(0)\}^2} \Leftrightarrow e_{S,t} \leq 1$$

which occurs when the t-test is better than the Sign test.

$$\bar{\Delta} = \frac{1}{N}\sum_{i=1}^{N} Z_i$$

so we have

$$\sqrt{N}\frac{\bar{\Delta} - \Delta}{\tau} \rightsquigarrow \mathcal{N}(0, 1)$$

by the Central Limit Theorem so

$$\sqrt{N}(\bar{\Delta} - \Delta) \rightsquigarrow \mathcal{N}(0, \tau^2)$$

For

$$\tilde{\Delta} : Z_{(1)} \leq \dots \leq Z_{(N)}$$

(assume N is odd for simplicity) $\tilde{\Delta} - \Delta$ has a continuous distribution. (Assignment 2)

$$\tilde{\Delta} = Z_{\left(\frac{N+1}{2}\right)}$$

Observe: $\forall i \in \{1, \ldots, N\}$

$$\begin{aligned}
Z_{(i)} \leq a &\Leftrightarrow \text{there are at least } i \ Z_i\text{'s} \leq a \\
&\Leftrightarrow \text{there are at most } N - i \ Z_i\text{'s} > a \\
&\Leftrightarrow \#\{j : Z_j - a > 0\} \leq N - i \\
&\Leftrightarrow \sum_{j=1}^{N} \mathbb{1}(Z_j - a > 0) \leq N - i \\
&\Leftrightarrow S_n(Z - a) \leq N - i
\end{aligned}$$

so we have that $\forall x \in \mathbb{R}$

$$\mathbb{P}\left(\sqrt{N}(\tilde{\Delta} - \Delta) \leq x)\right) \to \Phi(x 2 l(0))$$

since it goes to a normal distribution with mean 0 and variance $\frac{1}{4\{l(0)\}^2}$. But we also have that the left hand side is equal to

$$\mathbb{P}_\Delta\Bigg(\underbrace{\tilde{\Delta}}_{Z_{\left(\frac{N+1}{2}\right)}} \leq \underbrace{\frac{x}{\sqrt{N}} + \Delta}_{a} \Bigg) = \mathbb{P}_\Delta\Bigg(S_N\left(z - \frac{x}{\sqrt{N}} - \Delta\right) \leq \underbrace{N - \frac{N+1}{2}}_{\frac{N-1}{2}} \Bigg)$$

where

$$\begin{aligned}
p &= \mathbb{P}_\Delta\left(Z - \Delta - \frac{x}{\sqrt{N}}\right) \\
&= \mathbb{P}_\Delta\left(Z - \Delta > \frac{x}{\sqrt{N}}\right) \\
&= 1 - L\left(\frac{x}{\sqrt{N}}\right) \\
&\approx \Phi\left(\frac{\frac{N-1}{2} + \frac{1}{2} - Np}{\sqrt{Np(1-p)}}\right) \\
&= \Phi\left(\frac{\sqrt{N}\left(\frac{1}{2} - p\right)}{\sqrt{p(1-p)}}\right) \\
&\approx \Phi(2xl(0))
\end{aligned}$$

Taking $N \to \infty$, and $p \approx 1/2$, $p(1-p) \approx 1/4$, this means

$$\begin{aligned}
\frac{1}{2} - p &= \frac{1}{2} - 1 + L\left(\frac{x}{\sqrt{N}}\right) \\
&= L\left(\frac{x}{\sqrt{N}}\right) - L(0) \\
&\approx l(0)\frac{x}{\sqrt{N}}
\end{aligned}$$

Now if

$$\widehat{\Delta} : \frac{Z_i + Z_j}{2}$$

denote these averages by A so we have

$$A_{(1)} \leq \ldots \leq A_{\left(\frac{N(N+1)}{2}\right)}$$

We have that

$$\begin{aligned} A_{(i)} \leq a &\Leftrightarrow \text{At least } i \text{ of } A_k\text{'s are} \leq a \\ &\Leftrightarrow \text{At most } \frac{N(N+1)}{2} = 2 \text{ of } A_k\text{'s satisfy} A_k > a \\ &\Leftrightarrow \text{At most } \frac{N(N+1)}{2} = 2 \text{ of } A_k\text{'s satisfy} A_k - a > 0 \\ &\Leftrightarrow \sum_{i \leq j} \mathbb{1}\left(\frac{Z_i + Z_j - 2a}{2} > 0\right) \leq \frac{N(N+1)}{2} - i \\ &\Leftrightarrow \underbrace{\sum_{i \leq j} \mathbb{1}\left(Z_i - a + Z_j - a > 0\right)}_{V_S(Z-a)} \leq \frac{N(N+1)}{2} - i \end{aligned}$$

so

$$\begin{aligned} \mathbb{P}_\Delta\left(\widehat{\Delta} \leq \frac{x}{\sqrt{N}} + \Delta\right) &\underset{\frac{N(N+1)}{2}}{=} \mathbb{P}_\Delta\left(A_{\left(\frac{N(N+1)}{4} + \frac{1}{2}\right)} \leq \frac{x}{\sqrt{N}} + \Delta\right) \\ &= \mathbb{P}_\Delta\left(V_S\left(Z - \Delta - \frac{x}{\sqrt{N}}\right) \leq \frac{N(N+1)}{4} - \frac{1}{2}\right) \end{aligned}$$

16. November 5th, 2012

16.1. **Kruskal-Wallis Test (for several treatments).** In our test, the null hypothesis H_0 is that the treatments are equivalent, so that there is no difference among the treatments. We label our observations as

$$\begin{array}{cc} X_{11}, \ldots, X_{1n_1} & \text{Group } 1 \\ X_{21}, \ldots, X_{2n_2} & \text{Group } 2 \\ \vdots & \vdots \\ X_{s1}, \ldots, X_{sn_s} & \text{Group } s \end{array}$$

We pool the observations and then rank them.

$$R_{11}, \ldots, R_{n_1}, R_{22}, \ldots, R_{2n_2}, R_{s1}, \ldots, R_{sn_s}$$

Example 7. *Assume that our ranks are*

$$\begin{array}{ll} \textit{Group } 1 & 2, 4 \\ \textit{Group } 2 & 3, 5, 7 \\ \textit{Group } 3 & 1, 6 \end{array}$$

with no ties. The number of distinct sums of the ranks in the groups is

$$\frac{7!}{2!3!2!} = \binom{7}{n_1\ n_2\ n_3}$$

16.2. **Choosing an Alternative.** Our alternative is that each of the treatments is different from one another, so we look at the average of the ranks per group. Under the null, they should be close to each other.

The Kruskal-Wallis Statistic is given by

$$K = \frac{12}{N(N+1)} \sum_{i=1}^{s} n_i \left(\bar{R}_{i\bullet} - \frac{N+1}{2} \right)^2$$

which tests H_0 against H_1 that there is a difference in location. Equivalently, we may take

$$W_i = R_{i1} + \ldots + R_{in_i}$$

so we have

$$\begin{aligned} \left(\bar{R}_{i\bullet} - \frac{N+1}{2} \right)^2 &= \left(\frac{1}{n_i} W_i - \frac{N+1}{2} \right)^2 \\ &= \frac{W_i^2}{n_i^2} - (N+1)\frac{W_i}{n_i} + \frac{(N+1)^2}{4} \end{aligned}$$

so

$$\begin{aligned} K &= \frac{12}{N(N+1)} \sum_{i=1}^{s} \frac{W_i^2}{n_i} - \frac{12(N+1)}{N(N+1)} \sum_{i=1}^{s} W_i + \frac{12(N+1)^2}{4N(N+1)} \sum_{i=1}^{s} n_i \\ &= \frac{12}{N(N+1)} \sum_{i=1}^{s} \frac{W_i^2}{n_i} + \underbrace{\frac{12(N+1)}{\cancel{N(N+1)}} \frac{\cancel{N(N+1)}}{2} + \frac{12(N+1)^{\cancel{2}}}{4\cancel{N(N+1)}}\cancel{N}}_{-3(N+1)} \\ &= \frac{12}{N(N+1)} \sum_{i=1}^{s} \frac{W_i^2}{n_i} - 3(N+1) \end{aligned}$$

16.3. **Asymptotic Approximation of the Kruskal-Wallis Statistic.** Assume that $s = 2$. Then we have

$$K = \frac{12}{N(N+1)} \left(\frac{W_1^2}{n_1} + \frac{W_2^2}{n_2} \right) - 3(N+1)$$

We know that the W_2 is redundant since

$$W_1 + W_2 = \frac{N(N+1)}{2} \Leftrightarrow W_2 = \frac{N(N+1)}{2} - W_1$$

Now we know that

$$\frac{W_1^2}{n_1} + \frac{W_2^2}{n_2} = \frac{W_1^2}{n_1} + \frac{N^2(N+1)^2}{4n_2} - \frac{N(N+1)W_1}{n_2} + \frac{W_1^2}{n_2}$$

so plugging this in, we have

$$K = \frac{12}{N(N+1)n_1n_2} \left(W_1^2 \underbrace{(n_1+n_2)}_{=N} - W_1N(N+1) - n_1 + \frac{N^2(N+1)^2n_1 - (N+1)^2Nn_1n_2}{4} \right)$$

and we have the last term is

$$\frac{N^2(N+1)^2n_1 - (N+1)^2Nn_1n_2}{4} = \frac{(N+1)^2Nn_1 \overbrace{(N-n_2)}^{n_1}}{4}$$

$$= \frac{(N+1)^2 N n_1}{4}$$

So substituting this, we get

$$\begin{aligned} K &= \underbrace{\frac{12}{(N+1)n_1 n_2}}_{\frac{1}{\mathrm{Var}(W_S)}} \left(W_1 - \underbrace{\frac{(N+1)n_1}{2}}_{\mathbb{E}[W_1]} \right)^2 \\ &= \left(\frac{W_1 - \mathbb{E}[W_1]}{\sqrt{\mathrm{Var}(W_1)}} \right)^2 \\ &\sim \chi_1^2 \end{aligned}$$

For general s,

$$\begin{aligned} K &= \sum_{i=1}^{s} \frac{12}{N(N+1)} n_i \left(\frac{W_i}{n_i} - \frac{N+1}{2} \right)^2 \\ &= \sum_{i=1}^{s} \frac{12}{N(N+1)n_i} \left(W_i - \underbrace{\frac{n_i(N+1)}{2}}_{\mathbb{E}[W_i]} \right)^2 \\ &= \frac{1}{N} \sum_{i=1}^{s} (N - n_i) \left(\frac{W_i - \mathbb{E}[W_i]}{\sqrt{\mathrm{Var}(W_i)}} \right)^2 \\ &\rightsquigarrow \chi_{s-1}^2 \end{aligned}$$

Since knowing $s-1$ of the W_i's gives you the last one, they are redundant and hence not independent.

16.4. **Dealing with Ties.** Our altered statistic is given by

$$K^* = \frac{1}{N} \sum_{i=1}^{s} (N - n_i) \left(\frac{W_i^* - \mathbb{E}[W_i]}{\sqrt{\mathrm{Var}(W_i^*)}} \right)^2$$

where

$$\mathrm{Var}(W_i^*) = \frac{(N+1)(N-n_i)n_i}{12} \left(1 - \frac{\sum_{j=1}^{l} d_j^3 - d_j}{N^3 - N} \right)$$

so

$$\begin{aligned} K^* &= \frac{1}{N} \sum_{i=1}^{s} \frac{(N - n_i) \left(W_i^* - \frac{n_i(N+1)}{2} \right)^2}{\frac{(N+1)n_i(N-n_i)}{12} \cdot \left(1 - \frac{\sum_{j=1}^{l} d_j^3 - d_j}{N^3 - N} \right)} \\ &= \left(1 - \frac{\sum_{j=1}^{l} d_j^3 - d_j}{N^3 - N} \right) K \\ &= \left(1 - \frac{\sum_{j=1}^{l} d_j^3 - d_j}{N^3 - N} \right) \left(\frac{12}{N(N+1)} \sum_{i=1}^{s} \frac{W_i^2}{n_i} - 3(N+1) \right) \end{aligned}$$

$$K^* = \frac{1}{N}\sum_{i=1}^{s} \frac{(N-n_i)\left(W_i^* - \frac{n_i(N+1)}{2}\right)^2}{\frac{(N+1)n_i(N-n_i)}{12}\cdot\left(1-\frac{\sum_{j=1}^{l} d_j^3-d_j}{N^3-N}\right)}$$
$$= \frac{\frac{12}{N(N+1)}\sum_{i=1}^{s}\frac{(W_i^*)^2}{n_i} - 3(N+1)}{\left(1-\frac{\sum_{j=1}^{l} d_j^3-d_j}{N^3-N}\right)}$$

		++	+	0	
Tylenol	A	2	1	2	5
Advil	B	3	2	0	5

We could assign 2 to ++, 1 to + and 0 to 0 and use a Wilcoxon test for this. So our values are

$$A:\ 0\ 0\ 1\ 2\ 2$$
$$B:\ 2\ 2\ 2\ 1\ 1$$

and pooling the values, we have the pooled midranks given by

Values	0	0	1	1	1	2	2	2	2	2
Mid-ranks	1.5	1.5	4	4	4	8	8	8	8	8

We have in a standard contingency table, the standard notation

	"0"	"1"	
1			$n_{\bullet 1}$
$\vdots$			$\vdots$
i	n_{1i}	n_{2i}	$n_{\bullet i}$
$\vdots$			$\vdots$
t			$n_{\bullet t}$
	$n_{1\bullet}$	$n_{2\bullet}$	$n_{\bullet\bullet}$

Then

$$R_1^* = \frac{n_{1\bullet}+1}{2} \qquad R_2^* = n_{1\bullet} + \frac{n_{2\bullet}+1}{2}$$

All this stuff I should check since it's likely I made a typo

$$\left(1-\frac{n_{1\bullet}^3 - n_{1\bullet} - n_{2\bullet}^3 - n_{2\bullet}}{n_{\bullet\bullet}^3 - n_{\bullet\bullet}}\right) = \frac{n_{\bullet\bullet}^3 - n_{1\bullet}^3 - n_{2\bullet}^3}{n_{\bullet\bullet}^3 - n_{\bullet\bullet}}$$
$$= \frac{3n_{1\bullet}n_{2\bullet}\cancel{n_{\bullet\bullet}}}{\cancel{n_{\bullet\bullet}}(n_{\bullet\bullet}^2-1)}$$

and

$$n_{\bullet\bullet}^3 = (n_{1\bullet}+n_{2\bullet})^3$$
$$= n_{1\bullet}^3 + 3n_{1\bullet}^2 n_{2\bullet} + 3n_{1\bullet}n_{2\bullet}^2 + n_{2\bullet}^3$$

The Wilcoxon statistics are given by

$$\begin{aligned}
&W_i^* - \frac{n_{\bullet i}(n_{\bullet\bullet}+1)}{2} \\
&= n_{1i}\left(\frac{n_{1\bullet}+1}{2}\right) + n_{2i}\left(n_{1\bullet} + \frac{n_{2\bullet}+1}{2}\right) \\
&= \frac{n_{1i}n_{1\bullet} + n_{1i} + 2n_{2i}n_{1\bullet} + n_{2i}n_{2\bullet} + n_{2i} - n_{\bullet i} - n_{\bullet i}n_{\bullet\bullet}}{2} \\
&= \frac{n_{1i}n_{1\bullet} + 2n_{2i}n_{1\bullet} + n_{2i}n_{2\bullet} + n_{1i}n_{2\bullet} - n_{2i}n_{1\bullet} - n_{2i}n_{2\bullet}}{2} \\
&= \frac{n_{2i}n_{1\bullet}}{2} - \frac{n_{1i}n_{2\bullet}}{2}
\end{aligned}$$

$$\begin{aligned}
\bar{K}^* &= \frac{1}{n_{\bullet\bullet}}\sum_{i=1}^{t}(n_{\bullet\bullet} - n_{\bullet i})\frac{(n_{2i}n_{1\bullet} - n_{1i}n_{2\bullet})12}{4(n_{\bullet\bullet}+1)n_{i\bullet}\frac{3n_{1\bullet}n_{2\bullet}}{n_{\bullet\bullet}^2-1}} \\
&= \frac{n_{\bullet\bullet}-1}{n_{\bullet\bullet}}\sum_{i=1}^{t}\frac{(n_{2i}n_{1\bullet} - n_{1i}n_{2\bullet})^2}{n_{1\bullet}n_{2\bullet}n_{\bullet i}}
\end{aligned}$$

17.1. **General Formula.** We have that $B_i = n_i - A_i\ (n_{\bullet i} - n_{1i})$, so

$$\begin{aligned}
W_i^* &= R_i^* \\
&= A_i\left(\frac{m+1}{2} - m - \frac{n+1}{2}\right) + n_i\left(m + \frac{n+1}{2}\right) \\
&= A_i\left(-\frac{N}{2}\right) + n_i\left(\frac{m+N+1}{2}\right)
\end{aligned}$$

and so

$$\sum\frac{(W_i^*)^2}{n_i} = \frac{N^2}{4}\sum\frac{A_i^2}{n_i} - \frac{N}{2}(m+N+1)\underbrace{\sum_{i=1}^{t}A_i}_{m} + \left(\frac{m+N+1}{2}\right)^2\underbrace{\sum_{i=1}^{t}n_i}_{N}$$

18. November 12th, 2012

18.1. **The Jonckheere Test.** Suppose we have three diets we would like to compare

Diet A	133	139	149	160	184
Diet B	111	125	143	148	157
Diet C	99	114	116	127	146

If we wish to test

H_0 : no difference between the diets
H_1 : "general"

we must first rank them

Diet A	7	8	12	14	15
Diet B	2	5	9	11	13
Diet C	1	3	4	6	10

and we look at the Kruskal-Wallis statistic given by

$$K = \frac{12}{N(N+1)} \sum_{i=1}^{3} \underbrace{n_i}_{5} \left(\bar{R}_{i\bullet} - \frac{N+1}{2} \right)^2$$

We let

$$W_{XY} = \sum_{i=1}^{n} \left\{ \mathbb{1}(Y_i > X_j) + \frac{1}{2}\mathbb{1}(Y_i = X_j) \right\}$$

For Jonckheere's test, we must assume apriori that we know the ordering, that is

$$H_1^* : \quad A > B > C$$

We look at various Mann-Whitney statistics and expect, W_{BA}, W_{CA}, and W_{CB} to be large. Our statistic is defined as

$$W = W_{BA} + W_{CA} + W_{CB}$$

W_{BA}	$2+2+4+5+5=18$
W_{CA}	$4+4+5+5+5=23$
W_{CB}	$1+3+4+5+5=18$

Now to get the expected value, we look at

$$N = \sum_{i=1}^{s} n_i$$

which gives us

$$\begin{aligned} N^2 &= \left(\sum_{i=1}^{s} n_i \right)^2 \\ &= \sum_{i=1}^{s} n_i^2 + \underbrace{\sum_{\substack{i,j=1 \\ i \neq j}}^{s} n_i n_j}_{2\sum_{i<j} n_i n_j} \end{aligned}$$

but we have

$$\frac{1}{2} \sum_{i<j} n_i n_j = \frac{(N^2 - \sum n_i^2)}{4}$$

The variance? (check this) is given by

$$\frac{1}{72} \left(N^2(2N+3) - \sum_{i=1}^{s} n_i^2 (2n_i + 3) \right)$$

18.2. **Friedman's Test.**

	1	2	3	4	5
A	3	2	3	1	2
B	2	3	1	2	3
C	1	1	2	3	1

In each column, we have $3! = 6$ ways of permuting the ranks, so in total we have $(3!)^5$ possible configurations. Now we sum the ranks.

$$R_{1\bullet} = 3+2+3+1+2 = 11$$

$$R_{2\bullet} = 2+3+1+2+3 = 11$$
$$R_{3\bullet} = 1+1+2+3+1 = 8$$

and the averages are given by

$$\bar{R}_{1\bullet} = \frac{11}{5}$$
$$\bar{R}_{2\bullet} = \frac{11}{5}$$
$$\bar{R}_{3\bullet} = \frac{8}{5}$$

However, the sum of all these ranks is different. In the Kruskal-Wallis case, we had

$$\text{K-W}: \quad \frac{N(N+1)}{2N} = \frac{N+1}{2}$$

and

$$\frac{12}{N(N+1)} \sum n_i \left(\bar{R}_{i\bullet} - \frac{N+1}{2} \right)^2$$

but for the Friedman, we have

$$\text{F}: \quad \frac{ns(s+1)}{2ns} = \frac{s+1}{2}$$

and

$$\frac{12}{N(N+1)} \sum n_i \left(\bar{R}_{i\bullet} - \frac{s+1}{2} \right)^2$$

but we have that for both these tests, these two statistics under H_0 are $\chi^2_{(s-1)}$.

19. November 14th, 2012

Suppose we would like to investigate 3 tranquilizers. Within each column, the results are not independent since they are acting on the same body.

	1	2	3	4
A	$+++$	$++$	$+++$	$+++$
B	$++$	$+++$	$+$	$+$
C	$+$	$+$	$++$	$++$

Computing the ranks, we get

	1	2	3	4	
A	3	2	3	3	$R_{1\bullet} = 11$
B	2	3	1	1	$R_{2\bullet} = 7$
C	1	1	2	2	$R_{3\bullet} = 6$

This gives us that

$$\begin{aligned} Q &= \frac{12n}{s(s+1)} \sum_{i=1}^{s} \left(\bar{R}_{i\bullet} - \frac{s+1}{2} \right)^2 \\ &= \frac{12 \cdot 4}{3 \cdot 4} \left\{ \left(\frac{11}{4} - \frac{4}{2} \right)^2 + \left(\frac{7}{4} - \frac{4}{2} \right)^2 + \left(\frac{6}{4} - \frac{4}{2} \right)^2 \right\} \end{aligned}$$

The p-value is given by $\mathbb{P}_0(Q \geq q)$, which we compute using the χ^2 approximation.

If we look at our statistic, Q^*, we have that the numerator is

$$\frac{12}{sn(s+1)}\sum_{i=1}^{s}(R_{i\bullet}^*)^2 - 3n(s+1)$$

To get the denominator we first consider each block.

$$\begin{aligned}\mathrm{Var}(R_{\bullet j}^*) &= \mathrm{Var}(R_{1j}^* + \cdots + R_{sj}^*) \\ &= \frac{s^2-1}{12} - \frac{\sum_{i=1}^{l_j}(d_{ij}^3 - d_{ij}}{12\cdot s}\end{aligned}$$

We have

$$1 - \frac{1}{ns(s^2-1)}\sum_{j=1}^{n}\left\{s(s^2-1) - \mathrm{Var}(r_{\bullet j}^*)12s\right\}$$

$$= 1 - 1 + \frac{12s}{ns(s^2-1)}\sum_{j=1}^{n}\mathrm{Var}(R_{\bullet j}^*)$$

which gives us

$$Q^* = \frac{\frac{12n}{s(s+1)}\frac{1}{n^2}\sum_{i=1}^{s}\left(R_{1\bullet}^* - \frac{(s+1)n}{2}\right)^2}{\frac{12n}{ns(s^2-1)}\sum_{j=1}^{n}\mathrm{Var}(R_{\bullet j}^*)}$$

19.1. **A special case (s=2).** We have two treatments A and B and for each subject there is a preference, so we may have ranks 1 for A and 2 for B or 1 for B and 2 for A. Let A be the number of ones in row 1. Then there are $n - A$ twos in row 1. Similarly, the number of ones in row 2 is $n - A$ and the number of twos is A.

Let L_j be the number of successes in block j. This means there are $s - L_j$ zeroes. B_i is the total number of successes for treatment i. The mid-ranks for the zeros is given by

$$\frac{1 + \cdots + (s - L_j)}{s - L_j} = \frac{s - L_j + 1}{2} = \frac{s+1}{2} - \frac{L_j}{2}$$

For the ones, the mid-ranks is

$$\begin{aligned}\frac{(s - L_j + 1) + \cdots + s}{L_j} &= \frac{L_j(s - L_j)}{L_j} + \frac{1 + \cdots + L_j}{L_j} \\ &= s - L_j + \frac{L_j + 1}{2} \\ &= s + \frac{1}{2} - \frac{L_j}{2}\end{aligned}$$

This means that

$$\begin{aligned}R_{i\bullet}^* &= \sum_{j=1}^{n} R_{ij}^* \\ &= \sum_{j:\text{“0”}}\left(\frac{s+1}{2} - \frac{L_j}{2}\right) + \sum_{j:\text{“1”}}\left(s + \frac{1}{2} - \frac{L_j}{2}\right) \\ &= \frac{s+1}{2}(n - B_i) + \left(s + \frac{1}{2}\right)B_i - \frac{1}{2}\sum_{j=1}^{n} L_j \\ &= \sum_{i=1}^{n} B_i\end{aligned}$$

The numerator is equal to

$$\frac{3s}{n(s+1)}\sum_{i=1}^{n}B_i^2-\frac{3}{n(s+1)}\left(\sum_{i=1}^{n}B_i\right)^2$$

The denominator is

$$\begin{aligned}&1-\frac{1}{ns(s^2-1)}\sum_{j=1}^{n}\sum_{i=1}^{n}(d_{ij}^3-d_{ij})\\&=1-\frac{1}{ns(s^2-1)}\sum_{j=1}^{n}\left\{(s-L_j)^3-(s-L_j)+L_j^3-L_j\right\}\end{aligned}$$

20. November 19th, 2012

20.1. **Cochran and McNemar tests.** If we look at the Q^* statistic and consider $s=2$, then the numerator becomes

$$2B_1^2+2B_2^2-(B_1+B_2)^2=(B_1-B_2)^2$$

and the denominator is

$$\sum_{j=1}^{N}L_j(2-L_j)$$

If $s=2$ and the data are dichotomous, we only need to care about the number of the pairs $(0,0),(0,1),(1,0),(1,1)$, so we can summarize this in a 2×2 table.

		Tr 1	
		0	1
Tr 2	0	A	B
	1	C	D

The numerator is equal to

$$(\underbrace{C+D}_{B_1}-\underbrace{B-D}_{B_2})=(C-B)^2$$

and the denominator is equal to

$$\begin{aligned}A\cdot 0(2-0)+(B+C)1(2-1)+D\cdot 2(2-2)&=0+B+C+0\\&=B+C\end{aligned}$$

so

$$Q^*=\frac{(C-B)^2}{B+C}$$

Here we have $B+C=N_+=k$ which is equal to the number of non-zero differences. Now suppose we have

H_0 : no difference between the treatments

H_1 : Treatment 2 is better (one-sided)

or

H_1 : Treatment is either better or worse (two-sided)

Under H_0, $B\sim\text{Bin}(k,1/2)$. In the one-sided case, the p-value is

$$\mathbb{P}\left(B\geq B_{obs}\right)$$

and in the two-sided case, the p-value is

$$\mathbb{P}\left(\left|B - \frac{k}{2}\right| \geq \left|B_{obs} - \frac{k}{2}\right|\right)$$

Another way to look at this is

$$B - \frac{B+C}{2} = \frac{2B - B - C}{2} = \frac{B-C}{2}$$

so

$$\mathbb{P}\left(\frac{|B-C|}{k} \geq \frac{|B_{obs} - C_{obs}|}{k}\right) = \mathbb{P}\left(Q^* \geq Q^*_{obs}\right)$$

20.2. **Aligned Rank Tests.** With the Wilcoxon sign-ranked test, we had the assumption that under the null, the differences $Z_1, \ldots, Z_N$ were iid and symmetric about 0, that is $(X, Y) \stackrel{d}{=} (Y, X)$. For the sign-test, we required independence, not identically distributed.

	1	$\cdots$	N
Tr 1			
$\vdots$			
Tr s			

Now we have a slightly difference exchangeability condition. We have

$$H_0 : (X_{1i}, \ldots, X_{si}) \qquad \text{iid}$$

and our exchangeability condition is given by

$$(X_{1i}, \ldots, X_{si}) \stackrel{d}{=} (X_{\pi(1)i}, \ldots, X_{\pi(s)i})$$

for any permutation $\pi(1), \ldots, \pi(s)$ of $1, \ldots, s$.

20.3. **Testing for Trends.** Let us assume that we have

$$H_0 : \; X_1 \stackrel{d}{=} X_2 \stackrel{d}{=} \cdots \stackrel{d}{=} X_N \quad \text{and are independent}$$

In the case where $N = 5$.

1	2	3	4	5
R_1	R_2	R_3	R_4	R_5

The most upward trend occurs when $R_1 = 1, \ldots R_5 = 5$. There are $N!$ possible orderings. Let us look at a baby example, when $N = 3$.

1	2	3	D
1	2	3	0
1	3	2	2
2	1	3	2
2	3	1	6
3	1	2	6
3	2	1	8

where

$$D = \sum_{i=1}^{N} (R_i - i)^2$$

and so the distribution of D is

$$\mathbb{P}(D=d)=\begin{cases}0 & \text{with prob.}\frac{1}{6}\\ 2 & \text{with prob.}\frac{2}{6}\\ 6 & \text{with prob.}\frac{2}{6}\\ 8 & \text{with prob.}\frac{1}{6}\end{cases}$$

So if we observe 3 1 2 then our p-value is given by

$$\begin{aligned}\mathbb{P}(D\leq D_{obs})&=\mathbb{P}(D\leq 6)\\ &=\frac{5}{6}\end{aligned}$$

$$\begin{aligned}\mathbb{E}[R_i]&=\sum_{i=1}^{N} i\frac{1}{N}\\ &=\frac{1}{N}\frac{N(N+1)}{2}\\ &=\frac{N+1}{2}\end{aligned}$$

The variance is given by

$$\begin{aligned}\text{Var}(D)&=4\text{Var}\left(\sum_{i=1}^{N} iR_i\right)\\ &=4\left(\sum_{i=1}^{N} i^2\text{Var}(R_i)+\sum_{i\neq j} ij\text{Cov}(R_i,R_j)\right)\end{aligned}$$

To calculate the variance, we use

$$\begin{aligned}\mathbb{E}\left[R_i^2\right]&=\sum_{i=1}^{N}\frac{1}{N}i^2\\ &=\frac{N(N+1)(2N+1)}{6N}\\ &=\frac{(N+1)(2N+1)}{6}\end{aligned}$$

So we have

$$\begin{aligned}\text{Var}(R_i)&=\frac{(N+1)(2N+1)}{6}-\frac{(N+1)^2}{4}\\ &=\frac{(N+1)(2N+1-3N-3)}{12}\\ &=\frac{(N+1)(N-1)}{12}\end{aligned}$$

To compute the covariance, we have

$$\text{Cov}(R_i,R_j)=\underbrace{\mathbb{E}[R_iR_j]}_{\sum_{i\neq j} ij\frac{1}{N(N-1)}}-\underbrace{\mathbb{E}[R_i]\,\mathbb{E}[R_j]}_{\frac{(N+1)^2}{4}}$$

$$= -\frac{N+1}{12}$$

21. November 21st, 2012

$$D = \|(1,\dots,N) - (R_1,\dots,R_N)\|_2^2$$

Spearman's Footrule

$$\sum_{i=1}^{N} |R_i - i|$$

and

$$D = \sum_{i=1}^{N} (R_i - i)^2$$

with

$$\mathbb{E}[D] = \frac{N^3 - N}{6} \qquad \mathrm{Var}(D) = \frac{N^2(N+1)^2(N-1)}{36}$$

Consider the worse case for an increasing trend, that is when the ranks are $N, N-1, \dots, 1$.

$$\begin{aligned}
D &= \sum_{i=1}^{N} ((N-i+1) - i)^2 \\
&= \sum_{i=1}^{N} ((N+1) - 2i)^2 \\
&= (N+1)^2 N - 4(N+1)\sum_{i=1}^{N} i + 4\sum_{i=1}^{N} i^2 \\
&= (N+1)^2 N - 2(N+1)^2 N + 4\frac{N(N+1)(2N+1)}{6} \\
&= \frac{2N(N+1)(2N+1)}{3} - (N+1)^2 N \\
&= \frac{N(N+1)(4N+2-3N-3)}{3} \\
&= \frac{N(N+1)(N-1)}{3}
\end{aligned}$$

so D is small if

$$\frac{N(N^2-1)}{3} - D$$

is large, which is the same if

$$D' = \frac{N(N^2-1)}{6} - \frac{1}{2}D$$

is large.

$$R_j - 1 = \sum_{i<j} \mathbb{1}(R_i < R_j) + \sum_{i>j} \mathbb{1}(R_i < R_j)$$

$$= \sum_{i<j} U_{ij} + \sum_{i>j} U_{ij}$$

Multiplying by j and summing over all j,

$$\begin{aligned}
\sum_{j=1}^{N} j(R_j - 1) &= \sum_{j=1}^{N} \sum_{i=1}^{j-1} jU_{ij} + \sum_{j=1}^{N} \sum_{i=1}^{j-1} jU_{ij} \\
&= \sum_{j=1}^{N} \sum_{i=1}^{j-1} jU_{ij} + \sum_{i=1}^{N} \sum_{j=1}^{i-1} jU_{ij} \\
&= \sum_{j=1}^{N} \sum_{i<j} jU_{ij} + i \underbrace{U_{ji}}_{1-U_{ij}}
\end{aligned}$$

So this gives us

$$\sum_{j=1}^{N} jR_j - \frac{N(N+1)}{2} = \sum_{j=1}^{N} \sum_{i<j} (j-i)U_{ij} + \sum_{j=1}^{N} \sum_{i<j} i$$

but

$$\begin{aligned}
\sum_{j=1}^{N} \sum_{i<j} i &= \sum_{i=1}^{N} \sum_{j>i} i \\
&= \sum_{i=1}^{N} (N-i)i
\end{aligned}$$

$$B = \sum_{i<j} U_{ij}$$

Its expectation is given by

$$\begin{aligned}
\mathbb{E}[B] &= \sum_{i<j} \mathbb{E}[U_{ij}] \\
&= \sum_{i<j} \mathbb{E}[\mathbb{1}(X_i < X_j)] \\
&= \binom{N}{2} \mathbb{P}(X_1 < X_2) \\
&= \frac{N(N-1)}{4}
\end{aligned}$$

The variance is given by

$$\mathrm{Var}(B) = \sum_{i<j} \mathrm{Var}(\mathbb{1}(X_i < X_j)) + \sum_{\substack{i<j,k<l \\ (i,j)\neq(k,l)}} \mathrm{Cov}(\mathbb{1}(X_i < X_j), \mathbb{1}(X_k < X_l))$$

21.1. **Tests of Independence.**

$$\begin{aligned} B &= \sum_{i<j} \mathbb{1}(S_i < S_j) \\ &= \sum_{i<j} \mathbb{1}(S_i < S_j)\mathbb{1}(i < j) \\ &= \sum_{\substack{\text{all pairs} \\ (R_i,S_i),(R_j,S_j),i\neq j}} \mathbb{1}(S_i < S_j)\mathbb{1}(R_i < R_j) + \mathbb{1}(S_j < S_i)\mathbb{1}(R_j < R_i) \end{aligned}$$

Given (a_1, b_1) and (a_2, b_2), we say they are concordant if $a_1 < a_2$ and $b_1 < b_2$ or $a_2 < a_1$ and $b_2 < b_1$. They are discordant if $a_1 < a_2$ and $b_1 > b_2$ or $a_2 < a_1$ and $b_1 < b_2$.

$$\begin{aligned} \tau_N &= \text{Kendall's tau} \\ &= \frac{\sum_{\text{all pairs of points}} \mathbb{1}(\text{concordant}) - \mathbb{1}(\text{discordant})}{\binom{N}{2}} \\ &= \frac{B - C}{\binom{N}{2}} \\ &= \frac{B - \binom{N}{2} + B}{\binom{N}{2}} \end{aligned}$$

22. November 26th, 2012

$$D = \sum_{i=1}^{n} (R_i - S_i)^2$$

We have that

$$\mathbb{E}[D] = \frac{N^3 - N}{6} \qquad \text{Var}(D) = \frac{N^2(N+1)^2(N-1)}{36}$$

if $X \perp\!\!\!\perp Y$.

$$\begin{aligned} B &= \sum_{i<j} U_{ij} \\ &= \sum_{i<j} \mathbb{1}((R_i - R_j)(S_i - S_j) > 0) \\ &= \text{number of concordant pairs} \end{aligned}$$

Under $X \perp\!\!\!\perp Y$,

$$\mathbb{E}[B] = \frac{N(N-1)}{4} \qquad \text{Var}(B) = \frac{2N^3 - 3N^2 - 5N}{72}$$

$$\tau_N = \frac{2B}{\binom{N}{2}} - 1 = \frac{4B}{N(N-1)} - 1$$

Under the hypothesis of independent, we have

$$\mathbb{E}[\tau_N] = \frac{4}{N(N-1)}\mathbb{E}[B] - 1 = 0$$

and

$$\begin{aligned}\operatorname{Var}(\tau_N) &= \frac{16 \cdot \operatorname{Var}(B)}{N^2(N-1)^2} \\ &= \frac{16N(2N^2+3N-5)}{72N^2(N-1)^2} \\ &= \frac{2(2N+5)(N-1)}{9N(N-1)^2} \\ &= \frac{2(2N+5)}{9N(N-1)}\end{aligned}$$

To compare D to ρ_N, we note

$$\begin{aligned}\sum_{i=1}^{N}(R_i-\bar{R})^2 &= \sum_{i=1}^{N}\left(i-\frac{N+1}{2}\right)^2 \\ &= \frac{N^3-N}{12}\end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^{N} R_i S_i &= \frac{N(N+1)(2N+1)}{6} - \frac{D}{2} \\ S_N &= \frac{\frac{N(N+1)(2N+1)}{6} - \frac{D}{2} - \frac{N(N+1)^2}{4}}{\frac{N^3-N}{12}} \\ &= \frac{-6D}{N^3-N} + \frac{\frac{N(N+1)(4N+2-3N-3)}{12}}{\frac{N(N-1)(N+1)}{12}} \\ &= 1 - \frac{6D}{N^3-N}\end{aligned}$$

If we have $(X_1, Y_1) \perp\!\!\!\perp (X_2, Y_2) \sim H$, then

$$\mathbb{1}((X_1-X_2)(Y_1-Y_2)>0) = \begin{cases}1 & \text{concordance} \\ 0 & \text{discordance}\end{cases}$$

and

$$\mathbb{1}((X_1-X_2)(Y_1-Y_2)<0) = \begin{cases}1 & \text{discordance} \\ 0 & \text{concordance}\end{cases}$$

so

$$\begin{aligned}\tau &= 2\mathbb{P}\left(\text{concordance}\right) - 1 \\ &= \mathbb{P}\left(\text{concordance}\right) - \left(1 - \mathbb{P}\left(\text{concordance}\right)\right) \\ &= \mathbb{P}\left(\text{concordance}\right) - \mathbb{P}\left(\text{discordance}\right)\end{aligned}$$

To get the probability of concordance, we have

$$\begin{aligned}\mathbb{P}\left(\text{concordance}\right) &= \mathbb{P}\left(X_1 < X_2, Y_1 < Y_2\right) + \mathbb{P}\left(X_2 < X_1, Y_2 < Y_1\right) \\ &= 2\mathbb{P}\left(X_1 < X_2, Y_1 < Y_2\right) \\ &= 2\int H(x_2, y_2)\,\mathrm{d}H(x_2, y_2)\end{aligned}$$

When looking at Spearman's rho, to calculate I_n, we use the fact that

$$R_i = \sum_{j=1}^{N} \mathbb{1}(X_k \leq X_i)$$
$$S_i = \sum_{k=1}^{N} \mathbb{1}(Y_k \leq Y_i)$$

The value $\mathbb{P}(X_j \leq X_i, Y_j \leq Y_j)$ is one half of the probability of concordance.

$$\mathbb{E}[\rho_N] = \frac{12}{N(N+1)(N-1)} \Bigg\{ N(N-1)(N-2)\frac{\rho+3}{12} + 2N(N-1)\frac{1}{2} + N(N-1)\frac{\tau+1}{4} + N \Bigg\} - 3\frac{N+1}{N-1}$$

23. November 28th, 2012

The Pearson test is not good to use when normality does not hold. A good indication of this is when there are outliers.

$$\begin{aligned} \rho - &-3 + 12\mathbb{E}[F(X)G(Y)] \\ &= \operatorname{Corr}(F(X), G(Y)) \end{aligned}$$

and

$$\tau = -1 + 4\mathbb{E}[H(X,Y)]$$

where H is the underlying CDF of (X_i, Y_i), $(X,Y) \sim H$. In a bivariate CDF H of (X,Y),

$$F(x) = \mathbb{P}(X \leq x) = \lim_{y \to \infty} H(x,y)$$

We have that

$$H(x,y) = C_\theta(F(x), G(y))$$

where C_θ is some CDF on $[0,1]^2$ with $\mathcal{U}(0,1)$ margins. C_θ is called a "Copula".

Theorem 2 (Sklar's Representation Theorem)**.** *Let (X,Y) be a random pair with CDF H. Then there exists at least one copula C such that*

$$H(x,y) = C(F(x), G(y)) \quad x, y \in \mathbb{R}$$

where $F(x)$ and $G(y)$ are the margins of X and Y. If F and G are continuous, then C is unique and it happens to be the joint distribution function of $(F(X), G(Y))$.

$$\rho = -3 + 12 \int_0^1 \int_0^1 uv \, \mathrm{d}C(u,v)$$

and we would like to apply

$$W(u,v) \leq C(u,v) \leq M(u,v)$$

which are the Fréchet Hoeffding lower and upper bounds.

If we take $(U, V) \sim C$, where $U^* \sim \mathcal{U}(0,1)$, $V^* \sim \mathcal{U}(0,1)$ and $(U, V) \perp\!\!\!\perp U^* \perp\!\!\!\perp V^*$

$$\mathbb{P}(U^* \leq u, V^* \leq v) = \int_0^1 \int_0^1 uv \, \mathrm{d}C(u, v)$$

We see that

$$\begin{aligned}
\mathbb{P}(U^* \leq U, V^* \leq V) &= \mathbb{P}(U^* \leq U) + \mathbb{P}(U^* \leq U, V \leq V^*) \\
&= \underbrace{\mathbb{P}(U^* \leq U)}_{=1/2} - \underbrace{\mathbb{P}(V \leq V^*)}_{=1/2} + \mathbb{P}(U \leq U^*, V \leq V^*) \\
&= \mathbb{P}(U \leq U^*, V \leq V^*)
\end{aligned}$$

and so

$$\mathbb{P}(U \leq U^*, V \leq V^*) = \int_0^1 \int_0^1 C(u^*, v^*) \, \mathrm{d}u^* \, \mathrm{d}v^*$$

so we have that

$$\begin{aligned}
\rho &= -3 + 12 \int_0^1 \int_0^1 C(u, v) \, \mathrm{d}u \, \mathrm{d}v \\
&\leq -3 + 12 \int_0^1 \int_0^1 M(u, v) \, \mathrm{d}u \, \mathrm{d}v
\end{aligned}$$

and

$$\rho \geq -3 + 12 \int_0^1 \int_0^1 W(u, v) \, \mathrm{d}u \, \mathrm{d}v$$

Now for

$$\rho = \mathrm{Corr}(\underbrace{F(X)}_{U}, \underbrace{G(Y)}_{V})$$

if $U = V$, then $\mathrm{Corr}(U, U) = 1$ and if $V = 1 - U$, we have $\mathrm{Corr}(U, 1 - U) = -1$.

Consider $(X_1, Y_1), \ldots, (X_n, Y_n)$. If we knew F and G we could plot

$$(F(X_i), G(Y_i) \quad 1 \leq i \leq n$$

If F and G are unknown, plot

$$(F_n(X_i), G_n(Y_i))$$

instead

$$F_n(X_i) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{X_j \leq X_i} = \frac{R_i}{n}$$

Similarly,

$$G_n(X_i) = \frac{S_i}{n}$$

So we have that

$$(F_n(X_i), G_n(Y_i)) \approx \left(\frac{R_i}{n+1}, \frac{S_i}{n+1} \right)$$

24. December 3rd, 2012

Final:

- Chapter 1
- Chapter 2 (subsections 1-5)
- Chapter 3 (subsections 1-3)
- Chapter 4 (subsections 1-4)
- Chapter 5 (subsections 1-5)
- Chapter 6 (subsections 1-3)
- Chapter 7 (subsections 1-3)

Everything on copulas will **NOT** be on the final.

Sklar's Decomposition

$$H(x,y) = C(F(x), G(y))$$

C is a copula, which is a CDF with standard uniform margins. If F and G are continuous, then C is unique and it is the CDF of $(F(X), G(Y))$.

We have $(X_1, Y_1), \ldots, (X_n, Y_n)$ are from H.

$$X \perp\!\!\!\perp Y \Leftrightarrow C(u,v) = uv$$

If we knew the margins, we could look at

$$(F_n(X_1), G_n(Y_1)), \ldots, (F_n(X_n), G_n(Y_n))$$

We deduced last class that

$$F_n(X_i) = \frac{R_i}{n} \qquad G_n(X_i) = \frac{S_i}{n}$$

$$\tau = -1 + 4\int_0^1 \int_0^1 C(u,v)\,\mathrm{d}C(u,v)$$

$$\rho = -3 + 12\int_0^1 \int_0^1 C(u,v)\,\mathrm{d}u\,\mathrm{d}v$$

and

$$\sum_{i=1}^{n} \frac{R_i}{n+1}\frac{S_i}{n+1} = \sum_{i=1}^{n} J\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}\right)$$

where $J(u,v) = uv$. We have that our statistic is

$$\begin{aligned} V_n &= \sum_{i=1}^{N} \Phi^{-1}\left(\frac{R_i}{N+1}\right)\Phi^{-1}\left(\frac{S_i}{N+1}\right) \\ &= \sum_{i=1}^{N} \Phi^{-1}\left(\frac{i}{N+1}\right)\Phi^{-1}\left(\frac{Q_i}{N+1}\right) \end{aligned}$$

and

$$\mathbb{E}\left[V_n\right] = \sum_{i=1}^{N} \Phi^{-1}\left(\frac{i}{N+1}\right)\mathbb{E}\left[\Phi^{-1}\left(\frac{Q_i}{N+1}\right)\right]$$

For the FGM copula,

$$c_\theta(u,v) = 1 + \theta(1-2u)(1-2v)$$

and

$$\left.\frac{\partial c_\theta}{\partial \theta}\right|_{\theta=0} = (1-2u)(1-2v)$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(1 - 2\frac{R_i}{N+1} - 2\frac{S_i}{N+1} + 4\frac{R_i}{N+1}\frac{S_i}{N+1}\right)$$

but the middle two terms are irrelevant if we have no ties since the sum of the R_i and the sum of the S_i is constant given N. Hence this is equivalent to spearman's rho.

$$\left.\frac{\partial}{\partial \theta}\log c_\theta(u,v)\right|_{\theta=\theta_0} = \left.\frac{\frac{\partial}{\partial\theta}c_\theta(u,v)}{c_\theta(u,v)}\right|_{\theta=\theta_0}$$

We now want to find a statistic such that

$$S = 0 \Leftrightarrow X \perp\!\!\!\perp Y \text{ (no ties)}$$

We have $H_0 : C(u,v) = uv$ and we can test this by

$$B_n = \int_0^1\int_0^1 (C_n(u,v) - uv)^2 \,\mathrm{d}u\,\mathrm{d}v$$

where C_n is the empirical CDF of $\left(\frac{R_i}{N+1}, \frac{S_i}{N+1}\right)$. We could also look at

$$T_n = \sup_{u,v\in[0,1]} |C_n(u,v) - uv|$$

$$n\int_0^1\int_0^1 (C_n(u,v) - uv)^2 \,\mathrm{d}u\,\mathrm{d}v \rightsquigarrow \int_0^1\int_0^1 \{C(u,v)\}^2 \,\mathrm{d}u\,\mathrm{d}v$$

- Replicates of B_N under H_0.

$$B^*_{N,1}, \ldots, B^*_{N,M}$$

- Compute B_N from the sample

$$\mathbb{P}\left(B_N \geq B_{N,obs}\right) \approx \frac{1}{m}\sum_{i=1}^{M} \mathbb{1}(B^*_{N,i} \geq B_{N,obs})$$