MATH 503: Mathematical Statistics

Lecture 11: Nonparametric Tests

Reading: HMC Sections 10.2-10.4

Kimberly F. Sellers

Department of Mathematics and Statistics

What is Nonparametric Statistics?

- Model structure not specified a priori, but determined from data
- Number and nature of parameters are flexible and not fixed in advance
- Also called distribution free.
- Histogram: simple nonparametric probability distribution estimate

Today's Topics

- Sign Test
- Signed-Rank Wilcoxon Test
- Mann-Whitney-Wilcoxon Test
- Associated CIs for parameter of interest

Sign Test

- Denote $\theta =$ median
- Let $X_1, X_2, \dots, X_n$ random sample where $X_i = \theta + \epsilon_i$, ϵ_i's iid with cdf $F(x)$, pdf $f(x)$, median 0
- Consider $H_0: \theta = \theta_0$ vs $H_1: \theta > \theta_0$ and statistic,

  $$S = S(\theta_0) = \#\{X_i > \theta_0\} = \sum_{i=1}^{n} I(X_i > \theta_0)$$

  (called sign statistic)
- What do we expect if H_0 is true? If H_1 is true?

Let n = sample size.

H_0 true $\Rightarrow S(\theta_0) = n/2$; H_1 true $\Rightarrow S(\theta_0) > n/2$

Sign Test (cont.)

- Decision rule: Reject H_0 if $S \geq c$
- Under H_0, $S \sim$ Binomial(n, ½). Why?
  ① 2 outcomes : $X_i > \theta_0$ or $X_i \leq \theta_0$ $\forall i$
  ② indpt. events : X_i iid because random sample
  ③ common success probability : ½ because $\theta = \theta_0$ under H_0
- Level α test: find c s.t. $P_{H_0}(S \geq c) = \alpha$
  - For n small, exact Binomial test
  - For n large, use Central Limit Theorem

Example 1

DuBois (1960) conducted a study of the Shoshoni beaded baskets to see if the beaded rectangles contained within are "golden rectangles" (i.e. having a width-to-length ratio approximately equal to 0.618). Let X denote the ratio of width to length of a Shoshoni beaded basket, with sample size $n = 20$. The data are contained in **shoshoni.txt** on Canvas.

How do we proceed here?

The Data

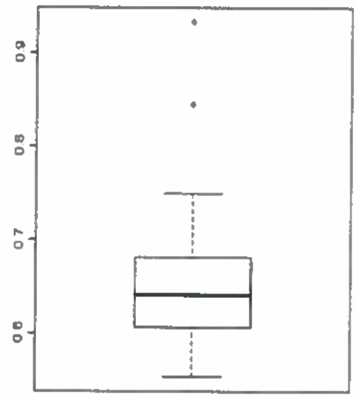
```
> stem(shoshoni$ratio)

  The decimal point is 1 digit(s) to the left of the |

  5 | 578
  6 | 01111135677799
  7 | 5
  8 | 4
  9 | 3
```

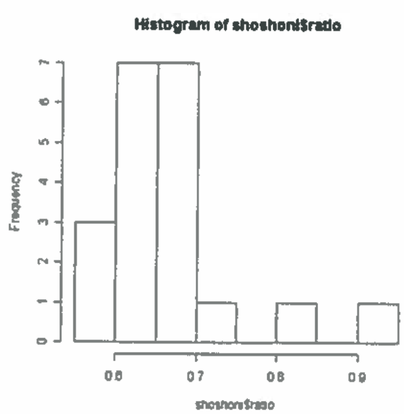
```
> summary(shoshoni$ratio)
  Min.  1st Qu.  Median   Mean 3rd Qu.   Max.
0.5530  0.6060  0.6410  0.6605  0.6765  0.9330
```

> boxplot(shoshoni$ratio)

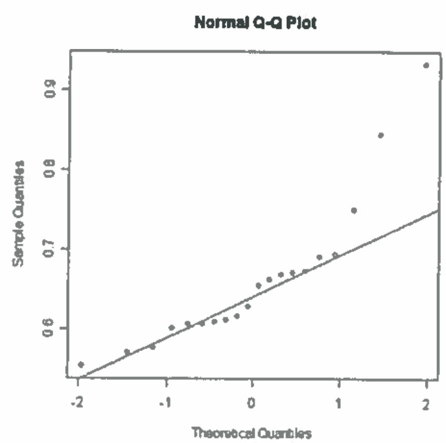


The Data (cont.)

> hist(shoshoni$ratio)



> qqline(shoshoni$ratio)



Implication: use nonparametric test, e.g. sign test

any of these exploratory data analytic approaches show that we don't have normal data. Thus, consider nonparametric statistics in lieu of classical hypothesis testing

Example 1 (cont.): The Test

- Consider hypothesis

  $H_0: \theta = 0.618$ vs $H_1: \theta \neq 0.618$
- Determine $S(\theta_0) = \#\{X_i > \theta_0\}$
- Decision rule: reject H_0 if $S(\theta_0) \leq c$ or $S(\theta_0) \geq n - c$, where c determined s.t.

  $$P(S(\theta_0) \leq c) = \frac{\alpha}{2}$$
- Using R with the command "qbinom(.025,20,.5)-1", $c = 5$

  $\overset{\shortparallel}{\frac{\alpha}{2}}$ $\overset{\shortparallel}{n}$ $\overset{\shortparallel}{p}$

because H_1 implies a two-sided test

Assume $\alpha = 0.05$

0.553
0.570
0.576
0.601
0.606
0.606
0.609
0.611
0.615
0.628
0.654
0.662
0.668
0.670
0.672
0.690
0.693
0.749
0.844
0.933

$\theta_0 = 0.618$

$S(\theta_0) = 11$ is not in the rejection region $\therefore$ fail to reject H_0

$n = 20$

$c = 5$

$\therefore n - c = 15$

Lemma 1

- Consider $H_0: \theta = \theta_0$ vs. $H_1: \theta > \theta_0$
- For every k, $P_\theta[S(0) \geq k] = P_0[S(-\theta) \geq k]$
  - $P_\theta[S(0) \geq k] = P_\theta[\#\{X_i > 0\} \geq k]$, X_i has median θ
  - $P_0[S(-\theta) \geq k] = P_0[\#\{X_i + \theta > 0\} \geq k]$, $X_i + \theta$ has median θ
- **Implication**: the power function of the sign test is monotone for one-sided tests

Formal thm on next slide

$\boxed{\text{Pf}}$ Without loss of generality (Wolog), let $\theta_0 = 0$, and let $\theta_1 < \theta_2$.
Show $\gamma(\theta_1) \leq \gamma(\theta_2)$.

$\theta_1 < \theta_2 \Rightarrow -\theta_1 > -\theta_2$, and $S(\theta_1) > S(\theta_2) \Rightarrow S(-\theta_1) < S(-\theta_2)$

$\gamma(\theta_1) = \mathbb{P}_{\theta_1}(S(0) > c_\alpha) = \mathbb{P}_0(S(-\theta_1) > c_\alpha)$ by Lemma 1

$\leq \mathbb{P}_0(S(-\theta_2) > c_\alpha)$ because $S(-\theta_1) < S(-\theta_2)$

$= \mathbb{P}_{\theta_2}(S(0) > c_\alpha)$ by Lemma 1

$= \gamma(\theta_2)$

Theorem 1

- Suppose model $X_i = \theta + \epsilon_i$ is true. Let $\gamma(\theta)$ be the power function of the sign test of level α for the hypotheses
$$H_0: \theta = \theta_0 \text{ vs. } H_1: \theta > \theta_0$$
Then $\gamma(\theta)$ is a nondecreasing function of θ.

- Implication: can extend decision rule to composite hypothesis, $H_0: \theta \leq \theta_0$ vs $H_1: \theta > \theta_0$

CI for the Median

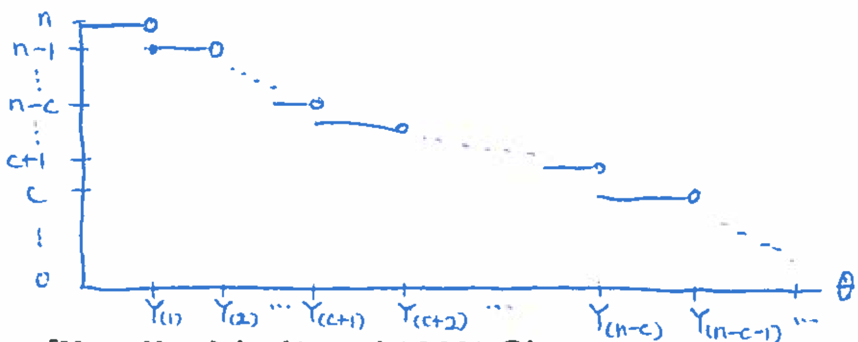
- Recall decision rule for two-sided test: reject H_0 if $S(\theta_0) \leq c$ or $S(\theta_0) \geq n - c$, where c determined s.t.
$$P(S(\theta_0) \leq c) = \alpha/2$$

- Confidence interval:
$$P(c < S(\theta) < n - c) = 1 - \alpha$$

- How do we "invert" this?

CI for the Median (cont.)

- Think about order statistics!



- $[Y_{c+1}, Y_{n-c})$ is $(1-\alpha)100\%$ CI
- Large sample approximation exists using CLT st.

$$c = \frac{n}{2} - \frac{z_{\alpha/2}\sqrt{n}}{2} - \frac{1}{2}$$

CI for the Median (cont.)

Derive the approximation, $c = \frac{n}{2} - \frac{z_{\alpha/2}\sqrt{n}}{2} - \frac{1}{2}$

Under H_0, $S(\theta_0) \sim Bin(n, \frac{1}{2}) \approx N\left(\mu = \frac{n}{2}, \sigma^2 = \frac{n}{4}\right)$

$\therefore \frac{\alpha}{2} = P(S(\theta_0) \le c) \approx P(S(\theta_0) \le c + \frac{1}{2})$ by continuity correction for normal approximation

$$= P\left(Z \le \frac{c + \frac{1}{2} - \frac{n}{2}}{\sqrt{n}/2}\right) = P\left(Z \le \underbrace{\frac{c - \left(\frac{n-1}{2}\right)}{\sqrt{n}/2}}_{-z_{\alpha/2}}\right)$$

$$\Rightarrow -z_{\alpha/2} = \frac{c - \frac{(n-1)}{2}}{\sqrt{n}/2} \Rightarrow c = -z_{\alpha/2}\frac{\sqrt{n}}{2} + \frac{n-1}{2}$$

Example 1 (cont.)

- Recall $H_0: \theta = 0.618$ vs. $H_1: \theta \neq 0.618$
- $n = 20$
- What is the sample median?
- $P_{H_0}(S \leq 5) = 0.021 \Rightarrow c = 5$

Notice: $c=6$ is too large (probability equals .058)

```
> pbinom(0:20,20,.5)
 [1] 9.536743e-07 2.002716e-05 2.012253e-04 1.288414e-03 5.908966e-03
 [6] 2.069473e-02 5.765915e-02 1.315880e-01 2.517223e-01 4.119015e-01
[11] 5.880985e-01 7.482777e-01 8.684120e-01 9.423409e-01 9.793053e-01
[16] 9.940910e-01 9.987116e-01 9.997988e-01 9.999800e-01 9.999990e-01
[21] 1.000000e+00
```

- $[Y_6, Y_{15}) = [0.606, 0.672)$ is 95.8% CI interval for θ

$[Y_{c+1}, Y_{n-c})$

- What do you conclude?

CI contains $\theta_0 = 0.618 \therefore$ fail to reject H_0

Signed-Rank Wilcoxon Test

- More efficient than sign test
- Let $X_1, X_2, \ldots, X_n$ random sample where $X_i = \theta + \epsilon_i$, where ϵ_i's iid with cdf $F(x)$, pdf $f(x)$, median 0
- Added assumption: let $f(x)$ be symmetric

Signed-Rank Wilcoxon Test (cont.)

- Consider $H_0: \theta = \theta_0$ vs $H_1: \theta > \theta_0$
- Test statistic:

$$T = \sum_{i=1}^{n} \operatorname{sgn}(X_i) R|X_i|$$

where $R|X_i|$ is rank of X_i among $|X_1|, \ldots, |X_n|$

- Decision rule: reject H_0 if $T \geq c$, where c determined for level α test

Theorem 2

Assume the model $X_i = \theta + \epsilon_i$, where ϵ_i's iid with cdf $F(x)$, pdf $f(x)$, median 0 is true for the random sample $X_1, \ldots, X_n$. Assume also that the pdf $f(x)$ is symmetric about 0. Then, under H_0,

- T is distribution free with a symmetric pdf
- $E_{H_0}(T) = 0$
- $\operatorname{Var}_{H_0}(T) = \frac{n(n+1)(2n+1)}{6}$
- $\frac{T}{\sqrt{\operatorname{Var}_{H_0}(T)}}$ has an asymptotically N(0,1) distribution

Notes

- Refer to applied nonparametric books, statistical software for exact T distribution
- Normal approximation is reasonable for $n \geq 10$
- Power function associated with signed-rank Wilcoxon test is nondecreasing wrt θ

Another Representation

- Note: sum of all ranks $= \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$
- $T = \sum_{i=1}^{n} \operatorname{sgn}(X_i) R|X_i| = \sum_{X_i>0} R|X_i| - \sum_{X_i<0} R|X_i|$

where $\sum_{i=1}^{n} R|X_i| = \sum_{X_i>0} R|X_i| + \sum_{X_i<0} R|X_i|$

$$\therefore T = \sum_{X_i>0} R|X_i| - \left(\sum_{i=1}^{n} R|X_i| - \sum_{X_i>0} R|X_i|\right)$$

$$= 2\underbrace{\sum_{X_i>0} R|X_i|}_{T^+} - \underbrace{\sum_{i=1}^{n} R|X_i|}_{= \sum_{i=1}^{n} i = \frac{n(n+1)}{2}}$$

$$\Rightarrow T = 2T^+ - \frac{n(n+1)}{2}$$

Another Representation

$\therefore T^+$ is a linear function of signed-rank test T.
What are $E_{H_0}(T^+)$ and $\text{Var}_{H_0}(T^+)$?

Recall: $T = 2T^+ - \frac{n(n+1)}{2} \Rightarrow T^+ = \frac{1}{2}\left(T + \frac{n(n+1)}{2}\right) = \frac{1}{2}T + \frac{n(n+1)}{4}$

$$\mathbb{E}_{H_0}(T^+) = \mathbb{E}_{H_0}\left(\frac{1}{2}T + \frac{n(n+1)}{4}\right) = \frac{1}{2}\cancel{\mathbb{E}_{H_0}(T)}^{0} + \frac{n(n+1)}{4} = \frac{n(n+1)}{4}$$

$$\text{Var}_{H_0}(T^+) = \text{Var}_{H_0}\left(\frac{1}{2}\left\{T + \frac{n(n+1)}{2}\right\}\right) = \frac{1}{4}\text{Var}_{H_0}\left(T + \frac{n(n+1)}{2}\right) = \frac{1}{4}\text{Var}_{H_0}(T)$$

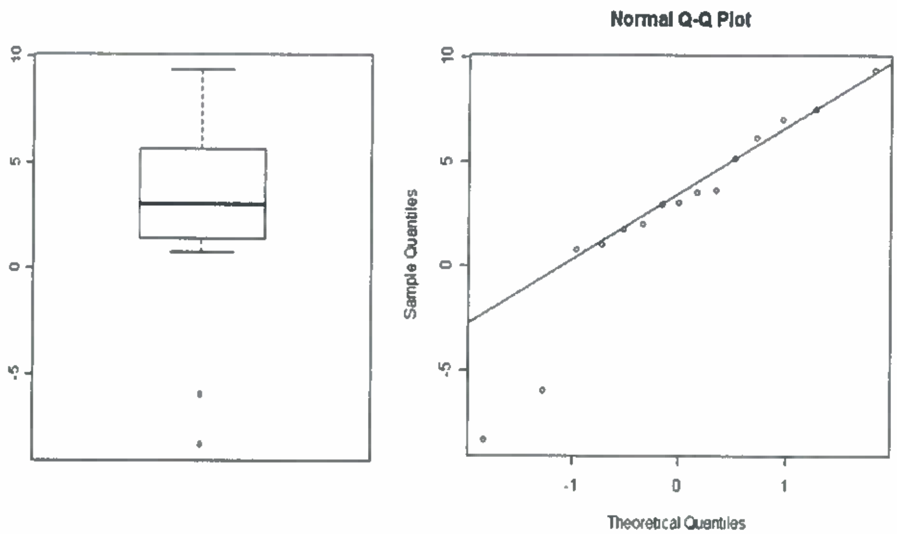
$$= \frac{1}{4}\left(\frac{n(n+1)(2n+1)}{6}\right)$$

$$= \frac{n(n+1)(2n+1)}{24}$$

Example 2

- Darwin (1878) recorded data on the heights of zea mays plants to determine what effect cross-fertilized or self-fertilized had on the height of zea mays. It is hypothesized that the cross-fertilized plants are generally taller than the self-fertilized plants. The data is provided in **zeamays.txt** in Canvas.
- $n = 15$ pots recorded
- (X_i, Y_i), $i = 1, \ldots, 15$ are heights of cross-fertilized and self-fertilized plants, respectively, in ith pot
- $W_i = X_i - Y_i$
- Which model is more appropriate? Parametric or nonparametric?

The Data



EDA shows non-normal data $\therefore$ consider non-parametric approach to model data & analyze

Example 2 (cont.)

- Consider nonparametric model: $W_i = \theta + \epsilon_i$, ϵ_i's iid with cdf $F(x)$, symmetric pdf $f(x)$, median 0
- Consider $H_0: \theta = 0$ vs. $H_1: \theta > 0$

W	Signed-Ranks
6.125	11
-8.375	-14
1.000	2
2.000	4
0.750	1
2.925	5
3.500	7
5.125	9
1.750	3
3.625	8
7.000	12
3.000	6
9.375	15
7.500	13
-6.000	-10

$T^+ \equiv \sum_{X_i > 0} R|X_i| = 96$

CLT applies because $n = 15 > 10$

$n = 15 > 10$

$E(T^+) = \frac{15(16)}{4} = 60$

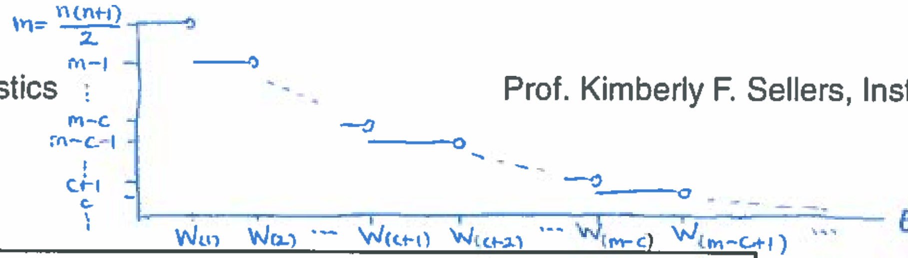
$V(T^+) = \frac{15(16)(31)}{24} = 310$

p-val $= \mathbb{P}(T^+ \geq 96)$

$\approx \mathbb{P}(T^+ \geq 95.5)$ by continuity correction

$= \mathbb{P}\left(Z \geq \frac{95.5 - 60}{\sqrt{310}}\right) = \mathbb{P}(Z \geq 2.016) = .022$

p-val $< \alpha = 0.05$ $\therefore$ reject H_0 at 5% significance level, i.e. the median of cross-fertilized plants is statistically significantly greater than that of the self-fertilized plants.



CI for the Median

- $T^+ = \#_{i \leq j}\{(X_i + X_j)/2 > 0\}$
- $W = (X_i + X_j)/2$ called Walsh averages
- $1 - \alpha = P_\theta[c_W < T^+(\theta) < m - c_W]$

  $= P_\theta[W_{c_W+1} \leq \theta < W_{m-c_W}]$, where $m = \frac{n(n+1)}{2}$
- $[W_{c_W+1}, W_{m-c_W})$ is the $(1-\alpha)100\%$ CI
- Large sample approximation exists using CLT st.

$$c_W = \frac{n(n+1)}{4} - z_{\alpha/2}\sqrt{\frac{n(n+1)(2n+1)}{24}} - \frac{1}{2}$$

Mann-Whitney-Wilcoxon Procedure

- Suppose you have two random samples:

  $X_i, i = 1, \ldots, n_1$ with continuous cdf $F(x)$, pdf $f(x)$

  $Y_j, j = 1, \ldots, n_2$ with continuous cdf $G(x)$, pdf $g(x)$
- Do the samples come from the same distribution or not?

  H_0: $F(x) = G(x)\ \forall x$

  vs. H_1: $G(x) \geq F(x)\ \forall x$, and $G(x) > F(x)$ for some x
- Note: H_1 defines X stochastically greater than Y

$X \overset{st}{>} Y \Rightarrow \mathbb{P}(X>t) \geq \mathbb{P}(Y \geq t)\ \forall t$ and $\mathbb{P}(X>t) > \mathbb{P}(Y>t)$ for some t

$\Rightarrow F(t) \leq G(t)\ \forall t$ and $F(t) < G(t)$ for some t

Mann-Whitney-Wilcoxon Procedure (cont.)

- Consider location model: $G(x) = F(x - \Delta)$ for some Δ
- Test becomes $H_0: \Delta = 0$ vs. $H_1: \Delta > 0$
- What does H_0 imply? $\Delta = 0 \Rightarrow F(x) = G(x)$

  $\therefore$ consider combined sample, $n = n_1 + n_2$
  - Under H_0, ranks are uniform between Xs and Ys
  - Under H_1, Ys will have larger ranks
- Let $W = \sum_{j=1}^{n_2} R(Y_j)$, where $R(Y_j)$ denotes ranks of Y_j in combined sample

Mann-Whitney-Wilcoxon Statistic

- W is Mann-Whitney-Wilcoxon (MWW) statistic
- Decision rule: reject H_0 if $W \geq c$
- No closed form for W's null distribution

Theorem 3

Suppose $X_1, \dots, X_{n_1}$ is a random sample from a distribution with a continuous cdf $F(x)$ and $Y_1, \dots, Y_{n_2}$ is a random sample from a distribution with a continuous cdf $G(x)$. Suppose H_0: $F(x) = G(x)$ for all x. If H_0 is true, then

- W is distribution free with a symmetric pmf
- $E_{H_0}(W) = \frac{n_2(n+1)}{2}$
- $\text{Var}_{H_0}(W) = \frac{n_1 n_2(n+1)}{12}$
- $\frac{W - [n_2(n+1)/2]}{\sqrt{\text{Var}_{H_0}(W)}}$ has an asymptotically N(0,1) distribution

How'd you get that?

Compute $E(W)$ under H_0.

$E_{H_0}(W) = E_{H_0}\left(\sum_{j=1}^{n_2} R(Y_j)\right) = \sum_{j=1}^{n_2} E_{H_0}(R(Y_j))$ where, under H_0, $R(Y_j)$ uniformly distributed throughout $\{1, 2, \dots, n\}$

$\therefore E_{H_0}(R(Y_j)) = \sum_{i=1}^{n} i\left(\frac{1}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} i = \frac{1}{\cancel{n}}\left(\frac{\cancel{n}(n+1)}{2}\right) = \frac{n+1}{2}$

$\Rightarrow E_{H_0}(W) = \sum_{j=1}^{n_2} \frac{n+1}{2} = \frac{n+1}{2}\sum_{j=1}^{n_2} 1 = \frac{(n+1)n_2}{2}$

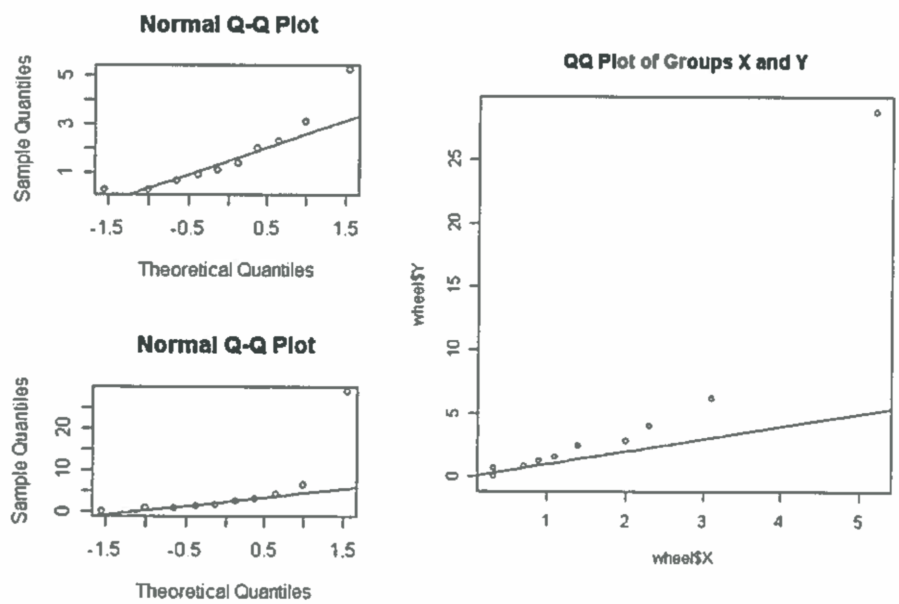
Example 3

Abebe et al. (2001) studied the number of wheel revolutions per minute of two groups of mice. Group 1 was a placebo group, while Group 2 were under the influence of a drug. Does the drug impact the performance of the mice? The data is contained in **wheel.txt** on Canvas.

X	2.3	0.3	5.2	3.1	1.1	0.9	2.0	0.7	1.4	0.3
Y	0.8	2.8	4.0	2.4	1.2	0.0	6.2	1.5	28.8	0.7

How do the data compare?

The Data



EDA show non-normality

Example 3 (cont.)

Consider H_0 vs. two-sided H_1.

X	2.3	0.3	5.2	3.1	1.1	0.9	2.0	0.7	1.4	0.3
$R(X)$	13	2.5	18	16	8	7	12	4.5	10	2.5
Y	0.8	2.8	4.0	2.4	1.2	0.0	6.2	1.5	28.8	0.7
$R(Y)$	6	15	17	14	9	1	19	11	20	4.5

$W = \sum_j R(y_j) = 6 + 15 + \ldots + 4.5 = 116.5$

What is the p-value?

$n_1 = n_2 = 10$

$n = n_1 + n_2 = 20$

$E_{H_0}(W) = \frac{n_2(n+1)}{2} = \frac{10(21)}{2} = 105$

$V_{H_0}(W) = \frac{n_1 n_2 (n+1)}{12} = \frac{10(10)(21)}{12} = 175$

$\mathbb{P}(W \geq 116.5) = \mathbb{P}\left(Z \geq \frac{116.5 - 105}{\sqrt{175}}\right) = \mathbb{P}(Z \geq .869) = \boxed{.1922}$

Another representation

- Without loss of generality, assume Y_j's ordered
- $R(Y_j) = \#_i\{X_i < Y_j\} + \#_i\{Y_i \leq Y_j\}$
- $W = \sum_{j=1}^{n_2} R(Y_j) = \sum_{j=1}^{n_2} \#_i\{X_i < Y_j\} + \sum_{j=1}^{n_2} \#_i\{Y_i \leq Y_j\}$

the rank of Y_j is determined by how many Xs are less than Y_j, along with how many Ys are less than or equal Y_j

$= \underbrace{\#_{i,j}\{Y_j > X_i\}}_{U} + \sum_{j=1}^{n_2} j$ — because there are j Ys that are less than or equal Y_j because Y_j is the jth Y

$= U + \frac{n_2(n_2+1)}{2}$

Another representation (cont.)

- $U = \#_{i,j}\{Y_j > X_i\}$
- Decision rule: reject H_0 if $U \geq c_2$
- By Theorem, U is distribution free with

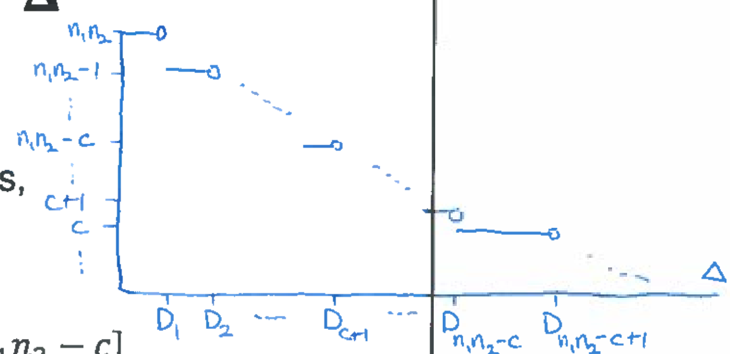
  $E(U) = \mathbb{E}(W) - \frac{n_2(n_2+1)}{2} = \frac{n_2(n+1)}{2} - \frac{n_2(n_2+1)}{2} = \frac{n_2}{2}(n+\cancel{1}-n_2-\cancel{1}) = \frac{n_1n_2}{2}$

  $\mathrm{Var}(U) = \mathrm{Var}\left(W - \frac{n_2(n_2+1)}{2}\right) = \mathrm{Var}(W) = \frac{n_1n_2(n+1)}{12}$

- Power function nondecreasing in Δ

CI for Δ

- More generally, denote $U(\Delta) = \#_{i,j}\{Y_j - X_i > \Delta\}$
- Consider ordered differences, $D_1 < \cdots < D_{n_1n_2}$



$$\Rightarrow 1-\alpha = P_\Delta[c < U(\Delta) < n_1n_2 - c]$$
$$= P_\Delta[D_{c+1} \leq \Delta < D_{n_1n_2-c}]$$

i.e., $[D_{c+1}, D_{n_1n_2-c})$ is $100(1-\alpha)\%$ CI for Δ

- Asymptotically, we can use CLT to approximate c:

$$c = \frac{n_1n_2}{2} - z_{\alpha/2}\sqrt{\frac{n_1n_2(n+1)}{12}} - \frac{1}{2}$$