# 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 <u>distribution free</u>.
- 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 $X_1, X_2, \ldots, X_n$ random sample where $X_i = \theta + \epsilon_i$, $\epsilon_i$'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?

# *Sign Test (cont.)*

- Decision rule: Reject $H_0$ if $S \geq c$
- Under $H_0$, $S \sim$ Binomial($n$, ½). Why?

- Level $\alpha$ 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

# *Lemma 1*

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

# *Theorem 1*

- Suppose model $X_i = \theta + \epsilon_i$ is true. Let $\gamma(\theta)$ be the power function of the sign test of level $\alpha$ for the hypotheses

$$H_0\colon \theta = \theta_0 \text{ vs. } H_1\colon \theta > \theta_0$$

  Then $\gamma(\theta)$ is a nondecreasing function of $\theta$.

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

# *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*

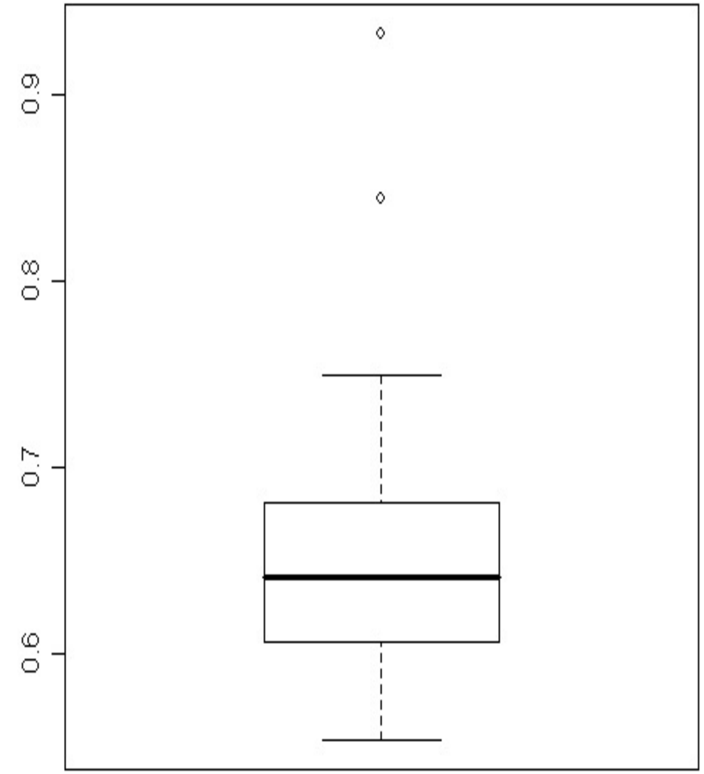
```
> stem(shoshoni$ratio)

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

  5 | 578
  6 | 011111135677799
  7 | 5
  8 | 4
  9 | 3
```

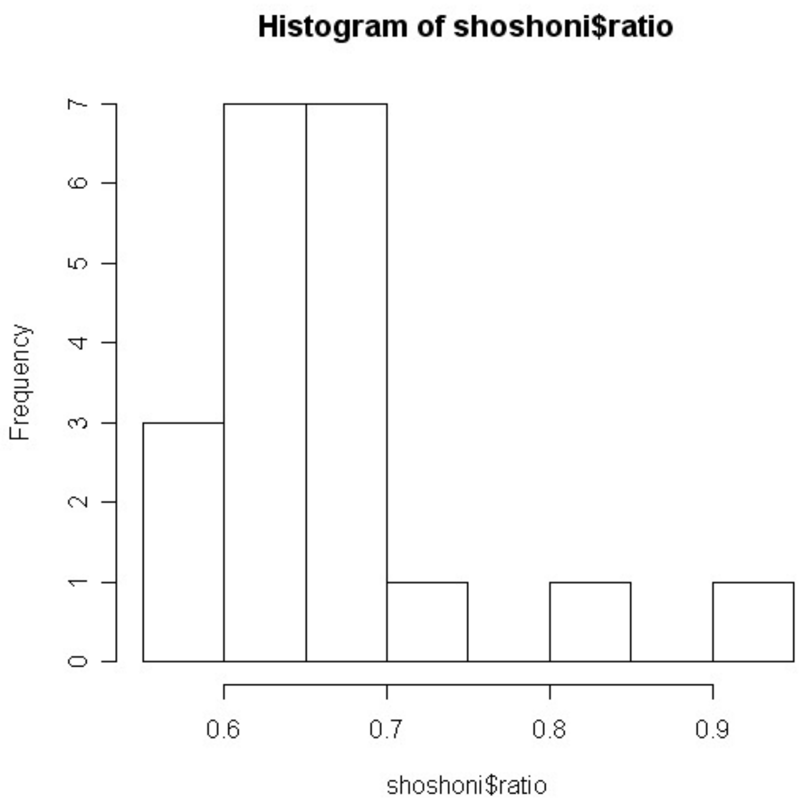
```
> summary(shoshoni$ratio)
   Min.  1st Qu.  Median    Mean 3rd Qu.    Max.
 0.5530  0.6060  0.6410  0.6605  0.6765  0.9330
```

> boxplot(shoshoni$ratio)

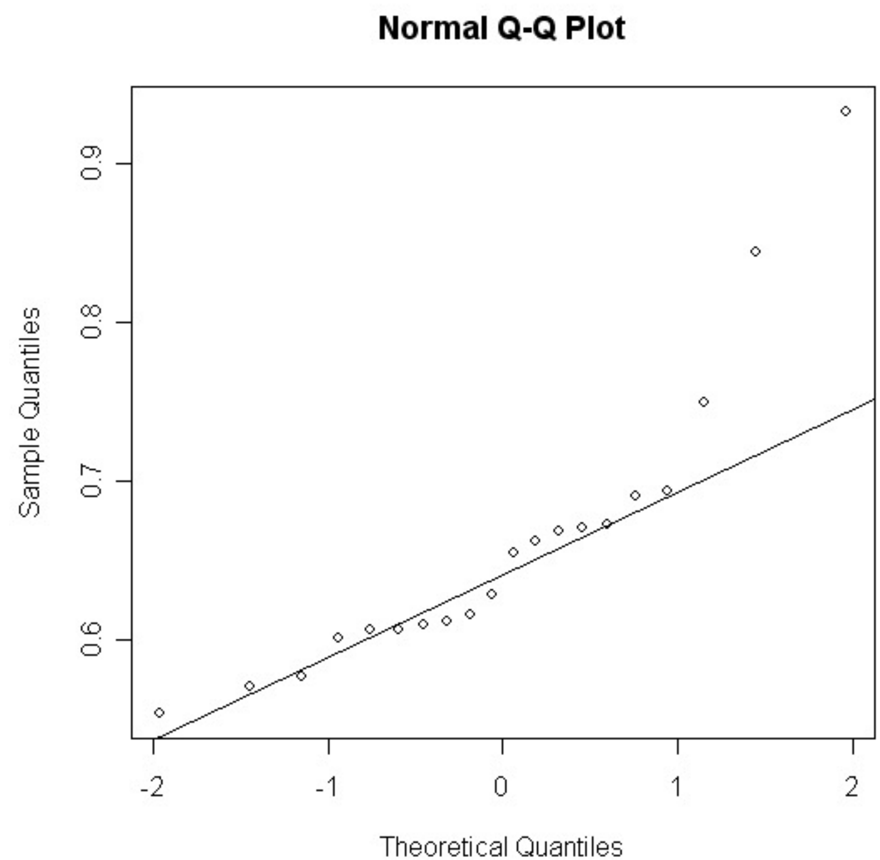


# The Data (cont.)

> hist(shoshoni$ratio)




> qqline(shoshoni$ratio)




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

# *Example 1 (cont.): The Test*

- Consider hypothesis

$$H_0: \theta = 0.618 \text{ vs } H_1: \theta \neq 0.618$$

- 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", $c = 5$

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

# *CI for the Median*

- Recall decision rule for two-sided test: reject $H_0$ if $S(\theta_0) \le c$ or $S(\theta_0) \ge n - c$, where $c$ determined s.t.

$$P(S(\theta_0) \le c) = \alpha/2$$

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

# *Example 1 (cont.)*

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

```
> 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 $\theta$
- What do you conclude?

# *Signed-Rank Wilcoxon Test*

- More efficient than sign test

- Let $X_1, X_2, \dots, X_n$ random sample where $X_i = \theta + \epsilon_i$, where $\epsilon_i$'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 $\alpha$ test

# *Theorem 2*

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

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

# *Another Representation*

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

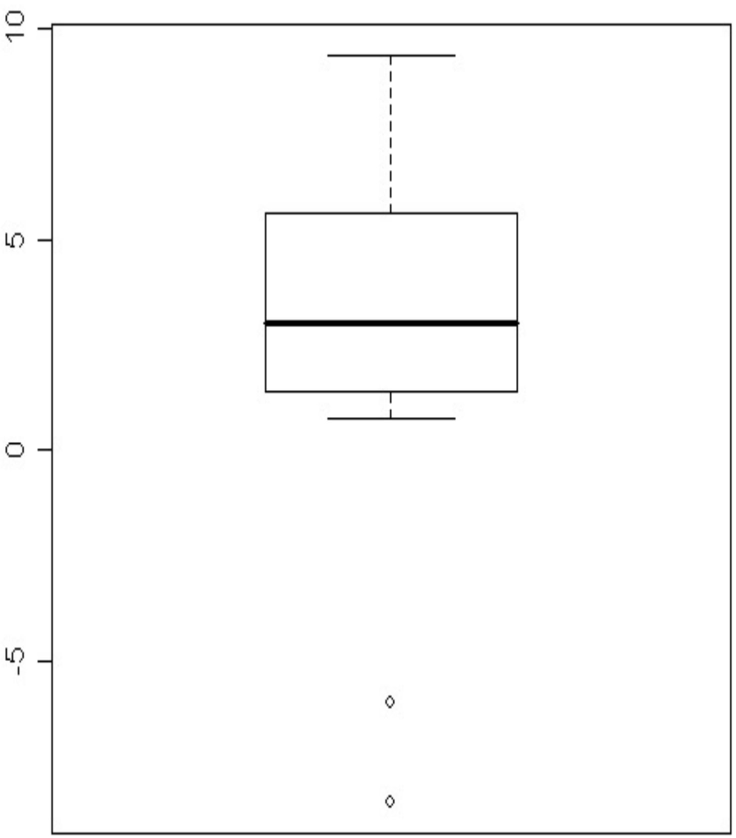
# Another Representation

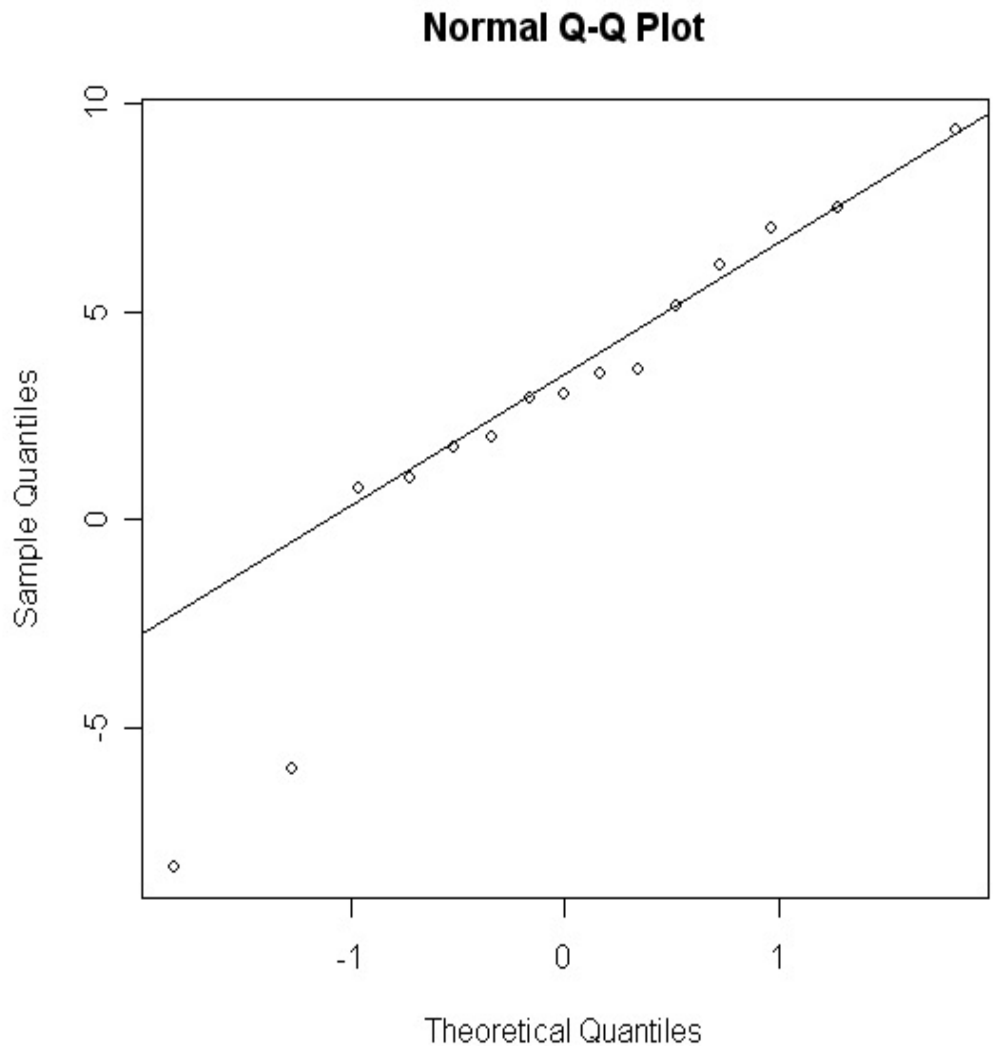
$\therefore T^+$ is a linear function of signed-rank test $T$.
What are $E_{H_0}(T^+)$ and $\mathrm{Var}_{H_0}(T^+)$?

# *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
- $(X_i, Y_i),\ \ i = 1, \dots, 15$ are heights of cross-fertilized and self-fertilized plants, respectively, in $i$th pot
- $W_i = X_i - Y_i$
- Which model is more appropriate? Parametric or nonparametric?

# The Data


10
5
0
-5


Normal Q-Q Plot
Sample Quantiles
10
5
0
-5
-1
0
1
Theoretical Quantiles

# Example 2 (cont.)

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

| W | Signed-Ranks |
|---|---|
| 6.125 | |
| -8.375 | |
| 1.000 | |
| 2.000 | |
| 0.750 | |
| 2.925 | |
| 3.500 | |
| 5.125 | |
| 1.750 | |
| 3.625 | |
| 7.000 | |
| 3.000 | |
| 9.375 | |
| 7.500 | |
| -6.000 | |

# CI for the Median

- $T^{+} = \#_{i \le j}\{(X_i + X_j)/2 > 0\}$
- $W = (X_i + X_j)/2$ called Walsh averages

- $1 - \alpha = P_\theta[c_W < T^{+}(\theta) < m - c_W]$

  $= P_\theta[W_{c_W+1} \le \theta < W_{m-c_W}]$, 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 $Y$

# *Mann-Whitney-Wilcoxon Procedure (cont.)*

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

- 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
- $E_{H_0}(W) = \frac{n_2(n+1)}{2}$
- $\mathrm{Var}_{H_0}(W) = \frac{n_1 n_2(n+1)}{12}$
- $\frac{W - \left[{n_2(n+1)}/{2}\right]}{\sqrt{\mathrm{Var}_{H_0}(W)}}$ has an asymptotically N(0,1) distribution

# How'd you get that?

Compute $E(W)$ under $H_0$.

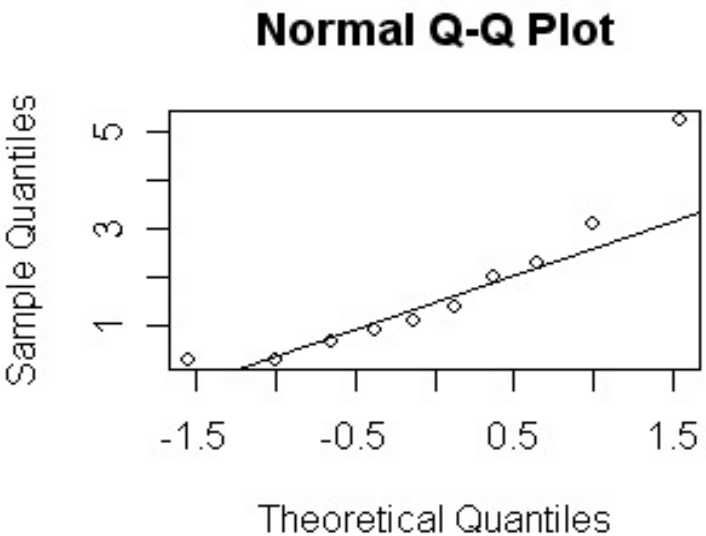
# *Example 3*

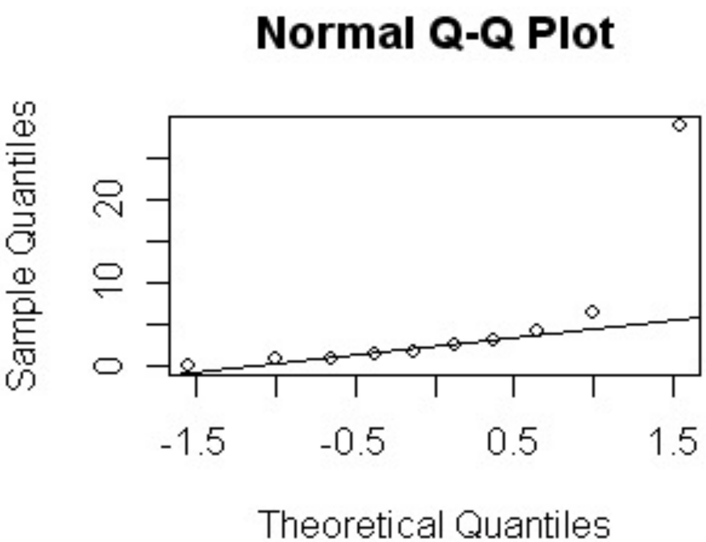
Abebe et al. (2001) studied the number of wheel revolutions per minute of two groups of mice.  Group 1 was a placebo group, while Group 2 were under the influence of a drug. Does the drug impact the performance of the mice?  The data is contained in **wheel.txt** on Canvas.

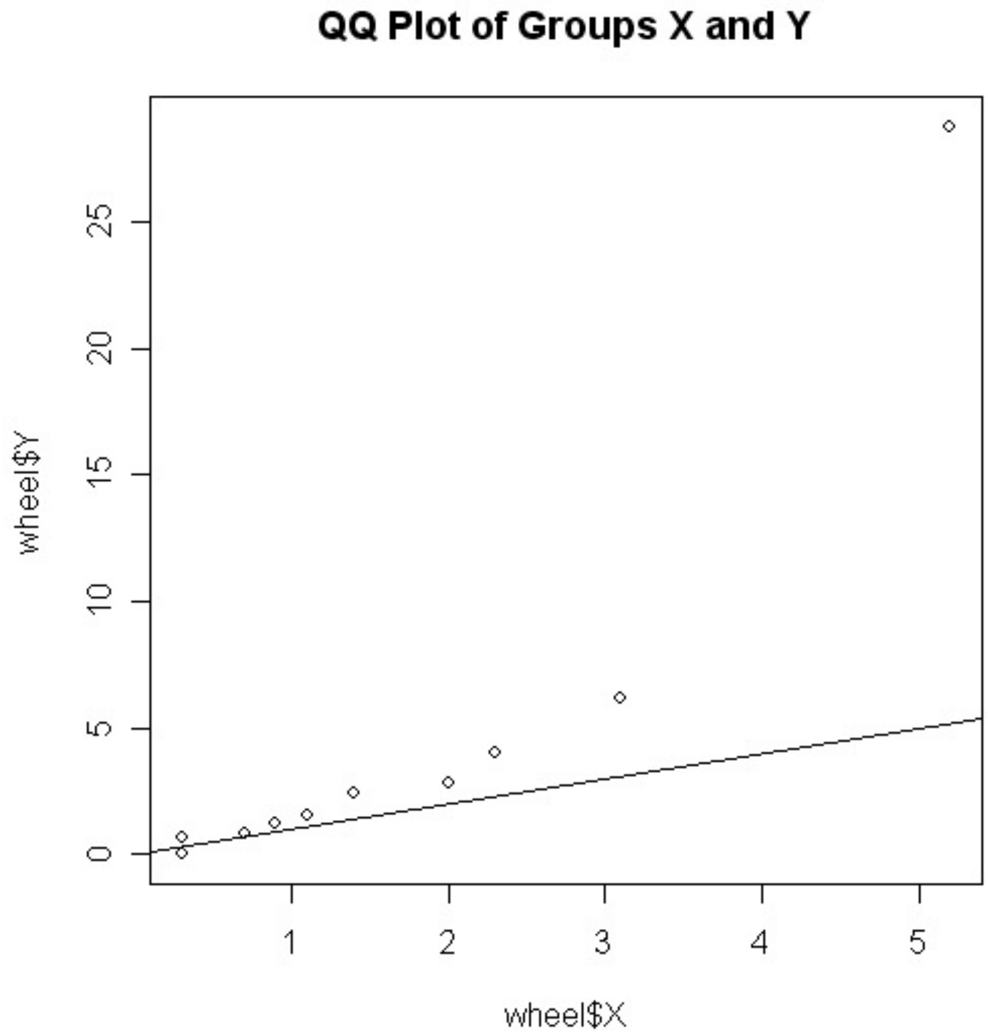
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| X | 2.3 | 0.3 | 5.2 | 3.1 | 1.1 | 0.9 | 2.0 | 0.7 | 1.4 | 0.3 |
| Y | 0.8 | 2.8 | 4.0 | 2.4 | 1.2 | 0.0 | 6.2 | 1.5 | 28.8 | 0.7 |

How do the data compare?

# *The Data*


Normal Q-Q Plot
Sample Quantiles
1
3
5
-1.5
-0.5
0.5
1.5
Theoretical Quantiles


Normal Q-Q Plot
Sample Quantiles
0
10
20
-1.5
-0.5
0.5
1.5
Theoretical Quantiles


QQ Plot of Groups X and Y
wheel$Y
0
5
10
15
20
25
1
2
3
4
5
wheel$X

# *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?

# *Another representation*

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

- $W \ = \sum_{j=1}^{n_2} R(Y_j)$

# *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

  $E(U) =$

  $\mathrm{Var}(U) =$

- Power function nondecreasing in $\Delta$

# *CI for Δ*

- More generally, denote

  $U(\Delta) = \#_{i,j}\{Y_j - X_i > \Delta\}$
- Consider ordered differences,

  $D_1 < \cdots < D_{n_1 n_2}$

$$\Rightarrow\ 1-\alpha = P_\Delta[c < U(\Delta) < n_1 n_2 - c]$$
$$= P_\Delta[D_{c+1} \le \Delta < D_{n_1 n_2 - c}]$$

  i.e., $[D_{c+1}, D_{n_1 n_2 - c})$ is $100(1-\alpha)\%$ CI for Δ
- Asymptotically, we can use CLT to approximate $c$:

$$c = \frac{n_1 n_2}{2} - z_{\alpha/2}\sqrt{\frac{n_1 n_2 (n+1)}{12}} - \frac{1}{2}$$