

7 Nonparametric test- binomial test

What if the normality assumption of the previous tests are in doubt?

7.1 Binomial test for proportion in count data (P.361-366)

Example: (Flu vaccine) A flu vaccine is known to be 20% effective in the second year after inoculation. To determine if a new vaccine is more effective 12 people are chosen at random and inoculated. If 5 of those receiving the new vaccine do not contact the virus in the second year after vaccination is the new vaccine superior to the old one?

What kind of test should be used if the data is binary (contact or not), not normal?

Suppose the *binary* outcomes X_1, X_2, \dots, X_n are the results of n independent trials with the success probability is p , that is,

$$X_i = \begin{cases} 1 & \text{for a success} \\ 0 & \text{for a failure} \end{cases}$$

Then the total number of success of these n trials

$$X = \sum_{i=1}^n X_i \sim \mathcal{B}(n, p), 0 < p < 1$$

is a binomial rv with a probability mass function (pmf)

$$\Pr(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, \dots, n.$$

The five steps to test if $p = p_0$ is:

1. **Hypothesis:** $H_0 : p = p_0$ vs $H_1 : p > p_0, p < p_0, p \neq p_0$
2. **Test statistic:** $T = X$.
3. **Assumption:** Trials are independent with the same probability of success p . Then $X \sim \mathcal{B}(n, p_0)$ under H_0 .
4. **P-value:**

$$\Pr(X \geq x) = \sum_{i=x}^n \binom{n}{i} p_0^i (1 - p_0)^{n-i} \text{ for } H_1 : p > p_0$$

$$\Pr(X \leq x) = \sum_{i=0}^x \binom{n}{i} p_0^i (1 - p_0)^{n-i} \text{ for } H_1 : p < p_0$$

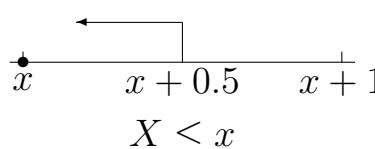
$$2 \Pr(X \geq x) \quad \text{if } x \geq n/2 \text{ or}$$

$$2 \Pr(X \leq x) \quad \text{if } x \leq n/2 \text{ for } H_1 : p \neq p_0.$$

5. **Decision:** reject H_0 if p -value $< \alpha$.

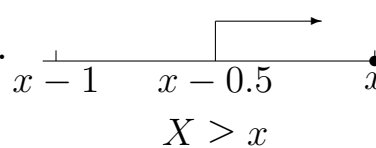
Remark:

1. The binomial table gives $\Pr(X \leq x)$ for $n = 2, \dots, 12$ and $p = 0.1, 0.2, \dots, 0.9$.
2. In R, use `pbinom(x,n,p0)` for $\Pr(X \leq x)$ and `pbinom(x-1,n,p0,lower.tail=F)` for $\Pr(X \geq x) = \Pr(X > x - 1) = 1 - P(X \leq x - 1)$.
3. For large n ($n \geq 20$), the central limit theorem assures that

$$\Pr(X \leq x) \approx \Phi \left(\frac{x + 0.5 - np_0}{\sqrt{np_0(1 - p_0)}} \right)$$


$X \leq x$

and

$$\Pr(X \geq x) \approx 1 - \Phi \left(\frac{x - 0.5 - np_0}{\sqrt{np_0(1 - p_0)}} \right)$$


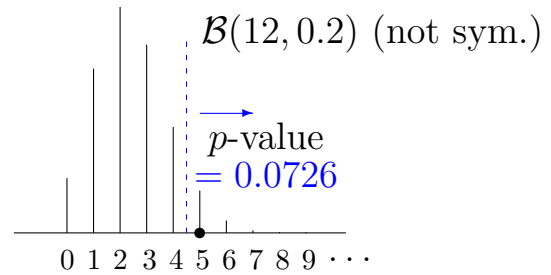
$X \geq x$

The half value added or subtracted is for the *continuity correction* but it may be omitted if sample size is large.

Example: (Flu vaccine)

Solution: Let X denote the number not getting the flu in the second year. We have $X = 5$ and $n = 12$. The *binomial test* on the effective rate in the second year after inoculation of a new vaccine is

1. **Hypotheses:** $H_0 : p = 0.2$ against $H_1 : p > 0.2$.
2. **Test statistic:** $T = X = 5$.
Large value of x will argue against H_0 in favor of H_1 .
3. **Assumption:** Independent trials with constant probability of success. Then $X \sim \mathcal{B}(12, 0.2)$ under H_0 .
4. **P-value:** $\Pr(X \geq 5) = 1 - \Pr(X \leq 4) = 1 - 0.9274 = 0.0726$
(bin. table; $n = 12, p = 0.2, x = 4$)
5. **Decision:** Since p -value is > 0.05 , we accept H_0 and conclude that the data is consistent with H_0 that the effective rate is 20%.



In R,

```
> n=12
> p0=0.2
> x=5
> binom.test(x,n,p0,alt="greater",0.95)
```

Exact binomial test

```
data:  x and n
number of successes = 5, number of trials = 12, p-value = 0.07256
alternative hypothesis: true probability of success is greater than 0.2
95 percent confidence interval:
 0.1810248 1.0000000
sample estimates:
probability of success
      0.4166667

> p.value=pbinom(x-1,n,p0,lower.tail=F)  # exact p-value
> p.value
[1] 0.0725555
```

Note:

Since the 95% CI for p include $p_0 = 0.2$, we accept H_0 .

`pbinom(x-1,n,p0,lower.tail=F)` gives $\Pr(X > x - 1) = \Pr(X \geq x)$
 whereas

`pbinom(x,n,p0,lower.tail=T)` or `pbinom(x,n,p0)` gives $\Pr(X \leq x)$.

Example: (Washers) A manufacturer of automatic washers offers a particular model in one of three colors : A, B or C. Of the first 100 washers sold, 40 were of color A. Would you conclude that customers have a preference for color A?

Solution: Let p denote the probability that a customer prefers color A, and X denote the number of customers who prefer color A in the first 100 customers. If customers have no preference for color A, then $p = 1/3$. Otherwise $p > 1/3$.

The test for the *proportion of customers who prefer color A* using the binomial test is

1. **Hypotheses:** $H_0 : p = 1/3$ against $H_1 : p > 1/3$.

2. **Test statistic:** $z_0 = \frac{X - 0.5 - np}{\sqrt{np(1-p)}} = \frac{39.5 - 100/3}{\sqrt{100(\frac{1}{3})(1 - \frac{1}{3})}} = 1.3081$.

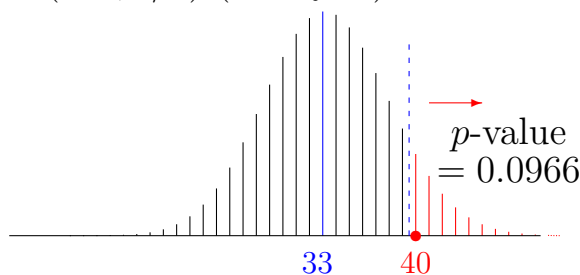
Large value of z_0 will argue against H_0 in favour of H_1 .

3. **Assumptions:** $X \sim \mathcal{B}(100, 1/3)$ under H_0 .

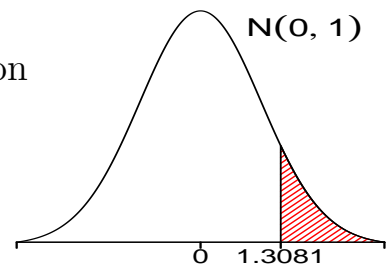
4. **P-value:** $\Pr(X \geq 40) \approx \Pr(Z \geq 1.3081) = 1 - 0.9046 = 0.0954$.

5. **Decision:** Since P -value is > 0.05 , there is not sufficient evidence in the data against H_0 . Customers have no preference for color A.

$\mathcal{B}(100, 1/3)$ (not sym.)



Normal
approximation
→



In R,

```
> n=100
> p0=1/3
> x=40
> binom.test(x,n,p0,alt="greater",0.95)
```

Exact binomial test

```
data:  x and n
number of successes = 40, number of trials = 100, p-value = 0.09662
alternative hypothesis: true probability of success is greater than 0.3333
95 percent confidence interval:
 0.317526 1.000000
sample estimates:
probability of success
                0.4
```

```
> z10=(x-0.5-n*p0)/sqrt(n*p0*(1-p0))  #above using normal approx.
> sp=x/n
> z20=(sp-p0)/sqrt(p0*(1-p0)/n)  #z.test using (*) below
> p.value.exact=pbinom(x-1,n,p0,lower.tail=F)  # exact p-value
> p.value.norm=pnorm(z10,lower.tail=F)  #normal approx.
> p.value.ztest=pnorm(z20,lower.tail=F)  #z.test
> c(sp,z10,z20)
[1] 0.400000 1.308148 1.414214
> c(p.value.exact,p.value.norm,p.value.ztest)
[1] 0.09662307 0.09541163 0.07864960
```

Remark:

1. Since the 95% CI for p include $p_0 = 1/3$, we accept H_0 .
2. The binomial test report exact p -value using binomial distribution which is different from 0.0954 using normal approximation.

3. When n is large ($n \geq 20$), the test statistic for z -test is

$$z_0 = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}} = \frac{\frac{x}{n} - p_0}{\sqrt{p_0(1 - p_0)/n}} \sim \mathcal{N}(0, 1) \quad (*)$$

by CLT *without continuity correction* where $\frac{x}{n}$ is the sample proportion.

8 Nonparametric test (P.655-656)

8.1 Sign test for mean μ (P.657-662)

Example: (Moisture retention) The following data are 15 measurements of moisture retention (%) using a new sealing system. The system is expected to be better (greater retention) than the previous system, for which the mean retention was 96%. Use the sign test to analyse the data.

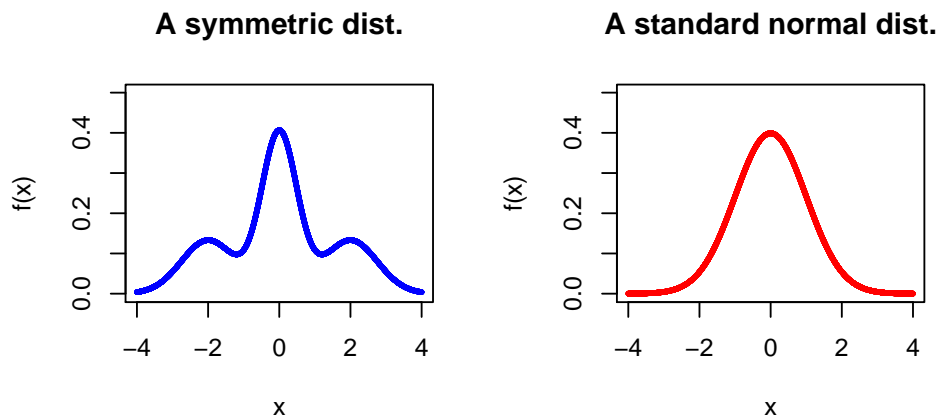
97.5 95.2 97.3 96.0 96.8 99.8 97.4 95.3
 98.2 99.1 96.1 97.6 98.2 98.5 99.4

What if we have a small sample of observations measured on a continuous range but fails the normality assumption?

Suppose a sample X_1, \dots, X_n is taken from a *continuous* distribution. We want to test $H_0 : \mu = \mu_0$. If the distribution is *symmetric* about μ_0 under H_0 , then $d_i = x_i - \mu_0$ should scatter around 0, equally likely to be positive or negative. Hence the probability p_+ , of getting a positive d_i is 0.5.

The *binomial* test reduces to a *sign test* of proportions. The sign test is a *non-parametric* test as no assumption on the data distribution is made.

Examples of symmetric distributions include:



1. **Hypotheses:** $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$ or $\mu < \mu_0$ or $\mu \neq \mu_0$
or $H_0 : p_+ = \frac{1}{2}$ vs $H_1 : p_+ > \frac{1}{2}$ or $p_+ < \frac{1}{2}$ or $p_+ \neq \frac{1}{2}$
2. **Test Statistics:** $X = \#(d_i > 0)$
3. **Assumptions:** X_1, X_2, \dots, X_n from a symmetric distribution.
 $X \sim \mathcal{B}(n, \frac{1}{2})$ under H_0 .
4. **P-value:**

$$\begin{aligned} &\Pr(X \geq x) \text{ for } H_1 : \mu > \mu_0; \\ &\Pr(X \leq x) \text{ for } H_1 : \mu < \mu_0; \\ &2 \Pr(X \geq x) \text{ for } H_1 : \mu \neq \mu_0 \text{ \& } x > \frac{n}{2}; \\ &2 \Pr(X \leq x) \text{ for } H_1 : \mu \neq \mu_0 \text{ \& } x < \frac{n}{2}; \\ &1 \text{ for } H_1 : \mu \neq \mu_0 \text{ \& } x = \frac{n}{2} \end{aligned}$$
5. **Decision:** If $p\text{-value} < \alpha$, there is evidence against H_0 .
If $p\text{-value} > \alpha$, the data are consistent with H_0 .

Remarks:

1. We should *drop* observations with $d_i = 0$ and *change* n accordingly as these *zeros* contains no information on the sign. However a few articles pointed out that it might be better to keep the zeros instead of cutting off them in terms of the power of the tests.
2. We only use the *sign* of the d_i and ignore their *magnitude* in the sign test. The test *ignores much information* in the sample but it can be applied in quite *general situations*. Hence the sign test is *more robust*, i.e. less affected by outlying large or small observations. However it may have a *lower power*, i.e. less likely to reject H_0 for a given sample information.

3. If the sample is believed to come from a normal population, you should use the more powerful t -test instead of a sign test.
4. We should check the assumption of symmetric data distribution using a *boxplot* with R command `boxplot(d)`.

Example: (Moisture retention) The 15 measurements are:

97.5 95.2 97.3 96.0 96.8 99.8 97.4 95.3
 98.2 99.1 96.1 97.6 98.2 98.5 99.4

Solution: The sign of differences $d_i = x_i - \mu_0 = x_i - 96$ are:

+ - + 0 + + + - + + + + + + +

Let p_+ be the probability of a positive difference. The *sign test* for the mean % of moisture retention using a new sealing system is

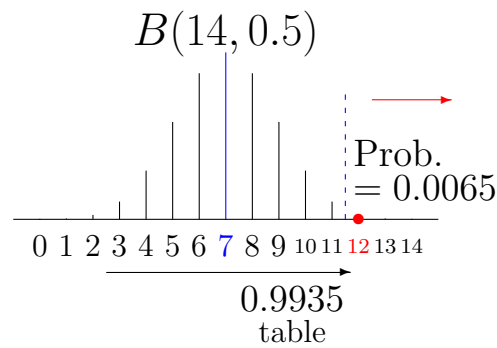
1. **Hypotheses:** $H_0 : p_+ = \frac{1}{2}$ against $H_1 : p_+ > \frac{1}{2}$.
2. **Test statistic:** $x = 12$ in the $m = 14$ (ignore 0 difference).

Large value of x will argue against H_0 in favour of H_1 .

3. **Assumption:** X_i follow a symmetric distribution. Then $X \sim \mathcal{B}(14, 0.5)$ under H_0 .
4. **P-value:** $X \sim \mathcal{B}(14, 0.5)$ under H_0 .

$$\begin{aligned}
 P(X \geq 12) &= \sum_{i=12}^{14} \binom{14}{i} 0.5^i 0.5^{14-i} \\
 &= 0.5^{14} \left[\binom{14}{12} + \binom{14}{13} + \binom{14}{14} \right] \\
 &= 0.00006104[14(13)/2 + 14 + 1] \\
 &= 0.0065. \quad (\text{or from R})
 \end{aligned}$$

5. **Decision:** Since P -value is < 0.05 , there is strong evidence in the data against H_0 . The retention rate is greater than 96%.



In R,

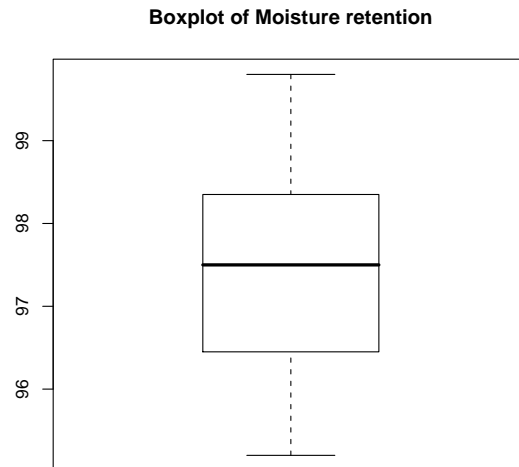
```
> y=c(97.5,95.2,97.3,96.0,96.8,99.8,97.4,95.3,98.2,99.1,96.1,97.6,
      98.2,98.5,99.4)
> mu0=96
> d=y-mu0
> d
[1]  1.5 -0.8  1.3  0.0  0.8  3.8  1.4 -0.7  2.2  3.1  0.1  1.6  2.2
     2.5  3.4
> n=length(d[d!=0])
> x=length(d[d>0])
> ps=x/n
> p0=0.5
> binom.test(x,n,0.5,alt="greater",0.95)
```

Exact binomial test

```
data:  x and n
number of successes = 12, number of trials = 14, p-value = 0.00647
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.6146103 1.0000000
sample estimates:
probability of success
 0.8571429
```

```
> boxplot(y) #check sym. dist.
> title("Boxplot of Moisture retention")
> p.value=pbinom(x-1,n,0.5,lower.tail=FALSE)
> p.value
[1] 0.006469727
```

```
> c(n,x,ps,p.value)
[1] 14.000000000 12.000000000 0.8571429 0.006469727
```

**Remark:**

1. From the boxplot, the assumption of symmetric distribution is satisfied. We don't check QQ plot as normality is not assumed.
2. Since the 95% CI for p_+ excludes $p_0 = 0.5$, the data are against H_0 .
3. The command `n=length(d[d!=0])` counts the number of non-zero differences where '`!=`' means ' \neq ' in R.

Example: (Smoking) Blood samples from 11 individuals before and after they smoked a cigarette are used to measure aggregation of blood platelets.

Before 25 25 27 44 30 67 53 53 52 60 28;

After 27 29 37 36 46 82 57 80 61 59 43.

Is the aggregation affected by smoking?

Solution: Our conclusion based on the paired t -test could be badly misguided if the sample differences $d_j = x_j - y_j$ are not from a normal distribution. The sign test avoid the normality assumption on the d_j .

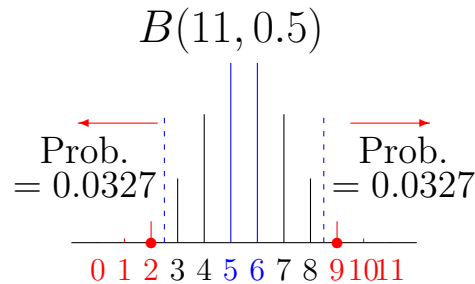
d_j	-2	-4	-10	8	-6	-15	-4	-27	-9	1	-15
Sign	-	-	-	+	-	-	-	-	-	+	-

The sign test for the mean of differences of the aggregation of blood platelets before and after smoking is

1. **Hypotheses:** $H_0 : m = 0$ against $H_1 : m \neq 0$.
2. **Test statistic:** $X = 2$.
3. **Assumption:** Distribution of d_i is symmetric. Then $X \sim \mathcal{B}(11, 0.5)$ under H_0 .
4. **P -value:**

$$\begin{aligned}
 2\Pr(X \leq 2) &= 2 \times 0.0327 = 0.0654 \text{ (bin. table; } n=11, p=0.5, x=2) \\
 &\stackrel{\text{or}}{=} 2 \times \sum_{i=0}^2 \binom{11}{i} 0.5^i 0.5^{11-i} \\
 &= 0.5^{11} \left[\binom{11}{0} + \binom{11}{1} + \binom{11}{2} \right] \\
 &= 0.000488[1 + 11 + 11(10)/2] \\
 &= 0.0654.
 \end{aligned}$$

5. **Decision:** There is just insufficient evidence against H_0 . Hence the aggregation is not affected by smoking.



But t -test shows strong evidence against H_0 (p -value=0.0157). A larger p -value show that there is less evidence against H_0 by just considering the sign. Hence sign test is less powerful, that is, it can detect less evidence against H_0 and hence has less chance to reject H_0 even if it is false. Normality and hence symmetry was checked.

In R,

```
> before=c(25,25,27,44,30,67,53,53,52,60,28)
> after=c(27,29,37,36,46,82,57,80,61,59,43)
> d=before-after
> d
[1] -2 -4 -10 8 -16 -15 -4 -27 -9 1 -15
> n=length(d[d!=0])
> x=length(d[d>0])
> c(n,x)
[1] 11 2
> binom.test(x,n,0.5,alt="two.sided",0.95)
```

Exact binomial test

data: x and n

number of successes = 2, number of trials = 11, p-value = 0.06543

alternative hypothesis: true probability of success is not equal to 0.5

95 percent confidence interval:

0.0228312 0.5177559

```
sample estimates:  
probability of success  
      0.1818182
```

```
> pvalue=2*pbinom(x,n,0.5)  
> pvalue  
[1] 0.06542969
```

Note that the 95% CI for p_+ include $p_0 = 0.5$. Hence we accept H_0 .

9 Wilcoxon sign-rank test (P.662-663,666-673)

While the normality assumption fails and sign test discards too much information, how to make use of the information of ordering or rank apart from the sign from a symmetric distribution?

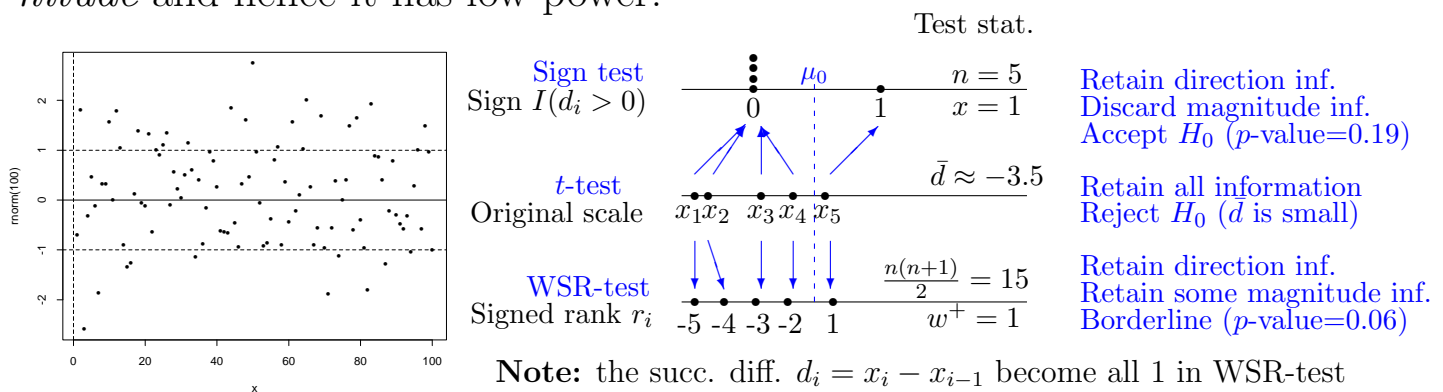
9.1 Introduction

Suppose the sample X_1, X_2, \dots, X_n are drawn from a population symmetric with respect to mean μ (or median). We test the hypotheses:

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0, \mu < \mu_0, \mu \neq \mu_0.$$

The t -test and z -test assume a normal population or at least a distribution without a long tail.

On the other hand, the *sign test* discards all data information on *magnitude* and hence it has low power.



Under the assumption of symmetric distribution and H_0 , *half* of the $d_i = x_i - \mu_0$ should be *negative* and *half positive* and the expected counts are both $n/2$.

The positive and negative d_i should be of *equal magnitude* and occur with *equal probability*. If we rank the *absolute values* of d_i in ascending order, the expected *rank sums* for the negative and positive d_i should be nearly equal.

Let R_1, \dots, R_n be the ranks of $|X_1 - \mu_0|, \dots, |X_n - \mu_0|$,

W^+ be the sums of the ranks R_i corresponding to positive $X_i - \mu_0$'s;

W^- be the sums of the ranks R_i corresponding to negative $X_i - \mu_0$'s;

$$W = \min(W^+, W^-).$$

The observed ranks are r_1, \dots, r_n for $|x_1 - \mu_0|, \dots, |x_n - \mu_0|$ and

$$w^+ = \sum_{i: x_i - \mu_0 > 0} r_i, \quad w^- = \sum_{i: x_i - \mu_0 < 0} r_i.$$

We should accept $H_1 : \mu > \mu_0$ ($H_1 : \mu < \mu_0$) if w^+ is substantially large (small) and accept $H_1 : \mu \neq \mu_0$ if w is substantially small.

The WSR test is

1. **Hypotheses:** $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0, \mu < \mu_0, \mu \neq \mu_0$
2. **Test Statistics:** W^+ for one-sided or
 $W = \min(W^+, W^-)$ for two-sided
3. **Assumptions:** X_1, X_2, \dots, X_n from symmetric distribution
4. **P-value:** $\Pr(W^+ \geq w^+)$ for $H_1 : \mu > \mu_0$;
 $\Pr(W^+ \leq w^+)$ for $H_1 : \mu < \mu_0$;
 $2 \Pr(W^+ \leq w)$ for $H_1 : \mu \neq \mu_0$;
5. **Decision:** If $p\text{-value} < \alpha$, there is evidence against H_0 .
If $p\text{-value} > \alpha$, the data are consistent with H_0 .

9.2 Calculation of p -value:

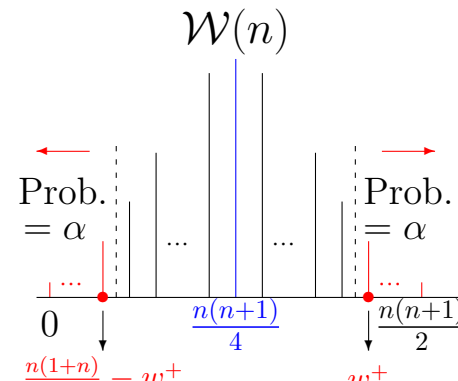
Case 1: Small sample size ($n \leq 20$), *no zeros* and *no ties* on the data $|x_1 - \mu_0|, \dots, |x_n - \mu_0|$.

The *exact* p -value $\Pr(W^+ \leq w^+)$ where w^+ lies in the *lower* range is given the *Wilcoxon Signed Rank Distribution* table.

Note the following conversion if w^+ lies in the *upper* range:

$$\Pr(W^+ \geq w^+) = \Pr(W^+ \leq n(n+1)/2 - w^+) \text{ (convert to lower value)}$$

Note:

$$\begin{aligned} W^+ + W^- &= 1 + 2 + \dots + n = n(1+n)\frac{1}{2} \\ \Rightarrow W^- &= n(1+n)\frac{1}{2} - W^+ \\ \Rightarrow E(W^+) &= n(1+n)\frac{1}{4} \end{aligned}$$


$\mathcal{W}(n)$

In R, the commands for the test are

```
wilcox.test(x, alternative="??", mu=mu0, exact=T, correct=F);
wilcox.test(x, y, alternative="??", mu=0, paired=T, exact=T, correct=F)
psignrank(w, n, lower.tail=T, log.p=F)
```

give the $\Pr(W \leq w)$ when the sample size is **n** with no ties and zeros.

Example: (Weight gain) Weights of twins on diets Y and X are

y_i	85	69	81	112	77	86
x_i	83	78	70	72	67	68
d_i	2	-9	11	40	10	18

Is there a weight gain in taking diet Y as compared with diet X?

Solution: Normal assumption may not be suitable for the outlier 40.

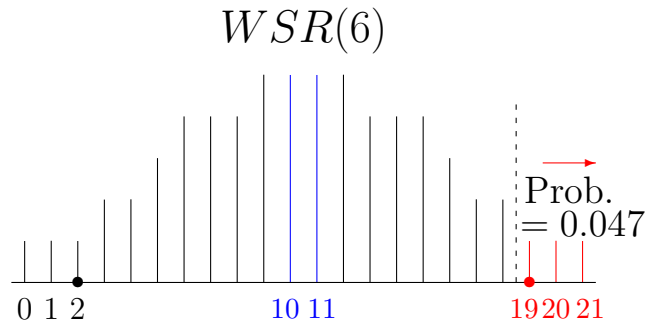
d_i	2	-9	11	40	10	18
$ d_i $	2	9	11	40	10	18
Rank $ d_i $	1	2	4	6	3	5
Sign r_i	1	-2	4	6	3	5

$n = 6$ and sum all $= \frac{n(n+1)}{2} = \frac{6(7)}{2} = 21$. The Wilcoxon sign-rank test for the mean of differences in weight between the two diets is

1. **Hypotheses:** $H_0 : \mu = 0$ against $H_1 : \mu > 0$.
2. **Test statistic:** $w^+ = 19$, $w^- = 2$, $w = 2$.
3. **Assumption:** D_i follow a symmetric dist. Then $W \sim WSR(n)$.
4. **P-value:**

$$\begin{aligned}
 p\text{-value} &= \Pr(W^+ \geq w^+) = \Pr(W^+ \geq 19) \\
 &= \Pr(W^+ \leq \frac{6(6+1)}{2} - 19 = 21 - 19 = 2) \\
 &= 0.047 \text{ (from Wilcoxon Signed Rank table; } n = 6, w^+ = 2)
 \end{aligned}$$

5. **Decision:** Since P -value is < 0.05 , we reject H_0 . There is weight gain in taking diet Y as compared with diet X.



In R,

```
> y=c(85,69,81,112,77,86)
> x=c(83,78,70,72,67,68)
> wilcox.test(y,x,alternative="greater",mu=0,paired=T,
  exact=T,correct=F)
```

Wilcoxon signed rank test

data: x and y

V = 19, p-value = 0.04688

alternative hypothesis: true location shift is greater than 0

```
> d=y-x    #checking only
> d
[1]  2 -9 11 40 10 18
> n=length(d)
> r = rank(abs(d))
> r
[1] 1 2 4 6 3 5
> sign.r=r*sign(d)
> sign.r
[1] 1 -2 4 6 3 5
> w.plus = sum(r[d>0])
> w.minus = sum(r[d<0])
```

```
> w = min(w.plus, w.minus)
> p.value=psignrank(w,n)
> c(n,w.plus,w.minus,w,p.value)
[1] 6.000000 19.000000 2.000000 2.000000 0.046875
```

Note that the sample size is too small to check for the symmetric distribution of d_i using boxplot.

The test stat. is $w^+ = 19$ out of $\frac{n(n+1)}{2} = 21$ whereas it is $x = 5$ out of $n = 6$ (p -value=0.1094) for the *sign test*.

Case 2: Large sample size ($n \geq 20$) or there are *ties* or there are *zeros* on the data $|x_1 - \mu_0|, \dots, |x_n - \mu_0|$.

We approximate W^+ by a normal distribution, *NOT the data* X_i . The p -value is approximately given by

$$p\text{-value} \approx \Pr \left(Z \geq \frac{w^+ - E(W^+)}{\sqrt{\text{Var}(W^+)}} \right) \quad \text{for } H_1 : \mu > \mu_0;$$

$$p\text{-value} \approx \Pr \left(Z \leq \frac{w^+ - E(W^+)}{\sqrt{\text{Var}(W^+)}} \right) \quad \text{for } H_1 : \mu < \mu_0;$$

$$p\text{-value} \approx 2 \Pr \left(Z \geq \left| \frac{w^+ - E(W^+)}{\sqrt{\text{Var}(W^+)}} \right| \right) \quad \text{for } H_1 : \mu \neq \mu_0.$$

where

$$\text{In general: } E(W^+) = \frac{1}{2} \sum_{i: x_i - \mu_0 \neq 0} r_i \quad \text{and} \quad \text{Var}(W^+) = \frac{1}{4} \sum_{i: x_i - \mu_0 \neq 0} r_i^2$$

$$\text{No ties \& zeros: } E(W^+) = \frac{1}{4} n(n+1) \quad \text{and} \quad \text{Var}(W^+) = \frac{1}{24} n(n+1)(2n+1)$$

Proof: Let $I_i = I(d_i > 0)$ be indicator which is 1 if $d_i > 0$ and 0 otherwise. I_i are binary variables which change from samples to samples.

$$\begin{aligned} E(W^+) &= E\left(\sum_{i=1}^n R_i I_i\right) \\ &= \sum_{i=1}^n R_i E(I_i) \quad \text{since } I_i \sim \text{Ber}\left(\frac{1}{2}\right) \text{ under } H_0, \quad E(I_i) = \frac{1}{2} \\ &= \sum_{i=1}^n R_i \frac{1}{2} = \frac{1}{2}(1 + 2 + \cdots + n) = \frac{1}{4}n(n+1) \end{aligned}$$

Moreover, since $I_i \sim \text{Ber}(\frac{1}{2})$, $\text{Var}(I_i) = p(1-p) = \frac{1}{2}(1 - \frac{1}{2}) = \frac{1}{4}$.

$$\begin{aligned} \text{Var}(W^+) &= \sum_{i=1}^n R_i^2 \text{Var}(I_i) = \sum_{i=1}^n R_i^2 \frac{1}{4} \quad \text{since } \text{Var}(I_i) = \frac{1}{4} \\ &\stackrel{\text{no ties}}{=} \frac{1}{4}(1^2 + 2^2 + \cdots + n^2) = \frac{1}{4} \times \frac{1}{6}n(n+1)(2n+1) \\ &= \frac{1}{24}n(n+1)(2n+1) \end{aligned}$$

With zeros or ties, $\text{Var}(W^+)$ will be smaller.

In R, commands for the test are

```
wilcox.test(x, alternative="??", mu=mu0, exact=F, correct=F);  
wilcox.test(x, y, alternative="??", mu=0, paired=T, exact=F, correct=F)
```

Remarks:

1. **paired**: if **TRUE**, the *Wilcoxon signed rank test* is computed. The default is **FALSE** and gives the Wilcoxon rank sum test.
2. **alternative**: **greater**, **less** or **two.sided**
3. **mu**: the location shift for the distribution of x .
4. **exact**: if **TRUE** the *exact* distribution for the test statistic is used to compute the p -value if possible. This refers to *table value*. *With*

ties, you should write **exact=F**. Otherwise with **exact=T**, a warning message is given which states that **exact=T** is impossible and normal approximation with **exact=F** is adopted instead.

5. **correct**: if **TRUE** a *continuity correction* ($w^+ \pm 0.5$ and $w + 0.5$) is applied to the *normal approximation* for calculating p -value. However, unlike X which is an integer in **binom.test**, W can be non-integer, say 3.25, and hence continuity correction does not apply in this *normal approximation*. Hence we should have **correct=F**.

Example: (Smoking) Blood samples from 11 individuals before and after they smoked a cigarette are used to measure aggregation of blood platelets.

Before(x_i)	25	25	27	44	30	67	53	53	52	60	28
After(y_i)	27	29	37	36	46	82	57	80	61	59	43
$d_i (x_i - y_i)$	-2	-4	-10	8	-16	-15	-4	-27	-9	1	-15

Is the aggregation affected by smoking?

Solution: Let μ be the aggregation difference.

d_i	-2	-4	-10	8	-16	-15	-4	-27	-9	1	-15
$ d_i $	2	4	10	8	16	15	4	27	9	1	15
Ranks $ d_i $	2	3.5	7	5	10	8.5	3.5	11	6	1	8.5
Sign r_i	-2	-3.5	-7	5	-10	-8.5	-3.5	-11	-6	1	-8.5

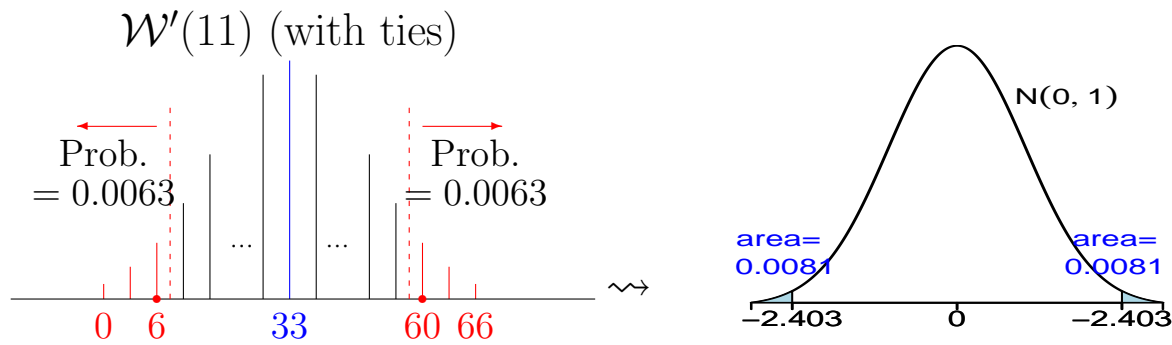
$n = 11$ and $\text{sum all} = \frac{11 \cdot 12}{2} = 66$. The Wilcoxon sign-rank test for the mean of diff. in the agg. of blood platelets before & after smoking is

- Hypotheses:** $H_0 : \mu_d = 0$ against $H_1 : \mu_d \neq 0$.
- Test statistic:** $w^+ = 5 + 1 = 6$, $w^- = 66 - 6 = 60$, $w = 6$.
- Assumption:** X_i follow a symmetric distribution.
- P-value:** $E(W^+) = \frac{n(n+1)}{4} = \frac{11(11+1)}{4} = 33$.

$$\text{Var}(W^+) = \frac{1}{4} \sum_{i=1}^{11} r_i^2 = \frac{1}{4} [(-2)^2 + \dots + (-8.5)^2] = \frac{506}{4} = 126.25.$$

$$\begin{aligned} p\text{-value} &= 2 \Pr(W^+ \leq 6) = 2 \Pr\left(Z \leq \frac{6 - 33}{\sqrt{126.25}}\right) \\ &= 2 \Pr(Z \leq -2.403) = 2(0.008131) = 0.0163 \end{aligned}$$

- Decision:** Since P -value is < 0.05 , we reject H_0 . There is strong evidence against H_0 and hence the aggregation is affected by smoking, which is consistent with the conclusion using t -test.



Note: Exact prob. condition on the ranks $\{1, 2, 3.5, 3.5, 5, 6, 7, 8.5, 8.5, 10, 11\}$ is

$$2Pr(W^+ \leq 6) = \frac{2 \times 13}{2^{11}} = 2 \times 0.006347656 = 0.01269531.$$

The 13 choices of rank sum W^+ such that $W^+ \leq 6$ are

$$\{0, 1, 2, 1 + 2, 3.5, 3.5, 1 + 3.5, 1 + 3.5, 5, 2 + 3.5, 2 + 3.5, 6, 1 + 5\}.$$

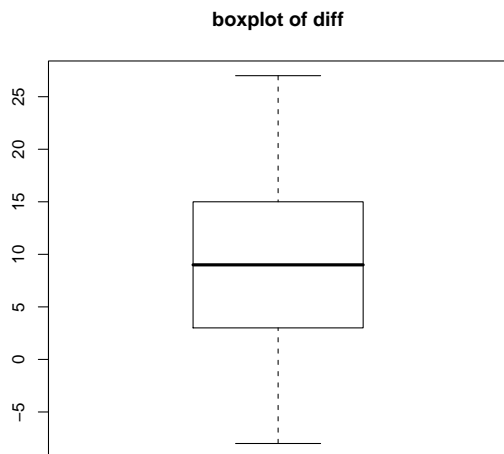
```
> x=c(25,25,27,44,30,67,53,53,52,60,28)
> y=c(27,29,37,36,46,82,57,80,61,59,43)
> wilcox.test(x,y,alternative="two.sided",mu=0,paired=T,exact=F,correct=F)
```

Wilcoxon signed rank test

```
data: x and y
V = 6, p-value = 0.01626
alternative hypothesis: true location shift is not equal to 0
> d=x-y #checking only
> d
[1] -2 -4 -10 8 -16 -15 -4 -27 -9 1 -15
> r = rank(abs(d))
> r
[1] 2.0 3.5 7.0 5.0 10.0 8.5 3.5 11.0 6.0 1.0 8.5
> sign.r=r*sign(d)
> sign.r
[1] -2.0 -3.5 -7.0 5.0 -10.0 -8.5 -3.5 -11.0 -6.0 1.0 -8.5
> w.plus = sum(r[d>0])
> w.minus = sum(r[d<0])
> w = min(w.plus, w.minus)
```

```
> ew.plus = sum(r[d!=0])/2
> varw.plus = sum((r[d!=0])^2)/4
> c(w.plus,w.minus,w,ew.plus,varw.plus) #output to check lower or upper
[1]  6 60  6 33 126.25000000
> z0=(w.plus-ew.plus)/sqrt(varw.plus)
> p.value=2*pnorm(z0)
> c(z0,p.value)
[1] -2.40296846  0.01626259
> boxplot(d) #check symmetric data distribution
```

1. The test stat. is $w^+ = 6$ out of total sum of rank $\frac{n(n+1)}{2} = 66$ whereas it is $x = 2$ out of count $n = 11$ (p -value=0.0654) for *sign test*. This shows that WSR test using sign and ranks can detect stronger evidence from the data.
2. Since the boxplot is symmetric, the assumption of symmetric data distribution is satisfied.



Summary of one sample tests on *mean*,

Assume:	Parametric <i>Normal</i> data distribution		Non-parametric <i>Symmetric</i> data distribution	
Type:	<i>t</i> -test (σ^2 unknown)	<i>z</i> -test (σ^2 known)	sign test	WSR test
Test stat:	$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$	$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$	$X \sim \text{Bin}(n, \frac{1}{2})$	$W \sim \text{WSR}(n)$ or $\frac{W - E(W)}{\sqrt{\text{Var}(W)}} \stackrel{\text{ties}}{\sim} \mathcal{N}(0, 1)$
Power:	most (use all data)	most (use all data)	least (use sign)	middle (use sign & rank)
Nor ass:	sensitive	sensitive	most robust	robust

10 Transformation of data

Many powerful tests are based on normal or symmetric distribution assumption. What if the data are not even symmetric?

10.1 Transformation of data to symmetry

A *right skewed* (left skewed) distribution is one where most values cluster around the lower (upper) end of the scale and there is a scattering of large (small) values.

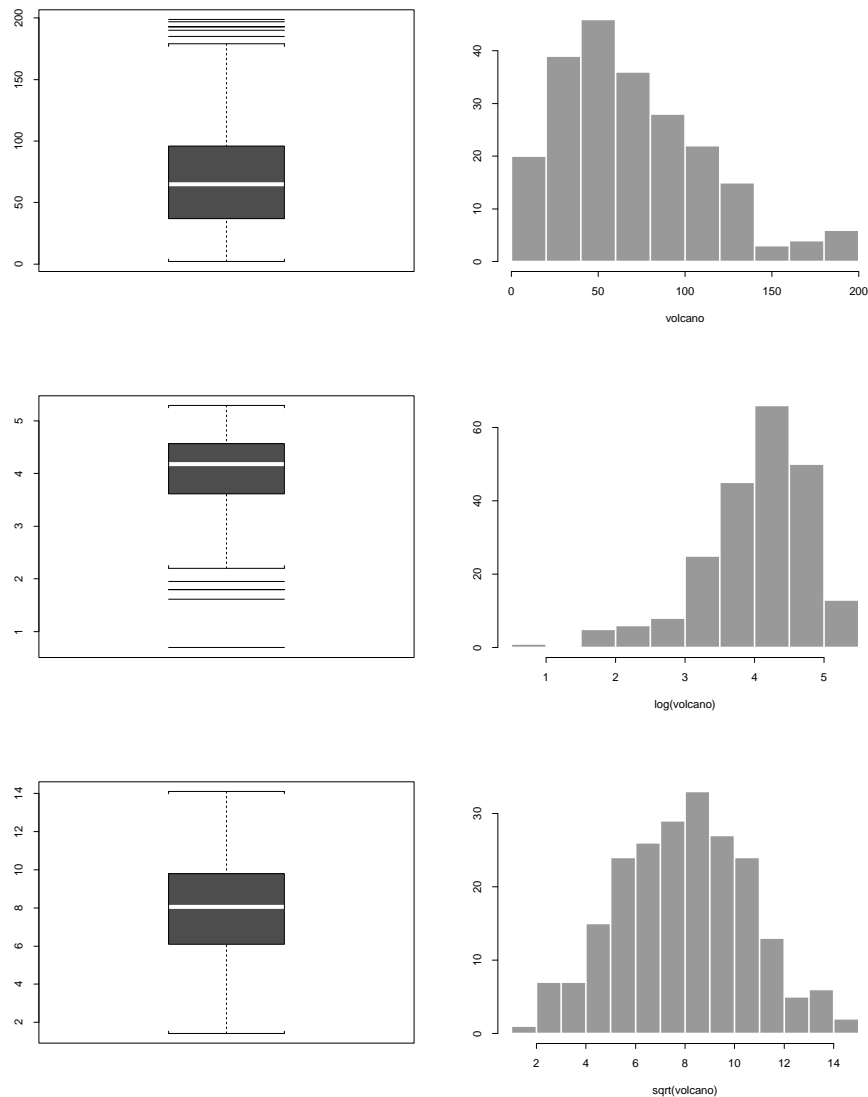
Example: Right skewed data

```
8 : 122223335679
9 : 000123344556779
10 : 0112233445689
11 : 0112334669
12 : 11244456
13 : 03478
14 : 00
15 : 667
16 : 25
17 : 29
18 : 5
19 : 03379
```

Example: Left skewed data

```
2 : 22333
2 : 4
2 : 6
2 : 88899
3 : 0000111
3 : 222223333333
3 : 4444445555
3 : 6666666667777777
3 : 8888888889999999999999
4 : 0000000000000000000111111
4 : 22222222222222222222333333333333
4 : 4444444444444444555555555555
4 : 6666666666666666777777777777
4 : 88888888888899999999
5 : 001111
5 : 2223333
```

Volcano



One of the following transformations or others may be useful to make *right skewed* data more symmetric:

1. $y = x^a$, where $0 < a < 1$;
2. $y = \log x$;
3. $y = -1/x^a$, where $a > 0$;

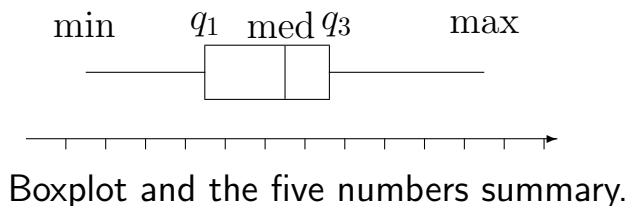
so that large value becomes less large.

Remarks:

1. We may sometimes need to adjust the origin of data, i.e., we may use $y = (x - c)^a$, etc, for transformation.
2. For left skewed data, the sign can be changed to make it right skewed.
3. There is no single unique best transformation to make perfect symmetry.
4. The effect of a transformation can be examined by a five number summary or a boxplot. The ratio

$$\frac{q_3 - \text{med}}{\text{med} - q_1} \quad \text{or} \quad \frac{\text{max} - \text{med}}{\text{med} - \text{min}}$$

should be close to unity if a suitable transformations has been achieved.



10.2 Test for median

Suppose that independent observations X_1, X_2, \dots, X_n are drawn from a population with a distribution function $F(x)$. Let m be a median. We want to test hypothesis

$$H_0 : m = m_0 \quad \text{vs} \quad H_1 : m > m_0, \text{ or} \\ m < m_0, \text{ or} \\ m \neq m_0.$$

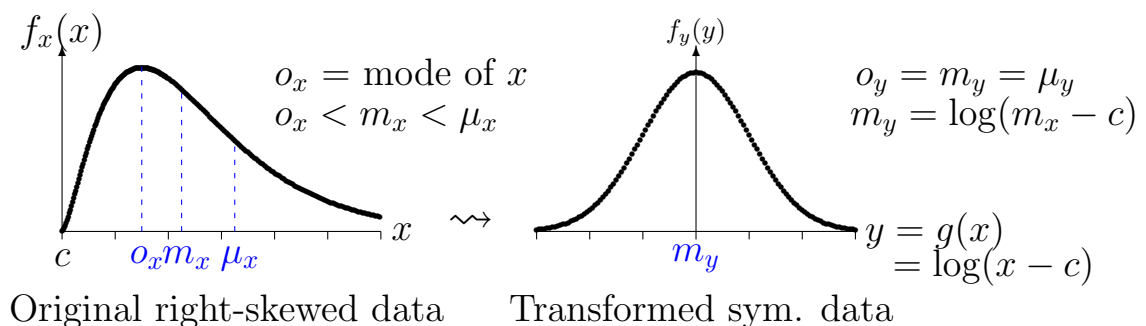
If there exists a positive strictly increasing function $g(x)$ such that $\{g(x_i)\}$ is more symmetric, we may construct a powerful test for m using $\{g(m)\}$ instead.

Indeed, to test $H_0 : m = m_0 \quad \text{vs} \quad H_1 : m > m_0,$

is equivalent to test $H_0 : g(m) = g(m_0) \quad \text{vs} \quad H_1 : g(m) > g(m_0),$

using the data $\{g(x_i)\}$.

Since $\{g(x_i)\}$ is more symmetric, the Wilcoxon sign-rank test is reliable in this case. By noting that median is close to mean for a symmetric data, we may even use t -test for the mean instead of the median.



High: 25.500 26.500 27.333 28.500 28.583 29.083 30.667 30.750
32.667 32.750 35.500 35.833 36.583 39.750 41.583 43.833
44.250 70.417 73.000

Suppose we want to test

$$H_0 : m = 18.6 \quad \text{vs} \quad H_1 : m > 18.6,$$

We should *accept* H_0 because the median is really 18.6.

```
> x = survey$Age
```

By sign test,

```
> length(x[x!=18.6]) # 5 18.6 in leaf plot are in 3 dec. pl.
[1] 237
> length(x[x>18.6]) # about half of 237 as median=18.6
[1] 116
> bi = binom.test(116, 237, p=0.5, alt="greater")
> bi$p-value
p-value = 0.6516 #accept H0. OK !!!
```

By Wilcoxon sign-rank test,

```
> wi = wilcox.test(x,alt="greater",mu=18.6,exact=F,correct=F)
> wi$p-value
p-value = 0.0085 #reject H0. Wrong !!!
```

By t -test,

```
> t = t.test(x,alt="greater",mu=18.6)
> t$p-value
p-value = 0 #reject H0. Wrong !!!
```

Sign test is least powerful so it often gives insignificant result. Wilcoxon sign-rank test assumes a symmetric distribution whereas t -test further assumes a normal distribution. As these *assumptions fail*, both tests give *wrong* result.

After transformation using $g(x) = \log(x - c)$ where c is certain origin taken to be 16 and 16.7 (since the min=16.7),

By Wilcoxon sign-rank test using log transformation at different center,

```
>wi.tr1=wilcox.test(log(x-16),alt="greater",  
                    mu=log(18.6-16),exact=F, correct=F)
```

```
>wi.tr$p-value  
p-value = 0.2396    #accept H0.  OK !!!
```

```
>wi.tr2=wilcox.test(log(x-16.7),alt="greater",  
                    mu=log(18.6-16.7),exact=F, correct=F)
```

```
>wi.tr$p-value  
p-value = 0.5686    #accept H0.  OK !!!
```

By t -test,

```
>t.tr1=t.test(log(x-16),alt="greater",mu=log(18.6-16))
```

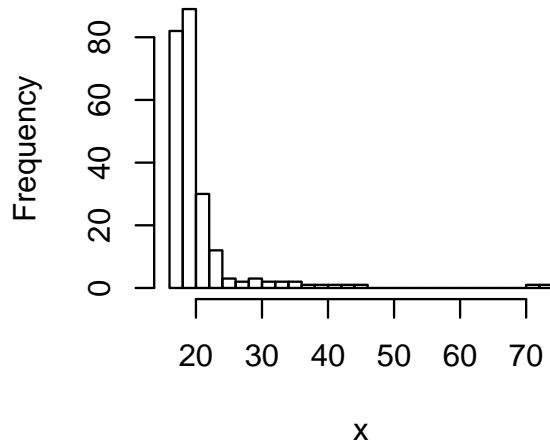
```
>t.tr1$p-value  
p-value = 0.0129    #still reject H0.  Wrong !!!
```

```
>t.tr2=t.test(log(x-16.7),alt="greater",mu=log(18.6-16.7))
```

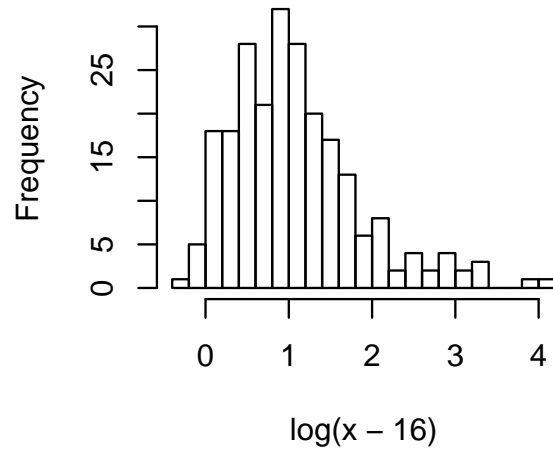
```
>t.tr2$p-value  
p-value = 0.3791    #accept H0.  OK !!!
```

With transformation to symmetry, Wilcoxon sign-rank test gives valid result regardless of the center whereas t -test is more sensitive to distribution assumption and hence outliers.

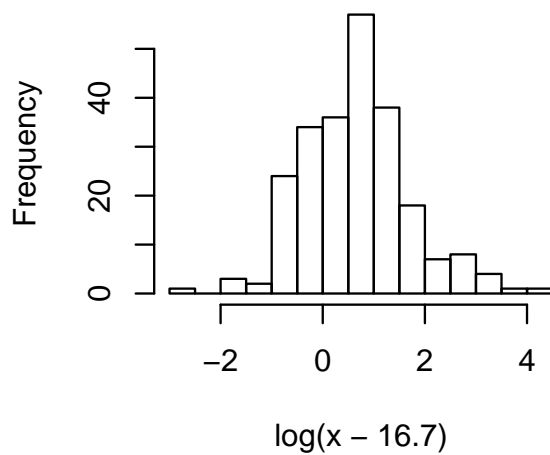
Histogram of x



Histogram of $\log(x - 16)$



Histogram of $\log(x - 16.7)$



11 Two sample t and z test

11.1 Two sample t test (P.457-467,481-483)

Example: (Height comparison) A fourth grade class has 10 girls and 13 boys. The children's heights are recorded on their 10th birthday as follows:

Boys: 135.3, 137.0, 136.0, 139.7, 136.5, 137.2, 138.8,
139.6, 140.0, 137.7, 135.5, 134.9, 139.5

Girls: 140.3, 139.8, 138.6, 137.1, 140.0, 136.2,
138.7, 138.5, 134.9, 141.0

Is there evidence that girls are taller than boys on their 10th birthday?

What if there are two independent samples?

We wish to test the population mean difference.

One sample: tests on the population *mean* (*median*) for the data (x_1, \dots, x_{n_2}) have been discussed under different assumptions.

1. Student's t -test, z -test, sign-test and Wilcoxon sign-rank test.
2. Fixed level tests (critical value, rejection region)
3. Power and confidence intervals.

Paired sample: the test on the *difference* $d_i = x_i - y_i$ for the paired data $(x_1, y_1), \dots, (x_n, y_n)$ is similar to the *one sample* test.

In paired observations $(x_1, y_1), \dots, (x_n, y_n)$, if x_i and y_i are the before and after observations from the same individual, they are *dependent*.

Two independent sample: x_i and y_i come from different individuals and hence are *independent*.

Example: (Height comparison)

Solution: Let μ_x and μ_y be the boys' and girls' average heights respectively. The question can be answered by testing:

$$H_0 : \mu_x = \mu_y \quad \text{vs} \quad H_1 : \mu_x < \mu_y.$$

using the two samples t -test:

1. **Hypotheses:** $H_0 : \mu_x = \mu_y$ vs $H_1 : \mu_x > \mu_y$ or $\mu_x < \mu_y$ or $\mu_x \neq \mu_y$

2. **Test statistic:**
$$t_0 = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where
$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

$$= \frac{\sum_{i=1}^{n_x} x_i^2 - \frac{1}{n_x} \left(\sum_{i=1}^{n_x} x_i \right)^2 + \sum_{i=1}^{n_y} y_i^2 - \frac{1}{n_y} \left(\sum_{i=1}^{n_y} y_i \right)^2}{n_x + n_y - 2}.$$

3. **Assumptions:** X_1, \dots, X_{n_x} are iid $\mathcal{N}(\mu_X, \sigma^2)$,
 Y_1, \dots, Y_{n_y} are iid $\mathcal{N}(\mu_Y, \sigma^2)$ and
 X_i 's are indept. of Y_i 's. Hence $t_0 \sim t_{n_x+n_y-2}$.

4. **P -value:** $\Pr(t_{n_x+n_y-2} \leq t_0)$ for $H_1 : \mu_x < \mu_y$
 $\Pr(t_{n_x+n_y-2} \geq t_0)$ for $H_1 : \mu_x > \mu_y$
 $2 \Pr(t_{n_x+n_y-2} \geq |t_0|)$ for $H_1 : \mu_x \neq \mu_y$

5. **Decision:** If p -value $< \alpha$, there is evidence against H_0 .
 If p -value $> \alpha$, the data are consistent with H_0 .

where the two samples x_1, \dots, x_{n_1} and y_1, \dots, y_{n_2} of the boys' and girls'

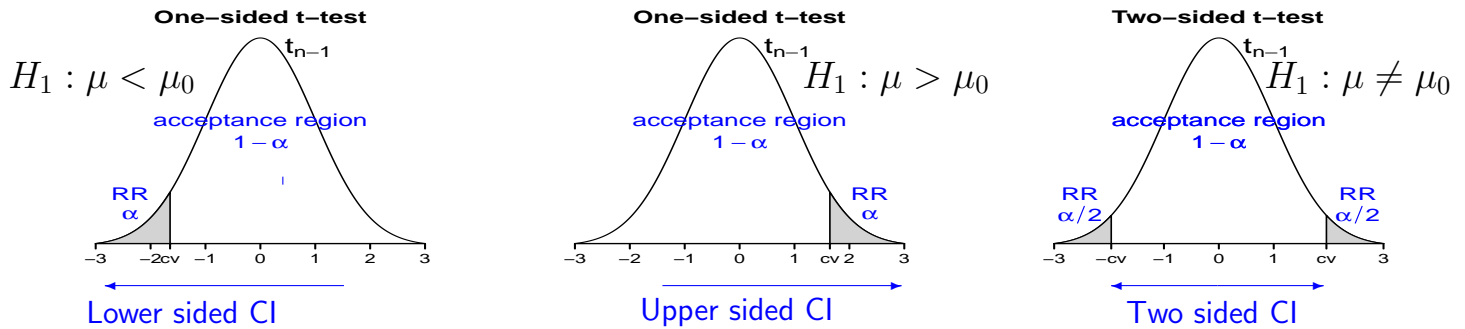
heights respectively are from distinct populations. The rejection regions for $\bar{x} - \bar{y}$ in the two samples t -test

$$\begin{aligned}
 \bar{x} - \bar{y} &\leq -t_{n_x+n_y-2,\alpha} s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \text{ for } H_1 : \mu_x < \mu_y; \\
 \bar{x} - \bar{y} &\geq t_{n_x+n_y-2,\alpha} s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \text{ for } H_1 : \mu_x > \mu_y; \\
 \bar{x} - \bar{y} &\leq -t_{n_x+n_y-2,\alpha/2} s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \text{ or} \\
 \bar{x} - \bar{y} &\geq t_{n_x+n_y-2,1-\alpha/2} s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \text{ for } H_1 : \mu_x \neq \mu_y \quad (11.1),
 \end{aligned}$$

and the CIs for $\mu_0 = \mu_x - \mu_y$ are

$$\begin{aligned}
 &(-\infty, \bar{x} - \bar{y} + t_{n_x+n_y-2,\alpha} s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}) \quad \text{for } H_1 : \mu_x < \mu_y; \\
 &(\bar{x} - \bar{y} - t_{n_x+n_y-2,\alpha} s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}, \infty) \quad \text{for } H_1 : \mu_x > \mu_y; \\
 &(\bar{x} - \bar{y} - t_{n_x+n_y-2,\alpha/2} s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}, \bar{x} - \bar{y} + t_{n_x+n_y-2,\alpha/2} s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}) \\
 &\quad \text{for } H_1 : \mu_x \neq \mu_y \quad (11.2).
 \end{aligned}$$

If $\mu_0 = 0$ lies outside the CI, H_0 should be rejected.



Remarks

1. We need to check the *normality* and *equality of variances* assumptions using *qq-plot* and *boxplot* respectively. If the *spread* (i.e. *ranges*; NOT symmetry!) of X_i and Y_i are similar, equality of variances assumption is satisfied.
2. The assumption of *equality of variances* ($\sigma_x = \sigma_y = \sigma$) is made to reduce the number of parameter when the sample sizes, n_x or n_y or both, are small in which case, s_x and s_y as the estimates of σ_x and σ_y may not be reliable.
3. If the observations are not normally distributed, the test statistic still distributes approximately as $t_{n_x+n_y-2}$ when n_x, n_y are large enough ($n_x, n_y \geq 10$), unless the distributions are very skewed.
4. The true variance of $\bar{X} - \bar{Y}$ should be $\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}$. It is not close to

$\sigma^2 \left(\frac{1}{n_x} + \frac{1}{n_y} \right)$ if σ_x^2 and σ_y^2 differ greatly. Moreover, if the sample sizes n_x and n_y are both large, we should use

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \quad (11.3)$$

as a test statistic whose distribution is approximately a Student- t

with

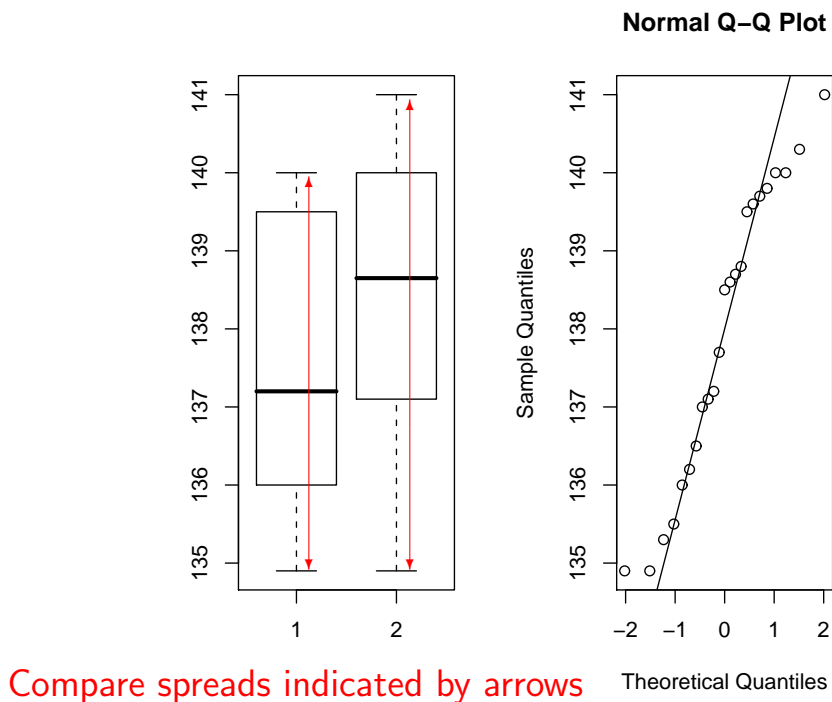
$$\frac{[s_x^2/n_x + s_y^2/n_y]^2}{\frac{(s_x^2/n_x)^2}{n_x-1} + \frac{(s_y^2/n_y)^2}{n_y-1}} \quad \text{as the degrees of freedom.} \quad (11.4)$$

5. Equations (11.1) for the rejection regions of $\bar{x} - \bar{y}$ and (11.2) for the CIs of $\mu_x - \mu_y$ still apply when the standard deviation (sd) and degree of freedom (df) are changed to (11.3) and (11.4) respectively.

Example: (Height comparison)

Solution: We have $n_x = 13$, $n_y = 10$, $\bar{x} = 137.5154$, $\bar{y} = 138.51$, $s_x^2 = 3.368077$ and $s_y^2 = 3.743222$.

We should check the *equality of variance* assumption using *two boxplots* and the *normality* of data assumption using the *qq-plot*.



The plots indicate that the *spreads* are approximately the same for boys and girls, and that the points except 3 outliers are close to normal qq line. Hence the assumptions are only approximately satisfied.

The *2 samples t-test* for the *difference in height between boys and girls at their 10th birthday* is

1. **Hypotheses:** $H_0 : \mu_x = \mu_y$ against $H_1 : \mu_x < \mu_y$.

2. **Test statistic:**
$$t_0 = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} = \frac{137.52 - 138.51}{1.8785 \sqrt{\frac{1}{13} + \frac{1}{10}}} = -1.2588.$$

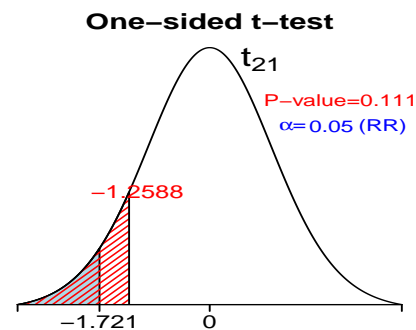
$$\begin{aligned}
 s_p^2 &= \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x + n_y - 2)} \\
 &= \frac{(13 - 1)3.3681 + (10 - 1)3.7432}{(13 + 10 - 2)} = 1.8785^2
 \end{aligned}$$

3. **Assumptions:** $X_i \sim \mathcal{N}(\mu_x, \sigma^2)$ & $Y_i \sim \mathcal{N}(\mu_y, \sigma^2)$. X_i and Y_i are independent.

4. **P-value:**

$$p\text{-value} = \Pr(t_{21} < -1.2588) \in (0.1, 0.25) \quad (0.1110, \text{from R}).$$

5. **Decision:** Since p -value is > 0.05 , the data are consistent with H_0 that the heights of the girls and boys on their 10th birthday are the same.



The rejection region is

$$\begin{aligned}
 \bar{x} - \bar{y} &< -t_{n_x+n_y-2, \alpha} \times s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \\
 &= -1.721 \times 1.8785 \sqrt{\frac{1}{13} + \frac{1}{10}} = -1.359829
 \end{aligned}$$

Since $\bar{x} - \bar{y} = 137.52 - 138.51 = -0.99 > -1.359829$, we accept H_0 .

The 95% one-sided CI for $\mu_x - \mu_y$ is

$$\begin{aligned} & \left(-\infty, (\bar{x} - \bar{y}) + t_{n_x+n_y-2, \alpha} s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \right) \\ &= \left(-\infty, (137.52 - 138.51) + 1.721 \times 1.8785 \sqrt{\frac{1}{13} + \frac{1}{10}} \right) \\ &= (-\infty, 0.3650) \end{aligned}$$

Since $\mu_x - \mu_y = 0 \in (-\infty, 0.3698)$, we accept H_0 .

In R,

```
> x=c(135.3,137.0,136.0,139.7,136.5,137.2,138.8,139.6,140.0,137.7,
      135.5,134.9,139.5)
> y=c(140.3,139.8,138.6,137.1,140.0,136.2,138.7,138.5,134.9,141.0)
> par(mfrow=c(1,2))
> boxplot(x,y)      # joint boxplots of x and y
> qqnorm(c(x,y))     # qqplot of the combined data
> qqline(c(x,y))
> t.test(x,y,alternative="less",mu=0,var.equal=T)
```

Two Sample t-test

```
data:  x and y
t = -1.2588, df = 21, p-value = 0.1110
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
 -Inf 0.3650281
sample estimates:
mean of x mean of y
 137.5154  138.5100

>#just for comparison. The Welch test should NOT be used!
>
> t.test(x,y,alternative="less",mu=0,var.equal=F)
```

Welch Two Sample t-test

```
data:  x and y
t = -1.2497, df = 18.958, p-value = 0.1133
alternative hypothesis: true difference in means is less than 0
90 percent confidence interval:
    -Inf 0.06216388
sample estimates:
mean of x mean of y
 137.5154  138.5100
```

Note that R code for the two-Sample-t-test is

```
t.test(x,y,alternative="??",mu=mu0,paired=F,var.equal=T,conf.level=.95)
```

11.2 Two sample z -test (P.468-470)

Example: (High-fiber cereal) High-fiber cereal manufacturers claim that people who eat high-fiber cereal for breakfast will consume less calories for lunch and hence result in weight reduction for dieters. In the test, 150 people were randomly selected and asked whether they regularly eat high-fiber cereal for breakfast. The number of calories consumed at lunch was also measured. 43 people belongs to the consumer group. The sample means are 604.01 and 633.23 respectively for the consumer and nonconsumer groups. The true standard deviations are known to be 64.05 and 103.29 respectively. Can the manufacturers conclude at 5% significance level that their claims are correct?

Should we use the information of true variances in the test?

The five-steps of the z -test is

1. **Hypotheses:** $H_0 : \mu_x = \mu_y$ vs $H_1 : \mu_x < \mu_y, \mu_x > \mu_y, \mu_x \neq \mu_y$
2. **Test statistic:**
$$z_0 = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$
3. **Assumptions:** X_i are iid $\mathcal{N}(\mu_x, \sigma_x^2)$, Y_i are iid $\mathcal{N}(\mu_y, \sigma_y^2)$, where σ_x^2 and σ_y^2 are known, and the X_i 's are independent of the Y_i 's. Then $Z_0 \sim \mathcal{N}(0, 1)$ under H_0 .
4. **P-value:**
$$\begin{aligned} \Pr(Z \geq z_0) & \quad \text{for } H_1 : \mu_x > \mu_y, \\ \Pr(Z \leq z_0) & \quad \text{for } H_1 : \mu_x < \mu_y \\ 2 \Pr(Z \geq |z_0|) & \quad \text{for } H_1 : \mu_x \neq \mu_y. \end{aligned}$$
5. **Decision:** If $p\text{-value} < \alpha$, there is evidence against H_0 .
If $p\text{-value} > \alpha$, the data are consistent with H_0 .

Remarks:

1. Even if the observations are not normally distributed, the test statistic will be approximately distributed as $\mathcal{N}(0, 1)$ when n_x and n_y are large enough ($n_x, n_y \geq 10$), unless the distributions are particularly long tailed.
2. Equations (11.1) for the rejection regions of $\bar{x} - \bar{y}$ and (11.2) for the CIs of $\mu_x - \mu_y$ still apply when the standard deviation (sd) and the distribution are changed to $\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$ and normal respectively.

Example: (High-fiber cereal)

Solution: We have $n_x = 43$, $n_y = 150 - 43 = 107$, $\bar{x} = 604.01$, $\bar{y} = 633.23$, $\sigma_x = 64.05$ and $\sigma_y = 103.29$. The two samples z -test is

1. **Hypothesis:** $H_0: \mu_x - \mu_y = 0$ vs $H_1: \mu_x - \mu_y < 0$

2. **Test statistic:** $z_0 = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} = \frac{604.01 - 633.23}{\sqrt{\frac{64.05^2}{43} + \frac{103.29^2}{107}}} = -2.09$

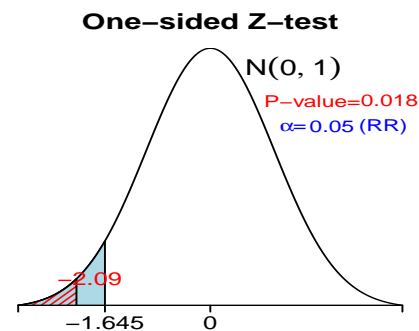
3. **Assumption:** X_i and Y_i are independent. No normality assumption for X_i and Y_i as the sample sizes are large.

4. **P-value:** $\Pr(Z < -2.09) = 1 - 0.9817 = 0.0183$

5. **Decision:** Since $p\text{-value} < 0.05$, there is strong evidence in the data against H_0 . There is weight reduction in calories intake at lunch after eating the high-fiber cereal.

In R,

```
> n1=43
> n2=107
> meanx=604.01
> meany=633.23
> sdx=64.05
> sdy=103.29
> z0=(meanx-meany)/sqrt(sdx^2/n1+sdy^2/n2)
> z0
[1] -2.091880
> p.value=pnorm(z0)
> p.value
[1] 0.01822464
```



12 Wilcoxon rank-sum-test (P.822-826)

12.1 Introduction

Example: (Two methods) The following data yield measurements by two different methods.

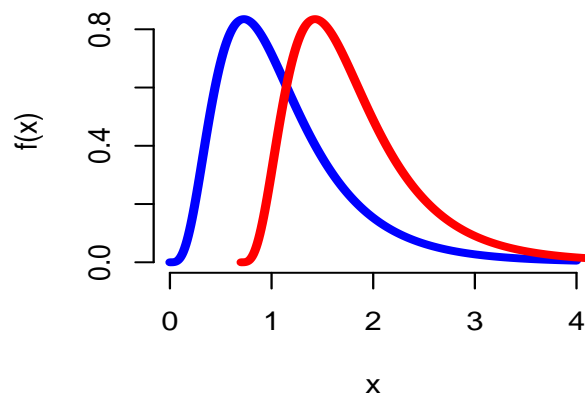
A:	32	29	35	28	
B:	27	31	26	25	30

If the normality assumptions are in doubt, does the data present sufficient evidence to indicate a difference in the methods A and B?

When the normality assumptions fail, should the two-sample test make use of the order information from certain assumed distributions?

The Wilcoxon rank-sum-test is a *non-parametric test* to compare mean based on *two independent samples*. If the boxplots look too skew, a non-parametric test like this test should be used because *it releases normality and even symmetry distribution assumptions*.

Suppose the samples X_1, X_2, \dots, X_{n_x} and Y_1, Y_2, \dots, Y_{n_y} are taken from two distinct populations that follow the same kind of distribution but differ in location, that is, $Y_j = X_j + \theta$ and $\mu_y = \mu_x + \theta$.



Two distributions differ in locations

Let μ_x and μ_y be two population means respectively. We want to test the hypothesis:

$$\begin{aligned} H_0 : \theta = 0 \ (\mu_x = \mu_y) \quad \text{vs} \quad H_1 : \theta > 0 \ (\mu_x > \mu_y), \text{ or} \\ \theta < 0 \ (\mu_x < \mu_y), \text{ or} \\ \theta \neq 0 \ (\mu_x \neq \mu_y). \end{aligned}$$

Let R_1, R_2, \dots, R_N ($N = n_x + n_y$) be the ranks of *combined sample*:

$$X_1, X_2, \dots, X_{n_x}, Y_1, Y_2, \dots, Y_{n_y},$$

For one sample, the ranks are summed over positive side of difference whereas for two samples, the ranks are summed over one of the sample, usually the *smaller* sample (*for using WSR table*), i.e.

$$W = R_1 + R_2 + \dots + R_{n_x},$$

which is the sum of the ranks of the X_j 's.

If H_0 is true, then W should be close to its expected value

$$E(W) = \text{Prop.} \times \text{Total rank} = \frac{n_x}{N} \times \frac{N(N+1)}{2} = \frac{n_x(N+1)}{2}.$$

If W is essentially small (large), we expect $\mu_x < \mu_y$ ($\mu_x > \mu_y$).

The five steps of the test are

1. **Hypothesis:** $H_0 : \mu_x = \mu_y$ vs $H_1 : \mu_x > \mu_y, \mu_x < \mu_y, \mu_x \neq \mu_y$
2. **Test statistic:** $W = R_1 + R_2 + \dots + R_{n_x}$.
3. **Assumption:** X_i and Y_i follow the same kind of distribution differ by a shift.
4. **P-value:** $\Pr(W \geq w)$ for $H_1 : \mu_x > \mu_y$,
 $\Pr(W \leq w)$ for $H_1 : \mu_x < \mu_y$,
 $2 \Pr(W \geq w)$ if $w > \frac{n_x(N+1)}{2}$
 $2 \Pr(W \leq w)$ if $w < \frac{n_x(N+1)}{2}$ for $H_1 : \mu_x \neq \mu_y$.
5. **Decision:** If $p\text{-value} < \alpha$, there is evidence against H_0 .
If $p\text{-value} > \alpha$, the data are consistent with H_0 .

12.2 Calculations of p -value:

Note that mid-ranks are used if there are ties. Let $x = (x_1, \dots, x_{n_x})$ and $y = (y_1, \dots, y_{n_y})$ be observations from two populations. In R,

```
w=sum(rank(c(x,y))[1:nx]).
```

Case 1: There are no ties on the data.

The exact p -value $\Pr(W \leq w)$ is given by the Wilcoxon rank-sum table $(n_x, n_y \leq 8)$, or in R,

```
pwilcox(w-min, nx, ny)
```

where $\min(W) = \underbrace{1 + \dots + n_x}_{n_x} = \frac{n_x(n_x + 1)}{2}$ (NOT 0!) and

$$\max(W) = \underbrace{n_y + 1 + \dots + N}_{n_x} = \frac{n_x(N + n_y + 1)}{2}$$

Note: the distribution of W is symmetric with respect to

$$E(W) = \frac{n_x(N + 1)}{2}$$

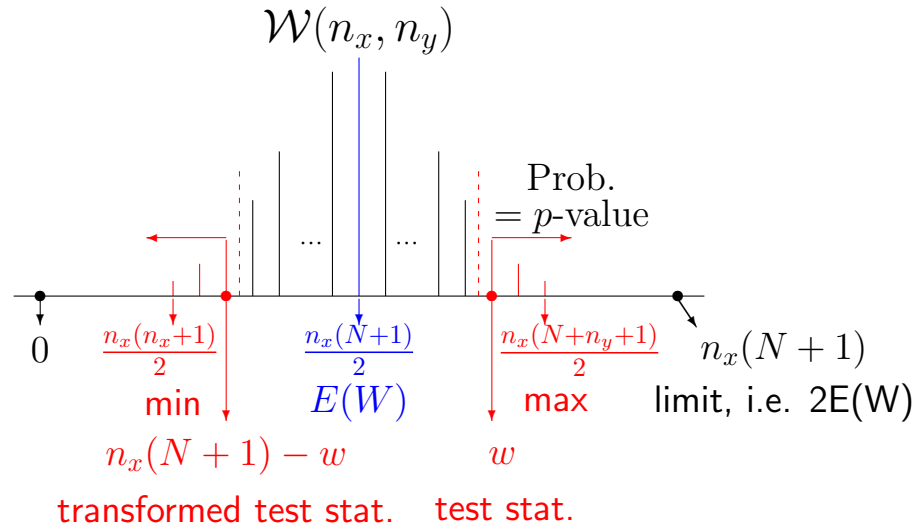
between 0 and $n_x(N + 1)$.

W is the sum of ranks from a smaller sample or sample with smaller ranks if the sample size are smaller since the WRS table is only for $n_x \leq n_y$ and $\frac{n_x(n_x + 1)}{2} \leq w \leq \frac{n_x(N + 1)}{2}$ in the *lower* range of W .

For $\frac{n_x(N + 1)}{2} \leq w \leq \frac{n_x(N + n_y + 1)}{2}$ in the upper range, we need to do *transformation*:

$$\Pr(W \geq w) = 1 - \Pr(W \leq w - 1), \quad (\text{to lower area})$$

$$\Pr(W \geq w) = \Pr(W \leq n_x(N + 1) - w). \quad (\text{to lower quantile})$$



Example: (Two methods)

Solution: Let μ_x and μ_y denote the average of measurements by using methods A and B respectively. The ranks corresponding to methods A and B are given as follows:

A:	32	29	35	28	B:	27	31	26	25	30
Ranks:	8	5	9	4		3	7	2	1	6

We have $n_x = 4$, $n_y = 5$ and $N = 9$. We should sum the ranks from a *smaller* sample and from a sample with lower ranks if the sizes are the same. This avoids having to convert the sum from upper half to lower half. The Wilcoxon rank-sum test is

- Hypotheses:** $H_0 : \mu_x = \mu_y$ vs $\mu_x \neq \mu_y$.
- Test statistic:** $W = W_x = 8 + 5 + 9 + 4 = 26$ because we sum smaller sample ($W_y = 3 + 7 + 2 + 1 + 6 = 19$).
- Assumption:** X_i and Y_i follow the same kind of distribution differ

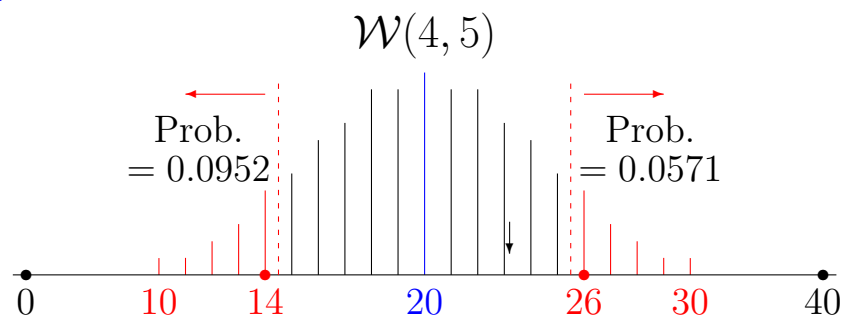
by a shift in location.

4. ***P*-value:**

$$\begin{aligned} p\text{-value} &= 2 \Pr(W \geq 26) \\ &= 2 \Pr(W \leq n_x(N+1) - 26) = 2 \Pr(W \leq 4(9+1) - 26) \\ &= 2 \Pr(W \leq 14) = 2 \times 0.0952 = 0.1905 \end{aligned}$$

(From WRS table with $n_1 = 4$, $n_2 = 5$, $w = 14$)

5. **Decision:** Since the p -value > 0.05 . The data is consistent with H_0 . There are no differences between measurements using methods A and B.



Note: This is the exact distribution from WRS table.

$$E(W) = \frac{n_x(N+1)}{2} = \frac{4 \times (9+1)}{2} = 20,$$

$$\text{Limit or } 2E(W) = n_x(N+1) = 4 \times (9+1) = 40,$$

$$\text{Min}(W) = \frac{n_x(n_x+1)}{2} = \frac{4(5)}{2} = 10 \text{ and}$$

$$\text{Max}(W) = \frac{n_x(N+n_y+1)}{2} = \frac{4(9+5+1)}{2} = 30.$$

In R,

```
> A=c(32,29,35,28)
> B=c(27,31,26,25,30)
> wilcox.test(A,B,alternative="two.sided",mu=0,exact=T,correct=F)
```


Wilcoxon rank sum test

data: A and B

W = 16, p-value = 0.1905

alternative hypothesis: true location shift is not equal to 0

```
> nx=length(A) #checking only
> ny=length(B)
> N=nx+ny
> c(nx,ny,N)
[1] 4 5 9
> rank=rank(c(A,B))
> rank
[1] 8 5 9 4 3 7 2 1 6
> rankA=rank(c(A,B))[1:nx]
> rankA
[1] 8 5 9 4
> rankB=rank(c(A,B))[(nx+1):N]
> rankB
[1] 3 7 2 1 6
> w=sum(rankA)
> min=nx*(nx+1)/2
> max=nx*(N+ny+1)/2
> Ew=nx*(N+1)/2
> wt=2*Ew-w #transform to lower sided
> w0=w-min #test stat in Wilcox.test W-min=26-10=16
> c(w,wt,w0,Ew,min,max)
[1] 26 14 16 20 10 30
> p.value=2*pnwilcox(wt-min,nx,ny)
> p.value
[1] 0.1904762
```

Case 2: There are *ties* in the data.

The p -value can be calculated using normal approximation to the distribution of test statistic W' or derive the exact distribution of W' based the observed set of ranks.

Central limit theorem should be applied and the test statistic is:

$$\frac{W - E(W)}{\sqrt{Var(W)}} \sim \mathcal{N}(0, 1), \quad \text{approximately,}$$

where $E(W) = \frac{n_x(N+1)}{2}$, and

$$Var(W) = \frac{n_x n_y}{N(N-1)} \left(\sum_{i=1}^N r_i^2 - \frac{N(N+1)^2}{4} \right).$$

(Proof is given in Tutorial 5.) Then the approximate p -value is

$$p\text{-value} \approx \Pr \left(Z > \frac{w - E(W)}{\sqrt{Var(W)}} \right) \quad \text{for } H_1 : \mu_x > \mu_y;$$
$$p\text{-value} \approx \Pr \left(Z < \frac{w - E(W)}{\sqrt{Var(W)}} \right) \quad \text{for } H_1 : \mu_x < \mu_y;$$
$$p\text{-value} \approx \Pr \left(Z > \left| \frac{w - E(W)}{\sqrt{Var(W)}} \right| \right) \quad \text{for } H_1 : \mu_x \neq \mu_y.$$

Note:

1. We use normal approximation for the *test statistic* W NOT the *data* X_i, Y_i .
2. As we do not consider sign, zero measurements should be ranked in the same way as other measurements.

Example: (latent heat of fusion) Two methods labelled A and B are used to measure the latent heat of fusion of ice (data available in R under the names `icea`, `iceb`). Does the data support the assumption that A gives larger results?

A	79.98	80.04	80.02	80.04	80.03	80.03	80.04	79.97
	80.05	80.03	80.02	80.00	80.02			
B	80.02	79.94	79.98	79.97	79.97	80.03	79.95	79.97

Solution: Let μ_x and μ_y denotes the average latent heat of fusion of ice corresponding to methods A and B respectively. The ranks are

rankA 7.5 19.0 11.5 19.0 15.5 15.5 19.0 4.5 21.0 15.5 11.5 9.0 11.5
rankB 11.5 1.0 7.5 4.5 4.5 15.5 2.0 4.5

We have $n_x = 13$, $n_y = 8$ and $N = 21$. The Wilcoxon rank-sum test for the difference between methods A and B is

- Hypotheses:** $H_0 : \mu_x = \mu_y$ vs $\mu_x > \mu_y$.
- Test statistic:** $W = W_x = 180$,
 $W_y = \frac{N(N+1)}{2} - W_x = \frac{21 \times 22}{2} = 231 - 180 = 51$
- Assumption:** X_i and Y_i follow the same kind of distribution, differ by a shift.
- P-value:** With normal approximation to the test statistic W :

$$E(W) = \frac{n_x(N+1)}{2} = \frac{13 \times (13+8+1)}{2} = 143$$

$$g = \frac{N(N+1)^2}{4} = \frac{(13+8)(13+8+1)^2}{4} = 2541$$

$$\begin{aligned} \text{Var}(W) &= \frac{n_x n_y}{N(N-1)} \left(\sum_{i=1}^N r_i^2 - \frac{N(N+1)^2}{4} \right) \\ &= \frac{13(8)(3293.5 - 2541)}{21(20)} \quad \text{since } 7.5^2 + \dots + 4.5^2 = 3293.5 \\ &= 186.33 \\ p\text{-value} &= \Pr(W \geq w) = \Pr\left(Z > \frac{w - E(W)}{\sqrt{\text{Var}(W)}}\right) \\ &= \Pr\left(Z > \frac{180 - 143}{\sqrt{186.33}}\right) = \Pr(Z > 2.710544) \\ &= 0.00336 \end{aligned}$$

5. **Decision:** Since the p -value < 0.05 . There is strong evidence in the data against H_0 . Measurements using method A gives larger results than those from method B.

Note: when the WRS table is not used, it does NOT matter whether the ranks from a smaller or larger sample are summed or whether the sum is in the lower or upper range. When the smaller sample is used,

$$\begin{aligned} E(W_y) &= \frac{n_y(N+1)}{2} = \frac{8 \times (13 + 8 + 1)}{2} = 88, \\ p\text{-value} &= \Pr\left(Z \leq \frac{51 - 88}{\sqrt{186.33}}\right) = \Pr(Z < -2.710544) = 0.00336 \end{aligned}$$

Note: the variance and p -value are the same as using the larger sample. Moreover sum of all ranks is $E(W_x) + E(W_y) = 88 + 143 = 231$, NOT twice of 143 because the sample sizes are unequal.

In R,

```
> icea=c(79.98,80.04,80.02,80.04,80.03,80.03,80.04,79.97,80.05,80.03,
          80.02,80.00,80.02)
> iceb=c(80.02,79.94,79.98,79.97,79.97,80.03,79.95,79.97)
> wilcox.test(icea,iceb,alternative="greater",mu=0,exact=F,correct=F)
```

Wilcoxon rank sum test

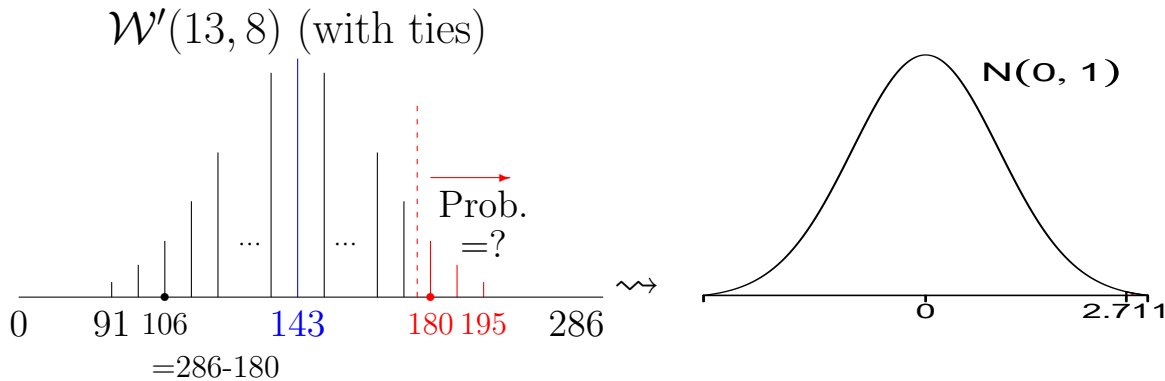
data: icea and iceb

W = 89, p-value = 0.003359

alternative hypothesis: true location shift is greater than 0

```
> rank=rank(c(icea,iceb))    #checking only
> nx=length(icea)
> ny=length(iceb)
> N=nx+ny
> c(nx,ny,N)
[1] 13 8 21
> rankA=rank(c(icea,iceb))[1:nx]
> rankA
[1]  7.5 19.0 11.5 19.0 15.5 15.5 19.0  4.5 21.0 15.5 11.5  9.0 11.5
> rankB=rank(c(icea,iceb))[(nx+1):N]
> rankB
[1] 11.5  1.0  7.5  4.5  4.5 15.5  2.0  4.5
> sum(rankB)
[1] 51
> w=sum(rankA)
> EW=nx*(nx+ny+1)/2
> sumsqrnk=sum(rank^2)
> g=N*(N+1)^2/4
> varW=nx*ny*(sumsqrnk-g)/(N*(N-1))
> z0=(w-EW)/sqrt(varW)
> p.value=1- pnorm(z0)
> min=nx*(nx+1)/2
> max=nx*(N+ny+1)/2
> w0=w-min
```

```
> c(w,min,max,w0,EW) #w0=180-91 is reported in WRS test
[1] 180 91 195 89 143
> round(c(sumsqrnk,g,varW,z0,p.value),digit=5)
[1] 3293.50000 2541.00000 186.33333 2.71054 0.00336
```



This is a rough sketch because, for example, there are bars for non-integral rank sums.

If we swap the order of A and B and change “greater” to “less” for alternate we have:

```
> wilcox.test(iceb,icea,alternative="less",mu=0,exact=F,correct=F)
```

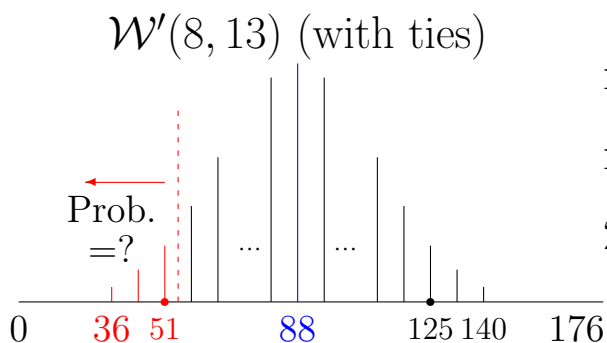
Wilcoxon rank sum test

data: iceb and icea

W = 15, p-value = 0.003359

alternative hypothesis: true location shift is less than 0

We get the same p -value. The distribution of W_y is



$$\min(W_y) = \frac{n_y(n_y+1)}{2} = \frac{8 \times 9}{2} = 36$$

$$\max(W_y) = \frac{n_y(N+n_x+1)}{2} = \frac{8(21+13+1)}{2} = 140$$

$$2E(W_y) - W_y = 88 \times 2 - 51 = 125$$