Math 152 4/~~8~~14

Comparing Two independent Samples.

Boxplots

ordered data are plotted vertically on a line.

1. horizontal lines are drawn at median and upper and lower quartiles
   Joined to make a box.
2. vertical line(s) from top (bottom) of box to most extreme observation within 1.5 IQR of top (bottom).
   short horizontal lines (whiskers) mark the ends
3. data points beyond the whiskers are marked with an asterisk.

computer examples.

A non-parametric method (Mann-Whitney test).

m+n experimental units assigned randomly to treatment (m) and control (n) groups.

H_0: treatment has no effect.

If H_0 holds then differences between the two groups are due to the randomization.

Test Statistic

1. group all $m+n$ observations and rank them in order of increasing size. (assume at first that there are no ties)

2. Calculate the sum R of the ranks of observations from the control group.

Reject the null if R is too large or too small.

2.5 If there are ties, assign averaged ranks to the tied observations.

e.g. $m = n = 2$.

Observations

Treatment	Control
1	6
3	4

Ranks.

Treatment	Control
1	4
2	3

$R = 7$.

What is the Null distribution of R.?

Under the null, each possible assignment of ranks, ($\binom{4}{2}$ of them), is equally likely

Ranks	R
{1, 2}	3
{1, 3}	4
{1, 4}	5
{2, 3}	5
{2, 4}	6
{3, 4}	7.

r	3	4	5	6	7
Prob (R=r)	1/6	1/6	1/3	1/6	1/6.

Prob (R=7) = 1/6.

In general, under H_0, each of

$\binom{m+n}{m}$ assignments of ranks to the control group is equally likely

For each of these assignments are can compute R and ∴ know the null dist. of R.

Notice that there is no assumption on the distribution of the data. The randomization ~~~~ in assignment to treatment and control allows us to analyze the test statistic probabilistically.

A table of the distribution of R

We reject H_0 for extreme (large or small) values of R.

Some tables:

Since the sum of all ranks is

$$\frac{(m+n)(m+n+1)}{2}.$$

knowing one rank sum tells us the other.

tables:

sometimes:
- smaller of the two groups
- smaller of two rank sums.

Table 8 Appendix B uses symmetry:

Suppose n_1 is the smaller of the two sample sizes and let R be the sum of the ranks of that sample.

~~We would do reject for small values~~

If we inverted the ranking going from largest to smallest instead, a rank

R (smallest to largest) would be changed to.

$$m+n-k+1.$$

(distance to top ranking in the 1st ordering).

and

$$\sum_{k \text{ in smaller group}} m+n-k+1 = n_1(m+n+1) - R$$

let $R^* = \min(R, R')$.

We will reject for small values of R^*.

e.g. latent heat of fusion of ice.

$n = 13 \quad m = 8$

method →	A	B
ranks →	7.5	11.5
	19.0	1.0
	11.5	7.5
	19.0	4.5
	15.5	4.5
	15.5	15.5
	19.0	2.0
	4.5	4.5
	21.0	
	15.5	$R = 51.$
	11.5	
	9.0	
	11.5	

ranks 3, 4, 5, 6 were tied at 79.87.

$$\frac{3+4+5+6}{4} = 4.5.$$

ranks 18, 19, 20 were tied.

$$\frac{18+19+20}{3} = 19.$$

$$R' = 8(8+13+1) - R = 125.$$

$R^* = 51.$, with $n_1 = 8$, $n_2 = 13$

we reject at $\alpha = .05$ for $R^* \leq 60$

at $\alpha = .01$ for $R^* < 53$.

When we do have distributions in mind, we think of.

control values $X_1, \ldots, X_n$ from a c.d.f F.

test values $Y_1, \ldots Y_m$ from a c.d.f G.

and $H_0: F = G$.

Let T_Y = sum of ranks of $(Y_1, \ldots, Y_m)$.

Theorem:

If $F = G$ then

$$E(T_Y) = \frac{m(m+n+1)}{2}$$

$$Var(T_Y) = \frac{mn(m+n+1)}{12}.$$

Pf: Under H_0

T_y is the sum of a random sample of size m taken without replacement from a population $\{1, \ldots, m+n\}$.

T_y then is just $m \cdot$ (sample mean).

So we know (chpt 7).

$$E(T_y) = m\mu.$$

$$\mathrm{Var}(T_y) = \frac{m^2\sigma^2}{m}\left(\frac{N-m}{N-1}\right) = m\sigma^2\left(\frac{N-m}{N-1}\right)$$

where $N = m+n$

$$\mu = \frac{1}{m+n}\sum_{k=1}^{m+n} k.$$

$$\sigma^2 = \frac{1}{m+n}\sum_{k=1}^{m+n} k^2 - \left(\frac{1}{m+n}\sum_{k=1}^{m+n} k\right)^2$$

We have $$\sum_{k=1}^{N} k = \frac{N(N+1)}{2}$$

$$\sum_{k=1}^{N} k^2 = \frac{N(N+1)(2N+1)}{6}$$

So $\mu = \frac{N+1}{2}$

and

$$\sigma^2 = \frac{(N+1)(2N+1)}{6} - \left(\frac{N+1}{2}\right)^2$$

$$= \frac{N+1}{2}\left(\frac{2N+1}{3} - \frac{N+1}{2}\right)$$

$$= \frac{N+1}{2}\left(\frac{4N+2}{6} - \frac{3N+3}{6}\right)$$

$$= \frac{(N+1)(N-1)}{12} = \frac{N^2-1}{12}.$$

Plugging these in we get the claimed values.

An alternative approach to Mann-Whitney

X's sampled from F (c.d.f)

Y's sampled from G (c.d.f).

$\longrightarrow \pi = P(X < Y)$

not the famous constant, just an unknown.

can be considered a measure of the effect of treatment

e.g. X, Y lifetimes of components manufactured in two different ways.

$$\pi = P(X < Y)$$

is the prob that a Y component lasts longer.

How can we estimate π?

For all possible pairs i, j

$$1 \le i \le n, \quad 1 \le j \le m.$$

let

$$Z_{ij} = \begin{cases} 1 & \text{if } X_i < Y_j \\ 0 & \text{else.} \end{cases}$$

and

let

$$\hat{\pi} = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij}.$$

This is the proportion of all pairs s.t $(X < Y)$.

It's convenient to order the samples and take

$$V_{ij} = \begin{cases} 1 & \text{if } X_{(i)} < Y_{(j)} \\ 0 & \text{else} \end{cases}$$

then

$$\sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{m} V_{ij}.$$

Now

$$\sum_{i=1}^{n} \sum_{j=1}^{m} V_{ij} = \#(X\text{'s less than } Y_{(1)}) + \#(X\text{'s less than } Y_{(2)}) + \vdots + \#(X\text{'s less than } Y_{(m)}).$$

Let R_{yk} = rank of $Y_{(k)}$ in the combined sample.

then

$\#(X\text{'s less than } Y_{(1)})$ is

$$R_{y1} - 1.$$

$$\#(X\text{'s less than } Y_{(2)}) = R_{y2} - 2$$

$$\vdots$$

$$\#(X\text{'s less than } Y_{(m)}) = (R_{ym} - m)$$

So $$\sum_{i=1}^{n}\sum_{j=1}^{m} V_{ij} = (R_{y1}-1) + \dots + (R_{ym}-m)$$

$$= \sum_{i=1}^{m} R_{yi} - \sum_{i=1}^{m} i$$

$$= T_y - \frac{m(m+1)}{2}.$$

Let $U_y = mn\hat{\pi} = \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m} V_{ij}$

$$= \#\left((i,j) : X_i < Y_j\right).$$

then.

$$U_y = T_y - \frac{m(m+1)}{2}.$$

and

$$E(U_r) = \frac{mn}{2}.$$

$$Var(U_r) = \frac{mn(m+n+1)}{12}.$$

If m, n are moderately large. ($\geq m, n \geq 10$).

Null
the distribution of U_r is approximately normal (even though the Z_{ij} are not independent, they are approximately so).

and we have.

$$\frac{U_r - E(U_r)}{\sqrt{Var(U_r)}} \sim N(0,1).$$

This gives a method for approximating the distribution of the rank sum.

e.g. (our previous example).

$n = 13$, $m = 8$ (close enough).

$$E(T) = \frac{8(8+13+1)}{2} = 88.$$

$$\sigma_T = \sqrt{\frac{8 \cdot 13 \cdot (8+13+1)}{12}} \simeq 13.8.$$

$$T = 51.$$

$$\frac{T - E(T)}{\sigma_T} = -2.68.$$

p-value $\simeq .007$

for a two sided test.

($X \sim N(0,1)$) $P(X < -2.68) + P(X > +2.68)$

$P(X > 2.68) = 1 - P(X \leq 2.68)$.

$= .0037$

So the Null is rejected at $\alpha = .01$.

Remark:

The standard error of $\hat{\pi}$ can be estimated and an approximate confidence interval for π constructed by the bootstrap method.

X's from F

Y's from G.

Sample from F
from G to simulate $\hat{\pi}$.

Not knowing F, G.

uses $F_{(n)}, G_{(m)}$ instead.