Math 152 4/16/09.

Last time:

Mann-Whitney test

$X_1, \ldots X_n$ a sample from F (c.d.f)

$Y_1, \ldots Y_m$ a sample from G (c.d.f).

$H_0: F = G$

Assign ranks to the $m+n$ pooled observations

Let R be the sum of the ranks of the (say) smaller group.

R or related quantities

$(R' = n_1(m+n+1) - R$

$n_1 = \min(m,n))$

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

is used as a test statistic for H_0.

Last time we saw that with

T_Y = sum of ranks of the Y's

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

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

Considering the X's and Y's to come from a treatment group and a control group, we are interested in

$$\pi \equiv P(X < Y).$$

as a measure of the effect of the treatment.

and we estimate π from the sample by

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

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

We may as well (and it is more convenient to) consider

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

then

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

and

$$\sum_{i=1}^{n} \sum_{j=1}^{m} V_{ij} = \left(\#\ X\text{'s} < Y_{(1)}\right) + \left(\#\ X\text{'s} < Y_{(2)}\right) + \ldots + \left(\#\ X\text{'s} < Y_{(m)}\right)$$

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

Then $\left(\#\ X\text{'s} < Y_{(1)}\right) = R_{y1} - 1$

$\left(\#\ X\text{'s} < Y_{(2)}\right) = R_{y2} - 2$

$\vdots$

$\left(\#\ X\text{'s} < Y_{(m)}\right) = R_{ym} - m$

So

$$\sum_{i=1}^{n} \sum_{j=1}^{m} V_{ij} = (R_{y_1} - 1) + (R_{y_2} - 2) + \ldots + (R_{y_m} - m)$$

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

$$= \sum_{i=1}^{m} R_{y_i} - \frac{m(m+1)}{2}$$

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

and

$$\hat{\pi} = \frac{1}{mn}\left(T_y - \frac{m(m+1)}{2}\right).$$

With $U_y = \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij}$ we

have (under $H_0: F = G$).

$$E(U_y) = \frac{m(m+n+1)}{2} - \frac{m(m+1)}{2} = \frac{mn}{2}$$

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

Since $U_Y = \sum_i \sum_j Z_{ij}$

and the Z_{ij} are identically distributed and approximately independent

we have.

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

Since T_Y is a linear function of U_Y we also have.

$$\frac{T_Y - E(T_Y)}{\sigma_{T_Y}} \overset{(approx)}{\sim} N(0,1).$$

Our example from last time on latent heat of fusion of ice.

Method A		Method B	
79.98	7.5	80.02	11.5
80.04	19.0	79.94	1.0
80.02	11.5	79.98	7.5
80.04	19.0	79.97	4.5
80.03	15.5	79.97	4.5
80.03	15.5	80.03	15.5
80.04	19.0	79.95	2.0
79.97	4.5	79.97	4.5
80.05	21.0		
80.03	15.5		
80.02	11.5		
80.00	9.0		
8.02	11.5		

$m = 8$, $n = 13$

T = sum of ranks of method B.

Last time we computed.

$$R = 51 \quad (= T)$$

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

$$R^* = \min(R, R') = 51$$

and used table 8 appendix B to see that 53 is the rejection value for $\alpha = .01$.

We can use "R" instead to get a p-value $\approx .005633$

The normal approximation just derived gives, with

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

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

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

$$\text{Prob}\left(\left| \frac{T - E(T)}{\sigma_T} \right| \geq 2.68 \right) \approx .007.$$