# Lec 4/10

Monday, April 10, 2017 15:01

From Friday

**Theorem 16.2** Under Assumptions Required by U test $H_0$,

$$\mathbb{E}U_1 = \mathbb{E}U_2 = \frac{n_1 n_2}{2}$$

$$Var(U_1) = Var(U_2) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

Proof:

$W_1$ = Sum of ranks in 1st sample, $U_1 = W_1 - \frac{n_1(n_1-1)}{2}$

$$W_1 = \sum_{i=1}^{n_1+n_2} i \cdot \mathbb{1}_{\{\text{rank } i \text{ is in sample } 1\}}$$

$$\Rightarrow \mathbb{E}W_1 \sum_{i=1}^{n_1+n_2} i \, \underbrace{\mathbb{P}(\text{rank } i \text{ is in sample } 1)}_{\frac{n_1}{n_1+n_2}} = \frac{n_1}{n_1+n_2} \sum_{i=1}^{n_1+n_2} i = \frac{n_1(n_1+n_2+1)}{2}$$

$$\Rightarrow \mathbb{E}U_1 = \frac{n_1(n_1+n_2+1)}{2} - \frac{n_1(n_1-1)}{2} = \frac{n_1 n_2}{2}$$

(same for $U_2$).

Purpose of the theorem: find $\mu$ and $\sigma^2$ of Normal approx to $U_1$ or $U_2$ under $H_0$

Book says if $n_1, n_2 > 8$ then $U_1, U_2 \sim$ Normal under $H_0$.

Ex: Burning flares

Comparing 2 brands of flares A and B.

$H_0$: distributions are the same

$H_1$: $\mu_1 < \mu_2$

$W_1 = 69$

$U_1 = 69 - \frac{n_1(n_1+1)}{2} = 24$

approx P-value: $z = \frac{U_1 - \mathbb{E}U_1}{\sqrt{Var(U_1)}} \sim N(0,1)$

Plugging in observed values:

$$Z = \frac{24 - \frac{9+10}{2}}{\sqrt{\frac{9\cdot 10\cdot (9+10+1)}{12}}} = -1.714$$

p-value $\approx 0.04 \Rightarrow$ strong evidence in favor of $\mu_1 < \mu_2$.

## §16.5: H-test (Kruskal-Wallis test) AKA Rank-sum test.

Generalization of Rank-sum test to $K > 2$ samples.

Want to test whether samples are from same population.

(sample sizes are $n_1, \dots, n_k$, $\sum_{i=1}^{k} n_i = n$

Steps:

1: Pool all observations & rank from smallest (1) to largest (n)

2: Count ranks in each sample $R_i$ = sum of ranks in sample $i$.

3: Compute a test statistic $H = \left[\frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i}\right] - 3(n+1)$

4: Reject $H_0$: all samples come from same population $(\mu_1 = \mu_2 = \dots = \mu_k)$

$H_1$: $\mu_i \neq \mu_j$ for some $i \neq j$

When H is large.

Since the distribution of H under $H_0$ depends on $n_1, n_2, \dots, n_k$, it can be impractical to tabulate it. Instead, we'll use a large sample approximation when n is large. $H \sim \chi^2_{K-1}$

$$\left\{\frac{12}{n+1} \sum_{i=1}^{k} \frac{n_i}{n} \left(\frac{R_i}{n_i} - \frac{n+1}{2}\right)^2\right.$$

Recall: $Var(A) = E(A^2) - (EA)^2$

$\frac{R_i}{n_i}$: average rank of sample $i$

$\frac{n+1}{2}$: average rank of all data.

weighted avg of squared difference of these things

$$= \frac{12}{n+1}\left\{\left(\sum_{i=1}^{k} \frac{n_i}{n}\frac{R_i^2}{n_i^2}\right) - \left(\sum_{i=1}^{k} \frac{n_i}{n}\frac{R_i}{n_i}\right)^2\right\}$$

$$= \frac{12}{n+1}\left\{\left(\sum_{i=1}^{k}\frac{n_i}{n}\frac{R_i^2}{n_i^2}\right) - \left(\sum_{i=1}^{k}\frac{n_i}{n}\frac{R_i}{n_i}\right)^2\right\}$$

$$= \frac{12}{n+1}\left\{\frac{1}{n}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - \frac{1}{n^2}\left(\frac{n(n+1)}{2}\right)^2\right\}$$

$$= \left[\frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right] - 3(n+1)$$

So $H$ is a weighted avg of squared diffs and is $\geq 0$.

Example: Time until failure of 3 types of stopwatches (thousands of cycles)

| Type 1 | Type 2 | Type 3 |
|---|---|---|
| ⋮ | ⋮ | ⋮ |
| $R_1 = 76.5$ | $R_2 = 78$ | $R_3 = 55.5$ |
| $n_1 = 9$ | $n_2 = 6$ | $n_3 = 5$ |

$H = 2.15$.

Under $H_0$, $H \sim \chi^2_2$ so the p-value is $> 0.05$

so don't reject $H_0$.