10.13. Test for homogeneity in populations

Example. Humans red ~~blood~~ blood cells can be divided into 4 hereditary blood-types: A, B, AB and 0. An international health organization wishes to find out if the relative occurrence is the same in 3 particular ethnic groups: G1, G2 and G3. Random samples of sizes 400, 200 and 100 persons are drawn from the 3 groups and their blood type is decided for each person.

Contingency Table

Blood type \ Group		G1	G2	G3	Total
B1	A	176 (192)	112 (96)	48 (48)	336
B2	B	41 (37.14)	16 (18.57)	8 (9.29)	65
B3	AB	19 (16.57)	6 (8.29)	4 (4.14)	29
B4	0	164 (154.29)	66 (77.14)	40 (38.57)	270
Total		400	200	100	700

Define $p_{ij} = P(BI \cap GJ)$, $i = 1,2,3,4$, $j = 1,2,3$

Homogenity means $p_{ij} = p_i$, $j = 1,2,3$ for all i

$\hat{p}_1 = \frac{336}{700}$, $\hat{p}_2 = \frac{65}{700}$, $\hat{p}_3 = \frac{29}{700}$ and $\hat{p}_4 = \frac{270}{700} = 1 - \hat{p}_1 - \hat{p}_2 - \hat{p}_3$

Let m_{ij} be estimated expected number in cell ij,

$$m_{ij} = \underbrace{n_j}_{\text{total for column } j} \cdot \frac{n_i}{700} = n_j \cdot \hat{p}_i$$

$H_0: p_{i1} = p_{i2} = p_{i3} = p_i \qquad i = 1, 2, \ldots 4$

Degrees of freedom: $(r-1)c - (r-1) = (r-1)(c-1) = 3 \cdot 2 = 6$

$$Q = \frac{(176-192)^2}{192} + \frac{(41-37.14)^2}{37.14} + \frac{(19-16.57)^2}{16.57} + \frac{(164-154.29)^2}{154.29}$$

$$+ \frac{(112-96)^2}{96} + \frac{(16-18.57)^2}{18.57} + \frac{(6-8.29)^2}{8.29} + \frac{(66-77.14)^2}{77.14}$$

$$+ \frac{(48-48)^2}{48} + \frac{(8-9.29)^2}{9.29} + \frac{(4-4.14)^2}{4.14} + \frac{(40-38.57)^2}{38.57} = 8.2,$$

$\chi^2_{0.05}(6) = 12.59 \Rightarrow$ no reason to reject H_0 on a 5% level.

Nonparametric Statistics

Nonparametric tests.

Tests where no or very little knowledge of the distributions are assumed. These are often formulated on the median, $\tilde{\mu}$, instead of the mean.

Note. The t-test and the F-test is robust to slight deviations from normality.

The sign test (Assumes only continuous distributions)

Example.

A rectangle where the ratio between the height and the length is 0.618 (the golden ratio) are called a golden rectangle. Did the Shoshoni indians know about this?
Below is given 8 such ratios found on leather goods made of Shoshoni indians.

x_i: 0.693 0.662 0.690 0.606 0.570 0.749 0.672 0.628

Want to test the following hypothesis

$H_0: \tilde{\mu} = 0.618 \qquad H_1: \tilde{\mu} \neq 0.618$

x_i	$x_i - 0.618$	signs of difference
0.693	0.075	+
0.662	0.044	+
0.690	0.072	+
0.606	-0.012	−
0.570	-0.048	−
0.749	0.131	+
0.672	0.054	+
0.628	0.008	+

Under H_0: $P(X < 0.618) = P(X > 0.618) = 0.5$

Let Y be the number of rectangles with ratio less than 0.618. $Y \sim B(8, 0.5)$. $Y_{obs} = 2$.

The p-value is $2\,P(Y \le 2) = \underline{0.289}$

<u>Assume the distribution is continuous and symmetric</u>

Then we can use the Wilcoxon signed-rank test.

We then rank the observations $(x_i - \tilde{\mu}_0)$ according to their absolute values.

We get:

	0.008	0.012	0.044	0.048	0.054	0.072	0.075	0.131
Rank:	1	2	3	4	5	6	7	8
Sign:	+	−	+	−	+	+	+	+

Let W_+ be the rank of those with a positive sign and

W_- — " — — " — negative sign

and let W be the smallest of W_+ and W_-

<u>For testing</u>:

H_0	H_1		
$\tilde{\mu} = \tilde{\mu}_0$	$\tilde{\mu} < \tilde{\mu}_0$	use	W_+
	$\tilde{\mu} > \tilde{\mu}_0$	use	W_-
	$\tilde{\mu} \neq \tilde{\mu}_0$	use	W

In this case $W_- = 6$ and $W_+ = 30 \Rightarrow W = W_-$

There are ($2^8 = 256$) possible ways to place out signs

$P(W_- = 0) = \frac{1}{2^8}$ $\quad P(W_- = 3) = \frac{2}{2^8}$ $\quad P(W_- = 6) = \frac{3}{2^8}$

$P(W_- = 1) = \frac{1}{2^8}$ $\quad P(W_- = 4) = \frac{2}{2^8}$ $\quad \Rightarrow P(W_- \le 6) = \frac{13}{2^8} = \frac{13}{256}$

$P(W_- = 2) = \frac{1}{2^8}$ $\quad P(W_- = 5) = \frac{3}{2^8}$

For a two sided test. this gives a p-value of $\frac{26}{256} > 0.05$

$\Rightarrow$ no reason to reject H_0 at a 5% level.

In order to avoid a lot of tedious calculations the critical values are given in Table A.16.

It says that H_0 should be rejected if. W (or W_-) ≤ 4.

<u>Approximation to the normal distribution</u>

Now. let $U = W_+ - W_-$

and define $I_j = \begin{cases} 1 & \text{if. } X_j - \tilde{\mu}_0 > 0 \\ -1 & \text{if } X_j - \tilde{\mu}_0 < 0 \end{cases}$

Then $U = \sum_{j=1}^{n} \underbrace{R_j}_{\text{Rank of } |X_j - \tilde{\mu}_0|} I_j = \sum_{j=1}^{n} j \, I_j$ in distribution.

Under H_0 $P(I_j = 1) = P(I_j = -1) = \frac{1}{2}$ and we get.

$E[I_j] = 0$, $Var[I_j] = 1$

and $E[U] = 0$ and $Var[U] = \sum_{j=1}^{n} j^2 \cdot 1 = \frac{n(n+1)(2n+1)}{6}$.

$$W_- + W_+ = \sum_{j=1}^{n} j = \frac{n(n+1)}{2}.$$ and since $W_- = W_+ - U$

we get $2W_+ = U + \frac{n(n+1)}{2}$.

Therefore $E[W_+] = \frac{n(n+1)}{4}$ and $Var[W_+] = \frac{n(n+1)(2n+1)}{24}$.

For $n \geq 15$ ~~the distribution~~ of W_+ (and W_-) $\approx N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$,