NON-PARAMETRIC METHODS

Dr. Noah Mutai

Introduction

- In statistical inference (theory of estimation and hypothesis testing) we assume that the observations come from a distribution for which the exact form is known even though the values of some parameters are unknown e.g.

(i) We might assume that the observations form a random sample from a Poisson distribution for which the mean is unknown i.e. $x_1, x_2 \ldots x_n \sim \text{Poisson}(\theta)$

(ii) It might be assumed that the observations come from two Normal distributions for which the means and variances are unknown

- In the above cases we have assumed that the observations come from certain parametric family of distributions and statistical inference must be made about the values of the parameter defining the family.
- In many of the problems to be discussed i.e. non-parametric methods, we shall not assume that the available observations come from a particular parametric family of distributions.

Examples

1) We may simply assume that the observations form a random sample from a continuous distribution without specifying the form of this distribution and further we might investigate the possibility that it is a normal distribution.

2) We may be interested in making inference about the value of the median of the distribution from which the sample was taken and we may assume that this is a continuous distribution.

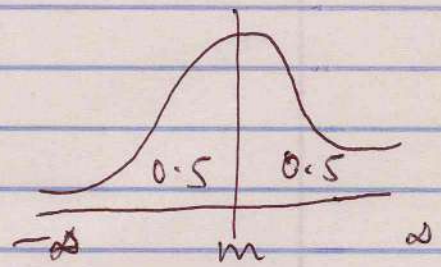
3) We might be interested that two independent ~~so~~ samples come from same distribution and we might only assume that both distributions are continuous.

Definition: Non-parametric methods: Statistical analyses which does not depend upon the knowledge of the distribution and parameters of the population are called non-parametric or distribution free methods.

The sign test

- It is the non-parametric alternative to t-test for hypothesis testing. It is so named because it counts +ve and -ve to test hypothesis.
- This test is performed when we wish to analyse two sets of data that were gathered independently. ~~use from~~
- To describe the test we assume that the pdf is continuous density and assume that the median denoted by m is uniquely defined by

$$\int_m^{\infty} f(x)\,dx = \tfrac{1}{2}$$



Note: $\int_{-\infty}^{\infty} f(x)\,dx = 1$

$$\Rightarrow \int_{-\infty}^{m} = \int_m^{\infty} = \frac{1}{2}$$

- Let $x_1, x_2, \ldots x_n$ be a random sample of X and consider testing the hypothesis $H_0: m = m_0$

Vs $H_1: m > m_0$. From the definition of median it follows that, when H_0 is true then $P(X > m_0) = P(X - m_0 > 0) = \frac{1}{2}$ and therefore $P(x_i - m_0 > 0) = 1/2 \quad i = 1, 2 \cdots n$

Let $z_i = \begin{cases} 1 & \text{if } x_i - m_0 > 0 \\ 0 & \text{if } x_i - m_0 < 0 \end{cases}$

- We note that z_i has a Bernoulli distribution with $p = 1/2$ when H_0 is true. Since the z_i's are independent, the sum $u = \sum^n z_i$ will be a Binomial random variable corresponding to n-independent Bernoulli trials for which $p = 1/2$ when H_0 is true.

- Under H_0 most x_i will tend to be larger than m_0 and the variance will tend to exceed the value to be expected when H_0 is true.
- As a result the right tail of the Binomial distribution should be chosen as the critical region of the test when H_0 is true i.e

$$z_i = \begin{cases} 1 & \text{with } p = 1/2 \\ 0 & \text{with } p = 1/2 \end{cases}$$

Recall $E[x] = \sum x \cdot f(x)$ for discrete dist

$$\Rightarrow E[z_i] = \sum z_i f(x)$$

$$= 0 \cdot 1/2 + 1 \cdot 1/2$$

$$= 1/2$$

$$E[x_i^2] = \sum x_i^2 \cdot f(x)$$

$= 0^2 \cdot 1/2 + 1^2 \cdot 1/2 = 1/2$

Now $u = z_1 + z_2 + \cdots + z_n$ i.e. $u = \sum z_i$

$E[u] = E[z_1] + E[z_2] + \cdots + E[z_n]$

$= 1/2 + 1/2 + \cdots + 1/2$ } n of them

$= n/2$

$Var(u) = Var(z_1) + Var(z_2) + \cdots + Var(z_n)$

But $Var(z_i) = E[z_i^2] - (E[z_i])^2$

and $E[z_i^2] = 1/2$,

$= 1/2 - (1/2)^2$

$= 1/2 - 1/4 = 1/4$

Therefore $Var(u) = 1/4 + 1/4 + \cdots + 1/4$

$= n/4$

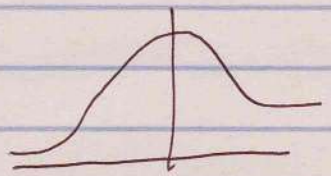
Thus $E[u] = n/2$ and $Var(u) = n/4$ when H_0 is true

- For very small values of n, it is necessary to calculate the right probabilities until a total probability of approximately α has been obtained to get the ~~initial~~ critical region for the test.
- For $p = 1/2$ the Binomial distribution is approximated well by ~~if~~ the Normal distribution for fairly small values of n.
- Therefore it usually suffices to use the Normal approximation for this purpose.

- We use the standard normal variables.

$$z = \frac{u - 1/2 - n/2}{\sqrt{n/4}}$$ for a right tail critical region (>)

and $$z = \frac{u + 1/2 - n/2}{\sqrt{n/4}}$$ for left tail critical region (<)



-ve left right

Example

1) The following data was obtained from testing the breaking strength of ceramic tiles manufactured by a new cheaper process

20, 42, 18, 21, 23, 35, 19, 18, 26, 20, 24
32, 22, 20, 24

Suppose that experience with the old process provided a median of 25, test the hypothesis $H_0: m = 25$ Vs $H_1: m < 25$

Solution

In testing $H_0: m = 25$ Vs $H_1: m < 25$ we use the test statistic

$$z = \frac{u + 1/2 - n/2}{\sqrt{n/4}}$$ and reject H_0 for small values of z where

$u = \sum_{i=1}^{n} z_i$ and $z_i = x_1 - m_0$, $z_2 = x_2 - m_0 \cdots$

so we substract 25 from each of the values to get

$-5, 17, -7, -4, -3, 10, -6, -7, 1, -5, -4, 7, -3, -5, -1$

- The corresponding z-values are:

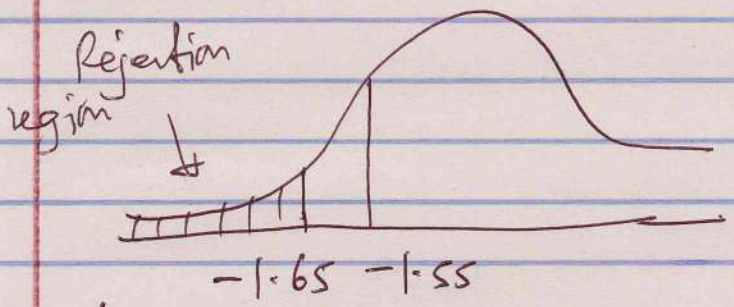
$0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0$

thus $u = \sum_{i=1}^{n} z_i = 4$

u is Binomial with $n = 15$, and $p = 1/2$, thus when H_0 is true

$$z = \frac{4 + 1/2 - 15/2}{\sqrt{15/4}} = \frac{-3}{\sqrt{3.75}} = -1.549$$

$$z = -1.55$$



$\alpha = 0.05$

From z-tables we find $P(Z \leq -1.65) = 0.05$. Since $-1.55 > -1.65$ we do not reject H_0 and conclude the median equals 25.

Note:

(i) Calculations by means of Binomial distribution with $n = 15$ and $p = 1/2$ will give $P(X=x) = \binom{n}{x} p^x q^{n-x}$

$$\binom{15}{5}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{15-x}$$

$f(0) = p(u=0) = 0.00003$
$f(1) = p(u=1) = 0.0005$
$f(2) = p(u=2) = 0.0032$
$f(3) = p(u=(3)) = 0.0139$
$f(4) = p(u=4) = 0.0417$

Then $p(u \leq 4) = f(0) + f(1) + f(2) + f(3) + f(4)$

$= 0.0593$

which demonstrates that the Normal approx with $1/2$ correction factor is excellent

(ii) Instead of getting the values of z_i in the above case we can record the size of the numbers $x_i - 25$ rather than the z-values in which case we would get

$- + - - - + - - + - - + - - -$

then the total number of +ve signs gives the value of u. It is for this reason that the test is called the sign test.

Example 2

1) The following data in tonnes are the amounts of Sulphur oxides emitted by a large industrial process plant in 40 days.

17	15	20	29	19	18	22	25	27	9
24	20	17	6	24	14	15	23	24	26
19	23	28	19	16	22	24	17	20	13
19	10	23	18	31	13	20	17	24	14

Test the null hypothesis $H_0: u = 21.5$ vs $H_1: m < 21.5$ at $\alpha = 0.01$

Solution

Let $z_i = \begin{cases} 1 & \text{if } x_i - 21.5 \geq 0 \\ 0 & \text{if } x_i - 21.5 < 0 \end{cases}$

The values of z are

0	0	0	1	0	0	1	1	1	0
1	0	0	0	1	0	0	1	1	1
0	1	1	0	0	1	1	0	0	0
0	0	1	0	1	0	0	0	1	0

$u = \sum z_i = 16$

−	−	−	+	−	−	+	+	+	−
+	−	−	−	+	−	−	+	+	+
−	+	+	−	−	+	+	−	−	−

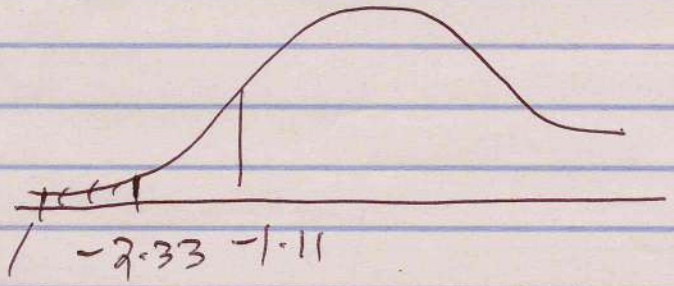
(9)

$$Z = \frac{u + 1/2 - n/2}{\sqrt{n/4}} = \frac{16 + 1/2 - 40/2}{\sqrt{40/4}}$$

$$= -1.11$$

from the normal tables we have

$$P(Z \leq -2.33) = 0.01 (\alpha)$$



Rejection region

Therefore we do not reject H_0 since $-1.11 > -2.33$ i.e -1.11 is not in the rejection region.

The signed-Rank test (Wilcoxon signed-rank test)

- As we have seen the sign test is very easy to perform but since we utilize only the signs of the differences between the observations M_0 (mean) in one sample case, it tends to be wasteful of information.
- An alternative non-parametric test ie the Wilcoxon signed rank test is less wasteful in that it takes into account the magnitude of the difference.
- In this test we rank the difference without regard to the signs assigning rank one to the smallest differences in magnitude in absolute value, rank two to the second smallest differences in magnitude absolute value and rank n to the largest difference in absolute value.
- Zero differences are ignored. If the absolute value of two or more differences are the same we assign each one the mean of the ranks, which they jointly occupy
- The signed rank test is based on T^+ the sum of the ranks assigned to the +ve difference, T^- the sum of ranks assigned to the -ve difference
- Therefore $T = \min(T^+, T^-)$
- We note that T^+ and T^- takes the values in the interval 0 to $\frac{n(n+1)}{2}$.
- T^- and T^+ are symmetric about $n(n+1)/4$. Assignment: Find out how.
- The critical ~~values~~ regions for various tests is given below;

Alternative hypothesis	critical region - rejection
$m \neq m_0$	$T \leq T_\alpha$
$m > m_0$	$T^- \leq T_{2\alpha}$
$m < m_0$	$T^+ \leq T_{2\alpha}$

Example

1) The following are 15 measurements of the octane rating of a certain kind of gasoline

97.5, 95.2, 97.3, 96.0, 96.8, 100.3, 97.4, 97.4, 95.3 93.2, 99.1, 96.1 97.6 98.2 98.5, 94.9

Find the signed-rank test at $\alpha = 0.05$ level of significance to determine whether or not the mean ~~of~~ octane rating of the given kind of gasoline is 98.5.

solution

We wish to test $H_0: \mu = 98.5$ Vs $H_1: \mu \neq 98.5$. Therefore we reject H_0 if $T \leq T_{0.05}$ for the appropriate value of n. We subract 98.5 from each value and rank the absolute difference i.e (In the next page)

$T^+ = 2 + 8 = 10$ (sum of ranks assigned to the +ve differences)

$T^- = 1 + 3 + 4 + 5 + 6 + 9 + 10 + 11 + 12 + 13 + 14 = 95$

$T = \min(T^+, T^-) = \min(10, 95)$

(continue in next page)

Measurements	$x_i - \mu$	rank	
97.5	−1	4	
95.2	−3.3	12	
97.3	−1.2	6	
96.0	−2.5	10	
96.8	−1.7	7	
100.3	+1.8	8	+ difference
97.4	−1.1	5	
95.3	−3.2	11	
93.2	−5.3	14	
99.1	+0.6	2	+ difference
96.1	−2.4	9	
97.6	−0.9	3	
98.2	−0.3	1	
98.5	0.0	—	disregard - zero
94.9	−3.6	13	

Note: — Rank absolute value starting from the smallest difference.
— Disregard zero difference

$n = 14$, $T_{0.05} = 21$ (from critical regions for Wilcoxon-sign rank test)

Since $10 < 21$, we reject H_0.

Note: $n = 14$, disgard one value (zero diff)
Rejection criterion is determined by the alternative hypothesis.

Signed-Rank test for paired data

- We have the observations $(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$
- Let μ_1 be the mean of x's and μ_2 be the mean of y's.
- We want to test $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$ or $H_1 : \mu_1 > \mu_2$ or $H_1 : \mu_1 < \mu_2$. We use the table below.

Alternative hypothesis	Reject H_0 if
$\mu_1 \neq \mu_2$	$T \leq T\alpha$
$\mu_1 > \mu_2$	$T^- \leq T2\alpha$
$\mu_1 < \mu_2$	$T^+ \leq T2\alpha$

For $n \geq 15$, it is considered reasonable to assume the distribution of T^+ is approximately normal. To perform the sign-rank test based on this assumption we need the following results

Theorem: The mean and variance of T^+ are

$$E[T^+] = n(n+1)/4$$

$$Var(T^+) = \frac{n(n+1)(2n+1)}{24}$$

Proof: If the sample of size n, the ranks to be assigned are $1, 2, 3, \cdots, n$.

- For each rank the probability it will be assigned a +ve or -ve difference are both $1/2$ when H_0 is true

$$Z_i = \begin{cases} 1 & \text{if } x_i > y_i \\ 0 & \text{if } x_i < y_i \end{cases} \quad \text{for } i = 1, 2 \cdots n$$

Thus $T^+ = 1 \cdot Z_1 + 2 \cdot Z_2 + 3 \cdot Z_3 + \cdots + n \cdot Z_n$

$Z_1, Z_2 \cdots Z_n$ are independent random variables having a Bernoulli distribution with $p = 1/2$ since $E[Z_i] = 1/2$ and $\text{var}(Z_i) = 1/4$ for $i = 1, 2 \cdots n$

Thus $E[T^+] = E[Z_i]$

$$= E[1 \cdot Z_1 + 2 \cdot Z_2 + \cdots + n \cdot Z_n)$$

$$= 1\,E[Z_1] + 2\,E[Z_2] + \cdots + n\,E[Z_n]$$

$$= 1 \cdot 1/2 + 2 \cdot 1/2 + \cdots + n \cdot 1/2$$

$$= \frac{1 + 2 + \cdots + n}{2} = \frac{n(n+1)}{4}$$

Normally $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$

$$\Rightarrow \frac{n(n+1)}{2} \cdot 1/2 = \frac{n(n+1)}{4}$$

Likewise

$$\text{Var}(T^+) = \text{Var}(1 \cdot Z_1 + 2 \cdot Z_2 + \cdots + n \cdot Z_n)$$

$$= 1^2 \text{var}(Z_1) + 2^2 \text{var}(Z_2) + n^2 \text{Var}(Z_n)$$

$$= 1^2 \cdot 1/4 + 2^2 \cdot 1/4 + \cdots + n^2 \cdot 1/4$$

$$= \frac{1^2 + 2^2 + 3^2 + \cdots + n^2}{4}$$

$$1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\Rightarrow \frac{n(n+1)(2n+1)}{6} \cdot 1/4$$

$$= \frac{n(n+1)(2n+1)}{24}$$

Example

The following are the weights in pounds before and after 16 persons who stayed on a certain weight reducing diet for 4 weeks

Before	After	Before	After
147.0	137.9	166.8	158.5
183.5	178.2	131.9	130.4
232.1	219.0	150.3	149.3
161.6	163.8	197.2	189.1
197.5	193.5	159.8	159.1
206.3	207.4	171.7	173.2
177.0	180.6		
215.4	203.2		
147.7	149.0		
208.1	195.4		

Use the sign-rank test to test at $\alpha = 0.05$ level of significance whether the weight reducing balanced diet is effective.

Solution

We are testing $H_0: \mu_1 = \mu_2$ Vs $H_1: \mu_1 > \mu_2$

We reject H_0 if $Z_{cale} \geq Z_{0.05} = 1.645$

Where $Z = \dfrac{T^+ - E(T^+)}{\sqrt{Var(T^+)}}$

The differences between the respective pairs are given on the next page

pair	diff	rank	
1	9.1	13	+
2	7.3	10	+
3	13.1	16	+
4	-2.2	5	
5	4.0	8	+
6	4.9	9	
7	-3.6	7	
8	12.2	14	+
9	-1.3	3	
10	12.7	15	+
11	8.3	12	+
12	-2.5	6	
13	1.0	2	+
14	8.1	11	+
15	0.7	1	+
16	-1.5	4	

NB: Ranks in absolute values

$n = 16$, $T+ = 13 + 10 + 16 + 8 + 9 + 14 + 15 + 12 + 12 + 11 + 1)$

$= 111$

$$E(T+) = \frac{n(n+1)}{4} = \frac{16(17)}{4} = 68$$

$$Var(T+) = \frac{n(n+1)(2n+1)}{24} = \frac{16(17)(33)}{24}$$

$= 374$

$$Z = \frac{111 - 68}{\sqrt{374}} = 2.22$$

$z_\alpha = z_{0.05} = 1.645$

$2.22 > 1.645$, thus we reject H_0

Alternatively: Reject H_0 if $T^- \leq T_{2\alpha}$

$T^- = 25$ $(5+7+3+6+4)$

$T_{0.10} = 35$

$T^- = 25 < T_{0.10} = 35$, thus reject H_0.

Rank-sum test – U-test

- The U-test is used for two samples. The other names are;
  - (i) Wilcoxon test
  - (ii) Man-Whitney test
- We are going to test the hypothesis that we are sampling identical continous populations against the alternative that the two populations have unequal means.
- To illustrate the procedure, suppose we want to compare the kinds of emergency flares on the basis of the following burning times.

Brand A	14.9	11.3	13.2	16.6	17.0	14.1	15.4	13.0
	16.9							
Brand B	15.2	19.8	14.7	18.3	16.2	24.2	18.9	12.2
	15.3	19.4						

Steps

1. Arrange the values in ascending order
2. Rank them jointly

Brand A	11.3	13.0	13.2	14.1	14.9	15.4	16.6
	(1)	(3)	(4)	(5)	(7)	(10)	(12)
	16.9	17.0					
	(13)	(14)					
Brand B	12.2	14.7	15.2	15.3	16.2	18.3	18.9
	(2)	(6)	(8)	(9)	(11)	(15)	(16)
	19.4	19.8	24.2				
	(17)	(18)	(19)				

→ The values of brand A occupy ranks 1, 3, 4, 5, 7, 10, 12, 13, and B occupy 2, 6, 8, 9, 11, 15, 16, 17, 18 and 19.

→ For ties we assign the mean of the ranks which they jointly occupy

- If there is an appropriate difference between the means of the two observations, most of the lower ranks are likely to go to the values of one sample, while most of the higher ranks will go to the values of the other sample
- The test is based on the value of W_1 which is the
- sum of the ranks of the first sample or W_2 which is the sum of the ranks of the second sample.
- It doesn't matter whether we use W_1 or W_2 for if there are n_1 values in the first sample and n_2 values in the second sample $W_1 + W_2$ is the sum of the first $n_1 + n_2$ integers i.e. $W_1 + W_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}$

- For any pair of values of W_1 and W_2. Therefore, the test based on W_1 and W_2 are equivalent. In actual practice we seldom use the test based on the statistics W_1 and W_2 instead we use the related statistics

$$U_1 = W_1 - \frac{n_1(n_1+1)}{2}$$

$$U_2 = W_2 - \frac{n_2(n_2+1)}{2}$$

or $U = \min(U_1, U_2)$

- The resulting tests are all equivalent to the one based on W_1 and W_2 but have the advantage that they label themselves more readily to the tables of

critical values i.e

$$u_1 + u_2 = W_1 - \frac{n_1(n_1+1)}{2} + W_2 - \frac{n_2(n_2+1)}{2}$$

$$= \frac{(n_1+n_2)(n_1+n_2+1)}{2} - \frac{n_1(n_1+1)}{2} - \frac{n_2(n_2+1)}{2}$$

$$= n_1^2 + n_1 n_2 + n_1 + n_2 n_1 + n_2 + n_2^2 - n_1^2 - n_1 - n_2^2 - n_2$$

$$= 2(u_1 + u_2) = 2n_1 n_2$$

$\Rightarrow u_1 + u_2 = n_1 n_2$ i.e the sum of $u_1 + u_2$ is always $n_1 n_2$ and both of these random variables take on the same range of the values from 0 to $n_1 n_2$.

- We use the following table for testing $H_0 : \mu_1 = \mu_2$ vs the various alternatives.

Alternative hypothesis	Reject H_0 if
$\mu_1 \neq \mu_2$	$u \leq u_\alpha$
$\mu_1 > \mu_2$	$u_2 \leq u_{2\alpha}$
$\mu_1 < \mu_2$	$u_1 \leq u_{2\alpha}$

where the level of significance is α for each test

- The critical values of u are such that u_α is the largest value for which $u \leq u_\alpha$ does not

exceed α.

Example

1) Two brands of flares have the following burning times

Brand A	14.9	11.3	13.2	16.6	17.0	14.1
	15.4	13.0	16.9			
Brand B	15.2	19.8	14.7	18.3	16.2	21.2
	18.9	12.2	15.3	19.4		

Test at $\alpha = 0.05$ level of significance whether the two samples come from identical continuous populations ~~of~~ or whether the mean burning times of B flares i.e outshines A.

Solution

We want to test $H_0: \mu_1 = \mu_2$ Vs $H_1: \mu_1 < \mu_2$
Arrange the values in ascending order (rank them jointly)

Brand A	11.3	13.0	13.2	14.1	14.9	15.4	16.6
	(1)	(3)	(4)	(5)	(7)	(10)	(12)
	16.9	17.0					
	(13)	(14)					
Brand B	12.2	14.7	15.2	15.3	16.2	18.3	18.9
	(2)	(6)	(8)	(9)	(11)	(15)	(16)
	19.4	19.8	21.2				
	(17)	(18)	(19)				

We reject H_0 if $U_1 \leq U_{2\alpha}$

thus $W_1 = 1+3+4+5+7+10+12+13+14$
$= 69$

$\Rightarrow u_1 = W_1 = \frac{n_1(n_1+1)}{2} = 69 - \frac{9(10)}{2}$

$= 24$

for $n_1 = 9,\ n_2 = 10$

$u_{2\alpha} = u_{0.10} = 24$

Thus we conclude since $24 \leq 24$ We reject H_0 and conclude that on average brand A flakes as a mean value time that is less than that of brand B.

- When both n_1 and n_2 are greater than 8, it is considered reasnable to assume that the distribution of ~~u_1 and u_2~~ u_1 and u_2 can be approximated closely by the normal distribution. To perform the given rank sum test on the basis of this assumption, we need the following result

Theorem: Under the null hypothesis, the mean and variance of ~~u_1 and u_2~~ u_1 and u_2 are

1) $E[u_1] = E[u_2] = \frac{n_1 n_2}{2}$ and

2) $\text{var}(u_1) = \text{var}(u_2) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$

Example

The following are the number of minutes it took random samples of 15 men and 12 women to complete a written test given for their renewal of their driving licences

Men: 9.9, 7.4, 8.9, 9.1, 7.7, 9.7, 11.8, 9.2, 10.0, 10.2, 9.5, 10.8, 8.0, 11.0, 7.5

Women: 8.6, 10.9, 9.8, 10.7, 9.4, 10.3, 7.3, 11.5, 7.6, 9.3, 8.8, 9.6

Use the U-test based on the normal approximation to test $H_0: \mu_1 = \mu_2$ Vs $H_1: \mu_1 \neq \mu_2$ where μ_1 and μ_2 are the average amount of time it takes men and women to complete the test respectively

Solution

Re-arrange the data in ascending order.

7.3	7.4	7.5	7.6	7.7	8.0	8.6	8.8	8.9
①	②	③	④	⑤	⑥	⑦	⑧	⑨
9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9
⑩	⑪	⑫	⑬	⑭	⑮	⑯	⑰	⑱
10.0	10.2	10.3	10.7	10.8	10.9	11.0	11.5	11.8
⑲	⑳	㉑	㉒	㉓	㉔	㉕	㉖	㉗

The underlined values are for men

$n_1 = 15$

$n_2 = 12$

(7)

Ranks for men $\Rightarrow W_1 = 2+3+5+6+9+10+11+14+16+18+19+20+23+25+27 = 208$

Ranks for Woman $\Rightarrow W_2 = 1+4+7+8+12+13+15+17+21+22+24+26$

$= 170$

$$U_1 = W_1 - \frac{n_1(n_1+1)}{2}$$

$$= 208 - \frac{15(16)}{2} = 88$$

$$U_2 = W_2 - \frac{n_2(n_2+1)}{2}$$

$$= 170 - \frac{12(13)}{2} = 92$$

$\Rightarrow U = \min(U_1, U_2) = \min(88, 92)$

$= 88$

Thus $E[U] = \frac{n_1 n_2}{2} = \frac{15 * 12}{2} = 90$

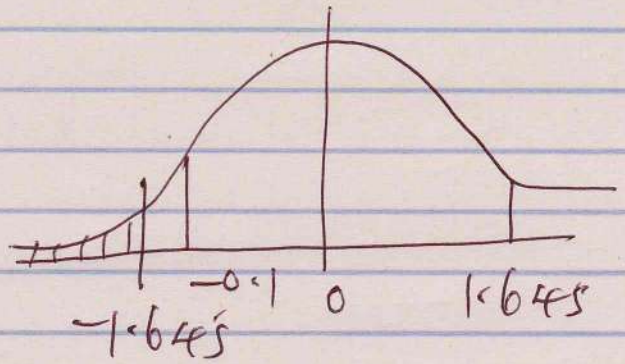
$$Var(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

$$= \frac{15 * 12 (15 + 12 + 1)}{12}$$

$= 420$

$$Z = \frac{u - E[u]}{\sqrt{Var(u)}} = \frac{88-90}{\sqrt{420}} = -0.098 \simeq -0.1$$

$Z_{\alpha = 2 \times 0.05} = 1.645$, $H_0 : \mu = \mu_2$
$H_1 : \mu_1 \neq \mu_2 \Rightarrow$ 2 sided



since $-1.645 < -0.1$ we do no reject the null hypothesis (-0.1 is not in the rejection region)

Exercises

1) The following are the amount of time in mins which took a random sample of 20 technicians to perform a task

18.1, 20.3, 18.3, 15.6, 22.5, 16.8, 17.6, 16.9, 18.2, 17.0, 19.3, 16.5, 19.5, 18.6, 20.0, 18.8, 19.1, 17.6, 18.5, 18.0

Use the sign test at $\alpha = 0.05$ to test the null hypothesis that the measurement constitute a random sample from a continuous pdf with mean $\mu = 19.4$ mins against the 2 sided alternative $\mu \neq 19.4$ mins. Use the Binomial test.

2) Rework problem 1 using the normal approximation to the Binomial. Use ~~random~~ sign-rank test also.

Rank sum test : The H-test (Kruskal-Wallis test)

- The H-test / Kruskal-Wallis test is a generalization of the rank-sum test to the case where we test the null hypothesis that K-samples come from identical continuous distributions.
- As in the U-test the data are ranked jointly from low to high as though they constitute one sample.
- Let R_i be the sum of the ranks of the values of the i^{th} sample.
- We base the test on the statistic

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

Therefore the H-statistic is proportional to the weighted mean of the squared difference $\left(R_i/n_i - \frac{n+1}{2} \right)^2$ where R_i/n_i is the mean rank for sample i and $(n+1)/2$ is the mean rank of all the data. It follows that the null hypothesis must be rejected for large values of H.

- Expression 1 can be simplified as

$$\left(R_i/n_i - \frac{n+1}{2} \right)^2 = \frac{R_i^2}{n_i^2} - 2\left(R_i/n_i \right)\left(\frac{n+1}{2} \right) + \left(\frac{n+1}{2} \right)^2$$

$$= \frac{R_i^2}{n_i^2} - \frac{R_i}{n_i}(n+1) + \frac{(n+1)^2}{4}$$

$$= n_i \left(R_i/n_i - \frac{n+1}{2} \right)^2 = \frac{R_i^2}{n_i} - R_i(n+1) + \frac{n_i(n+1)^2}{4}$$

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

$$= \sum_{i=1}^{K} R_i^2 / n_i - (n+1) \sum_{i=1}^{K} R_i + \frac{(n+1)^2}{4} \sum_{i=1}^{K} n_i$$

But

$$\sum_{i=1}^{K} n_i = n$$

$$= \sum_{i=1}^{K} \frac{R_i^2}{n_i} - (n+1) \cdot n \frac{(n+1)}{2} + \frac{(n+1)^2}{4} n \quad \text{—} \otimes$$

$$= \sum_{i=1}^{K} R_i^2 / n_i - \frac{n(n+1)^2}{4}$$

Substituting in ① we find that

$$H = \frac{12}{n(n+1)} \left[\sum_{i=1}^{K} R_i^2 / n_i - n \frac{(n+1)^2}{4} \right]$$

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{K} R_i^2 / n_i - 3(n+1) \quad \text{—} ②$$

- For every small values of K and n_i the test of the null hypothesis may be based on special tables but since the sampling distribution of H depends on the values of n_i, it is impossible to tabulate it in compact form; hence the test is usually based on the large sample theory, that the distribution of H can be approximately closely with a χ^2 distribution with $K-1$ degrees of freedom.

Example

1) A sampling of the acidity of rain on 10 randomly selected rainfall was recorded at 3 different locations i.e, 1, 2, 3.

1	2	3
4.45	4.60	4.55
4.02	4.27	4.31
4.13	4.31	4.84
3.51	3.88	4.67
4.42	4.49	4.28
3.89	4.22	4.95
4.18	4.54	4.72
3.95	4.76	4.63
4.07	4.36	4.36
4.29	4.21	4.47

Use the Kruskal-Wallis test at $\alpha = 0.05$ to test whether or not there is a difference in the average acidity of rain in the 3 locations.

Solution

$k = 3$

$n = n_1 + n_2 + n_3 = 10 + 10 + 10$

$\Rightarrow n = 30$

$\sum_{i=1}^{k} n_i = n$

For each location arrange the values in ascending order.

1	2	3
3.51	3.88	4.28
3.89	4.21	4.31
3.95	4.22	4.36
4.02	4.27	4.47
4.07	4.31	4.55
4.13	4.36	4.63
4.18	4.49	4.67
4.29	4.54	4.72
4.42	4.60	4.84
4.45	4.76	4.95

- Rank the values jointly, we get

3.51, 3.88, 3.89, 3.95, 4.02, 4.07, 4.13, 4.18
① ② ③ ④ ⑤ ⑥ ⑦ ⑧

4.21, 4.22, 4.27, 4.28, 4.29, 4.31, 4.31, 4.36,
⑨ ⑩ ⑪ ⑫ ⑬ ⑭ ⑮ ⑯

4.36, 4.42, 4.45, 4.47, 4.49, 4.54, 4.55, 4.60
⑰ ⑱ ⑲ ⑳ ㉑ ㉒ ㉓ ㉔

4.63, 4.67, 4.72, 4.76, 4.84, 4.95
㉕ ㉖ ㉗ ㉘ ㉙ ㉚

Note: 4.31, 4.31
⑭ ⑮ $14+15 = 29/2 = 14.5$

4.36, 4.36 $16+17 = 33/2 = 16.5$
⑯ ⑰

Ranks for location I: $1+3+4+5+6+7+8+13+18+19$

II: $24+11+14.5+2+21+10+22+28+16.5+9$

III : 13 + 14.5 + 29 + 26 + 12 + 30 + 27 + 25 + 16.5 + 20

$R_1 = 84$,
$R_2 = 158$
$R_3 = 223$

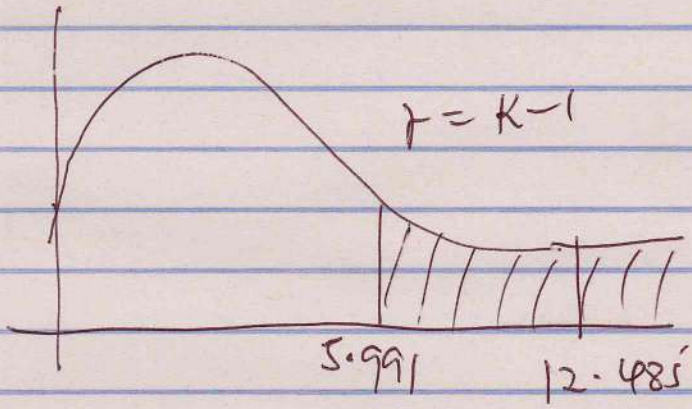
Therefore $H = \frac{12}{n(n+1)} \sum_{i=1}^{K} R_i^2 / n_i - 3(n+1)$

$= \frac{12}{30(31)} \sum_{i=1}^{3} R^2 / 10 - 3(n+1)$

$= \frac{12}{30(31)} \left(\frac{84^2}{10} + \frac{158^2}{10} + \frac{223^2}{10} \right) - 3(31)$

$= \frac{12}{930} \left(\frac{84^2}{10} + \frac{158^2}{10} + \frac{223^2}{10} \right) - 93$

$\Rightarrow H = 12.4835$



$n = 30$, $K = 3$, $r = 3 - 1 = 2$

$\Rightarrow \chi^2_{0.05}(2) = \chi^2_{0.05}(2) = 5.991$

$12.4835 > 5.991$, we reject H_0.

Exercise

1) In an example to determine which of the 3 missiles systems is preferable, the propellant burning rate is measured. The data after coding are given in the table below. Use the Kruskal-Wallis test at $\alpha = 0.05$ to test the hypothesis that the propellant burning rate are the same for the three missile systems.

Propellant burning rate		
1	2	3
24.0	23.2	18.4
16.7	19.8	19.1
22.8	18.1	17.3
19.8	17.6	19.7
18.9	20.2	18.9
	17.8	18.8
		19.3

2) The following data represents the operating time in hours for 3 types of scientific pocket calculators before a recharge is required

Calculator		
A	B	C
4.9	5.5	6.4
6.1	5.4	6.8
4.3	6.2	5.6
4.6	5.8	6.5
5.3	5.5	6.5
	5.2	6.3
	4.8	6.6

Use Kruskal-Wallis test at $\alpha = 0.01$ to test the hypothesis that the operating times for all 3 calculators are the same

The run's test

- We look at a test for testing randomness of observed data on the basis of the order of which they were obtained.

Definition: Run – A run is a succession of identical letters (or other kinds of symbols) which is preceded and followed by different letters or no letters at all.

- Consider the following arrangement of effective (d) and non-effective (n) pixels produced in the given order by a certain machine.

n n n n n d d d d n n n n n
(1) (2) (3)

n n n n n d d n n d d d
(4) (5) (6)

d n d d nn
7 2 (9)

hence 9 runs.

- The total number of runs appearing in an arrangement of this kind is often a good indicator of a possible lack of randomness
- If there are too few runs, we might suspect a definite grouping or clustering or perhaps a trend.
- If there are too many runs, we might suspect some sort of repeated alternating pattern
- In our illustration, there seems to be some sort of clustering (few runs)

- If we have n_1 letters of some kind and n_2 letters of another kind, there are:

$\binom{n_1+n_2}{n_1}$ possible arrangements of these letters

- All arrangements are regarded to be equally likely.

Example

1) If we have 2 a's and b's, then the no of possible arrangements of these letters is

$$\binom{2+2}{2} = \binom{4}{2} = \frac{4!}{2!\,2!} = \frac{4\times3\times2\times1}{2\times2\times1\times1} = 6$$

These arrangements are:

aa bb (2r) a b a b (4r) a b b a (3r) b a b a (4r) bb aa (2r)

b aa b (3r)

2) We have five letters e's which are divided into 3 runs using vertical bars to represent the 5 letters into 3 runs. We find that there are 6 possibilities i.e.

e | e | eee e | ee | ee ee | ee | ee

eee | e | e ee | e | ee e | eee | e

(6)

- Let there be n_1 letters of one kind and n_2 letters of another kind.
- Let u be the number of runs formed by these $n_1 + n_2$ letters
- If $u = 2k$, where k is a positive integer then we have k runs of n_1 letters and k runs of the n_2 letters.
- Number of ways of which n_1 letters can form k run is $\binom{n_1 - 1}{k-1}$ and number of ways in which n_2 letters can for k runs is $\binom{n_2 - 1}{k-1}$

$$P(u = 2k \text{ runs}) = F(u) = \frac{2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}}{\binom{n_1+n_2}{n_1}}$$

if $u = 2k+1$

$$F(u) = P(u_1 = k+1, u_2 = k) + P(u_1 = k, u_2 = k+1)$$

$$= \binom{n_1-1}{k}\binom{n_2-1}{k-1} + \binom{n_1-1}{k-1}\binom{n_2-1}{k}$$

$$= \frac{\binom{n_1-1}{k}\binom{n_2-1}{k-1} + \binom{n_1-1}{k-1}\binom{n_2-1}{k}}{\binom{n_1+n_2}{n_1}}$$

When n_1 and n_2 are small the test of randomness is based on u (the number of runs) are usually performed with the use of special tables. We reject the null hypothesis of randomness at α level of significance if $u < u_{\alpha/2}$ or

$u \geq u_{\alpha/2}$

where $u_{\alpha/2}$ is the largest value for which $p(u \leq u_{\alpha/2})$ does not exceed $\alpha/2$ and $u_{\alpha/2}$ is the smallest value for which $p(u \geq u_{\alpha/2})$ does not exceed $\alpha/2$

Example

Checking on the palm tree that were planted many years ago along a country road a county official obtained the following arrangements of health (H) and diseased (D) trees i.e.

H H H H D D D H H H H H H H

D D H H D D D D

Test at $\alpha = 0.05$ whether this arrangement can be regarded to be random

Solution

We wish to test

H_0: Arrangement is random

H_1: Arrangement is not random

No of H's, $n_1 = 13$

No of B's, $n_2 = 9$

$u_{\alpha/2} = u_{0.05/2} = 6$

We reject H_0 if $u \leq 6$ or $u \geq 17$

From the data $u = 6$, since $6 \leq 6$ we reject H_0 and conclude that the arrangement of healthy and diseased tree is not random.

- If n_1 and n_2 are both greater or equal to 10, it is considered reasonable to assume that the distribution of u can be closely approximated by the normal.
- To perform the test on the basis of this assumption we need the following results.

Theorem: Under the null hypothesis of randomness the mean and variance of u are

$$E[u] = \frac{2n_1n_2}{n_1+n_2} + 1$$

$$Var(u) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)}$$

Example

The following series of boys and girls were selected haphazardly in a core education school

G BB G G B G BBB G G G G B G B G B B G B G G G G
B B G B G G B G B G B B G

Test for randomness at $\alpha = 0.01$

Solution

We want to test H_0: Arrangement is random

H_1: Arrangement is not random

(6)

If u is the number of runs, then

$$E[u] = \frac{2n_1n_2}{n_1+n_2} + 1$$

$$Var(u) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{n_1}$$

We use the test statistic

$$Z = \frac{(u \pm 1/2) - E[u]}{\sqrt{Var(u)}}$$

We reject H_0 if $Z \leq -z_{0.005} = -2.575$ $= -\alpha/2$

or if $Z \geq z_{0.005} = 2.575$

No of girls $G = n_1 = 20$

No of boys $B = n_2 = 18$

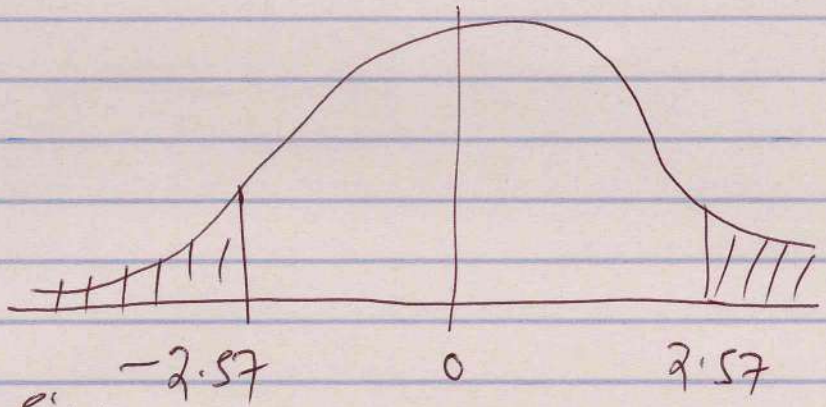
$u = 25$ (runs)

$$E[u] = \frac{2(20)(18)}{20+18} + 1 = 19.947$$

$$Var(u) = \frac{2(20)(18)[2(20)(18) - 20 - 18]}{(20+18)^2(20+18-1)}$$

$$= 9.190$$

$$Z = \frac{25 - 1/2 - 19.947}{\sqrt{9.190}}$$



Since

$-2.57 < u < 2.57$

$-2.57 < 1.50 < 2.57$

~~we accept~~ ~~H_0~~ ~~and~~

We do not reject H_0 and say the arrangement is random (1.50 is in the acceptance region)

~~Exercise~~

1) The following are the amounts of money in dollars spend by 16 persons at a certain amusement bar

10.15 9.85 13.75 8.63 11.09 15.63

6.65 9.27 8.80 11.45 10.29 9.51 13.80

10.0 7.48 9.11

(a) Use the sign-test at $\alpha = 0.05$ to test the null hypothesis that on average a person spends 9 dollars at the bar.

(b) Remark the preceding problem using the sign-rank test based on the normal approximation to the distribution of the test statistic

Rank correlation coefficient

- Since the assumptions underlying the significance test for the correlation coefficient are rather stringent, it is sometimes preferable to use a non-parametric alternative.
- Most popular among the non-parametric measures of association is the rank correlation coefficient also called Spearman's rank correlation coefficient.
- For a given set of data $(x_1, y_1), (x_2, y_2), (x_3, y_3) \cdots (x_n, y_n)$,
- The rank correlation is obtained by ranking the x_i's among themselves and the y's both from low to high or high to low, the rank correlation coefficient is given by:

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} b_i^2}{n(n^2-1)}$$

where b_i is the difference between the ranks assigned to x_i and y_i. Where there are ties in the ranks we assigned the tied observations the mean of the ranks which they jointly occupy.

Examples

Calculate the Spearman's coefficient rank correlation from the following.

x	73	198	191	107	95	81	79	75
y	147	125	132	137	130	140	139	145

Calculate r.

x	r_x	y	r_y	d_i	d_i^2
73	1	147	8	−7	49
118	8	125	1	7	49
115	7	132	3	4	16
101	6	137	4	2	4
95	5	130	2	3	9
81	4	140	6	−2	4
79	3	139	5	−2	4
75	2	145	7	−5	25

$\sum d_i^2 = 160$

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6(160)}{8(63)}$$

$$= 1 - \frac{120}{63}$$

$$= -0.9048$$

- For small values of n, i.e $n \leq 10$ the test of H_0 of no correlation may be based on special tables determined from the exact sampling distribution of r_s.
- Most of the time though we use the fact that the distribution of random variables can be closely approximated to the normal distribution

Theorem: Under the null hypothesis of no correlation the mean and variance of random variables are:

$E[r_s = 0]$ and $Var(r_s) = 1/n-1$

Note: Strictly speaking the theorem applies when there are no ties, but the result can be used as an approximation unless the number of ties is large

Example

1) For the data in the previous example test at $\alpha = 0.09$ whether the rank correlation coefficient is significant

Solution

We want to test

H_0: There is no correlation

H_1: There is correlation

The test statistic is:

$$Z = \frac{r_s - E[r_s]}{\sqrt{Var(r_s)}} = \frac{r_s - 0}{\sqrt{1/n-1}} = r_s\sqrt{n-1}$$

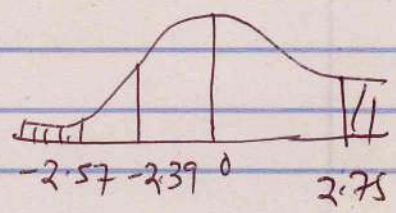
We reject H_0 if $Z \leq -2.75$ or $Z \geq 2.75$

i.e $Z_{\alpha/2} = Z_{0.09/2} = Z_{0.005} = \pm 2.75$

$r_s = -0.9048$, $n = 8$

$\Rightarrow Z = -0.9048\sqrt{8-1}$

$= -2.3938$



$Z = -2.3938 > -2.575$ We do reject H_0 and conclude that the correlation is not significant

Note: When there are no ties in ranks, r_s actually equals the correlation coefficient r, calculated for the ranks.

Let r_i and s_i be the ranks of x_i and y_i respectively.

- We find that

$$\sum_{i=1}^{n} r_i = \sum_{i=1}^{n} s_i = \frac{n(n+1)}{2} \quad (1+2+3+\cdots+n)$$

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} s_i^2 = \frac{n(n+1)(2n+1)}{6} \quad (1^2+2^2+\cdots+n^2)$$

Let $d_i = r_i - s_i$

$$d_i^2 = (r_i - s_i)^2 = r_i^2 - 2r_i s_i + s_i^2$$

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} r_i^2 - 2\sum_{i=1}^{n} r_i s_i + \sum_{i=1}^{n} s_i^2$$

$$\Rightarrow 2\sum_{i=1}^{n} r_i s_i = \sum_{i=1}^{n} r_i^2 + \sum_{i=1}^{n} s_i^2 - \sum_{i=1}^{n} d_i^2$$

$$= \frac{2n(n+1)(2n+1)}{6} - \sum_{i=1}^{n} d_i^2$$

$$r = \frac{\sum r_i s_i - \frac{1}{n}(\sum r_i)(\sum s_i)}{\sqrt{\sum (r_i^2 - n\bar{r}^2)\sum (s_i^2 - n\bar{s}^2)}}$$

$$= \frac{n(n+1)(2n+1) - 1/2 \sum d_i^2 - 1/n \left(\frac{n(n+1)}{2}\right)\left(\frac{n(n+1)}{2}\right)}{\sqrt{\left[\left(\frac{n(n+1)(2n+1)}{6}\right) - \frac{n(n+1)^2}{2}\right]\left[\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)}{2}\right]^2}}$$

$$= \frac{\frac{n(n+1)(2n+1)}{6} - 1/n\left(\frac{n(n+1)}{2}\right)^2 - 1/2 \sum d_i^2}{\frac{n(n+1)(2n+1)}{6} - 1/n\left(\frac{n(n+1)}{2}\right)^2}$$

Note: $\sum r_i^2 - n\bar{r}^2$, $\bar{r} = 1/n \sum r_i^2$

$$= 1/n\left(\frac{n(n+1)}{2}\right)$$

$$n\bar{r}^2 = n\left(1/n^2\left(\frac{n(n+1)}{2}\right)^2\right)$$

$$= \frac{n(n+1)^2}{2}$$

$$= 1 - \frac{1/2 \sum_{i=1}^{n} d_i^2}{\frac{n(n+1)(2n+1)}{6} - 1/n\left(\frac{n(n+1)}{2}\right)^2} \quad \text{--- (*)}$$

$$\frac{n(n+1)(2n+1)}{6} - 1/n\left(n\left(\frac{n+1}{2}\right)\right)^2 = \frac{(n+1)}{12}\left[2n(2n+1) - 3n(n+1)\right]$$

$$= \frac{n+1}{12}\left[4n^2 - 2n - 3n^2 - 3n\right]$$

$$= \frac{n+1}{12}\left(n^2 - n\right)$$

$$= \frac{n(n+1)(n-1)}{12}$$

$$= \frac{n(n^2-1)}{12}$$

Replacing in eqn (*) we find that

$$r = 1 - \frac{1/2 \sum d_i^2}{\frac{n(n^2-1)}{12}}$$

$$\frac{\sum d_i^2}{2} \times \frac{12}{n(n^2-1)} = \frac{6\sum d_i^2}{n(n^2-1)}$$

$$r = 1 - \frac{6\sum d_i^2}{n(n^2-1} = r_s$$

Check expansion of

$$\frac{n(n+1)(2n+1)}{6} - \frac{1}{n}\left(\frac{n(n+1)}{2}\right)^2$$

$$\downarrow$$

$$1/6\left\{n(2n^2+3n+1)\right\} = 1/6\left\{2n^3+3n^2+n\right\}$$

(7)

$$\frac{1}{n}\left(\frac{n^2+n}{2}\right)^2 = \frac{1}{4n}\left(n^4+2n^3+n^2\right)$$

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

$$\Rightarrow \frac{2n^3+3n^2+n}{6} - \left(\frac{n^3+2n^2+n}{4}\right)$$

$$= \frac{4n^3+6n^2+2n-3n^3-6n^2-3n}{12}$$

$$= \frac{n^3-n}{12} = \frac{n\left(n^2-1\right)}{12}$$

Testing for goodness-of-fit

- Hypothesis testing procedures are designed for problems in which the probability distribution is known and hypothesis involve the parameters of the distribution.
- Another kind of hypothesis is often encountered.
- We do not know the underlying distribution of the population and we wish to test the hypothesis that particular distribution will be satisfactory as a population model.
- We describe a formal goodness-of-fit test procedure based on the chi-square distribution.
- The test procedure requires a random sample of size n from the population whose probability distribution is unknown.
- These n observations are arranged in a frequency histogram having k-class intervals.
- Let O_i be the observed frequency of the i^{th} class.
- From the hypothesized probability distribution we compute the expected frequency denoted by E_i
- The test statistic is $\chi_0^2 = \frac{(O_i - E_i)^2}{E_i}$

which has an approximately χ^2_{k-p-1} distribution with $k-p-1$ degrees of freedom where p is the number of parameters of the hypothesized distribution estimated by sample statistics.

- This approximation improves as n increases.
- We will reject ~~H0~~ the hypothesis that the distribution of the population is the hypothesized if the calculated value of the test statistic $\chi_0^2 > \chi^2_{\alpha, k-p-1}$

Example

The number of defects in a printed circuit board is hypothesized to follow a Poisson distribution. A random sample of $n = 60$ printed boards has be collected and the following number of defects observed

No of defects	Observed frequency (O_i)
0	32
1	15
2	9
3	4

The mean of the assumed Poisson distribution in this example is unknown and must be estimated from the sample data. The estimate of the mean is

$$(32 + (15 \times 1) + (9 \times 2) + (4 \times 3))/60 = 0.75$$

From the Poisson distribution with parameter 0.75 we may compute p_i the theoretical hypothesized probability associated with the ith class interval. p_i are found as follows.

$$p_1 = P(X=0) = \frac{e^{-0.75} 0.75^0}{0!} = 0.472$$

$$p_2 = P(X=1) = (e^{-0.75} 0.75^1)/1! = 0.354$$

$$p_3 = P(X=2) = \frac{e^{-0.75} 0.75^2}{2!} = 0.133$$

$$p_4 = P(X=3) = \frac{e^{-0.75} 0.75^3}{3!} = 0.041$$

Note:

$\Rightarrow p_4 = 1 - p(X \leq 3)$

$= 1 - (p_1 + p_2 + p_3)$

The E_i are computed by multiplying the sample size $n = 60$ times p_i i.e $E_i = np_i$

No of defects	p_i	E_i
0	0.472×60	28.32
1	0.354×60	21.24
2	0.133×60	7.98
3 (or more)	0.041×60	2.46

Since the expected frequency in the last cell is less than 3, we combine the last 2 cells

No of defects	O_i	E_i
0	32	28.32
1	15	21.24
2 (or more)	13	10.44

The degree of freedom are $K - p - 1 = 3 - 1 - 1 = 1$

H_0: The form of the dist of defects is Poisson
H_1: The form of the dist of defects is not a Poisson

$\alpha = 0.05$

$$\chi_0^2 = \sum^{K} \frac{(O_i - E_i)^2}{E_i}$$

Reject H_0 if $\chi^2_{0\ 0.05,1} = 3.84$

$$\chi_0^2 = \frac{(32-28.32)^2}{28.32} + \frac{(15-21.24)^2}{21.24} + \frac{(13-10.44)^2}{10.44}$$

$$= 2.94$$

Since $\chi_0^2 = 2.94 < \chi^2_{0.05} = 3.84$ We are unable to reject H_0 that the distribution of defects printed on the circuit board is Poisson.

Contigency table tests

- It is used to test statistical independence of two categorical variables classified in a table.
- A contigency table is a two-way classification of data i.e a company has to choose among 3 pension plans. The management wish to know whether the preference of plans is independent of job classification using $\alpha = 0.05$
- The opinion of a random sample of 500 employees are shown below

Job classification	Pension plan 1	2	3	Total
Salary workers	160	140	40	340
Hourly workers	40	60	60	160
Total	200	200	100	500

H_0: preference is independent of salaried workers versus hourly job classification

H_1: preference is not independent of salaried versus hourly job classification.

$$\chi^2_{calculated} = \sum_{i=1}^{m} \sum_{j=1}^{p} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

m = rows
p = cols

Where O_{ij} is the observed frequency in row i column j

E_{ij} is the expected frequency in row i column j and is given by

$$E_{ij} = \frac{i^{th} \text{ row total} \times j^{th} \text{ column total}}{\text{sample size}}$$

$$= \frac{R_i \cdot C_j}{n}$$

The critical value is given by $\chi^2_{crit} = \chi^2_{(m-1)\times(p-1)}{}^{\alpha}$. Reject H_0 iff $\chi^2_{calc} > \chi^2_{crit}$

$$E_{ij} \Rightarrow E_{11} = \frac{340 \times 200}{500} = 136$$

$$E_{12} = \frac{340 \times 200}{500} = 136$$

$$E_{13} = \frac{340 \times 100}{500} = 68$$

(b)

$$E_{21} = \frac{160 \times 200}{500} = 64$$

$$E_{22} = \frac{160 \times 200}{500} = 64$$

$$E_{23} = \frac{160 \times 100}{500} = 32$$

Job classification	Pension 1	2	3	Total
Salaried Workers	160 (136)	140 (136)	40 (68)	340
Hourly Workers	40 (64)	60 (64)	60 (32)	160
				500

$$\chi^2_{cal} = \frac{(160-136)^2}{136} + \frac{(140-136)^2}{136} + \frac{(40-68)^2}{68} + \frac{(40-64)^2}{64} + \frac{(60-64)^2}{64} + \frac{(60-32)^2}{32}$$

$$= 49.63$$

$$\chi^2_{crit} = \chi^2_{(m-1)(p-1)} = \chi^2_2(0.05) = 5.991$$

Conclusion: Since $\chi^2_0 = 49.63 > \chi^2_2(0.05) = 5.991$ We reject H_0 and conclude preference for pension plans is not dependent on job classification

Exercise

1) Grades in statistics and OR courses taken simultaneously were as follows for a group of students.

	OR grades				
Statistics grades	A	B	C	Others	Total
A	25	6	7	13	51
B	17	16	15	6	54
C	18	4	8	10	50
Others	10	8	11	20	49
Total	70	34	51	49	204

Are the grades in statistics and OR related? $\alpha = 0.01$ in teaching your conclusion.

KOlmogorov-Smirnov test

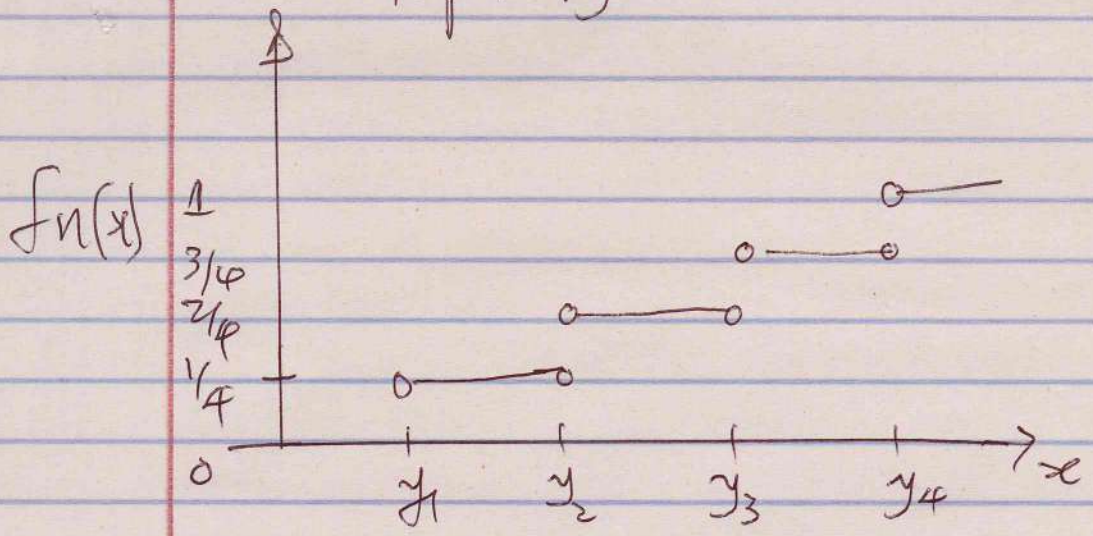
- Suppose the random variables $x_1, x_2 \ldots x_n$ is from a continuous random sample and let $x_1, x_2 \ldots x_n$ be the observed values of $X_1, X_2 \ldots X_n$
- Since the observations come from a continuous distribution there is a probability of zero (0) that any of the observed values $x_1, x_2 \ldots x_n$ will be equal. Thus we shall assume for simplicity that all the n-values are different.
- We consider a function $f_n(x)$ which is constructed from the values $x_1, x_2 \ldots x_n$ as follows.
- For each number x $(-\infty < x < \infty)$ the value of $f_n(x)$ is defined as the proportion of observed values in the sample which are less than or equal to x
- In other words if exactly K of the observed values in the sample are less than or equal to x, then $f_n(x) = K/n$.

- The function $f_n(x)$ defined this way is called the sample distribution function (sdf).
- Sometimes $f_n(x)$ is called the empirical distribution function. If we let $y_1 < y_2 < \ldots y_n$ denote the value of the order statistics of the sample, then $f_n(x) = 0$ for $x < y_1$

- $f_n(x)$ jumps to the value $1/n$ at $x = y_1$ and remains at $1/n$ for $y_1 \leq x < y_2$.

- $f_n(x)$ jumps to $2/n$ at $x = y_2$ and remains at $2/n$ for $y_2 \leq x < y_3$ and so on.

Graphically



- Now let $f(x)$ denote the distribution function of the distribution from which the sample was drawn.
- For a given number x $(-\infty < x < \infty)$ the probability that any particular x_i is less than or equal to x is $F(x)$.
- Therefore it follows that from the law of large numbers (LLN) that as $n \to \infty$ the proportion of $f_n(x)$ of the observations in the sample which are less than or equal to x will converge to $f(x)$ in probability.

$$P\left(\lim_{n \to \infty} f_n(x)\right) = f(x) \quad \text{---} \quad (4)$$

for $-\infty < x < \infty$

Where n is the sample size of $x_1, x_2 \cdots x_n$.
$f_n(x)$ is the sample distribution function.
$f(x)$ is the distribution function

- The relation (4) expresses the fact that at each point x, the sample distribution function $f_n(x)$ will converge to the actual $f(x)$ of the distribution from which the random sample was drawn.

Let $D_n = \sup_{-\infty < x < \infty} |f_n(x) - f(x)|$

- Suppose we now wish to test the simple hypothesis H_0: the unknown distribution function $f(x)$ is actually a particular continuous distribution $f^*(x)$ Vs the general alternative that the actual distribution function is not $f^*(x)$ i.e

$$\left.\begin{aligned} H_0 &: f_n(x) = f^*(x) \quad -\infty < x < \infty \\ H_1 &: \text{the hypothesis } H_0 \text{ is not true} \end{aligned}\right\} \quad (2)$$

- This is a non-parametric problem because the unknown distribution from which the random sample was taken might be any continuous distribution.
- Let $f_n(x)$ denote the sample distribution function (sdf) and let

$$D_n = \sup_{-\infty < x < \infty} |f_n(x) - f^*(x)|$$

- In other words D_n is the maximum difference between the sdf $f_n(x)$ and the hypothesized $f^*(x)$
- When H_0 in eqn (2) is true the probability distribution of D_n will be a certain distribution which is the same for any possible continuous distribution
- $F^*(x)$ does not depend on a particular distribution function being studied in a specific problem.
- Tables of this distribution for various values of n (sample size) have been developed and are tabulated in collection of statistical tables. The tables gives the values of D_n^{α} such that;

$$P(D_n \leq D_n^{\alpha}) = 1 - \alpha$$

$$\Rightarrow P(D_n > D_n^{\alpha}) = \alpha$$

- Below are some critical values for the Kolmogorov–Smirnov test

n \ α	0.20	0.10	0.05	0.01
5	0.45	0.51	0.56	0.67
10	0.32	0.37	0.41	0.49
15	0.27	0.26	0.34	0.40
20	0.23	0.24	0.29	0.36
25	0.21	0.22	0.27	0.32
30	0.19	0.20	0.24	0.29
35	0.18	0.19	0.23	0.27
40	0.17	0.18	0.21	0.25
45	0.16	0.17	0.2	0.24
50	0.13	0.7	0.9	0.23
$n>50$	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.03}{\sqrt{n}}$

Example

Test the hypothesis by the Kolmogorov Smirnov test that the following sample came from a standard normal distribution

−1.23	1.64	−0.21	0.70	1.40	0.44	−0.07	−0.02
−0.15	1.76	1.62	0.40	−2.11	−0.99	−0.42	0.81
1.47	−2.46	0.88	1.39	0.42	0.27	−0.39	−0.10
1.07							

Solution

We want to test

H_0: $f(x) = f^*(x)$ for $-\infty < x < \infty$

H_1: the hypothesis is not true.

where $f^*(x) = \Phi(x)$ the cdf of the standard normal distribution at $\alpha = 0.05$

$n = 25$, $D_{25}^{0.05} = 0.27$ (from tables)

We reject H_0 if $D_n \geq D_{25}^{0.05} = 0.27$

where $D_{25} = \sup_{-\infty < x < \infty} |f_{25}(x) - f^*(x)|$

Arrange the sample values in ascending order and for each value determine $f_n(x)$ and $f^*(x)$.

i	x	$F_n(x)$	$f^*(x)$	$\lvert f_n(x) - f^*(x) \rvert$
1	−2.46	1/25 = 0.04	0.007	0.033
2	−2.11	2/25 = 0.08	0.017	0.063
3	−1.23	3/25 = 0.12	0.109	0.011
4	−0.99	4/25 = 0.16	0.161	0.001
5	−0.42	0.20	0.337	0.137 → sup
6	−0.39	0.24	0.348	0.108
7	−0.24	0.28	0.417	0.137 → sup
8	−0.15	0.32	0.440	0.120
9	−0.10	0.36	0.460	0.10
10	−0.07	0.40	0.472	0.072
11	−0.02	0.44	0.492	0.052
12	0.27	0.48	0.606	0.126
13	0.40	0.52	0.655	0.135
14	0.42	0.56	0.663	0.103
15	0.44	0.60	0.670	0.070
16	0.70	0.64	0.758	0.114
17	0.81	0.68	0.791	0.111

i.	x	$F_n(x)$	$F^*(x)$	$F_n(x) - F^*(x)$
18	0.88	0.72	0.811	0.091
19	1.07	0.76	0.858	0.098
20	1.39	0.80	0.918	0.118
21	1.40	0.84	0.919	0.079
22	1.47	0.88	0.929	0.049
23	1.62	0.92	0.947	0.027
24	1.64	0.96	0.950	0.010
25	1.76	$\frac{25}{25} = 1.0$	0.961	0.039

Getting $f^*(x)$

$$f^*(x)(-2.46) = p(x \leq -2.46) = p\left(\frac{x-\mu}{\sqrt{\sigma^2}}\right)$$

$$= \frac{-2.46 - \mu}{\sqrt{\sigma^2}}$$

$$\Rightarrow p\left(z \leq \frac{-2.46-0}{1}\right)$$ for Standard normal

$Z \sim N(0, 1)$

$\mu = 0, \sigma^2 = 1$

$$\Rightarrow p(z \leq -2.46) = 1 - p(z \leq -2.46)$$

$$= 1 - \phi(2.46)$$

$$= 1 - 0.9931$$

$$= 0.0069$$

$$\simeq 0.007$$

$D_{25} = 0.137$ and $D_{25}^{0.05} = 0.27$

$D_{25} = 0.137 < D_{25}^{0.05} = 0.27$ hence we do not

Reject A_0 and conclude that the sample came from a standard normal distribution.

Exercise

Test the hypothesis by Kolmogorov-Smirnov test that the following sample values came from a normal distribution with mean 2 and variance 4 at $\alpha = 0.01$

2.72	3.84	0.88	5.72	5.48	3.12	0.10	2.48
1.70	0.52	2.64	3.64	3.40	1.80	−0.52	−0.12
2.30	3.10	1.04	1.02				