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. $x_1, x_2, \dots x_n \sim \text{Poisson}(\lambda_i)$

  (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 independant ~~ra~~ samples come from same distribution and we might only assume that both distributions are continuous.

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

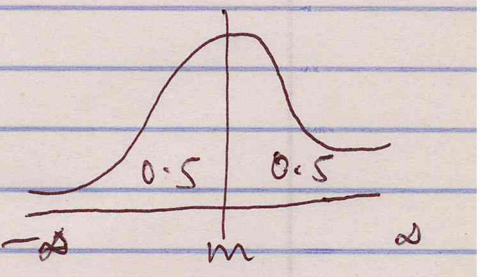
The sign test

- 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.

~~i.e 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

$$\int_{m}^{\infty} f(x)\, dx = \frac{1}{2}$$



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 $\underset{\sim}{X}$ 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 $P(X \geq m_0] = P(X' - m_0 \geq 0) = \frac{1}{2}$ and therefore $P(X_i - m_0 \geq 0) = \frac{1}{2} \quad i = 1, 2 \dots n$

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

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

$$Z_i = \begin{cases} 1 & \text{with } p = \frac{1}{2} \\ 0 & \text{with } p = \frac{1}{2} \end{cases}$$

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

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

$$= 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2}$$

$$= \frac{1}{2}$$

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

$$= 0^2 \cdot \tfrac{1}{2} + 1^2 \cdot \tfrac{1}{2} = \tfrac{1}{2}$$

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

$$\left.\begin{aligned} E[u] &= E[z_1] + E[z_2] + \cdots + E[z_n] \\ &= \tfrac{1}{2} + \tfrac{1}{2} + \cdots + \tfrac{1}{2} \\ &= \tfrac{n}{2} \end{aligned}\right\} \text{n of them}$$

$$\mathrm{var}(u) = \mathrm{var}(z_1) + \mathrm{var}(z_2) + \cdots + \mathrm{var}(z_n)$$

But $\mathrm{var}(z_i) = E[z_i^2] - (E[z_i])^2$

and $E[z_i^2] = \tfrac{1}{2}$,

$$= \tfrac{1}{2} - \left(\tfrac{1}{2}\right)^2$$

$$= \tfrac{1}{2} - \tfrac{1}{4} = \tfrac{1}{4}$$

Therefore $\mathrm{var}(u) = \tfrac{1}{4} + \tfrac{1}{4} + \cdots + \tfrac{1}{4}$

$$= \tfrac{n}{4}$$

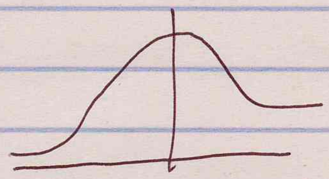
Thus $E[u] = \tfrac{n}{2}$ and $\mathrm{var}(u) = \tfrac{n}{4}$ when H_0 is true

- 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 $p = \tfrac{1}{2}$ the Binomial distribution is approximated well by ~~it~~ 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 $(<)$



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

Suppose that experience with the old process produced 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

(6)

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

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:

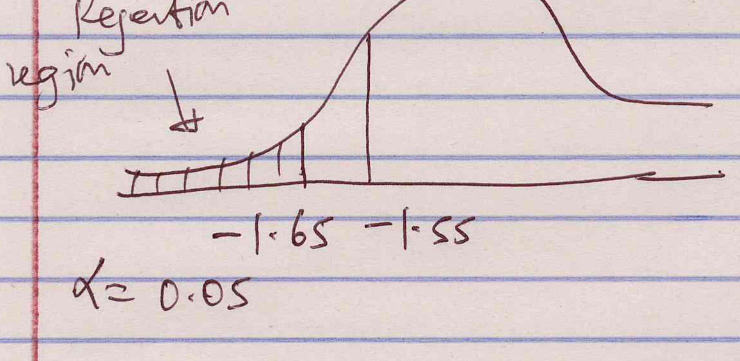
$0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0,$

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

$$\left.\begin{aligned} f(0) &= p(u=0) = 0.00003 \\ f(1) &= p(u=1) = 0.0005 \\ f(2) &= (p=2) = 0.0032 \\ f(3) &= (p=(3)) = 0.0139 \\ f(4) &= p(\cancel{u}=4) = 0.0417 \end{aligned}\right\} \quad \begin{aligned} &p(X=x) = \\ &\binom{n}{x} p^{x} q^{n-x} \\ &\binom{15}{5}\left(\tfrac{1}{2}\right)^{x}\left(\tfrac{1}{2}\right)^{15-x} \end{aligned}$$

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 <u>sign test</u>

<u>Example 2</u>

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 $A_1: m < 21.5$ at $\alpha = 0.01$

Solutions

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

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

$u = \Sigma z_i = 16$

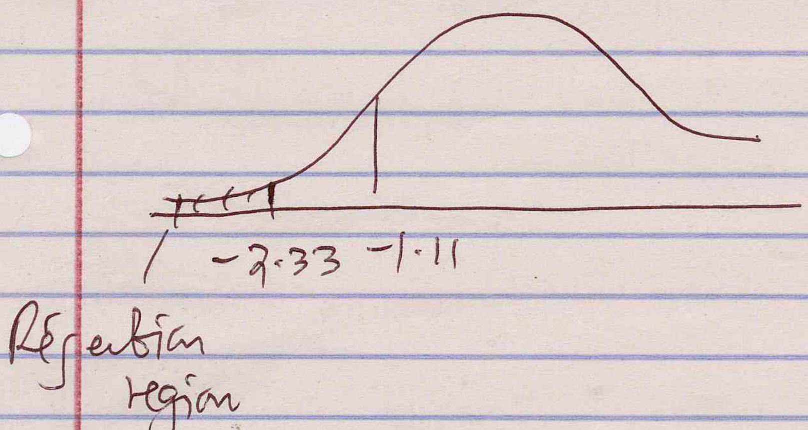
− − − + − − + + + −
+ − − − + − − + + +
− + + − − + + − − −

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

$$= -1.11$$

from the normal tables we have

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



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