Non-Parametric Inference

Suman Paul

(1) Suppose that each of 13 randomly chosen female registered-voters was asked to indicate if she was going to vote for Candidate A or B in an upcoming election. The result shows that 9 of the subjects preferred A. Is this sufficient evidence to conclude that candidate A is prefered to B by female voters? Draw the power curve taking at least 8 points.

We have a population of 13 female registered voters.

In an upcoming election, they have option to vote for Candidate A or B.

Let us consider p is the prob to vote for candidate A

We test $H_0 : p = \frac{1}{2}$ vs $H_1 : p > \frac{1}{2}$

We have test statistic, $S = 9$

A test is constructed as,

$$\phi(s) = \begin{cases} 1 & s > K_\alpha \\ \gamma & s = K_\alpha \\ 0 & s < K_\alpha \end{cases}$$

under H_0,

$S_{H_0} \sim Bin(13, \frac{1}{2})$

γ and K_α are to be determined from the size condition.

$$E_{H_0}(\phi(s)) = \alpha = 0.05$$

$$\Rightarrow P_{H_0}(s > K_\alpha) + \gamma P_{H_0}(s = K_\alpha) = 0.05$$

$$\Rightarrow \gamma = \frac{0.05 - P_{H_0}(s > K_\alpha)}{P_{H_0}(s = K_\alpha)}$$

$$\Rightarrow \gamma = \frac{P_{H_0}(s \le K_\alpha) - 0.95}{P_{H_0}(s = K_\alpha)}$$

We construct the table as follows under H_0, $s \sim Bin(13, \frac{1}{2})$

K_α	$P(s = K_\alpha)$	$P(s \le K_\alpha)$
0	0.00012	0.00012
1	0.00159	0.00171
2	0.00952	0.01123
3	0.03491	0.04614
4	0.08728	0.13342
5	0.1571	0.29052
6	0.20947	0.4999
7	0.20947	0.70946
8	0.1571	0.86656
9	0.08798	0.95384 → the point to be randomized.
10	0.03491	0.98875

We take $K_\alpha = 9$ for randomization and we see

$$\gamma = \frac{P(s \le K_\alpha) - 0.95}{P(s = K_\alpha)} = \frac{P(s \le 9) - 0.95}{P(s=9)} = 0.044.$$

The test is considered as,

$$\phi(s) = \begin{cases} 1 & s > 9 \\ 0.044 & s = 9 \\ 0 & s < 9 \end{cases}$$

Since $s = 9$. ⇒ we reject the Null Hypothesis with rejection probability $\gamma = 0.044$. Hence, the evidence is not so conclusive.

We used R-programming for the power curve.

The Educational Testing Service (ETS) reports that the 75th percentile for scores of the GRE examinations is 693. in a certain year. A random sample of 15 freshmen majoring in statistics report their GRE scores as 690, 750, 680, 700, 660, 710, 720, 730, 650, 670, 740, 730, 660, 750, 690

Are the scores of students majoring in Statistics consistent with the 75th percentile value?

→

ETS reports that 75th percentile scores of GRE exams is 693

Let p be the prob. that GRE scores lies in the percentile range

We test, $H_0: p = \frac{3}{4}$ vs $H_1: p \neq \frac{3}{4}$

We have a sample of GRE scores of size 15.

Under H_0, the no of observations satisfying $P(x_i < 693) = \frac{3}{4}$

$$S_{H_0} \sim Bin\left(15, \frac{3}{4}\right)$$

$$\therefore P_{H_0}(S = s) = \binom{15}{s}\left(\frac{3}{4}\right)^{s}\left(\frac{1}{4}\right)^{15-s}$$

Now, we have 8 such observations where (observation − 693) > 0

$\therefore$ The no of positive differences = the value of test statistic $\boxed{S = 8}$

Fix α = (level of significance) = 0.1

Let $K_{\alpha/2}$ and $K'_{\alpha/2}$ be two constants constructing rejection regions.

From the size conditions

$$\sum_{s=0}^{K_\alpha} \binom{15}{s} \left(\frac{3}{4}\right)^s \left(\frac{1}{4}\right)^{15-s} \leq 0.05$$

When $K_{\alpha/2} = 7$ we find that $P_{H_0}(s \leq 7) \leq 0.05$

Also, $$\sum_{s=K'_{\alpha/2}}^{15} \binom{15}{s} \left(\frac{3}{4}\right)^s \left(\frac{1}{4}\right)^{15-s} \leq 0.05$$

finding $K'_{\alpha/2} = 14$ satisfies the inequality

Hence the test is constructed by $\phi(s) = \begin{cases} 1 & \text{if } s \leq 7 \text{ or } s \geq 1 \\ 0 & \text{o.w.} \end{cases}$

As we have $s = 8$, we can conclude that students majoring in statistics has scored are consistent with the percentile value (accepting H_0).

In a marketing research test, 15 adult males were asked to shave one side of their face with a brand 'A' razor and the other side of their face with a brand 'B' razor and state their ~~perfect~~ preferred razor. 12 men preferred brand A. Find the p-value for the alternative for prefering brand A is greater than 0.5.

→

Scanned with CamScanner

The Null Hypothesis is,

H_0: A and B are equally preferable $\approxeq \pi = 1/2$

vs H_1: A is more preferable $\approxeq \pi > 1/2$

Let S be the sample statistic ie no of adults preferring brand A ie $S=12$

Under H_0 $S \sim Bin(15, 1/2)$

Now, p value $= P[S \geq 12 | H_0]$.

Consider μ_0 to the median of the population preferring A

then $F_{\mu_0}(x) = 1/2$ under $H_0: \mu = \mu_0$

When Brand preference of A is more than 50% median should be shifted to the right of μ_0.

Then we have $H_1: \mu > \mu_0$

Therefore, the p-value is,

$$p \text{ value} = \sum_{x=12}^{15} \left(\frac{1}{2}\right)^{n} \binom{15}{x}$$

$$= \binom{15}{12}\left(\frac{1}{2}\right)^{15} + \binom{15}{13}\left(\frac{1}{2}\right)^{15} + \binom{15}{14}\left(\frac{1}{2}\right)^{15} + \binom{15}{15}\left(\frac{1}{2}\right)^{15}$$

$= 0.0176$

$\therefore$ The p-value for the alternative that the probability of preferring brand A is greater than 0.05 is 0.0176

Scanned with CamScanner

(1) A study of 5 years ago reported that median amount of sleep by American adult is 7.5 hours out of 24 hours. A current sample of 8 adults reported their avg amount of sleep per 24 hours are 7.2, 8.3, 5.6, 7.4, 7.8, 5.2, 9.1 and 5.8 hrs. Use the most appropriate test to determine whether american adults sleep less today than 5 years ago.

$\Rightarrow$ Let us assume that the data is coming from a Continuous distribution $F_X(x)$ with median μ_x.

We have to test the Hypothesis $H_0: \mu_x = 7.5$

vs $H_1: \mu_x < 7.5$

X	$D = X - \mu_x$	$\lvert D \rvert$	Rank($\lvert D \rvert$)
7.2	−0.3	0.3	2.5
8.3	0.8	0.8	4
5.6	−1.9	1.9	7
7.4	−0.1	0.1	1
7.8	0.3	0.3	2.5
5.2	−2.3	2.3	8.
9.1	1.6	1.6	5
5.8	−1.7	1.7	6

Now $T^+ =$ Sum of ranks of +ve obs $= 4 + 2.5 + 5 = 11.5$

$T^- =$ " " " −ve " $=$ [illegible]

at $n = 8$, $\alpha = 0.01$; $T_\alpha = 2$

Here $T^+ > T_\alpha \Rightarrow$ we fail to reject the Null Hypothesis

$\therefore$ American Adults Sleep equally today than they did 5 years ago.

A large Company was ~~distributed~~ disturbed about the no of person hour lost per month due to accident and institutional ~~ext~~ an extensive industrial safety program. The data below show the number of person-hours lost in a month at each of 8 different plants before and after the safety program was implemented. Has the safety program been effective in reducing time lost from accident.

Plant	Before	After
1	51.2	45.8
2	46.5	41.3
3	24.1	15.8
4	10.2	11.1
5	65.3	58.5
6	92.1	70.3
7	30.3	31.6
8	49.2	35.4

→ Suppose person hour lost before and after safety program is denoted by a bivariate random variable (X,Y)

Assume, (X,Y) is coming from a continuous distn function $F_{X,Y}(x,y)$

We are to test $H_0: \mu_x = \mu_y$
vs $H_1: \mu_x > \mu_y$

Take the transformation $D = X - Y$, Assume μ_D to be the median of distribution of D then ~~the~~ we have,

$H_0: \mu_D = 0$ vs $H_1: \mu_D > 0$

→

Scanned with CamScanner

Now we constructed the following table for ~~each~~ Calculation

Plant	Before	After	$D = x - y$	$\lvert D\rvert$	Rank ($\lvert D\rvert$)
1	51.2	45.8	5.4	5.4	4
2	46.5	41.3	5.2	5.2	3
3	24.1	15.8	8.3	8.3	6
4	10.2	11.1	-0.9	0.9	1
5	65.3	58.5	6.8	6.8	5
6	92.1	70.3	21.8	21.8	8
7	30.3	31.6	-1.3	1.3	2
8	49.2	35.4	13.8	13.8	7

Under H_1 rank of +ve obs will be higher resulting T^+ larger and T^- smaller simultaneously.

Therefore we reject H_0 if $T^- < T_\alpha$

where T_α being the tabular value

$$T^- = 1 + 2 = 3$$

here $T_\alpha = 2$ at $\alpha = 0.01$

So, $T_1^- > T_\alpha$ as $3 > 2$

We fail to reject.

$\therefore$ Safety program is not effective.

Reducing high blood pressure by diet requires reduction of sodium intake. Listed below are the avg. sodium contents of 5 ordinary foods in processed form and natural form for equivalent quantities. Do you see any difference in the median of processed food and natural food?

Natural Food		Processed Foods	
Corn of the Cob	2	Canned Corn	251
Chicken	63	Fried Chicken	1220
Ground Surloin	60	All beef biscoti	461
Beans	3	Can beans	300
Fresh tuna	40	Canned tuna	409

⟹ Let us denote the processed food and natural food by a bivariate random variable (X,Y) with a continuous dist. $F_{X,Y}(x,y)$. Consider (M_x, M_y) be the median of $F_{X,Y}(x,y)$.

We want to test, $H_0: M_x = M_y$ vs $H_1: M_x \neq M_y$

Take $D = X - Y$ and assume M_D to be the distribution of D. Then we have,

$H_0: M_D = 0$ vs $H_1: M_D \neq 0$

We form the following table to ease our calculation

⟶

Scanned with CamScanner

Natural Food (X)	Processed Food (Y)	$D=X-Y$	$\lvert D \rvert$	Rank($\lvert D \rvert$)
2	251	-249	249	1
63	1220	-1157	1157	5
60	461	-401	401	4
3	300	-297	297	2
40	400	-360	360	3

$\because$ This is a both sided test we reject Ho if

$T^+ \leq t_{\alpha/2}$ or $T^- \leq t_{\alpha/2}$ here $T^+=0$ and $T^-=15$

using R-studio we can see that the p-value for this two sided test is p-value $= 0.0625$

and Since p value > 0.05, the null hypothesis Ho should be accepted. ie There are a difference between the median of processed and natural food.

The 2000 census statistics for Alabama given the Percentage changes between 1990 and 2000 for each of the 67 countries. There are two types of countries that rural and non-rural. According to the population size < 25000. Below is the data of 9 rural and 7 non-rural countries on percentage of population change.

Rural: 1.1, −21.7, −16.3, −11.3, −10.4, −7.0, −2.0, 1.9, 6.2

Non-Rural: −2.4, 9.9, 14.2, 18.4, 20.1, 23.1, 70.4

Use Mann-Whitney test for testing the Null Hypothesis of equal population change.

→ Let the population change of Rural County came from a continuous distribution with CDF $F_Y(y)$ where median is M_Y. Similarly for Non-Rural County the CDF is $F_X(x)$ with median M_X.

We want to test $H_0: M_X = M_Y$ vs

$H_1: M_X \neq M_Y$

Now, the arranged combined sample is, in ascending order,

−21.7	−16.3	−11.3	−10.4	−7.0	−2.4	−2.0
Y	Y	Y	Y	Y	X	Y

1.1	1.9	6.2	9.9	14.2	18.4	20.1	23.1	70.4
Y	Y	Y	X	X	X	X	X	X

The test statistic $U = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} D_{ij}$ = # Y obs precedes X obs

$= 9+9+9+9+9+9+5$

$= 59.$

→

Similarly $U' = \#$ X obs. precedes Y obs $= 4$

At $n_1 = 9$, $\alpha = 0.02$, $U_{tab} = 9$

Now Since $U' < U_{tab}$ we reject the Null Hypothesis of equal population change.

Consider 10 students take a test and their scores are as follows : 95, 80, 40, 52, 60, 80, 82, 58, 65, 50.

Test the Null Hypothesis that the CDF of the proportion of Right answer a student gets on the test is

$$F_0(x) = \begin{cases} x^2(3-2x) & , 0 \le x < 1 \\ 1 & , x \geq 1 \end{cases}$$

→ Let us create the following table for calculation

we want to test $H_0: F_X(x) = F_0(x)$ vs $H_1: F_X(x) \neq F_0(x)$

X	proportion	increasing order	$F_X(x)$	$F_0(x)$	$\lvert F_X(x) - F_0(x) \rvert$
95	0.95	0.4	0.1	0.352	0.252
80	0.80	0.50	0.2	0.5	0.3
40	0.40	0.52	0.3	0.529	0.229
52	0.52	0.58	0.4	0.618	0.218
60	0.6	0.60	0.5	0.648	0.148
80	0.8	0.65	0.6	0.718	0.118
82	0.82	0.80	0.7	0.896	0.096
58	0.58	0.80	0.8	0.896	0.096
65	0.65	0.82	0.9	0.914	0.014
50	0.5	0.95	1	0.992	0.008

where, $F_X(x) = \# \text{ obs} \leq x / n$

Now, the test statistic KS is $D = \max_x |F_X(x) - F_0(x)| = 0.3$

At $\alpha = 0.01$, $n = 10$, $D_\alpha = 0.489$

Now, Since $D_{tab} > D_{cal}$. ie $0.489 > 0.3$ we accept the Null Hypothesis that the obs are coming from $F_0(x)$

A random sample of 12 persons are interviewed to estimate median annual ~~growth~~ gross income in a certain economically depressed town. Use the most appropriate test for the null hypothesis that income data is standard normally distributed.

9800, 10200, 9300, 8700, 15200, 6800, 8600, 9600,

11600, 7200, 12200, 15500.

We want to test the hypothesis $H_0: F_X(x) = F_0(x)$
vs $H_1: F_X(x) \neq F_0(x)$

where $F_0(x) \sim N(0,1)$.

We need to convert ~~correct~~ the given data into Standard Normal

The mean is given by $\bar{x} = \frac{1}{12}(9800 + \cdots + 15500) = 10391.67$

and $S_x^2 = \frac{1}{n-1} \sum_{i=1}^{12} (x_i - \bar{x})^2 = 7755378.788$.

X	$Z = (x-\bar{x})/s_n$	$F_x(Z)$	$\Phi(Z)$	$\lvert F_x(Z) - \Phi(Z) \rvert$
9800	−0.212	1/12	0.42	0.33
10200	−0.069	2/12	0.47	0.3
9300	−0.392	3/12	0.35	0.1
8700	−0.607	4/12	0.27	0.06
15200	1.727	5/12	0.96	0.54
6800	−1.288	6/12	0.09	0.41
8600	−0.643	7/12	0.26	0.32
9600	−0.284	8/12	0.38	0.28
11600	0.434	9/12	0.66	0.09
7200	−1.146	10/12	0.12	0.71
12200	0.649	11/12	0.74	~~0.17~~
15500	1.834	1	0.96	0.04

The KS test statistic is $D = \underset{x}{Max} \lvert F_x(z) - F_0(z) \rvert$

$= 0.71$

at $\alpha = 0.01$, $n = 12$, $D_\alpha = 0.449$

here $D_{cal} > D_\alpha$ we reject Null hypothesis ie the data is $N(0,1)$.

(10) 2 Mutually indept random samples of each size 8 are generated one from the $N(0,1)$ dist and another from χ^2_{18} dist. The resulting data are as follows:–

$N(0,1)$:	-1.91	-1.22	-0.96	-0.72	$+0.14$	0.82	0.45	1.86
χ^2_{18} :	4.90	7.25	8.04	14.10	18.3	21.21	23.1	28.12

Do you believe they are coming from the same distribution?

→ We need to convert the obs from χ^2_{18} to its standardized form. We know that $E(\chi^2_{18}) = 18$

$\& \ V(\chi^2_{18}) = 36$

$\therefore$ Standard χ^2 dist is $\dfrac{X-18}{\sqrt{36}} = \chi^2_{std.}$

Now, standard χ^2 obs are; $-2.18, -1.75, -1.66, -0.65,$

$+0.05, 0.535, 0.85, 1.685$

We test the hypothesis $H_0: F_1(x) = F_2(x)$

vs $H_1: F_1(x) \neq F_2(x)$.

where $F_1(x)$ is $N(0,1)$ & $F_2(x)$ is χ^2_{std}

We combine the 2 samples and arrange increasingly

⟶

Combined Sample	$F_{n_1}(x)$	$F_{n_2}(x)$	$\lvert F_{n_1}(x) - F_{n_2}(x)\rvert$
-2.18 (2)	0	1/8	1/8
-1.91 (1)	1/8	1/8	0
-1.71 (2)	1/8	2/8	1/8
-1.22 (1)	1/8	3/8	0
-1.66 (2)	2/8	3/8	1/8
-0.96 (1)	3/8	3/8	0
-0.72 (1)	4/8	3/8	1/8
-0.65 (2)	4/8	4/8	0
0.05 (2)	4/8	5/8	1/8
0.14 (1)	5/8	5/8	0
0.535 (2)	5/8	6/8	1/8
0.82 (1)	6/8	6/8	0
0.835 (2)	6/8	7/8	1/8
0.45 (1)	7/8	7/8	0
1.686 (2)	7/8	1	1/8
1.86 (1)	1	1	0

Now the KS test statistic is given by $D_{n_1,n_2} = \max_x \lvert F_{n_1}(x) - F_{n_2}(x)\rvert$

$n_1 = 8,\ n_2 = 8$ $\qquad = 2/8 = 0.25$

Now, for $n_1 n_2 D_\alpha = 8 \cdot 8 \times \frac{2}{8} = 16$

from the table we have, for $n_1 n_2 D = 32$; pvalue $= 0.283$

$n_1 n_2 D_\alpha = 16$; pvalue ≥ 0.283

now, $D_{n_1,n_2} = 0.25 \leq n_1 n_2 D_\alpha = 16$

$\therefore$ We accept the Null Hypothesis that the obs are coming from the same dist ($\because D_{cal} < D_{tab}$).