Chapter 4. Nonparametric Statistics

Reading Assignment

Chapter 15 (skip Sections 7, 8, 10, 11).
Sections 14.1, 14.2, 14.3.

Parametric Methods: Distributions are known apart from the values of a finite number of parameters.

For example,

1. Population distribution $N(\theta, \sigma^2)$, with θ and σ unknown.

$$H_0 : \theta = 0, \quad H_a : \theta \neq 0.$$

2. Population distribution $B(n; p)$, with p unknown.

$$H_0 : p = 0.5, \quad H_a : p > 0.5.$$

Non-Parametric Methods: Far weaker assumptions on the underlying distributions – family of distributions cannot be labelled by finitely many parameters.

For example,

1. Population cumulative distribution function $F(x)$ unknown.

$$H_0 : F(x) = \bar{F}(x), \quad H_a : F(x) \neq \bar{F}(x).$$

2. Two-sample shift model. Population 1 distribution $F(x)$, and Population 2 distribution $G(x) = F(x - \theta)$ for some θ. F is unspecified.

   (a) Estimate θ.

   (b) Test $H_0 : \theta = 0$, $H_a : \theta \neq 0$ [or $\theta > 0$, $\theta < 0$].

Matched pair experiment

1. Sign test. This test does *not* require quantitative measurement.

   (a) Example: Two ice creams are were made with different flavors but otherwise similar. Out of a panel of 8 industry experts, all but one ranked flavor A as perferred. Is there any significant evidence that flavor A is superior? [significance level $\alpha = 5\%$]

Judge	a	b	c	d	e	f	g	h
Flavor A	1	1	1	2	1	1	1	1
Flavor B	2	2	2	1	2	2	2	2
$+$ or $-$	$+$	$+$	$+$	$-$	$+$	$+$	$+$	$+$

H_0 : A and B have same taste, H_a : A is superior to B.

Rewrite the hypothesis as

$$H_0 : P(+) = P(-) = 0.5, \qquad H_a : P(+) > 0.5 > P(-).$$

Test statistics

$$M \doteq \text{number of ``+''}.$$

Under H_0, it has distribution $B(8, 0.5)$.

$$\begin{aligned} \text{P-value} &= P(B(8,0.5) \geq \text{observed value of } M) \\ &= P(B(8,0.5) \geq 7) \\ &= P(B(8,0.5) = 7) + P(B(8,0.5) = 8) \\ &= 8 \times (0.5)^8 + (0.5)^8 = 0.035 \end{aligned}$$

Reject H_0 and accept H_a.

Remark: What is the P-value if the alternative hypothesis is H_a : A and B have different taste, *i.e.*, $H_a : P(+) \neq P(-)$.

$$\begin{aligned} \text{P-value} &= P(B(8, 0.5) \geq 7) + P(B(8, 0.5) \leq 8 - 7) \\ &= 2P(B(8, 0.5) \geq 7) \\ &= 0.070. \end{aligned}$$

Remark: When $n > 25$, one can use normal approximation that under H_0 [i.e., $P(+) = 0.5$]

$$Z = \frac{M - n/2}{\sqrt{n}/2} \approx N(0, 1).$$

(b) In a matched pair experiment of $n = 9$ pairs, the actual differences are as follows.

$$-1, 1, 2, 3, 4, 4, 6, 7, 10.$$

Compare the results from the [two-sided] sign test and the t-test.

Sign test: P-value $= 2P(B(9, 0.5) \geq 8) = 0.039$.

t-test: $\bar{x} = 4$, $s = 3.32$.

$$T = \frac{\bar{x} - 0}{s/\sqrt{n}} = 3.62$$

$$\text{P-value} = 2P(t(n-1) \geq |T|) = 0.007.$$

(c) In a matched pair experiment of $n = 9$ pairs, the actual differences are as follows.

$$1, 1, 2, 3, 4, 4, 6, 7, -10.$$

Compare the results from the [two-sided] sign test and the t-test.

Sign test: P-value $= 2P(B(9, 0.5) \geq 8) = 0.039$.

t-test: $\bar{x} = 2$, $s = 4.95$.

$$T = \frac{\bar{x} - 0}{s/\sqrt{n}} = 1.21$$

$$\text{P-value} = 2P(t(n-1) \geq |T|) = 0.26.$$

2. Wilcoxon's signed-rank test. This test requires the measurements to be quantitative. Read the textbook Section 15.4 for more details.

Independent samples from two populations

Rank-sum test. Two populations, I and II. Sample size n_1 and n_2 respectively, and $n \doteq n_1 + n_2$. Wilcoxon [1945], Mann-Whitney [1947].

Determine the rank sum. Put all the observations into a single array in increasing order, the smallest observation given rank 1.

$$W_1 = \text{sum of ranks from Sample I.}$$
$$W_2 = \text{sum of ranks from Sample II.}$$

Remark:

1. If several observations tie, the ranks will be averaged.
2. W_1 is equivalent to W_2 in that $W_1 + W_2 = n(n+1)/2$.

Computation of P-value

Suppose population I has distribution $F_1(x)$ and population II has distribution $F_2(x) = F_1(x - \theta)$ [two-sample shift model]

P-value for one-sided test $H_0 : \theta = 0$, $H_a : \theta > 0$ [or $\theta < 0$]. Let $(r_1, r_2, \ldots, r_{n_1})$ be ranks with $r_1 < r_2 < \cdots < r_{n_1}$. Then

$$\text{P-value} = \frac{\#\text{ of } (r_1, r_2, \ldots, r_{n_1}) \text{ such that } r_1 + r_2 + \cdots + r_{n_1} \leq W_1 \text{ [or } \geq W_1].}{\binom{n_1 + n_2}{n_1}}$$

P-value for two-sided test $H_0 : \theta = 0$, $H_a : \theta \neq 0$. Let $(r_1, r_2, \ldots, r_{n_1})$ be distinct ranks such that $r_1 < r_2 < \cdots < r_{n_1}$. Then

$$\text{P-value} = 2 \cdot \frac{\#\text{ of } (r_1, r_2, \ldots, r_{n_1}) \text{ such that } r_1 + r_2 + \cdots + r_{n_1} \leq \min\{W_1, W_2\}}{\binom{n_1 + n_2}{n_1}}$$

Mann-Whitney Statistics: Note that one must identify sample I as the sample with smaller sample size in order to use the tables. That is, $n_1 \le n_2$.

$$U_1 \doteq n_1 n_2 + \frac{n_1(n_1+1)}{2} - W_1$$
$$U_2 \doteq n_1 n_2 + \frac{n_2(n_2+1)}{2} - W_2$$

Test statistics U_0. [Notation is different than textbook's]

1. $H_0: \theta = 0, H_a: \theta \neq 0$. $U_0 \doteq \min\{U_1, U_2\}$. P-value $= 2P(U \le U_0)$.
2. $H_0: \theta = 0, H_a: \theta > 0$. $U_0 \doteq U_2$. P-value $= P(U \le U_0)$.
3. $H_0: \theta = 0, H_a: \theta < 0$. $U_0 \doteq U_1$. P-value $= P(U \le U_0)$.

Examples

1. Weight gains for female rats under two diets.

$$\text{High protein}: \quad 134, 146, 104, 119, 124, 161, 107, 83, 113, 129$$
$$\text{Low protein}: \quad 70, 118, 101, 85, 107, 132, 94.$$

Is there any evidence to indicate that the two diets have different effects [use significance level $\alpha = 5\%$]?

Solution: $n_1 = 7, n_2 = 10$. Ranks

$$\text{High protein [sample II]}: \quad 15, 16, 6, 11, 12, 17, 7.5, 2, 9, 13$$
$$\text{Low protein [sample I]}: \quad 1, 10, 5, 3, 7.5, 14, 4.$$
$$W_1 = 44.5, \quad U_1 = 53.5; \quad W_2 = 108.5, \quad U_2 = 16.5$$
$$\text{Test statistics}: U_0 = \min\{U_1, U_2\} = 16.5.$$

$$\text{P-value} = 2P(U \leq U_0) = 2P(U \leq 16.5) = 2 \times 0.351 = 0.702.$$

2. The survival times for cats and rabbits, under anoxic conditions.

$$\text{Cats [sample I]}: \quad 25, 33, 43, 45$$
$$\text{Rabbits [sample II]}: \quad 16, 17, 20, 22, 22, 23, 28, 28, 30, 30$$

Is there any evidence to indicate that cats' survival times exceed those of Rabbits [use significance level $\alpha = 5\%$]?

Solution: $n_1 = 4, n_2 = 10$. Ranks

$$\text{Cats [sample I]}: \quad 7, 12, 13, 14$$
$$\text{Rabbits [sample II]}: \quad 1, 2, 3, 4.5, 4.5, 6, 8.5, 8.5, 10.5, 10.5$$

$$W_1 = 46, \quad U_1 = 4; \quad W_2 = 59, \quad U_2 = 36$$
$$\text{Test statistics}: U_0 = U_1 = 4.$$

$$\text{P-value} = P(U \leq U_0) = 0.012.$$

Remark: When $n_1 > 10$ and $n_2 > 10$, one can use normal approximation to compute P-value. That is,

$$Z = \frac{U_1 - (n_1 n_2)/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}} \approx N(0,1)$$

1. $H_0 : \theta = 0, H_a : \theta \neq 0.$ P-value $= 2P(N(0,1) \geq |Z|)$.
2. $H_0 : \theta = 0, H_a : \theta > 0.$ P-value $= P(N(0,1) \geq Z)$.
3. $H_0 : \theta = 0, H_a : \theta < 0.$ P-value $= P(N(0,1) \leq Z)$.

Test of population distribution

Sample: $X_1, X_2, \ldots, X_n$ from continuous population distribution $F(x)$ [the cumulative distribution function].

$$H_0 : F(x) = \bar{F}(x), \qquad H_a : F(x) \neq \bar{F}(x).$$

1. Kolmogorov-Smirnov test.

Test statistics D_n. Use empirical distribution $F_n(x)$, and let

$$D_n \doteq \sup_{x \in \mathbb{R}} |F_n(x) - \bar{F}(x)|.$$

Under the null hypothesis H_0, the distribution of D_n is *independent* of $\bar{F}$ [distribution free]. Denote this distribution G_n.

$$\text{P-value} = P(G_n \geq D_n).$$

2. χ^2 goodness-of-fit test.

Digression to the χ^2-test for multinomial distributions. Suppose $Y = (Y_1, Y_2, \ldots, Y_k)$ with $Y_1 + Y_2 + \cdots + Y_k = n$.

$$H_0 : Y \text{ is multinomial } (n; p_1, p_2, \ldots, p_k).$$

Test statistics:

$$X^2 \doteq \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Under H_0, when n large, X^2 is approximately $\chi^2(k-1)$.

$$\text{P-value} = P(\chi^2(k-1) \geq X^2).$$

Example. A course has 4 sections given at the same time but with different instructors. There are total 120 students, and the number of students in each section is

$$22, 27, 26, 45$$

Is there any evidence to indicate that the 4 sections are equally preferred by the students?

Solution: The number of students follows multinomial distribution with $(120; p_1, p_2, p_3, p_4)$.

$$H_0 : p_1 = p_2 = p_3 = p_4 = 1/4.$$

$$X^2 = \frac{(22-30)^2}{30} + \frac{(27-30)^2}{30} + \frac{(26-30)^2}{30} + \frac{(45-30)^2}{30} = 10.47$$

$$\text{P-value} = P(\chi^2(3) \geq 10.47) = 0.015.$$

χ^2 goodness-of-fit test for

$$H_0 : F(x) = \bar{F}(x), \quad H_a : F(x) \neq \bar{F}(x).$$

Procedure:

(a) Divide the real line by numbers $a_1 < a_2 < \cdots < a_{k-1}$

(b) Define intervals

$$I_1 = (-\infty, a_1],\ I_2 = (a_1, a_2], \cdots,\ I_k = (a_{k-1}, \infty).$$

(c) Let Y_i = number of $\{X_1, \ldots, X_n\}$ that fall into interval I_i.

(d) The hypotheses change to

$$H_0 :\ Y = (Y_1, \ldots, Y_k) \text{ is multinomial } (n; p_1^*, \ldots, p_k^*)$$

with

$$p_1^* = \bar{F}(a_1),\ p_2^* = \bar{F}(a_2) - \bar{F}(a_1), \cdots,\ p_k^* = 1 - \bar{F}(a_{k-1}).$$

(e) Form test statistics

$$X^2 = \sum_{i=1}^{k} \frac{(Y_i - np_i^*)^2}{np_i^*}$$

(f) When n large, X^2 is approximately $\chi^2(k-1)$.

(g)

$$\text{P-value} = P(\chi^2(k-1) \geq X^2).$$

Remark: In general, $\bar{F}(x)$ may not be completely prespecified, and may be partially determined by the sample itself. And the d.f. of the χ^2 distribution is

$$\text{d.f.} = k - 1 - \text{number of fitted parameters}$$

See examples below.

Examples

1. The 100 random samples from population distribution $F(x)$.

$$H_0 : F(x) \text{ is uniform on } [0,1].$$

Class limits	Observed	Expected	p_i^*
0-0.1	10	10	0.1
0.1-0.2	11	10	0.1
0.2-0.3	12	10	0.1
0.3-0.4	15	10	0.1
0.4-0.5	8	10	0.1
0.5-0.6	15	10	0.1
0.6-0.7	8	10	0.1
0.7-0.8	10	10	0.1
0.8-0.9	8	10	0.1
0.9-1	3	10	0.1
Total	100	100	1

$$X^2 = \frac{(10-10)^2}{10} + \cdots + \frac{(3-10)^2}{10} = 11.6$$

$$\text{P-value} = P(\chi^2(9) \geq 11.6) = 0.24$$

2. Heights of 1052 mothers with sample mean $\bar{X} = 62.49$ in. and sample standard deviation $s = 2.435$ in.

$$H_0 : \text{Height is normally distributed.}$$

Solution: $\bar{F}$ is not completely specified, except that it is normal. But $\bar{F}(x)$ should be approximately $N(62.49, 2.435^2)$. [2 parameters (μ, σ^2) use fitted values from sample].

Class limits	Observed	Expected	p_i^*
0-55	3	1.1	0.00105
55-57	9	11.6	0.01102
57-59	52	67.1	0.06381
59-61	215	204.5	0.19441
61-63	346	328.9	0.31265
63-65	277	279.6	0.26573
65-67	120	125.5	0.11932
67-69	24	29.7	0.02825
69-∞	6	3.9	0.00375
Total	1052	1052	1

$$X^2 = \frac{(3-1.1)^2}{1.1} + \cdots + \frac{(6-3.9)^2}{3.9} = 11.18$$

$$\text{d.f.} = 9 - 1 - 2 = 6.$$

$$\text{P-value} = P(\chi^2(6) \geq 11.18) = 0.083.$$