TECHNISCHE UNIVERSITÄT WIEN

Introduction to Statistics
Nonparametric Statistical Inference

LV Nr. 105.692

Summer Semester 2021

Nonparametric Statistical Inference

- Nonparametric statistical inference is a collective term given to inferences that are valid under less restrictive assumptions than with classical (parametric) statistical inference.
- The assumptions that can be relaxed include specifying the probability distribution of the population from which the sample was drawn and the level of measurement required of the sample data. For example, we may have to assume that the population is symmetric, which is much less restrictive than assuming the population is the normal distribution.
- The data may be ratings or ranks, i.e., measurements on an ordinal scale, instead of precise measurements on an interval or ratio scale. Or the data may be counts.
- In nonparametric inference, the null distribution of the statistic on which the inference is based does not depend on the probability distribution of the population from which the sample was drawn. In other words, the statistic has the same sampling distribution under the null hypothesis, irrespective of the form of the population distribution.

The Chi-square Goodness-of-Fit Test

The compatibility of a set of observed sample values with a normal or any other distribution can be checked with a goodness-of-fit type of test.

1. A random sample of size n is drawn from a population with unknown cdf F_X.
2. We want to test the hypotheses

$$H_0 : F_X(x) = F_0(x)\ \forall x \quad \text{vs.} \quad H_1 : F_X(x) \neq F_0(x) \text{ for some } x$$

   where F_0 is completely specified.

The Chi-square Goodness-of-Fit Test

- Assume the n observations have been grouped into k mutually exclusive categories
- Denote the observed and expected frequencies for the ith class by O_i and e_i, respectively, $i = 1, \ldots, k$.
- The test statistic suggested by Pearson (1900) is

$$X^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i} \tag{1}$$

- A large value of X^2 reflects an incompatibility between the observed and expected frequencies, supporting rejecting the null.
- The exact pdf of (1) is quite complicated, but for *large* samples,

$$X^2 \approx_{H_0} \chi^2(k-1)$$

with

$$p - value = \mathbb{P}(\chi^2(k-1) > X^2)$$

Justification

- Let n_i denote the random variables for class frequencies, $i = 1, \ldots, k$. Then,
$$(n_1, \ldots, n_k) \sim Multinomial(n, \theta_1, \ldots, \theta_k)$$
where θ_i is the probability associated with class i.
- The likelihood of the sample is
$$L(\theta_1, \ldots, \theta_k) = \prod_{i=1}^{k} \theta_i^{o_i}, \; o_i = 0, 1, \ldots, n, \; \sum_i \theta_i = 1, \; \sum_i o_i = n$$
- Then the null can be expressed as
$$H_0 : \theta_i = \theta_{i0} \quad \text{for } i = 1, \ldots, k$$

Justification

- The likelihood ratio statistic for this hypothesis is

$$\lambda = \frac{L(\hat{\theta}_1, \ldots, \hat{\theta}_k)}{L(\theta_{10}, \ldots, \theta_{k0})} = \prod_{i=1}^{k} \left(\frac{\hat{\theta}_i}{\theta_{i0}} \right)^{o_i}$$

  where $\hat{\theta}_i = o_i/n$ is the MLE of θ_i.
- The distribution of $-2 \log \lambda$ can be approximated by $\chi^2(k-1)$, since $\sum_i \theta_i = 1$ leaves $k-1$ free parameters to be estimated.

Justification

- **HW:** Show that

$$-2\log\lambda = -2\sum_{i=1}^{k} o_i \left(\log\theta_{i0} - \log(o_i/n)\right)$$

  is asymptotically equivalent to (1) using a Taylor series expansion.

- Then the distribution of X^2 converges to that of $-2\log\lambda$, which is $\chi^2(k-1)$ for large samples.

- Large samples: $e_i \geqslant 5 \; \forall i$

Example

A quality control engineer has taken 50 samples of size 13 each from a production process. The numbers of defectives for these samples are recorded below. Thest the null hypothesis that the number of defectives follows a Poisson distribution at $\alpha = 0.05$.

Number of defectives	**Number of samples**
0	10
1	24
2	10
3	4
4	1
5	1
6 or more	0

Example (ctd.)

- Since no parameters are specified, they must be estimated from the data.
  - Poisson pmf is $f(x) = e^{-\mu}\mu^x/x!$, $x = 0, 1, 2, \ldots$, where μ is the mean number of defectives in a sample of size 13.
  - The MLE of μ is the mean number of defectives in the 50 samples:

$$\hat{\mu} = \frac{0 \times 10 + 1 \times 24 + \ldots + 5 \times 1}{50} = \frac{65}{50} = 1.3$$

  - We use this value to compute $\hat{\theta}_i = o_i/n$ and $\hat{e}_i = 50 \times \hat{\theta}_i$

Example (ctd)

Number of defectives	o_i	$\hat{\theta}_i$	$\hat{e}_i$
0	10	0.2725	13.625
1	24	0.3543	17.715
2	10	0.2303	11.515
3	4	0.0998	4.990
4	1	0.324	1.620
5 or more	1	0.0107	0.535

We put together the last two categories of 4 and 5 or more to obtain,

$$X^2 = 3.6010 \approx_{H_0} \chi^2(3)$$

where we started with $k - 1 = 4$ and lose another for estimating θ.

$$p - value = \mathbb{P}(\chi^2(3) > 3.601) = 0.3078$$

so there is little evidence against the null the distn is Poisson.

The Sign Test and Confidence Interval for the Median

1. $X_1, \ldots, X_n$ iid F_X
2. F_X is assumed to be continuous and strictly increasing, at least in the vicinity of the unknown median M; i.e.

$$F_X^{-1}(M) = 0.5$$

3. We want to test $H_0 : M = M_0$, where M_0 is a specified value.
4. Since we assumed the median is unique, the null is equivalent to

$$H_0 : \theta = \mathbb{P}(X > M_0) = \mathbb{P}(X < M_0) = 0.5$$

The Sign Test and Confidence Interval for the Median

5. If the sample data are consistent with the hypothesized median value, on average half of the sample observations will lie above M_0 and half below.
6. Thus the number of observations larger than M_0, denoted by K, can be used to test the validity of the null.
7. Now, the observations are a random sample from the Bernoulli with parameter $\theta = \mathbb{P}(X > M_0)$, regardless of the population cdf F_X, so that

$$K \sim Bin(n, \theta)$$

   and $\theta = 0.5$ when the null holds.
8. Since K is the number of plus signs among the n differences $X_i - M_0$, $i = 1, \ldots, n$, the nonparametric test based on K is called the sign test.

The Sign Test and Confidence Interval for the Median

9. The rejection region for $H_1 : M > M_0$ or $\theta > 0.5$ is

$$K \geqslant C_\alpha$$

where C_α is chosen to be the smallest integer that satisfies

$$\mathbb{P}(K \geqslant C_\alpha \mid H_0) = \sum_{j=C_\alpha}^{n} \binom{n}{i} 0.5^n \leqslant \alpha$$

10. The rejection region for $H_1 : M < M_0$ or $\theta < 0.5$ is

$$K \leqslant C'_\alpha$$

where C_α is chosen to be the largest integer that satisfies

$$\mathbb{P}(K \leqslant C'_\alpha \mid H_0) = \sum_{j=0}^{C'_\alpha} \binom{n}{i} 0.5^n \leqslant \alpha$$

The Sign Test and Confidence Interval for the Median

11. The rejection region for $H_1 : M \neq M_0$ or $\theta \neq 0.5$ is

$$K \leqslant C'_{\alpha/2} \quad \text{or} \quad K \geqslant C_{\alpha/2}$$

where $C_{\alpha/2}$ and $C_{\alpha/2}$ are, respectively, the smallest and the largest integers that satisfie

$$\sum_{j=C_{\alpha/2}}^{n} \binom{n}{j} 0.5^n \leqslant \alpha/2 \quad \text{and} \quad \sum_{j=0}^{C'_{\alpha/2}} \binom{n}{j} 0.5^n \leqslant \alpha/2$$

12. Note that since the Binomial at $\theta = 0.5$ is symmetric,

$$C_{\alpha/2} = n - C'_{\alpha/2}$$

The Sign Test

H_1	RR at α level	Exact p-value
$\theta > 0.5$	$K \geqslant C_\alpha$	$\sum_{j=K_0}^{n} \binom{n}{j} 0.5^n$
$\theta < 0.5$	$K \leqslant C'_\alpha$	$\sum_{j=0}^{K_0} \binom{n}{j} 0.5^n$
$\theta \neq 0.5$	$K \leqslant C'_{\alpha/2}$ or $K \geqslant C_{\alpha/2}$	$2\times$ smaller of the one-tailed p-values

where K_0 is the observed value of K.

Large sample test for the median

For large sample sizes, the appropriate rejection regions and p-values based on the normal approximations to the binomial distribution with a continuity correction are:

H_1	RR at α level	Exact p-value
$\theta > 0.5$	$K \geqslant 0.5n + 0.5 + 0.5z_\alpha\sqrt{n}$	$1 - \Phi\left(\frac{K_0 - 0.5n - 0.5}{0.5\sqrt{n}}\right)$
$\theta < 0.5$	$K \leqslant 0.5n - 0.5 - 0.5z_\alpha\sqrt{n}$	$\Phi\left(\frac{K_0 - 0.5n + 0.5}{0.5\sqrt{n}}\right)$
$\theta \neq 0.5$	Both above with $z_{\alpha/2}$	$2\times$ smaller of the one-tailed p-values

Confidence Interval for the Median

- The acceptance region of the sign test yields a two-sided $100(1-\alpha)\%$ CI for an unknown population median:

$$C'_{\alpha/2} + 1 \leqslant K \leqslant C_{\alpha/2} - 1$$

where K is the number of positive differences among $X_i - M_0$, $i = 1, \ldots, n$, and $C_{\alpha/2}$ and $C'_{\alpha/2}$ are, respectively, the smallest and the largest integers that satisfy

$$\sum_{j=C_{\alpha/2}}^{n} \binom{n}{j} 0.5^n \leqslant \alpha/2 \quad \text{and} \quad \sum_{j=0}^{C'_{\alpha/2}} \binom{n}{j} 0.5^n \leqslant \alpha/2$$

Example

Suppose that each of 13 randomly selected female registered voters was asked to indicate if she was going to vote for candidate A or B in an upcoming election. The results show that 9 of the subjects preferred A. Is this sufficient evidence to conclude that A is preferred to B by female voters?

Answer: With this kind of data, the sign test is one of the few statistical tests that is valid and applicable.
Let θ be the true probability candidate A is preferred over B. The null is $H_0 : \theta = 0.5$ and the alternative $H_1 : \theta > 0.5$.
The test statistic is $K = 9$ with

$$p - value = \sum_{j=9}^{n} \binom{13}{j} 0.5^{13} = 0.1334$$

There is not sufficient evidence to conclude the female voters prefer A over B at as high a significant level as 0.10.

Example (ctd.)

In R: To verify results we can use the `binom.test()` from base `R`

```
> binom.test(9, n = 13, p = 0.5, alternative = "greater")

        Exact binomial test

data:  9 and 13
number of successes = 9, number of trials = 13, p-value = 0.1334

alternative hypothesis: true probability of success
                                        is greater than 0.5
95 percent confidence interval:
 0.4273807 1.0000000
sample estimates:
probability of success
             0.6923077
```

HW

Some researchers claim that susceptibility to hypnosis can be acquired or improved through training. To investigate this claim six subjects were rated on a scale of 1-20 according to their initial susceptibility to hypnosis and then given 4 weeks of training. Each subject was rated again after the training period. In the ratings below, higher numbers represent greater susceptibility to hypnosis. Do these data support the claim?

Subject	Before	After
1	10	18
2	16	19
3	7	11
4	4	3
5	7	5
6	2	3

Comparing two population distributions

- Suppose that two independent samples of sizes n_1 and n_2 are drawn from two **continuous** populations
- We want to test the null hypothesis of identical distributions against the location alternative that the populations have the same form but a different measure of central tendency.
- This can be expressed as

$$H_0 : F_Y(x) = F_X(x)\,\forall x \quad vs. \quad H_1 : F_Y(x) = F_X(x - \theta)\,\forall x \text{ and some } \theta \neq 0$$

  with
  - $\theta < 0$: distn of Y shifted to the left of distn of X
  - $\theta > 0$: distn of Y shifted to the right of distn of X

The Wilcoxon Signed-Rank Test

- Wilcoxon (1945) proposed a test where we accept
  - the one-sided location alternative $H_1 : \theta < 0$ ("X>Y ") if the sum of the ranks of the X's is too large, or
  - $H_1 : \theta > 0$ ("X<Y ") if the sum of the ranks of the X's is too small, and
  - the two-sided alternative $H_1 : \theta \neq 0$ ("X≠Y ") if the sum of the ranks of the X's is either too small or too large.
- The test statistic for the Wilcoxon test is based on the ranks for each of the two samples (rank sums):

$$W_n = \sum_{i=1}^{n} i Z_i$$

$n = n_1 + n_2$, and $Z_i = 1$ if the ith random variable in the combined ordered sample is an X and 0 if it is a Y.

The Wilcoxon Signed-Rank Test

- If there are no ties, the exact mean and variance under the null are

$$\mathbb{E}(W_n) = \frac{n_1(n+1)}{2}, \quad \mathbb{V}\mathrm{ar}(W_n) = \frac{n_1 n_2(n+1)}{12}$$

  [HW]

- If $n_1 \leqslant n_2$, W_n has a minimum value of

$$\sum_{i=1}^{n_1} i = \frac{n_1(n_1+1)}{2}$$

  and a maximum of

$$\sum_{i=n-n_1+1}^{n} i = \frac{n_1(2n-n_1+1)}{2}$$

- We can also show that the test statistic is symmetric about its mean under $\theta = 0$ (see Gibbons and Chakraborti)

The Wilcoxon Signed-Rank Test

- The exact null probability distribution can be obtained by enumeration using these properties.
  - For example, if $n_1 = 3, n_2 = 4$, there are $\binom{7}{3} = 35$ possible distinguishable configurations of 1's and 0's in the vector $\mathbf{Z}$.
  - W_n will range between 6 and 18, symmetric about 12. For example, $W_n = 18$ if the 3 X values are 5,6,7, and $W_n = 17$ if the X values are 4,6,7. Thus $\mathbb{P}(W_n \geqslant 17) = 2/35$.
- Ties: Assign tied measurements the average of the ranks they would receive if they were unequal but occurred in successive order. For example, if the third-ranked and fourth-ranked measurements are tied, assign each a rank of $(3+4)/2 = 3.5$.

The Wilcoxon Signed-Rank Test: $n_1 \leqslant n_2 \leqslant 10$

H_1	RR at α level	p-value
$\theta < 0$	$W_n \geqslant w_\alpha$	$\mathbb{P}(W_n \geqslant w_0)$
$\theta > 0$	$W_n \leqslant w'_\alpha$	$\mathbb{P}(W_n \leqslant w_0)$
$\theta \neq 0$	$W_n \geqslant w_{\alpha/2}$ or $W_n \leqslant w'_{\alpha/2}$	$2\times$ smaller of above

where w_0 is the observed value of W_n. The critical values are given in tables for $n_1 \leqslant n_2 \leqslant 10$.

The Wilcoxon Signed-Rank Test: Large samples

- For larger sample sizes, the rejection regions and p-values are based on the normal approximation with a continuity correction as follows.

H_1	RR at α level	p-value
$\theta < 0$	$W_n \geqslant \frac{n_1(n+1)}{2} + 0.5 + z_\alpha \sqrt{\frac{n_1 n_2 (n+1)}{12}}$	$1 - \phi\left(\frac{w_0 - 0.5 - n_1(n+1)/2}{\sqrt{\frac{n_1 n_2 (n+1)}{12}}}\right)$
$\theta > 0$	$W_n \leqslant \frac{n_1(n+1)}{2} - 0.5 - z_\alpha \sqrt{\frac{n_1 n_2 (n+1)}{12}}$	$\phi\left(\frac{w_0 + 0.5 - n_1(n+1)/2}{\sqrt{\frac{n_1 n_2 (n+1)}{12}}}\right)$
$\theta \neq 0$	Both above with $z_{\alpha/2}$	$2\times$ smaller of above

where w_0 is the observed value of W_n.

The Wilcoxon Signed-Rank Test Procedure

- Assign ranks, R_i, to the $n_1 + n_2$ sample observations
  - If unequal sample sizes, let n_1 refer to smaller-sized sample
  - Smallest rank value = 1
  - Average ties
- Sum the ranks for each sample
- Test statistic is W_n (smallest sample)

Example

You're a production planner. You want to see if the operating rates for two factories is the same.

For factory 1, the rates (% of capacity) are 85, 82, 94, and 97 ($n_1 = 4$)

For factory 2, the rates are 71, 82, 77, 92, and 88 ($n_2 = 5$).

Do the factory rates have the same probability distributions at the .10 level of significance?

Answer

1. We have to test

$$H_0 : F_Y(x) = F_X(x) \quad vs \quad H_0 = F_Y(x) = F_X(x - \theta),\ \theta \neq 0$$

2. From the Wilcoxon Rank Sum Table (Portion) for $\alpha = 0.05$ one-tailed, $\alpha = .10$ two-tailed the critical values are $w'_{0.05} = 12$ and $w_{0.05} = 28$.

3. Compute the ranks:

Factory 1		Factory 2	
Rate	Rank	Rate	Rank
85	5	71	1
82	3 **3.5**	82	4 **3.5**
94	8	77	2
97	9	92	7
		88	6
Rank Sum	25.5		19.5

Answer

4. Test Statistic is based on the smaller sample $n_1 = 4$:

$$W_n = 5 + 3.5 + 8 + 9 = 25.5$$

5. Since $W_n > 12$ and $W_n < 28$, we cannot reject the null that the two distributions are identical.

5. From Table J in Gibbons and Chakraborti (pp. 575),

$$p - value = 2 \times .143 = .286$$

Descriptive Statistics

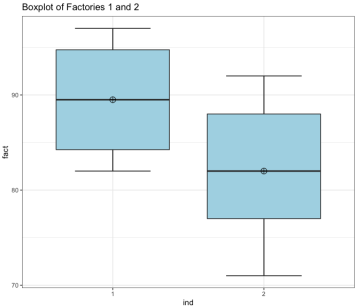
```
library("gmodels")
library("car")
library("DescTools")
library("ggplot2")
library("qqplotr")
library("dplyr")

fact1=c(85,82,94,97)
fact2=c(71,82,77,92,88)

dat=data.frame(cbind(fact,ind))
dat$ind<-as.factor(dat$ind)
```

Descriptive Statistics

```
#Produce descriptive statistics by group
dat %>% select(ind, fact) %>% group_by(ind) %>%
  summarise(n = n(),
            mean = mean(fact, na.rm = TRUE),
            sd = sd(fact, na.rm = TRUE),
            stderr = sd/sqrt(n),
            LCL = mean - qt(1 - (0.05 / 2), n - 1) * stderr,
            UCL = mean + qt(1 - (0.05 / 2), n - 1) * stderr,
            median = median(fact, na.rm = TRUE),
            min = min(fact, na.rm = TRUE),
            max = max(fact, na.rm = TRUE),
            IQR = IQR(fact, na.rm = TRUE),
            LCLmed = MedianCI(fact, na.rm=TRUE)[2],
            UCLmed = MedianCI(fact, na.rm=TRUE)[3])
# A tibble: 2 x 13
  ind       n  mean    sd stderr   LCL   UCL median   min   max
IQR LCLmed UCLmed
* <fct> <int> <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
<dbl>  <dbl>
1 1         4  89.5  7.14   3.57  78.1 101.    89.5    82    97
10.5   -Inf    Inf
2 2         5  82    8.40   3.75  71.6  92.4   82      71    92
11     -Inf    Inf
```

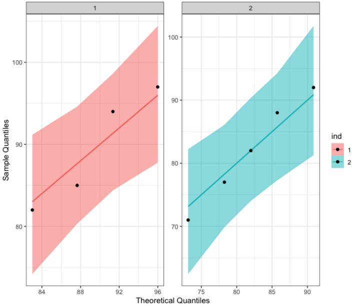
Box-plots

```
#Produce Boxplots
#and visually check for outliers

ggplot(dat, aes(x = ind, y = fact, fill = ind))
  stat_boxplot(geom ="errorbar", width = 0.5) +
  geom_boxplot(fill = "light blue") +
  stat_summary(fun=mean, geom="point", shape=10
  size=3.5, color="black") +
 ggtitle("Boxplot of Factories 1 and 2") +
  theme_bw() + theme(legend.position="none")
```



QQ Plots

```
#Perform QQ plots by group
#to check for normality

ggplot(data = dat, mapping = aes(sample = fact,
color = ind, fill = ind)) +
  stat_qq_band(alpha=0.5, conf=0.95,
   qtype=1, bandType = "boot") +
  stat_qq_line(identity=TRUE) +
  stat_qq_point(col="black") +
  facet_wrap(~ ind, scales = "free") +
  labs(x = "Theoretical Quantiles",
   y = "Sample Quantiles") +
  theme_bw()
```



Example: Children?'s recall of TV ads.

A study of children's recall of television advertisements appeared in the *Journal of Advertising* (2006), where two groups of children were shown a 60-second commercial for Sunkist FunFruit Rock-n-Roll Shapes. One group (the A/V group) was shown the ad with both audio and video; the second group (the video only group) was shown only the video portion of the commercial. The number of 10 specific items from the ad recalled correctly by each child is shown in the table. The researchers theorized that children who receive an audiovisual presentation will have the same level of recall as those who receive only the visual aspects of the ad. Test the researchers' theory using the Wilcoxon rank sum test.

A/V	0	4	6	6	1	2	2	6	6	4	1	2	6	1	3	0	2	5	4	5
Video only	6	3	6	2	2	4	7	6	1	3	6	2	3	1	3	2	5	2	4	6

Answer

1. We have to test

$$H_0 : F_Y(x) = F_X(x) \quad vs \quad H_1 = F_Y(x) = F_X(x-\theta),\ \theta \neq 0$$

2. $n_1 = n_2 = 20$, so we can use the asymptotic normal test.

3. Compute the ranks:

Ranks

	AV	Rank	Video	Rank
1	0.00	1.50	6.00	34.50
2	4.00	24.00	3.00	19.00
3	6.00	34.50	6.00	34.50
4	6.00	34.50	2.00	12.00
5	1.00	5.00	2.00	12.00
6	2.00	12.00	4.00	24.00
7	2.00	12.00	7.00	40.00
8	6.00	34.50	6.00	34.50
9	6.00	34.50	1.00	5.00
10	4.00	24.00	3.00	19.00
11	1.00	5.00	6.00	34.50
12	2.00	12.00	2.00	12.00
13	6.00	34.50	3.00	19.00
14	1.00	5.00	1.00	5.00
15	3.00	19.00	3.00	19.00
16	0.00	1.50	2.00	12.00
17	2.00	12.00	5.00	28.00
18	5.00	28.00	2.00	12.00
19	4.00	24.00	4.00	24.00
20	5.00	28.00	6.00	34.50
Sum		385.5		434.5

The p-value is 2 $\times$ the smaller of

$$1-\Phi\left(\frac{w_0-0.5-n_1(n+1)/2}{\sqrt{\frac{n_1n_2(n+1)}{12}}}\right)=1-\phi\left(\frac{385.5-0.5-(20\times 41)/2}{\sqrt{\frac{20\times 20\times 41}{12}}}\right)$$
$$=1-\Phi(-0.6762522)=0.7505597$$

or

$$\Phi\left(\frac{w_0+0.5-n_1(n+1)/2}{\sqrt{\frac{n_1n_2(n+1)}{12}}}\right)=\Phi\left(\frac{385.5+0.5-(20\times 41)/2}{\sqrt{\frac{20\times 20\times 41}{12}}}\right)$$
$$=\Phi(-0.6492021)=0.2581039$$

which gives

$$p-value=0.5162078$$

so we cannot reject the null that the two distributions are the same.

NB: The t-test does not find significant difference in the means as well:

```
> t.test(fruit$AV,fruit$VIDEO)

        Welch Two Sample t-test

data:  fruit$AV and fruit$VIDEO
t = -0.61955, df = 37.516, p-value = 0.5393
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.7075704  0.9075704
sample estimates:
mean of x mean of y
      3.3       3.7
```

Reference

The slides have been based on

Gibbons JD, Chakraborti S (2010). *Nonparametric statistical inference*, 5th edn. Taylor & Francis/CRC Press, Boca Raton.

and some problems are taken from *Statistics for Business and Economics* (ed. 12) by McClave, Benson and Sincich.