STATISTICAL INFERENCE (MA862)

Lecture Slides

Topic 6: Non-parametric Tests

Non-parametric Inference

- X has a CDF F with known functional form except perhaps some parameters. In this case, we need to find value of the unknown parameters based on a sample. This is known as parametric inference.
- X has a CDF F who's functional form is unknown. In this case, we need to estimate a parametric function or test a statistical hypothesis without known functional form of the CDF. This is known as non-parametric inference.
- In this course, we will mainly talk about non-parametric tests some practically meaningful statistical hypotheses.

Order Statistics

- Let $X_1, X_2, \ldots, X_n$ denote a random sample from a population with continuous CDF F.
- The probability of any two or more of these random variables have equal magnitude is zero.
- Let us define
  - $X_{(1)}$: smallest of $X_1, X_2, \ldots, X_n$.
  - $X_{(2)}$: second smallest of $X_1, X_2, \ldots, X_n$.

    $\vdots$
  - $X_{(n)}$: largest of $X_1, X_2, \ldots, X_n$.
- Then $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ denotes the original random sample after arrangement in increasing order of magnitude.
- These random variables are collectively termed the order statistics corresponding to the random sample $X_1, X_2, \ldots, X_n$.

Order Statistics

- For $r = 1, 2, \ldots, n$, the r-th smallest $X_{(r)}$ is called r-th order statistic.
- For odd n, the sample median is defined by $X_{\left(\frac{n+1}{2}\right)}$. For even n, it is any number between $X_{(\frac{n}{2})}$ and $X_{(\frac{n}{2}+1)}$. The sample median is a measure of central tendency.
- The sample midrange is defined by $\frac{X_{(1)}+X_{(n)}}{2}$. It is also a measure of central tendency.
- The sample range is defined by $X_{(n)} - X_{(1)}$. This is a measure of dispersion.

Joint Distribution of Order Statistics

Theorem 6.1: Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order statistics corresponding to a random sample of size n from a population having PDF $f_X(\cdot)$. Then the joint PDF of the order statistics is

$$f_{X_{(1)}, X_{(2)}, \ldots, X_{(n)}}(x_1, x_2, \ldots, x_n) = n! \prod_{i=1}^{n} f_X(x_i) \text{ if } x_1 < x_2 < \ldots < x_n.$$

Distribution of $X_{(r)}$

Theorem 6.2: Let $X_{(r)}$ be the r-th order statistic corresponding to a random sample of size n from a continuous CDF $F_X(\cdot)$. Then, the CDF of $X_{(r)}$ is

$$F_{X_{(r)}}(t) = \sum_{i=r}^{n} \binom{n}{i} [F_X(t)]^i \, [1 - F_X(t)]^{n-i} \;\text{ for } t \in \mathbb{R}.$$

Theorem 6.3: Let $X_{(r)}$ be the r-th order statistic corresponding to a random sample of size n from a continuous CDF $F_X(\cdot)$ with corresponding PDF $f_X(\cdot)$. Then, the PDF of $X_{(r)}$ is

$$f_{X_{(r)}}(t) = \frac{n!}{(r-1)!(n-r)!} [F_X(t)]^{r-1} \, [1 - F_X(t)]^{n-r} \, f_X(t) \text{ for } t \in \mathbb{R}.$$

Distribution of $X_{(r)}$

Corollary 6.1: For a random sample of size n from $U(0, 1)$ distribution, the CDF of the r-th order statistic is

$$F_{X_{(r)}}(x) = \sum_{i=r}^{n} \binom{n}{i} x^i (1-x)^{n-i} \text{ for } 0 < x < 1.$$

Corollary 6.2: For a random sample of size n from $U(0, 1)$ distribution, the r-th order statistic follows a $beta(r, n-r+1)$ distribution with PDF

$$f(x) = \frac{n!}{(r-1)!(n-r)!} x^{r-1} (1-x)^{n-r} \text{ for } 0 < x < 1.$$

Joint Distribution of Subset of Order Statistics

Theorem 6.4: Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order statistics corresponding to a random sample of size n from a population having PDF $f_X(\cdot)$ and CDF $F_X(\cdot)$. Then, for $1 \leq r_1 < r_2 < \ldots < r_k \leq n$ and $1 \leq k \leq n$, the joint PDF of $X_{(r_1)}, X_{(r_2)}, \ldots, X_{(r_k)}$ is

$$
\begin{aligned}
& f_{X_{(r_1)}, X_{(r_2)}, \ldots, X_{(r_k)}}(x_1, x_2, \ldots, x_k) \\
& = \frac{n!}{(r_1 - 1)!(r_2 - r_1 - 1)! \ldots (n - r_k)!} \\
& \quad \times [F_X(x_1)]^{r_1 - 1} [F_X(x_2) - F_X(x_1)]^{r_2 - r_1 - 1} \ldots [1 - F_X(x_k)]^{n - r_k} \\
& \quad \times f_X(x_1) f_X(x_2) \ldots f_X(x_k),
\end{aligned}
$$

for $x_1 < x_2 < \ldots < x_k$.

Probability-Integral Transform

Theorem 6.5: Let X be a random variable with CDF $F_X(\cdot)$. If $F_X(\cdot)$ is continuous, then $F_X(X) \sim U(0, 1)$.

Corollary 6.3: If $X_1, X_2, \ldots, X_n$ be a random sample from a continuous CDF $F_X(\cdot)$, then $F_X(X_1), F_X(X_2), \ldots, F_X(X_n)$ is a random sample from $U(0, 1)$ distribution.

Corollary 6.4: Let $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ be the order statistics corresponding to a random sample of size n from a population having continuous CDF $F_X(\cdot)$. Then, the distribution of

$$F_X(X_{(1)}) < F_X\left(X_{(2)}\right) < \ldots < F_X(X_{(n)})$$

is same as that of the order statistics corresponding to a random sample of size n from $U(0, 1)$ distribution. a random sample of

Quantile Function

Definition 6.1: Let X be a random variable with CDF $F_X(\cdot)$. The function $Q_X : (0,\, 1) \to \mathbb{R}$, defined by

$$Q_X(p) = F^{-1}(p) = \inf\left\{x \in \mathbb{R} : F_X(x) \geq p\right\}$$

is known as quantile function (QF) of the random variable X. For $0 < p < 1$, $Q_X(p)$ is known as p-th quantile of X.

Remark 6.1:

- The 0.5-th quantile is known as population median.
- The first quartile is 0.25-th quantile, the second quartile is 0.50-th quantile, and the third quartile is 0.75-th quantile.
- The CDF and QF provide similar information regarding the distribution of the random variable.
- Different moments can be expressed in terms of QF.

Empirical Distribution Function

Definition 6.2: For a random sample of size n from the distribution with CDF $F_X(\cdot)$, the empirical distribution function (EDF) , $S_n : \mathbb{R} \to [0, 1]$, is defined by

$$S_n(x) = \frac{\text{number of sample values } \leq x}{n}.$$

Remark 6.2: The EDF is most conveniently defined in terms of the order statistics as

$$S_n(x) = \begin{cases} 0 & \text{if } x < X_{(1)} \\ \frac{i}{n} & \text{if } X_{(i)} \leq x < X_{(i+1)},\ i = 1, 2, \ldots, n-1 \\ 1 & \text{if } x \geq X_{(n)}. \end{cases}$$

Some Properties of EDF

Theorem 6.6: For fixed $x \in \mathbb{R}$, $T_n(x) \sim Bin(n, F_X(x))$, where $T_n(x) = nS_n(x)$.

Corollary 6.5: For any fixed $x \in \mathbb{R}$, $E\left(S_n(x)\right) = F_X(x)$ and $Var(S_n(x)) = \frac{F_X(x)(1-F_X(x))}{n}$.

Corollary 6.6: For any fixed $x \in \mathbb{R}$, $S_n(x)$ is consistent estimator of $F_X(x)$.

Theorem 6.7: For any fixed $x \in \mathbb{R}$,

$$\frac{\sqrt{n}\left[S_n(x) - F_X(x)\right]}{\sqrt{F_X(x)\left[1 - F_X(x)\right]}} \xrightarrow{\mathcal{D}} Z \sim N(0, 1).$$

Theorem 6.8: (Glivenko-Cantelli Theorem) $S_n(\cdot)$ converges uniformly to $F_X(\cdot)$ with probability 1, *i.e.*,

$$P\left[\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left|S_n(x) - F_X(x)\right| = 0\right] = 1.$$

Test for Randomness

- 10 persons (M-5, F-5) waiting in a queue for movie tickets.
- The arrangement is M, F, M, F, M, F, M, F, M, F.
- Would it be considered as a random arrangement of genders?

Test for Randomness

- 10 persons (M-5, F-5) waiting in a queue for movie tickets.
- The arrangement is M, F, M, F, M, F, M, F, M, F.
- Would it be considered as a random arrangement of genders?
- F, F, F, F, F, M, M, M, M, M.
- M, M, M, M, M, F, F, F, F, F.
- M, M, F, F, F, M, F, M, M, F.

Test for Randomness

- A ordered sequence of two types of symbols (or objects).
- Length of the sequence is n.
- n_1 : number of Type-I symbol
- n_2 : number of Type-II symbol
- $n = n_1 + n_2$
- We want to test

$$H_0 : \text{the arrangement of the } n \text{ symbols is random}$$

against

$$H_1 : \text{the arrangement of the } n \text{ symbols is not random.}$$

Run

Definition 6.3: Given an ordered sequence of two type of symbols, a run is defined to be a succession of one type of symbols that are followed or preceded by a different symbol or no symbol at all.

Example 6.1:

- M, F, M, F, M, F, M, F, M, F — 10 runs (5 of M, 5 of F)
- F, F, F, F, F, M, M, M, M, M — 2 runs (1 of M, 1 of F)
- M, M, M, M, M, F, F, F, F, F — 2 runs (1 of M, 1 of F)
- M, M, F, F, F, M, F, F, M, M — 5 runs (3 of M, 2 of F)

Test based on Total Number of Runs

- A ordered sequence of two types of symbols (or objects).
- Length of the sequence is n.
- n_1 : number of Type-I symbol
- n_2 : number of Type-II symbol
- $n = n_1 + n_2$
- R_1 : Number of runs of Type-I symbol
- R_2 : Number of runs of Type-II symbol
- $R = R_1 + R_2$: Number of total runs
- H_0 is rejected if and only if R is too small or too large
- Need the null distribution of R

Exact Null Distribution of R

Lemma 6.1: The number of distinguishable ways of distributing n-like objects into r distinguishable cells with no cell empty is $\binom{n-1}{r-1}$, $n \geq r \geq 1$.

Theorem 6.9: Under H_0, the joint probability mass function of R_1 and R_2 is

$$f_{R_1, R_2}(r_1, r_2) = \frac{c\binom{n_1-1}{r_1-1}\binom{n_2-1}{r_2-1}}{\binom{n_1+n_2}{n_1}},$$

for $(r_1, r_2) \in \{(a, b) \in N : a = b \text{ or } a = b \pm 1\}$, where $N = \{1, 2, \ldots, n_1\} \times \{1, 2, \ldots, n_2\}$, $c = 2$ if $r_1 = r_2$ and $c = 1$ if $r_1 = r_2 \pm 1$.

Exact Null Distribution of R

Corollary 6.7: Under H_0, the marginal probability mass function of R_1 is

$$f_{R_1}(r_1) = \frac{\binom{n_1-1}{r_1-1}\binom{n_2+1}{r_1}}{\binom{n_1+n_2}{n_1}} \text{ for } r_1 = 1,\, 2,\, \ldots,\, n_1.$$

Corollary 6.8: Under H_0, the marginal probability mass function of R_2 is

$$f_{R_2}(r_2) = \frac{\binom{n_2-1}{r_2-1}\binom{n_1+1}{r_1}}{\binom{n_1+n_2}{n_2}} \text{ for } r_2 = 1,\, 2,\, \ldots,\, n_2.$$

Exact Null Distribution of R

Theorem 6.10: The probability mass function of R in a random sample is

$$f_R(r) = \begin{cases} \dfrac{2\dbinom{n_1-1}{\frac{r}{2}-1}\dbinom{n_2-1}{\frac{r}{2}-1}}{\dbinom{n_1+n_2}{n_1}} & \text{if } r \text{ is even} \\[3em] \dfrac{\dbinom{n_1-1}{\frac{r-1}{2}}\dbinom{n_2-1}{\frac{r-3}{2}}+\dbinom{n_1-1}{\frac{r-3}{2}}\dbinom{n_2-1}{\frac{r-1}{2}}}{\dbinom{n_1+n_2}{n_1}} & \text{if } r \text{ is odd,} \end{cases}$$

for $r = 2, 3, \ldots, n$.

Exact Null Distribution of R

Example 6.2: If $n_1 = 5$ and $n_2 = 4$, then under H_0

$$
\begin{aligned}
f_R(9) &= \frac{\binom{4}{4}\binom{3}{3}}{\binom{9}{4}} = \frac{1}{126} \approx 0.008, \\
f_R(8) &= \frac{2\binom{4}{3}\binom{3}{3}}{\binom{9}{4}} = \frac{8}{126} \approx 0.063, \\
f_R(3) &= \frac{\binom{4}{1}\binom{3}{0} + \binom{4}{0}\binom{3}{1}}{\binom{9}{4}} = \frac{7}{126} \approx 0.056, \\
f_R(2) &= \frac{2\binom{4}{0}\binom{3}{0}}{\binom{9}{4}} = \frac{2}{126} \approx 0.016.
\end{aligned}
$$

For a two-sided test that rejects the null hypothesis for $R \leq 2$ or $R \geq 9$, the exact significance level α is $\frac{3}{126} \approx 0.024$.

Moments of R under H_0

Theorem 6.11: The first two central moment of R under H_0 is

$$E(R) = 1 + \frac{2n_1n_2}{n},$$

$$Var(R) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{n^2(n-1)}.$$

Asymptotic Test

Theorem 6.12: Suppose that the total sample size n increases to ∞ such a way that $\frac{n_1}{n} \to \lambda$, where $0 < \lambda < 1$ is a fixed number. Then under H_0,

$$\frac{R - 2n\lambda(1-\lambda)}{2\sqrt{n}\lambda(1-\lambda)} \xrightarrow{\mathcal{D}} N(0,\, 1).$$

- Using the normal approximation, the null hypothesis of randomness would be rejected at level α if and only if

$$\left| \frac{R - 2n\lambda(1-\lambda)}{2\sqrt{n}\lambda(1-\lambda)} \right| > z_{\frac{\alpha}{2}}.$$

Tests of Goodness-of-Fit

- Want to know if the given sample compatible to a particular distribution or not.
- The null hypothesis is about the form the CDF of the parent distribution.
- Let $X_1, \ldots, X_n$ be a random sample from unknown CDF $F(\cdot)$.
- $H_0 : F(x) = F_0(x)$ for all $x \in \mathbb{R}$ against $H_1 : F(x) \neq F_0(x)$ for some $x \in \mathbb{R}$.
- Ideally, null hypothesis completely specifies the distribution.
- We hope to accept the null hypothesis.
- Rejection of null hypothesis does not provide much specific information.
- Two types of tests will be discussed:
  - Graphical test — Q-Q plot
  - Formal Statistical tests — χ^2 Goodness-of-Fit, KS test

The Chi-square Goodness-of-Fit Test

- The sample data must be grouped according to some scheme in order to form a frequency distribution.
- k: Number of categories.
- f_i : Frequency of the i-th category.
- $e_i = n \times P_{H_0}(\text{a random observation belongs to } i\text{-th category})$: Expected frequency of the i-th category.
- The test statistic is

$$Q = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}.$$

- For large sample, the distribution of Q under H_0 can be approximated by χ^2-distribution with d.f. $k-1$.
- Reject H_0 at level α if and only if $Q > \chi^2_{k-1,\alpha}$.

The Chi-square Goodness-of-Fit Test

- **Information:** In the context of LRT, $-2\ln\Lambda$ converges to $\chi^2_{k_1-k_2}$ distribution as $n \to \infty$, where k_1 and k_2 are, respectively, the dimension of the spaces $\Theta_0 \cup \Theta_1$ and Θ_0, $k_1 > k_2$.
- Using the above fact, the use of Chi-square test can be justified.
- If $F_0(\cdot)$ does not specify the distribution completely, one can use MLE of the unknown parameters (based on grouped data). In this case, H_0 is rejected at level α if and only if $Q > \chi^2_{k-1-s,\alpha}$, where s is the number of unknown parameters.

Kolmogorov-Smirnov Test

- $H_0 : F(x) = F_0(x)$ for all x ag. $H_1 : F(x) \neq F_0(x)$ for some x.
- It is assumed that $F_0(\cdot)$ is continuous.
- The test statistic is

$$D_n = \sup_x |S_n(x) - F_0(x)| .$$

- Large value of D_n implies disagreement with H_0.
- Thus, rejection region is of the form $D_n > k$.

Kolmogorov-Smirnov Test

Theorem 6.13: The statistic D_n is distribution-free for any specified continuous CDF $F_0(\cdot)$.

Theorem 6.14: (Exact null distribution of D_n) Let $F_0(\cdot)$ be continuous. Then under H_0, we have for $0 < v < \frac{2n-1}{n}$

$$P\left(D_n < \frac{1}{2n} + v\right)$$

$$= \int_{\frac{1}{2n}-v}^{\frac{1}{2n}+v} \int_{\frac{3}{2n}-v}^{\frac{3}{2n}+v} \cdots \int_{\frac{2n-1}{2n}-v}^{\frac{2n-1}{2n}+v} f\left(u_1,\, u_2,\, \ldots\, u_n\right) du_n du_{n-1} \ldots du_1,$$

where

$$f\left(u_1,\, u_2,\, \ldots,\, u_n\right) = \begin{cases} n! & \text{for } 0 < u_1 < u_2 < \ldots < u_n < 1 \\ 0 & \text{otherwise.} \end{cases}$$

The above probability is zero and one for $v \leq 0$ and $v \geq \frac{2n-1}{n}$, respectively.

Kolmogorov-Smirnov Test

Theorem 6.15: (Large sample null distribution of D_n) If $F_0(\cdot)$ is continuous, then under H_0 for every $d > 0$,

$$\lim_{n\to\infty} P\left(D_n \leq \frac{d}{\sqrt{n}}\right) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 d^2}.$$

CI for Population Quantile

- Let the underlying CDF is $F(\cdot)$.
- Assume that $F(\cdot)$ is continuous and strictly increasing.
- $\kappa_p = Q(p)$: The p-th quantile.
- We are interested to find confidence interval for κ_p based on $X_{(r)}$ and $X_{(s)}$ for $r < s$.
- To find $100(1-\alpha)\%$ CI for κ_p, we need to find two integers r and s with $1 \le r < s \le n$ such that

$$P\left(X_{(r)} \le \kappa_p \le X_{(s)}\right) = 1 - \alpha.$$

- Note that

$$P\left(X_{(r)} < \kappa_p < X_{(s)}\right) = P\left(X_{(r)} < \kappa_p\right) - P\left(X_{(s)} < \kappa_p\right).$$

CI for Population Quantile

- $P\left(X_{(r)} < \kappa_p\right) = P\left(U_{(r)} < p\right) = \int_0^p n\binom{n-1}{r-1}x^{r-1}(1-x)^{n-r}dx.$
- Thus, we need to find two integers r and s such that

$$\int_0^p n\binom{n-1}{r-1}x^{r-1}(1-x)^{n-r}dx \\ - \int_0^p n\binom{n-1}{s-1}x^{s-1}(1-x)^{n-s}dx = 1-\alpha.$$

- In general, two unknowns (r and s) cannot be uniquely found from one equation. We need to impose some other restrictions. For example, we may consider equal tail CI.

CI for Population Quantile

- Note that

$$P\left(X_{(r)} < \kappa_p\right) = P\left(U_{(r)} < p\right) = \sum_{i=r}^{n} \binom{n}{i} p^i (1-p)^{n-i}.$$

- Thus,

$$P\left(X_{(r)} < \kappa_p < X_{(s)}\right) = \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}.$$

- Thus, alternatively, we need to find r and s such that

$$\sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} = 1 - \alpha.$$

CI for Population Quantile

- For $n > 20$, it is difficult to use the previous method.
- In this case, one can use normal approximation to the binomial distribution with a continuity correction.
- Let $K \sim Bin(n, p)$.
- Then, for k in the support of K,

$$P\left(K \leq k\right) \approx \Phi\left(\frac{k + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right).$$

- Thus, for asymptotic equal tail CI of κ_p, we can take

$$r = \left\lfloor np + \frac{1}{2} - z_{\frac{\alpha}{2}}\sqrt{np(1-p)}\right\rfloor$$

and

$$s = \left\lceil np + \frac{1}{2} + z_{\frac{\alpha}{2}}\sqrt{np(1-p)}\right\rceil.$$

Hypothesis Testing for Population Quantile

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with CDF $F(\cdot)$, a continuous distribution function.
- We want to test $H_0 : \kappa_p = \kappa^0$ against $H_1 : \kappa_p \neq \kappa^0$.
- Let K be the number of observations greater than κ^0.
- Too big or too small observed value of K indicate evidence against H_0.
- Thus, H_0 is rejected if and only if $K \leq r$ or $K \geq s$, $0 \leq r < s \leq n$.
- For a level α test, r and s satisfy

$$\sum_{i=0}^{r} \binom{n}{i} (1-p)^i p^{n-i} \leq \frac{\alpha}{2}$$

and

$$\sum_{i=s}^{n} \binom{n}{i} (1-p)^i p^{n-i} \leq \frac{\alpha}{2}.$$

Hypothesis Testing for Population Quantile

- For large sample size ($n > 20$), one can use normal approximation.
- The critical region for a level α test is given by

$$K \leq n(1-p) + \frac{1}{2} - z_{\frac{\alpha}{2}} \sqrt{np(1-p)}$$

or

$$K \geq n(1-p) + \frac{1}{2} + z_{\frac{\alpha}{2}} \sqrt{np(1-p)}.$$

- Problem of zeros.