MANA130083.01 Nonparametric *Spring 2025*

Lecture 2

Instructor: Prof. Bowen Gang *Scribes: Yize Wang, Jingyi Zhou*

2.1 Interval for Next Observations

We want a random interval such that the probability that the next observation will be contained in the interval is high. Say you collect some data $X_1, X_2, ..., X_n$. Then you can calculate the interval based on the data. The most natural thing to do is to look at the order statistics.

$$Pr\left[Pr\left(X_{(r)} < X < X_{(s)}\right) \geqslant p\right] \geqslant \gamma \text{ (tolerance coefficient)}$$

where $Pr\left(X_{(r)} < X < X_{(s)}\right)$ is a probability with respect to X and $Pr\left[Pr\left(X_{(r)} < X < X_{(s)}\right) \geqslant p\right]$ is a probability with respect to $X_{(r)}$ and $X_{(s)}$. The proof is as follows.

$$\begin{aligned} Pr\left(X_{(r)} < X < X_{(s)}\right) &= Pr\left(X < X_{(s)}\right) - Pr\left(X < X_{(r)}\right) \\ &= F_X\left(X_{(s)}\right) - F_X\left(X_{(r)}\right) \\ &= U_{(s)} - U_{(r)} \text{ (which is distribution free)} \end{aligned}$$

where the distribution of $U_{(s)} - U_{(r)}$ is given by the following theorem.

Theorem. *If $U_1, U_2, ..., U_n \overset{i.i.d}{\sim} U(0,1)$, then $U_{(s)} - U_{(r)}, 1 \leqslant r < s \leqslant n$ is distributed as $U_{(s-r)}$.*

The proof is actually intuitive. We have derived the joint distribution of $U_{(r)}$ and $U_{(s)}$ last class. Then using the method of Jacobian we can know that $U_{(s)} - U_{(r)}$ and $U_{(s-r)}$ have the same distribution.

Corollary. $U_{(s)} - U_{(r)} \sim Beta(s - r, n - s + r + 1)$ *This corollary solves the one sample coverage problem: finding γ such that* $\Pr\left(U_{(r-s)} \geqslant p\right) \geqslant \gamma$.

2.2 Two Sample Coverage

Suppose there are two series of samples.

$$X_1, X_2, ..., X_m \overset{i.i.d}{\sim} F_X,\ Y_1, Y_2, ..., Y_n \overset{i.i.d}{\sim} F_Y$$

Also let's define several intervals by the order statistics of Y:

$$I_1 = (-\infty, Y_{(1)}],\ I_2 = (Y_{(1)}, Y_{(2)}],\ I_3 = (Y_{(2)}, Y_{(3)}], ..., I_n = (Y_{(n-1)}, Y_{(n)}], I_{n+1} = (Y_{(n)}, \infty)$$

Correspondingly we define the block frequency $B_i = \#\ of\ X\ in\ I_i$. So the question is: what is the joint distribution of B_i when $F_X = F_Y$?

Theorem. *The joint probability is distributed uniformly.*

$$Pr\left(B_1 = b_1, B_2 = b_2, ..., B_{n+1} = b_{n+1}\right) = \frac{1}{\binom{m+n}{n}}$$

The idea is actually intuitive: we have $m + n$ positions and n bars, which correspond to the intervals.

2.3 Ranks Block Frequencies and Placement

Definition. *Rank of X_i in a sample of m observations,*

$$\begin{aligned} rank\left(X_i\right) &= \sum_{j=1}^{m} I\left(X_j \leqslant X_i\right) \\ &= m\hat{F}_m\left(X_i\right) \\ rank\left(X_{(i)}\right) &= i \end{aligned}$$

When there are two samples,

$$\begin{aligned} X_1, X_2, ..., X_m \\ Y_1, Y_2, ..., Y_n \end{aligned}$$

rank is with respect to the combined sample.

$$rank\left(Y_{(j)}\right) = \underbrace{\sum_{i=1}^{m} I\left(X_i \leqslant Y_{(j)}\right)}_{r_1+r_2+...+r_j} + j$$

where r_j is the frequency of the i-th block $(Y_{(i-1)}, Y_{(i)}]$. We also define placement as follows.

$$\begin{aligned} P\left(Y_{(j)}\right) &= \#\ of\ X's \leqslant Y_{(j)} \\ &= \sum_{i=1}^{m} I\left(X_i \leqslant Y_{(j)}\right) \end{aligned}$$

2.4 Test of Randomness

2.4.1 Test Based on the Total Number of Runs

Suppose there is a sequence of only two types, say, M (male) and F (male).

$$MFMFMFMF \tag{2.1}$$
$$MMMMFFFF \tag{2.2}$$
$$MMFMFFFM \tag{2.3}$$

It is reasonable to say (3) is more random than (1) and (2). The intuitive idea is that (1) changes/flips too frequently and (2) is the opposite.

Definition. *Runs: A succession of one or more types of symbols which are followed and preceded by a different symbol or no symbol at all.*

In the previous example, there are 8, 2 and 5 runs in (1), (2) and (3). Either there are too many runs or too less runs will make the sequence seemingly less 'random'. Then a natural question is: under pure randomness, what is the distribution of runs?

Setting: Two types of elements,

$$\begin{aligned} n_1 &= \#\ of\ type\ 1\ elt \\ n_2 &= \#\ of\ type\ 2\ elt \\ r_1 &= \#\ or\ runs\ of\ type\ 1 \\ r_2 &= \#\ or\ runs\ of\ type\ 2 \\ r &= r_1 + r_2\ (total\ runs) \end{aligned}$$

For example, if we say M is type 1 and F is type 2, in (3) we have $r_1 = 3$ and $r_2 = 2$. So suppose n_1 and n_2 are given, how many rearrangements are there with $R_1 = r_1$ and $R_2 = r_2$? (The answer is $\frac{n!}{n_1!n_2!}$.

Lemma. *The # of distinguishable ways of distributing n objects into r cells with no cell empty is $\binom{n-1}{r-1}$, $n \geqslant r$.*

So we have $\binom{n_1-1}{r_1-1}$ ways to have r_1 runs of type 1 and $\binom{n_2-1}{r_2-1}$ ways to have r_2 runs of type 2. We claim that there are only 3 cases of the relationship between r_1 and r_2:

$$\begin{cases} r_1 = r_2 + 1 \\ r_1 = r_2 - 1 \\ r_1 = r_2 \end{cases}$$

Theorem.

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

where $C = 2$ if $r_1 = r_2$; $C = 1$ if $r_1 = r_2 \pm 1$.

Corollary. *The marginal distribution of R_1 is*

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

The proof is to directly calculate the summation $\sum_{r_2} f_{R_1,R_2}(r_1, r_2)$. The calculation is not difficult because there are only 3 cases of r_2. Next we are interested in the distribution of $r = r_1 + r_2$.

Theorem. *The pmf for R (total # of runs) of $n_1 + n_2$ objects is*

$$f_R(r) = \begin{cases} \frac{2\binom{n_1-1}{\frac{r}{2}-1}\binom{n_2-1}{\frac{r}{2}-1}}{\binom{n_1+n_2}{n_1}}\ ,\ r\ even \\ \frac{\binom{n_1-1}{\frac{r-1}{2}}\binom{n_2-1}{\frac{r-3}{2}}+\binom{n_1-1}{\frac{r-3}{2}}\binom{n_2-1}{\frac{r-1}{2}}}{\binom{n_1+n_2}{n_1}}\ ,r\ odd \end{cases}$$

So in theory, we can calculate exactly the p-value of the # of runs we observe. However, even though the pmf is known, it is still hard to work with because the calculation is sometimes too difficult. We want some approximation. We first look at $\mathbb{E}[R]$ and $Var[R]$. The trick here is to use the definition of runs:

$$R = 1 + I_2 + I_3 + ... + I_n$$

$$where\ I_k = \begin{cases} 1\ ,\ if\ k^{th}\ elt\ \neq\ (k-1)^{th}\ elt \\ 0\ ,\ otherwise \end{cases}$$

Then

$$\begin{aligned}\mathbb{E}[R] &= 1 + \mathbb{E}[I_2] + \mathbb{E}[I_3] + ... + \mathbb{E}[I_n] \\ &= 1 + (n-1)\frac{n_1 n_2}{\binom{n}{2}}\end{aligned}$$

where $I_k \sim Ber\left(\frac{n_1 n_2}{\binom{n}{2}}\right)$ and $\mathbb{E}[I_k] = \frac{n_1 n_2}{\binom{n}{2}}$. This is because there are $\binom{n}{2}$ combinations of $(k-1, k)$ and you want the two elements to be different, which corresponds to $n_1 n_2$. Here the good thing is that $\mathbb{E}[I_k] = \mathbb{E}[I_k^2]$. So

$$\begin{aligned}\mathbb{E}[R] &= 1 + \sum_{k=2}^{n} \mathbb{E}[I_k] \\ &= 1 + \frac{2n_1 n_2}{n_1 + n_2} \\ Var[R] &= Var[1 + \sum_{k=2}^{n} I_k] \\ &= Var[\sum_{k=2}^{n} I_k] \\ &= (n-1)Var[I_k] + \overbrace{\sum_j \sum_k}^{2 \leqslant j \neq k \leqslant n} Cov[I_j, I_k] \\ &= (n-1)\mathbb{E}[I_k^2] + \overbrace{\sum_j \sum_k}^{2 \leqslant j \neq k \leqslant n} \mathbb{E}[I_j I_k] - (n-1)^2 \left[\mathbb{E}[I_k]\right]^2\end{aligned}$$

Here we need to deal with $\mathbb{E}[I_k I_j]$. Let's first consider the case $j = k \pm 1$. In this case,

$$\mathbb{E}[I_k I_j] = \frac{n_1 n_2 (n_1 - 1) + n_2 n_1 (n_2 - 1)}{n(n-1)(n-2)}$$

Similarly for the remaining case,

$$\mathbb{E}[I_k I_j] = \frac{4n_1 n_2 (n_1 - 1)(n_2 - 1)}{n(n-1)(n-2)(n-3)} = \frac{n_1 n_2}{n(n-1)}$$

Substitute back, we get

$$Var[R] = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}$$

Then we can calculate the asymptotic null distribution. Here null means pure randomness, and asymptotic means $n \to \infty$, $n_1/n \to \lambda$, for fixed $\lambda \in (0, 1)$. It is intuitive that

$$\begin{aligned}\lim_{n \to \infty} \mathbb{E}[R]/n &= 2\lambda(1-\lambda) \\ \lim_{n \to \infty} Var[R/\sqrt{n}] &= 4\lambda^2(1-\lambda)^2\end{aligned}$$

Then we can standardize R:

$$\frac{R - \overbrace{2n\lambda(1-\lambda)}^{\mathbb{E}[R]}}{\underbrace{2\sqrt{n}\lambda(1-\lambda)}_{Var[R]}} \xrightarrow{D} N(0,1)$$

The proof starts with the pmf of R. Then we can use Stirling's formula to approximate factorials. This is really an interesting fact. Even though the topic is non-parametric statistics, we still get something related to standard normal, which does have exact expectation and variance. Another interesting thing which is good to know: if we have two distributions D_1 and D_2 and we know nothing about them. We only know that $\mathbb{E}[D_1] = \mathbb{E}[D_2]$. Then we can get $Entropy(D_1) \geqslant Entropy(D_2)$ for any D_2. if D_1 is a Gaussian distribution. i.e., Gaussian distribution is least modelled. If you assume something other than Gaussian, you are putting more restrictions in the model.

Based on the asymptotic distribution, we are able to calculate the p-value. So what kind of p-value should we use, one-sided or two-sided? It depends on the application.

2.4.2 Test Based on the Length of Longest Run

The above is one way to test randomness (based on the total number of runs). We also have other tests. For example, here is a test based on the length of the longest runs. Let's look at the following sequences:

$$xyyyyyxx \tag{1}$$
$$xyxyxyxy \tag{2}$$

Here (1) has the length of longest run equal to 5 and (2) is 1. They are either too long or too short, which make the sequence seem less random.

Let R_{ij} denote the number of runs of objects of type i which are of length j, where $i = 1, 2$, $j = 1, 2, ..., n_i$. Say we have a sequence:

$$xxyyyxxyyyx$$

Then $R_{12} = 2$ and $R_{23} = 2$. Then we have the following.

$$\sum_{j=1}^{n_i} j r_{ij} = n_i$$
$$\sum_{j=1}^{n_i} r_{ij} = r_i$$

So what is the joint pmf for R_{ij}? What is the # of arrangements in which there are exactly r_{ij} runs of type i and length j for each i and j? We notice that the total # of arrangements equal to $\binom{n_1+n_2}{n_1}$. In terms of each run of each type, or say R_{ij}, there are $\frac{r_1!}{\prod_{j=1}^{n_1} r_{1j}!}$ arrangements for type 1 and similarly for type 2 there are $\frac{r_2!}{\prod_{j=1}^{n_2} r_{2j}!}$. In this way, the total # of arrangements becomes $\frac{C r_1! r_2!}{\left(\frac{r_1!}{\prod_{j=1}^{n_1} r_{1j}!}\right)\left(\frac{r_2!}{\prod_{j=1}^{n_2} r_{2j}!}\right)}$, where $C = 2$ if $r_1 = r_2$ and $C = 1$ if $r_1 = r_2 \pm 1$. So the pmf is given by

$$f(r_{11}, r_{12}, ..., r_{1n_1}, r_{21}, r_{22}, ..., r_{2n_2})\big|_{r_1, r_2} = \frac{C r_1! r_2!}{\prod_{i=1}^{2}\prod_{j=1}^{n_i} r_{ij}! \binom{n_1+n_2}{n_1}}$$

So if we want to calculate the p-value, we simply calculate

$$Pr\left(length\ of\ longest\ type\ 1\ run\ \leqslant k\right) = \sum_{r_1}\sum_{r_{11}}\sum_{r_{12}}\cdots\sum_{r_{1k}} \frac{r_1!\binom{n_2+1}{r_1}}{\prod_{j=1}^{k} r_{1j}!\binom{n_1+n_2}{n_1}}$$

This is very complicated! So in practice what should we do? Let's look at this example: $n_1 = 20$, $n_2 = 15$, $L = 6$. We can use simulation to compute the p-value. (write the code here)

2.4.3 Test for Numerical Variable

The above tests are all for 0-1 variables. What if we have numerical variables which have more than two types? Say,

$$8, 13, 1, 3, 4, 7 \tag{1}$$

We can transform it into a 0-1 sequence. Say, $even \rightarrow 0$ and $odd \rightarrow 1$. This is our first approach, artificially dividing the numbers into two groups. However this is kind of weird and not convincing.

Second approach: Runs up and down

$$\underbrace{}_{-}\,8\,\underbrace{,}_{+}\,13\,\underbrace{,}_{-}\,1\,\underbrace{,}_{+}\,3\,\underbrace{,}_{+}\,4\,\underbrace{,}_{+}\,7 \tag{2}$$

But both of the two methods lose the information of the numerical values. We want a measure that is related to the numerical magnitude but has nothing to do with the distribution (because we do not want to make assumptions for the distribution). Remember we talked about rank earlier this lecture. Consider this:

$$NM = \sum_{i=1}^{n-1} \left[rank\left(X_i\right) - rank\left(X_{(i)}\right) \right]^2$$

If a sequence is exactly from smallest to largest, or there are some patterns:

$$X_{(1)}, X_{(2)}, X_{(3)}, ..., X_{(n)} \tag{3}$$

$$X_{(1)}, X_{(n)}, X_{(2)}, X_{(n-1)}, ... \tag{4}$$

The NM statistics will be either too large or too small. So what is the distribution of NM under the null? Standardize the NM statistic, we get

$$RVN = \frac{\sum_{i=1}^{n-1} \left[rank\left(X_i\right) - rank\left(X_{i+1}\right) \right]^2}{\sum_{i=1}^{n} (rank(X_i) - \underbrace{\frac{n+1}{2}}_{median\ rank})^2} \xrightarrow{D} N(2, \frac{20}{5n+7})$$

2.5 Test of Goodness of Fit

The basic question is: given $X_1, X_2, ..., X_n \overset{i.i.d}{\sim} F_X$,

$$H_0 : \ F_X = F_0$$
$$H_a : \ F_X \neq F_0$$

The Chi-square test is one example of the test of goodness of fit: We group n observations into k mutually exclusive categories, then denote the observed and expected frequency by $f_1, ..., f_k$ and $e_1, ..., e_k$. Then

$$Q = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \xrightarrow{d} \chi^2_{k-1}$$

In practice, we are interested in 'composite null'. Say, the null hypothesis might be: the variables follow some normal distribution but we do not know/specify the exact mean and variance. Then we can let θ_i be the probability of observing category i. Then we can estimate e_i using $n\hat{\theta}_i$, where $\hat{\theta}_i$ is the MLE of θ_i. Then

$$Q \xrightarrow{H_0} \chi^2_{k-1-s}$$

where s is the total # of estimated parameters. Note that here it is king of arbitrary and too artificial to classify categories. We should do the category classification before we get observations. We can also use the empirical CDF as discussed in lecture 1 to solve the problem. We look at the maximum distance between the null CDF and the empirical CDF:

$$D_n = \sup_x |F_X(x) - \hat{F}_n(x)|$$

Note that we have the Cantelli inequality telling us that $Pr\left(\sup_x |F_X(x) - \hat{F}_n(x)| > \epsilon\right)$ decreases exponentially. Is this distribution (of D_n) free under the null? It is intuitive that D_n will be large only at the 'discontinuous' points. So instead of looking at $F_X(x)$ we can simply look at $F_X(X_{(i)})$. Note that $F_X(X_{(i)}) \sim U(0,1)$ which is distribution free. In practice to estimate p-value we can simply take $F_X(x) = x$. (Kolmogorov-Smirnov test)