MANA130083.01 Nonparametric *Spring 2025*

Lecture 3

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

3.1 Review

Last class we talked about test of goodness of fit. The chi-square test is able to solve both simple null and composite null hypothesis testing, but it is arbitrary. So this lecture we will introduce some new tests though they are only for simple null $F = F_0$.

Theorem. *(Kolmogorov-Smirnov One Sample Statistic) Suppose we have samples $X_1, X_2, ..., X_n \sim F$, then we can calculate the following statistic:*

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

where $\hat{F}_n(x)$ is the empirical CDF.

Here we actually look at the difference between the null distribution and the empirical CDF. Under the null hypothesis ($H_0 : F = F_X$), the statistic should be very small. Also note that: D_n can only attain its value at sample points, which are the inflection points of the empirical CDF.

3.2 More on K-S Statistic

3.2.1 Distribution of K-S Statistic

So what is the distribution of D_n? Note that D_n is distribution free:

$$D_n = \sup_x |F_X(x) - \hat{F}_n(x)| = \max\left\{|F_X\left(X_{(1)}\right) - \hat{F}_n\left(X_{(1)}\right)|, ..., |F_X\left(X_{(n)}\right) - \hat{F}_n\left(X_{(n)}\right)|\right\}$$

where $F_X\left(X_{(i)}\right)$ is distributed as $U_{(i)}$ and $\hat{F}_n\left(X_{(i)}\right) = \frac{i}{n}$. The exact distribution of D_n is hard to compute but you can use either simulation or the asymptotic distribution to calculate the p-value. We have the following analytic formula:

Theorem. *If F_X is continuous then for all $d > 0$,*

$$\lim_{n\to\infty} Pr\left(D_n \leqslant \frac{d}{\sqrt{n}}\right) = L(d)$$

where $L(\cdot)$ is a very complicated function:

$$L(d) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-\frac{2i^2}{d^2}}$$

The proof for this theorem is tedious. Another theorem:

Theorem. *Let* $D_n^+ = \sum_x \left(\hat{F}_n(x) - F_X(x)\right)$. *If* F_X *is continuous then for* $d \geqslant 0$, *we have*

$$\lim_{n\to\infty} Pr\left(D_n^+ < \frac{d}{\sqrt{n}}\right) = 1 - e^{-2d^2}$$

This is a lot easier to compute than the previous one.

Corollary. *If* F_X *is continuous then for* $d \geqslant 0$,

$$4n(D_n^+)^2 \xrightarrow{D} \chi_2^2$$

Proof: We have $D_n^+ < \frac{d}{\sqrt{n}}$ if and only if $4n(D_n^+)^2 < 4d^2$. Denote $v = 4n(D_n^+)^2$. Then

$$\begin{aligned}\lim_{n\to\infty} Pr\left(v < 4d^2\right) &= \lim_{n\to\infty} Pr\left(D_n^+ < \frac{d}{\sqrt{n}}\right) \\ &= 1 - e^{-2d^2} \\ &= 1 - e^{-\frac{4d^2}{2}}\end{aligned}$$

That is to say,

$$\lim_{n\to\infty} Pr\left(v < c\right) = 1 - e^{-\frac{c}{2}}$$

where $1 - e^{-\frac{c}{2}}$ is exactly the CDF of χ_2^2.

3.2.2 Confidence Band for CDF

Then let's talk about the related confidence band. Let's first define $D_{n,\alpha}$ as the α quantile of D_n: $Pr\left(D_n > D_{n,\alpha}\right) = \alpha$. So,

$$Pr\left[\sup_x |\hat{F}_n(x) - F_X(x)| < D_{n,\alpha}\right] = 1 - \alpha$$

$$Pr\left[\hat{F}_n(x) - D_{n,\alpha} \leqslant F_X(x) \leqslant \hat{F}_n(x) + D_{n,\alpha}, \forall x\right] \geqslant 1 - \alpha$$

where you can get $\hat{F}_n(x)$ from the data and $D_{n,\alpha}$ is a fixed number. So we have a bound for $F_X(x)$. Let $L_n = \max\left[\hat{F}_n(x) - D_{n,\alpha}, 0\right]$ and $U_n = \min\left[\hat{F}_n(x) + D_{n,\alpha}, 1\right]$. Then they become the lower confidence band and upper confidence band.

What if we have composite null? For example, the null is $X_1, ..., X_n \sim \mathcal{N}$. We can use Lilliefor's test for normality, where the test statistic is given by $\sup |\hat{F}_n(x) - \hat{F}_0(x)|$ and $\hat{F}_0(x)$ is the CDF of $\mathcal{N}(\hat{\mu}_{\text{MLE}}, \hat{\sigma}^2_{\text{MLE}})$.

3.2.3 Problem with K-S Statistic

What if we are facing composite null? Say $H_0 : F_X \sim N(\mu, \sigma^2)$ where μ and σ^2 are unknown. Here we have the Lilliefor's test for normality. The test statistic is given by

$$D_n = \sup_x |\hat{F}_n(x) - \hat{F}_0(x)|$$

where $\hat{F}_0(x)$ is the CDF for $N(\bar{X}, s^2)$. Here we just estimate the mean and variance by the sample mean and sample variance. The only thing you can do then is to use simulation. Say have another null hypothesis $H_0 : F_X \sim exp(x)$ then $\hat{F}_0(x)$ becomes $1 - e^{\frac{x}{\bar{X}}}$.

3.2.4 More Extension on K-S

If you further look at $D_n = \sup_x |\hat{F}_n(x) - F_X(x)|$, basically D_n is a measure of distance between the empirical CDF and the null CDF. This is exactly the L-infinity distance $||\hat{F}_n(x) - F_X(x)||_\infty$. So if you change a distance then you get a different test. Say you look at another class of distance of this form:

$$||\hat{F}_n(x) - F_0(x)|| = \int_{-\infty}^{\infty} \left(\hat{F}_n(x) - F_0(x)\right)^2 w(x)dx$$

where $w(x)$ is a weight. If we let $w(x) = \frac{1}{F_0(x)[1-F_0(x)]}$ then you get the Anderson-Darling test. The question is: why this weight?

Note that $w(x)$ is big if $F_0(x)$ is small or $1 - F_0(x)$ is small. So it basically tells us that you have to up-weight the points at the extremes. Also it is the extremes that make t-distribution different from normal distribution.

For sure you can choose $w(x) = 1$ and this is then Cramer Von Mises test. Basically you don't weight the observations. In this way you are ignoring the extreme points.

3.3 Visual Analysis of Goodness of Fit

Consider the graph (QQ-plot): $\left(F^{-1}(p), G^{-1}(p)\right)$, where $0 < p < 1$. Suppose $F(x) = G\left(\frac{x-\mu}{\sigma}\right)$, which indicates that F and G are only different in a scale parameter and a location parameter. Then $F^{-1}(p) = \mu + \sigma G^{-1}(p)$, and the graph becomes a straight line. For example the null hypothesis is $X_1, ..., X_n \sim N(\mu, \sigma^2)$. You can simply plot $\left(F^{-1}(p), \hat{F}_n(p)\right)$ where F is the CDF for $N(0, 1)$. If the distribution is really normal, the graph would be a straight line.

3.4 Inference for a Population Quantile

3.4.1 Confidence Interval

For example you observe some samples $X_1, ..., X_n$ and you want to construct a confidence interval for the median. Let's discuss the general case. Denote k_p as the p-th quantile for F_X. To begin with, we need a point estimate for k_p, which is $X_{(r)}$, where

$$r = \begin{cases} np & , if\ np \in \mathbb{Z} \\ [np + 1] & , if\ np \notin \mathbb{Z} \end{cases}$$

This is quite intuitive. For example, the point estimate for population median is given by the sample median. For confidence interval, we want r, s such that

$$Pr\left(X_{(r)} < k_p < X_{(s)}\right) \geqslant 1 - \alpha$$

where $X_{(r)}$ and $X_{(s)}$ are random variables. We can first look at one side and then combine two sides step by step as follows.

$$Pr\left(X_{(r)} < k_p\right) = \sum_{i=r}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

$$Pr\left(X_{(r)} < k_p < X_{(s)}\right) = \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} \geqslant 1-\alpha$$

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

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

Using normal approximation, we get

$$r = np + 0.5 - z^{\alpha/2}\sqrt{np(1-p)}$$

where 0.5 is the continuity correction. Of course you can take the integer part of it. Similarly,

$$s = np + 0.5 + z^{\alpha/2}\sqrt{np(1-p)}$$

3.4.2 Hypothesis Testing

Let's state the hypothesis:

$$H_0 : k_p \leqslant k_p^0$$
$$H_1 : k_p > k_p^0$$

We reject H_0 if at most $r-1$ samples are smaller than k_p^0 for some r. The corresponding p-value is:

$$Pr\left(X_{(r)} > k_p^0 | H_0\right) = 1 - Pr\left(X_{(r)} \leqslant k_p^0 | H_0\right)$$

and we want this value to be equal to or smaller than α for some certain r. Same as we compute for confidence interval, let

$$k = \#\ of\ +\ signs\ among\ X_{(i)} - k_p^0$$

We can reject the null when $k \geqslant n - r + 1$. That is,

$$Pr\left(k \geqslant n-r+1 | H_0\right) = \sum_{i=n-r+1}^{n} \binom{n}{i} (1-p)^i p^{n-i} \leqslant \alpha$$

Just as before, we can still use normal approximation to find such r.

3.4.3 The Sign Test and Confidence Interval for the Median

This is just the special case of what we discussed in previous section. Let M be the median. Then the null hypothesis is $M = M_0$. Given $X_1, ...X_N$, consider $X_1 - M_0, X_2 - M_0, ..., X_N - M_0$. If there are either too many or too few plus(+) signs, we reject the null. This is why it is called a sign test. Let K be the # of '+' signs among $X_i - M_0$'s under H_0. Here we know the exact distribution of K:

$$K \sim Bin\left(N, 0.5\right) \Rightarrow K \sim \mathcal{N}(\frac{n}{2}, \frac{n}{4})$$

Then the calculation of p-value and confidence interval would be quite easy. That is to find r and s such that

$$Pr\left(X_{(r)} < M < X_{(s)}\right) \geqslant 1-\alpha$$
$$\sum_{i=0}^{r-1} \binom{N}{i} 0.5^N \leqslant \frac{\alpha}{2}$$
$$\sum_{i=s}^{N} \binom{N}{i} 0.5^N \leqslant \frac{\alpha}{2}$$

Consider zero differences: what if $X_i - M_0$ for some i? If you are dealing with some continuous distribution then you don't need to worry about that because the probability of $X_i - M_0$ happening is zero. However if you are dealing with discrete variables then it is actually possible. There are several seemingly naive strategies. First, you can simply ignore those X_i and reduce accordingly. Second, you can treat half of of 0's as + and half as -. The third one is slightly more sophisticated. It is a conservative strategy. That is to assign to all 0's such sign that is least conductive to the rejection of H_0. In this way, you really want to make sure that type I error is controlled. The fourth way is still naive: randomize + or -, for example, randomly flip a coin.

3.5 Rank-order Statistic

Definition. *$r(\cdot)$ is a rank order statistic if*

$$X_i \leqslant X_j \Rightarrow r(X_i) \leqslant r(X_j)$$

So basically $r(\cdot)$ is a monotone function of the observations. Why do we call it a 'rank' order statistic? The natural example is that the rank $1, 2, ..., n$, which orders the observations from smallest to largest, is itself a rank function.

The question is: what is the correlation between X and $r(X)$? Let $Y = r(X)$, then

$$\rho\left[X, Y\right] = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sigma_X \sigma_Y}$$

The marginal distribution of Y is discrete uniform distribution over $1, 2, \ldots, n$:

$$f_Y(j) = \frac{1}{N}$$

for $j = 1, ..., N$. Therefore,

$$\mathbb{E}\left[Y\right] = \sum_{j=1}^{N} j \cdot \frac{1}{N} = \frac{N+1}{2}$$
$$\mathbb{E}\left[Y^2\right] = \sum_{j=1}^{N} \frac{j^2}{N} = \frac{(N+1)(2N+1)}{6}$$

So we also get

$$Var\left[Y\right] = \frac{N^2-1}{12}$$

For $\mathbb{E}[XY]$, we can think of the joint pdf of X and $r(X) = Y$:

$$f_{X,Y}(x,j) = f_{X|Y=j}(x|j) \cdot f_Y(j) = \frac{f_{X_{(j)}}(x)}{N}$$

for $j = 1, ..., N$. Then

$$\mathbb{E}[XY] = \frac{1}{N} \int_{-\infty}^{\infty} \sum_{j=1}^{N} jx f_{X_{(j)}}(x) dx = \sum_{j=1}^{N} \frac{j\mathbb{E}\left[X_{(j)}\right]}{N}$$

Put everything back,

$$\rho[X, r(X)] = \frac{\left(\frac{12}{N^2-1}\right)^{\frac{1}{2}} \left\{\sum_{j=1}^{N} j\mathbb{E}\left[X_{(j)}\right] - \left[\frac{N(N+1)}{2}\right] \mathbb{E}[X]\right\}}{N\sigma_X}$$

Basically we cannot simplify it any further because $\mathbb{E}\left[X_{(j)}\right]$, $\mathbb{E}[X]$ and σ_X all depend on the distribution of X and they are not distribution free. If N goes to infinity,

$$\lim_{N\to\infty} \rho[X, r(X)] = \frac{2\sqrt{3}}{\sigma_X} \left\{\mathbb{E}[XF_X(x)] - \frac{1}{2}\mathbb{E}[X]\right\}$$

where $\mathbb{E}[XF_X(x)]$ is a covariance between X and its CDF and intuitively should be positive.

3.5.1 Problem With Rank-order Statistic: Ties in Observations

Ties exist when two or more observations are same. To explain it more precisely, in a set of N observations which are not all different, arrangement in order of magnitude produces a set of r groups of different numbers, the i-th different value occurring with frequency t_i, where $\sum t_i = N$. Any group of numbers with $t_i \geqslant 2$ comprises a set of tied observations. There are five strategies.

1 Randomization

Within each group, we can randomly assign ranks. There are $\prod t_i!$ possible assignments. For example,

$$3.0, \quad \underbrace{4.1, 4.1,}_{randomly\ assign\ 2\ and\ 3} \quad 5.2, \quad \underbrace{6.3, 6.3, 6.3,}_{randomly\ assign\ 5{,}6\ and\ 7} \quad 9$$

there are 2!(3!) or 12 possible assignments of the integer ranks 1 to 8 which the above sample could represent. Using this method, some theoretical properties of the rank statistic are preserved, since each assignment occurs with equal probability. in particular, the null probability distribution of the rank-order statistic, and therefore of the rank statistic, is unchanged, so that the test can be performed in the usual way. However, an additional element of chance is artificially imposed, affecting the probability distribution under alternatives.

2 Midranks

The midrank method assigns to each member of a group of tied observations the simple average of the ranks they would have if distinguishable. For example,

$$3.0, \underbrace{4.1, 4.1,}_{both\ 2.5} 5.2, \underbrace{6.3, 6.3, 6.3,}_{all\ 6} 9$$

Using this approach, tied observations are given tied ranks. The midrank method is perhaps the most frequently used, as it has much appeal experimentally. However, the null distribution of ranks is affected. Obviously, the mean rank is unchanged, but the variance of the ranks would be reduced.

3 Average Statistic

If one does not wish to choose a particular set of ranks as in the previous two methods, one may instead calculate the value of the test statistic for all the $\prod t_i$ assignments and use their simple average as the single sample value. For example, in the randomization case, we can calculate 12 possible test statistic values and take their simple average. Again, the test statistic would have the same mean but smaller variance.

4 Least Favorable Statistic

Having found all possible values of the test statistic, one might choose as a single value that one which minimizes the probability of rejection. This procedure leads to the most conservative test, i.e., the lowest probability of committing a type I error.

5 Omission of Tied Observations

The final and most obvious possibility is to discard all tied observations and reduce the sample size accordingly. This method certainly leads to a loss of information, but if the number of observations to be omitted is small relative to the sample size, the loss may be minimal. This procedure generally introduces bias toward rejection of the null hypothesis. Assume H_0 is true, i.e., $X_1, ..., X_n \overset{H_0}{\sim} F_X$. It is not very likely that ties exist in extreme values. It is only possible that ties exist near the highest part of p.m.f. Note that the possibility that ties exist in a continuous distribution is zero. So if we omit the ties, we are omitting the observations from the area which observations could actually lie in with most probability. We actually omit a good chunk of the null observation. So we will have more bias toward rejection of null.

3.5.2 Application of Rank-order Statistic

3.5.2.1 The Wilcox Signed-rank Test

Suppose we have a sample: $X_1, ..., X_N \sim F$. The question is: what is the median of F? The null hypothesis is given by: $H_0 : M = M_0$. We have discussed one naive idea, which is to calculate the total number of +/- signs of $X_i - M_0$'s. The total number of +/- signs follow binomial distribution. If the +/- ratio is far away from 50/50, we should reject the null.

However, there is a problem that the magnitudes of $X_1 - M_0, ..., X_N - M_0$ are ignored. Here we can use rank to reflect the magnitude. The most naive way is to rank from 1 to N with 1 corresponding to the smallest observation and N corresponding to the largest observation:

$$\underbrace{X_{(1)}}_{1} \leqslant \underbrace{X_{(2)}}_{2} \leqslant ... \leqslant \underbrace{X_{(n)}}_{n}$$

However, this naive way still does not incorporate the magnitudes of the data. It doesn't work. To solve the problem, let's consider $D_i = X_i - M_0$ and look at the the rank of absolute values of D_i's. To order $|D_i|$, let

$$s_i = I\left(D_i > 0\right)$$

and define

$$T^+ = \sum_{i=1}^{n} s_i r\left(|D_i|\right)$$

$$T^- = \sum_{i=1}^{n} (1 - s_i) r\left(|D_i|\right)$$

Under null hypothesis, $T^+ \approx T^-$. However if $|T| = |T^+ - T^-|$ is too large, we should reject the null. For example, we have

$$1, 2, 3, 4, 10, 11, 12, 13$$

and the null hypothesis is $H_0 : M = 5$. If you simply look at the +/- signs, you cannot reject the null. However if you look at the magnitude, it becomes

$$\begin{aligned} &|5-1|, |5-2|, |5-3|, |5-4|, |5-10|, |5-11|, |5-12|, |5-13| \\ \rightarrow &4, 3, 2, 1, 5, 6, 7, 8 \end{aligned}$$

We then sum the ranks of$4, 3, 2, 1$ and $5, 6, 7, 8$ separately and they may be far away from each other. Therefore according to Wilcox signed-rank test, we might be able to reject the null. If we change our null to $\mu = 7$, we get

$$\begin{aligned} &|7-1|, |7-2|, |7-3|, |7-4|, |7-10|, |7-11|, |7-12|, |7-13| \\ \rightarrow &6, 5, 4, 3, 3, 4, 5, 6 \end{aligned}$$

Then we should not reject the null because T^+ is close to T^-.

To sum up, the test procedure is as follows:

$$\begin{aligned} T^+ &= \sum s_i r\left(|D_i|\right) \Rightarrow \sum \mathbb{E}[s_i r(|D_i|)] = \sum \underbrace{\mathbb{E}[s_i]}_{1/2} \underbrace{\mathbb{E}[r(|D_i|)]}_{\frac{N(N+1)}{4}} \\ T^- &= \sum (1 - s_i) r\left(|D_i|\right) \\ T &= T^+ - T^- \ (test\ statistic) \end{aligned}$$

Under null, T^+ and T^- have the same distribution

$$\begin{aligned} \mathbb{E}\left[T\right] &= 0 \\ Var\left[T\right] &= \frac{N(N+1)(2N+1)}{6} \end{aligned}$$

We also have large sample approximation:

$$Pr\left[\frac{T}{\sqrt{Var\left(T\right)}} \leqslant t\right] \rightarrow \Phi(t)$$

where $\Phi(\cdot)$ is the CDF of standard normal distribution. Interestingly, we can compare the power of Wilcox signed-rank test and the test directly based on signs.

3.5.2.2 Paired Sample Procedure

Suppose our observations are paired: $(X_1, Y_1), ..., (X_n, Y_n)$. Similarly we calculate the difference in each pair: $D_i = Y_i - X_i$ and the question is to do inference for the median of D_i. For example, our null hypothesis can be $H_0 : M(D_i) = M_0$. Then we can use similar procedure as talked before to do the test. There is one point to emphasize here: the difference of the median is not equal to the median of difference. In other words,

$$M(Y_i) - M(X_i) \neq M(D_i)$$

We can justify this idea by the following example. Say, we have joint density function

$$f(x, y) = \begin{cases} \frac{1}{2} & , y - 1 \leqslant x \leqslant y, -1 \leqslant y \leqslant 1 \ or \ y + 1 \leqslant x \leqslant 1, -1 \leqslant y \leqslant 0 \\ 0 & , otherwise \end{cases}$$

In this case, we also have the marginal distribution of Y the same as that of X. Both marginal probability follows $U[-1,1]$ here. Therefore they also have same median, i.e., $M(Y)-M(X)=0$. However, $median(Y-X)$ is obviously not zero because $Pr(X<Y)=\frac{3}{4}$ and $Pr(X\geqslant Y)=\frac{1}{4}$ (which can be indicated from visualized jpdf).

3.5.2.3 Two Sample Problem (Wald-Wolfowitz Runs Test)

Suppose we have $X_1,...,X_m\sim F_X$ and $Y_1,...,Y_n\sim F_Y$. The hypothesis is given by

$$H_0: F_Y=F_X$$
$$H_1: F_X(x)\geqslant F_Y(x)$$

The alternative hypothesis is basically saying that Y is stochastically larger. Here 'stochastically larger' means $Pr(X\leqslant x)>Pr(Y\leqslant x)$ so Y is 'somehow' more likely to be larger than X. To be more specific, the alternative may suggest that F_Y and F_X differ in either location or scale parameter or both. All of the following three cases are possible.

$$H_1: F_Y(x)=F_X(x-\theta),\theta\neq 0$$
$$H_1: F_Y(x)=F_X(\theta x),\theta\neq 1$$
$$H_1: Pr(Y-\mu_Y\leqslant x)=Pr(X-\mu_X\leqslant \theta x)$$

The idea is that both $XXXYYYYY$ and $XXYYYYYX$ seem to indicate $F_X=F_Y$ is not true.

For the third case, it is equivalent to say that $Y-\mu_Y$ and $\frac{X-\mu_X}{\theta}$ are identically distributed. In order to test the hypothesis, here we use Wald-wolfowitz runs test. We can order the pooled observations from smallest to largest and calculate the number of runs. For example,

$$\underbrace{X_{(1)}}_{run1},\underbrace{Y_{(1)}}_{run2},\underbrace{X_{(2)}}_{run3},\underbrace{Y_{(2)},Y_{(3)}}_{run4},\underbrace{X_{(3)},X_{(4)}}_{run5},\underbrace{Y_{(4)},Y_{(5)}}_{run6},\underbrace{X_{(5)}}_{run7}$$

Here we reject the null only if there are too few runs. This is a very general test. Besides, we also have the Kolmogorov-Smirnov two sample test, which is to consider

$$D_{m,n}=\max_x|\hat{F}_X(x)-\hat{F}_Y(x)|$$

where the null hypothesis is same as before, $H_0: F_X=F_Y=F$. When $D_{m,n}$ is too big, we should reject the null. And here we can actually calculate the corresponding p-value with explicit formula, which will be discussed later.