Bruce M. Boghosian

Motivation

The sign test

Wilcoxon tests

Summary

Nonparametric Statistics

The sign test and the Wilcoxon sign test

Bruce M. Boghosian

Tufts University School of Arts and Sciences

Department of Mathematics

Tufts University

Outline

Bruce M. Boghosian

1 Motivation

2 The sign test

3 Wilcoxon tests

4 Summary

Motivation

- Virtually all of the tests that we have used to date require some a priori knowledge of the form of the probability distribution of the data.
- Most also assume that the data is independent and identically distributed (iid).
  - Example: iid Bernoulli random numbers
  - Example: iid Normally distributed random numbers
- If W is the test statistic with actual pdf $f_W(w \mid H_0)$ when H_0 is true, then

$$\alpha' = P(W \in C) = \int_C dw\ f_W(w \mid H_0)$$

will not be equal to the desired α, because a different pdf was used to make the estimate of region C.

Robustness

- One solution we have found for this problem is to find *robust* methods of statistical analysis.
- For large n, we have invoked the CLT to find and test Z statistics, assuming only finiteness of the variance of the underlying distribution.
- Even for smaller n, it has been noted that T tests are also reasonably robust.

Nonparametric statistics

- The above raises the question of whether or not it is possible to find statistical tests that are truly independent of the form of the underlying distribution.
- Such *nonparametric statistics* are an active area of statistical research.
- In this module, we begin examination of some of the more general methods of nonparametric statistical research.

The sign test: Estimating the median

- Most of our tests to date have centered on estimating the mean μ, or testing the hypothesis $\mu = \mu_0$.
- Suppose instead we wish to devise a test about the median $\tilde{\mu}$, defined so

$$P(W < \tilde{\mu}) = \int_{-\infty}^{\tilde{\mu}} dw\ f_W(w) = \int_{\tilde{\mu}}^{+\infty} dw\ f_W(w) = P(W > \tilde{\mu})$$

- If the null hypothesis $H_0 :\ \tilde{\mu} = \tilde{\mu}_0$ is true, then no matter what the underlying distribution, we have

$$P\left(Y \le \tilde{\mu}_0\right) = P\left(Y \le \tilde{\mu}_0\right) = {}^1\!/_2.$$

The sign test: Estimating the median

- So, if H_0 is true and we take n samples, then the number of observations X exceeding μ_0 should be binomially distributed with $p = {}^1\!/_2$, and hence

$$E(X) = \frac{n}{2}$$
$$\mathrm{Var}(x) = \frac{1}{2}\left(1 - \frac{1}{2}\right) n = \frac{n}{4}$$

- Hence we would expect the statistic $\frac{X - n/2}{\sqrt{n/4}}$ to have approximately a standard normal distribution by the CLT (Laplace-DeMoivre Theorem), if n sufficiently large.
- Values of X much smaller or larger than $n/2$ would be evidence that $\tilde{\mu} \neq \tilde{\mu}_0$.

Hypothesis testing for the median

- Let $y_1, \ldots, y_n$ be a random sample of size n from any continuous distribution having median $\tilde{\mu}$, where $n \geq 10$. Let k denote the number of y_i's greater than $\tilde{\mu}_0$, and let $z = \frac{k - n/2}{\sqrt{n/4}}$. Then
  - To test $H_0: \ \tilde{\mu} = \tilde{\mu}_0$ versus $H_1: \ \tilde{\mu} > \tilde{\mu}_0$, at the αth level of significance, reject H_0 if $z \geq +z_\alpha$.
  - To test $H_0: \ \tilde{\mu} = \tilde{\mu}_0$ versus $H_1: \ \tilde{\mu} < \tilde{\mu}_0$, at the αth level of significance, reject H_0 if $z \leq -z_\alpha$.
  - To test $H_0: \ \tilde{\mu} = \tilde{\mu}_0$ versus $H_1: \ \tilde{\mu} \neq \tilde{\mu}_0$, at the αth level of significance, reject H_0 if either $z \leq -z_{\alpha/2}$ or $z \geq +z_{\alpha/2}$.

Example of hypothesis testing for the median

- Alice says the median number of birds that visit her feeder per day is 30.
- Bob says it is greater than 30.
- Bob collected data over the course of 10 days

Day	1	2	3	4	5	6	7	8	9	10
Birds	27	31	33	37	32	34	21	36	37	35

- Null hypothesis is to believe Alice $H_0: \ \tilde{\mu} = \tilde{\mu}_0 = 30$, so

Day	1	2	3	4	5	6	7	8	9	10
Sign	-	+	+	+	+	+	-	+	+	+

- There are 8 $+$ values and 2 $-$ values

$$z = \frac{8 - 10/2}{\sqrt{10/4}} = 1.897$$

- If $\alpha = 0.05$, then $z > z_\alpha = 1.645$, so we reject H_0.

Small-sample version of the example

- Tabulate the binomial distribution for $n = 10$, $p = 1/2$

j	$\binom{10}{j}\left(\frac{1}{2}\right)^{10}$
0	0.000976563
1	0.00976563
2	0.0439453
3	0.117188
4	0.205078
5	0.246094
6	0.205078
7	0.117188
8	0.0439453
9	0.00976563
10	0.000976563

- Note that $P(X \geq 8) = \sum_{j=8}^{10} \binom{10}{j}\left(\frac{1}{2}\right)^{10} = 0.5469$
- This would indicate that you have failed to reject H_0 at confidence level $\alpha = 0.05$.

Sign test for paired data

Bruce M. Boghosian

Motivation

The sign test

Wilcoxon tests

Summary

- Suppose you have two streams of data, x_j and y_j, and you wish to examine pairs (x_j, y_j) for $j = 1, \ldots, n$.
- Let $p = P(X_j > Y_j)$ for $j = 1, \ldots, n$.
- Null hypothesis is that the two streams represent distributions with the same median, i.e., $H_0 : \ p = 1/2$.
- Let $W_j = \begin{cases} 0 & \text{if } X_j \leq Y_j \\ 1 & \text{if } X_j > Y_j \end{cases}$
- Measure $u = \sum j = 1^n w_j$ and compute statistic $\frac{u - n/2}{\sqrt{n/4}}$.
- If n sufficiently large, test this against z_α.

More on sign test for paired data

- We invoked the CLT above, so it may seem we are exploiting robustness, as in earlier Z tests, but note:
  - The distributions of X_i and Y_i do not need to be the same.
  - In fact, the distributions of X_i and X_j for $i \neq j$ do not need to be the same (and likewise for Y_i and Y_j).
  - None of these distributions needs to be symmetric.
  - All of these distributions could have different variances.
- The only essential requirements are that
  - X and Y have continuous pdfs (for continuous r.v.s).
  - The null hypothesis adds the requirement that $\tilde{\mu}_{X_i} = \tilde{\mu}_{Y_i}$ within each pair – i.e., for $i = 1, \ldots, n$.

From the median to the mean

- For a symmetric distribution, median and mean are equal.
- For a skewed distribution, they tend to be different.
- Example: Household wealth in the United States has...
  - a median of just under \$100,000.
  - a mean of about \$750,000.
- Fact that the mean is so much greater than the median is indicative of very wealthy households at the very top.
- In fact, the 400 wealthiest households in the US have as much wealth as the bottom 60% of the population – about 300,000,000 people – namely around \$3.5 trillion.

From the median to the mean

- If Elon Musk walked into our classroom,
  - the mean wealth would increase enormously.
  - the median wealth would hardly change.
- The median "cares" only about the number of households above and below its value:

$$\sum_{j=1}^{n} \mathrm{sgn}\left(X_i - \tilde{\mu}_X\right) = 0$$

- The mean "cares" about *by how much* those households are above or below its value:

$$\sum_{j=1}^{n} \left(X_i - \mu_X\right) = 0$$

Wilcoxon tests

- The only information retained by the sign test was whether data was greater than or less than the median.
- The amount by which the data was greater than or less than the median did not matter.
- The above approach worked for the median, but it will not work for the mean.
- If the statistic depended on the magnitude of the deviation from the median, on the other hand, it would require knowledge of the underlying distribution of the data.
- We can, however, allow the statistic to depend on the *rank* of the magnitude of the deviation from the median, in order to give more weight to higher deviations.

Example of Wilcoxon testing approach

- Suppose $n = 3$ and data is $y_1 = 6.0$, $y_2 = 4.9$, $y_3 = 11.2$.
- Objective is to test $H_0 : \mu = 10.0$ against $H_1 : \mu \neq 10.0$.
- Ranks are:

$$\begin{aligned} |y_1 - \mu_0| &= 4.0 & r_1 &= 2 \\ |y_2 - \mu_0| &= 5.1 & r_2 &= 3 \\ |y_3 - \mu_0| &= 1.2 & r_3 &= 1. \end{aligned}$$

- Signs of deviations from the mean are captured by

$$\begin{aligned} z_1 &= \frac{1}{2}\left(1 + \mathrm{sgn}(y_1 - \mu_0)\right) = 0 \\ z_2 &= \frac{1}{2}\left(1 + \mathrm{sgn}(y_2 - \mu_0)\right) = 0 \\ z_3 &= \frac{1}{2}\left(1 + \mathrm{sgn}(y_3 - \mu_0)\right) = +1 \end{aligned}$$

- *Wilcoxon rank statistic*

$$w = \sum_{j=1}^{n} r_j z_j = (2)(0) + (3)(0) = (1)(+1) = 1.$$

How does one turn this into a hypothesis test?

- If the null hypothesis $H_0: \ \mu = \mu_0$ is true, there should be as many zeros as ones among the z_js.
- **Thm.:** Let $y_1, \ldots, y_n$ be independent observations drawn from the continuous and symmetric (though not necessarily identical) pdfs, $f_{Y_i}(y)$ for $i = 1, \ldots, n$. Suppose that each of the $f_{Y_i}(y)$s has the same mean μ. If $H_0: \ \mu = \mu_0$ is true, the pdf of the data's signed rank statistic, $p_W(w)$ is given by

$$p_W(w) = P(W = w) = \frac{c(w)}{2^n},$$

where $c(w)$ is the coefficient of e^{wt} in the expansion of

$$\prod_{j=1}^{n} \left(1 + e^{jt}\right).$$

How does one turn this into a hypothesis test?

- **Pf.:** If H_0 is true, the pdf of the Wilcoxon rank statistic is equivalent to that of $U = \sum_j^n U_j$, where

$$U_j = \begin{cases} 0 & \text{with probability } 1/2 \\ j & \text{with probability } 1/2 \end{cases}$$

- Hence W and U have the same moment-generating function. Since the U_js are independent,

$$\begin{aligned} M_W(t) = M_U(t) &= \prod_{j=1}^n M_{U_j}(t) = \prod_{j=1}^n E\left(e^{U_j t}\right) \\ &= \prod_{j=1}^n \left(\frac{1}{2}e^{0t} + \frac{1}{2}e^{jt}\right) = \frac{1}{2^n}\prod_{j=1}^n \left(1 + e^{jt}\right) \end{aligned}$$

How does one turn this into a hypothesis test?

- At this point, we have

$$M_W(t) = \frac{1}{2^n}\prod_{j=1}^{n}\left(1+e^{jt}\right) = \frac{1}{2^n}\sum_{w=0}^{n(n+1)/2} c(w)e^{wt}.$$

since $1+2+\cdots+n = n(n+1)/2$.

- Next, by definition of generating functions,

$$M_W(t) = E\left(e^{Wt}\right) = \sum_{w=0}^{n(n+1)/2} p_W(w)e^{wt}$$

- Comparing the above, we see the desired result

$$p_W(w) = \frac{c(w)}{2^n}. \quad \square$$

Example of Wilcoxon test

- Suppose that $n = 4$ so $2^n = 16$ and $n(n+1)/2 = 10$, so

$$M_W(t) = \left(\frac{1+e^t}{2}\right)\left(\frac{1+e^{2t}}{2}\right)\left(\frac{1+e^{3t}}{2}\right)\left(\frac{1+e^{4t}}{2}\right)$$
$$= \frac{1}{16}\left(1 + e^t + e^{2t} + 2e^{3t} + 2e^{4t} + 2e^{5t} + 2e^{6t} + 2e^{7t} + e^{8t} + e^{9t} + e^{10t}\right)$$

- Hence, e.g., $P(W = 2) = {}^1\!/16$ and $P(W = 7) = {}^2\!/16$, etc.
- Note that $c(w)$ is the number of ways of adding subsets of the numbers $\{1, 2, \ldots, n\}$ to obtain w.

Example of application of Wilcoxon test

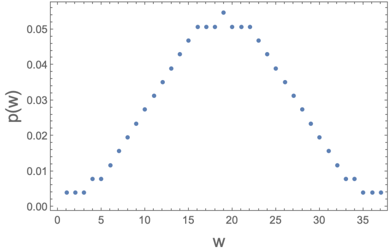
- Energy expenditures for women from heart rate (in kcal)
- During summer months and winter months
- Test difference D with $H_0: \mu = 0$ and $H_1: \mu \neq 0$

Subject	Summer, x_j	Winter, y_j	$d_j = y_j - x_j$	r_j	z_j
1	1458	1424	-34	1	0
2	1353	1501	148	5	1
3	2209	1495	-714	8	0
4	1804	1739	-65	2	0
5	1912	2031	119	4	1
6	1366	934	-432	7	0
7	1598	1401	-197	6	0
8	1406	1339	-67	3	0

- Wilcoxon rank statistic: $w = 5(1) + 4(1) = 9$

Example of application of Wilcoxon test

- Wilcoxon rank statistic: $w = 5(1) + 4(1) = 9$
- $\sum_{w=0}^{n(n+1)/2} p(w) e^{wt} = \frac{1}{2^n} \prod_{j=1}^{n} \left(1 + e^{jt}\right)$



- $\sum_{w=0}^{7} p(w) = \sum_{w=29}^{36} p(w) = \frac{19}{256} \approx 0.0742$
- Test is two-sided, so $2 \times \frac{19}{256} = \frac{19}{128} \approx 0.148$
- So for $\alpha = 0.15$, since $7 < w < 29$, we fail to reject H_0.

Example of application of Wilcoxon test

- $\sum_{w=0}^{n(n+1)/2} p(w) e^{wt} = \frac{1}{2^n} \prod_{j=1}^{n} \left(1 + e^{jt}\right)$
- Tables of cutoffs for $n = 4, \ldots, 12$ are in Larsen & Marx Appendix A, Table A.6.
- For $n = 8$ one can look up that

$$P(W \leq w_1^*) = P(W \geq w_2^*) = 0.074$$

corresponds to $w_1^* = 7$ and $w_2^* = 29$, as we calculated on the previous slide.

Summary

- We have motivated and discussed nonparametric statistics.
- We have learned about the sign test for estimations of medians $\tilde{\mu}$, and worked an example.
- We have learned about the Wilcoxon text for estimation of means μ, and worked an example.