Rank-Based Tests

Geoff Gordon
ggordon@cs.cmu.edu

December 6, 1995
The problem

Given two sets of samples, do they come from the same distribution?

*E.g.*, does the new drug change the expected lifetime of the patients, does the new EBL algorithm change the performance of our theorem prover?

Assume all samples are independent.
The framework

Given:
- sample $X_1\ldots X_n$
- indicators $Y_1\ldots Y_n$ (0 if sample $i$ from first set, 1 if from second)

Wish to check a null hypothesis such as

$$H_0: \text{The } X_i\text{'s all come from Gaussian distributions with the same mean and variance: } X_i \sim N(\mu, \sigma)$$

Evidence against $H_0$ strong $\Rightarrow$ reject $H_0$

Evidence weak $\Rightarrow$ provisionally accept $H_0$

Allow probability $\alpha$ (the significance level) of rejecting $H_0$ if it is true

No need to specify alternate hypothesis yet
Power

Suppose some alternate hypothesis, $H_1$, is true instead — e.g. the Gaussian location shift

$$H_1: (X_i - Y_i \theta) \sim N(\mu, \sigma)$$

Probability of rejecting $H_0$ if $H_1$ is true is the power of our test against $H_1$

If we choose a specific $H_1$ (e.g. $\theta = 1.8$), can look for most powerful test of $H_0$ v. that $H_1$

Or, could look for a good test against many different alternates — e.g., all $\theta > 0$, all $\theta \neq 0$

Such a test may not be most powerful against any one alternate
Power curves for one- and two-tailed \( t \)-tests, variance 1, 5% significance level, 50 samples in each group.

If alternates are parameterized by \( \theta \), can graph power vs. \( \theta \) — provides a concise summary.

For example, the point (.277, .5) means that the two-tailed \( t \)-test with this many samples can detect a difference of \( +.277 \) standard deviations half the time.

Want graph as high as possible at \( H_1 \), but no higher than \( \alpha \) at \( H_0 \).
Testing v. Estimation

Related problem: estimate \( E(S(X, Y)) \)

\( S \) is a statistic — some function of the data

Can choose \( S \) so null h. is \( E(S) = 0 \)

This \( S \) is called the test statistic

Observed value of \( S \) is evidence against null h.
Designing a parametric test

For a parametric test, assume we know how every sample depends on parameter of interest

That is, write $X_i \sim g_i$, where $g_i$ are known densities, each depending on parameter $\theta$

Want to estimate $\theta$ or test $H_0 : \theta = 0$
Maximum likelihood

To estimate $\theta$ by maximum likelihood:

$$\frac{d}{d\theta} \ln L(X, \theta) = \frac{d}{d\theta} \ln \prod_{i} g_i(x_i)$$

$$= \sum_{i} \frac{d}{d\theta} \ln g_i(x_i)$$

$$= \sum_{i} \frac{d}{d\theta} \frac{g_i(x_i)}{g_i(x_i)}$$

We say $\xi_i = \frac{d}{d\theta} \frac{g_i(x_i)}{g_i(x_i)}$ is the score for $X_i$

Can estimate $\theta$ by setting sum of scores to 0
ML example

If \( X_i \sim N(Y_i \theta, 1) \), then

\[
g_i(x) = \frac{1}{\sqrt{2\pi}} \exp \frac{-(x - y_i \theta)^2}{2}
\]

\[
\frac{d}{d\theta} g_i(x) = g_i(x) (x - y_i \theta) y_i
\]

\[
\xi_i = (x_i - y_i \theta) y_i
\]

So if \( Y_i \) is 0, \( i \)th score is 0, while if \( Y_i \) is 1, \( i \)th score is \( (x_i - \theta) \)

Suppose first \( m \) samples have \( Y_i = 1 \). Then sum of scores is \((\sum_i^m X_i - m\theta)\), and setting to 0 gives \( \theta_{ML} = \frac{1}{m} \sum_i^m X_i \).
Score statistic

Get ML estimate by setting total score to 0

How good an estimate is $\theta_0 \neq \theta_{ML}$ of $\theta$?

Sum scores for $\theta_0$, compare to 0

Called the score statistic
Score tests — I

Fact: locally most powerful test for $\theta = \theta_0$ can be based on the score statistic (consequence of Neyman-Pearson lemma)

Locally most powerful: nearly most powerful for alternates $\theta_1$ near $\theta_0$

Form of score test for $\theta_1 > \theta_0$:

- compute null distribution of score statistic
- pick cutoff $C$ so $P_0(\text{score} > C) = \alpha$
- reject if score $> C$

N-P doesn’t tell us null distribution
Score test example

Suppose \( X_i \sim N(Y_i\theta, 1) \) and \( H_0 : \theta = 0 \)

Score statistic at \( \theta = 0 \) is \( \xi = \sum_i^m X_i \)

Each \( X_i \sim N(0, 1) \) under \( H_0 \) so \( \xi \sim N(0, \sqrt{m}) \)

For \( \alpha = .05, \theta_1 \text{ +ve, reject if } \sum_i^m X_i > 1.65\sqrt{m} \)

Simple version of Student’s \( t \)-test
Parametric null hypotheses

$t$-test specifies a parametric null h.: statement about parameters of an assumed distribution

If it rejects $H_0$, know either

- $X \not\sim Y$, or
- $X \not\sim N(\mu, \sigma)$, or
- $Y \not\sim N(\mu, \sigma)$

If we’re not sure that $X$ and $Y$ are Gaussian, above conclusion is useless
Nonparametric null hypotheses

Nonparametric h. assumes no distribution: e.g.

\[ H_0: X_i \sim X_j \]

To assess power, can use any alternate h., parametric or nonparametric

Often choose a parametric alternate, to see whether our nonparametric test is less powerful than corresponding parametric test
Designing nonparametric tests

Test must not reject a true $H_0$ too often, no matter what distribution $X_i$s have

One way to ensure this: base test on a statistic whose null distribution doesn’t depend on distribution of $X_i$s

Fact: can transform any distribution with continuous c.d.f. to any other via a monotone transformation (if c.d.f.s are $F, G$ then transform is $G^{-1}(F(X))$)

$\Rightarrow$ test statistic must be invariant under monotone transforms
Rank tests

Define (1) to be index of smallest $X$, (2) next smallest, etc.

Rank vector $R = ((1), (2), \ldots, (N))$ is \textit{maximal invariant statistic} under monotone transforms

That is, any statistic unaffected by monotone transforms is a function of rank vector

$\Rightarrow$ test statistic must be a function of $R$
Rank scores

Suppose \( x_i \) has density \( g_i \)

Let \( A \) be the region where \( x_{(1)} < x_{(2)} < \ldots, \) i.e., where \( R \) is correct rank vector.

Score for \( R \) is then

\[
\frac{d}{d\theta} \ln L(R, \theta) = \frac{d}{d\theta} \ln \int_A \prod_{i}^{N} g_i(x_i) dX
\]

\[
= \frac{1}{L(R, \theta)} \int_A \frac{d}{d\theta} \prod_{i}^{N} g_i(x_i) dX
\]

\[
= \int_A \left( \sum_{i}^{N} \frac{\frac{d}{d\theta} g_i(x_i)}{g_i(x_i)} \right) \frac{\prod_{i}^{N} g_i(x_i)}{L(R, \theta)} dX
\]

\[
= \sum_{i}^{N} E_{\theta} \left( \frac{\frac{d}{d\theta} g_i(x_i)}{g_i(x_i)} \right)
\]
Properties of rank scores

Score for $X_i$ is $\xi_i = E_\theta \left( \frac{d}{d\theta} g_i(x_i) \right) / g_i(x_i)$

That is, rank-based scores are the expectation (over observations consistent with the rank vector) of the original scores

Above is true in general of partly-observed data

Even though we computed scores from assumed $g_i$s, $\xi$ is a function of ranks only and so does not depend on distribution of $X_i$s

$\Rightarrow$ test is nonparametric
Normal scores test

In the $t$-test, scores were 0 or $X_i$

For rank-based test, want 0 or $E(X_i|R) = E(X_{(j)})$

Call latter quantity $z_{jn}$ (a normal score)

E.g., $z_{3,17}$ is expectation of 3rd largest of 17 samples from a standard normal
Permutation distribution

What is distribution of $\xi$?

Under $H_0$, $X_i \sim X_j$ — so interchanging $X_i$ and $X_j$ leaves likelihood unchanged

So all $2^n$ permutations of $X_i$s are equally likely

So $\xi$ is the sum of $m$ numbers chosen w/o replacement from the set $z_1 \ldots z_n$

So $\xi$ is asymptotically normal with

$$E(\xi) = \frac{1}{n} \sum_i^n z_{in} = 0$$

$$V(\xi) = \frac{1}{n-1} \sum_i^n z_{in}^2 \sum_i^n (Y_i - \bar{Y})^2$$
Normal scores example

Suppose $X = (5, 1, 3, 2, 6)$ and $Y = (0, 0, 1, 0, 1)$

Normal scores for $n = 5$ are $-1.16, -.5, 0, .5, 1.16$

$\xi = 0 + 1.16$

$V(\xi) = \frac{1}{4}(1.35 + .25 + 0 + .25 + 1.35)(.36 + .36 + .16 + .36 + .16) = 1.12$

So $\xi$ is $\frac{1.16}{\sqrt{1.12}} = 1.09$ devs above mean, and $p = 14\%$, not enough to reject $H_0$
Wilcoxon test

Normal and logistic density functions

Logistic distribution has c.d.f. \( \frac{1}{1+\exp(-x)} \)

Similar to normal, but heavier tails (in graph, 13% higher std. dev.)

Logistic scores are \( w_{in} = \frac{2i}{n+1} - 1 \)

Corresponding test is Wilcoxon (also Kruskal-Wallis, Mann-Whitney, rank sum)
Comparison

\[ H_0 : X_i \sim X_j \text{ v. location } H_1 : (X_i - Y_i\theta) \sim g \]

If \( g \) is Gaussian:

- \( t \)-test is fully efficient
- normal scores asymptotically efficient
- Wilcoxon has asymptotic relative efficiency 0.955, \textit{i.e.}, about 5\% more samples for same power

If \( g \) is not Gaussian:

- \( t \)-test is invalid
- normal scores and Wilcoxon are still valid, but may be less than 100\% efficient
- Wilcoxon has ARE 1 for \( g \) logistic

Gaussian location-scale alternate: \( t \) is best
Paired tests

Two samples, $X_1 \ldots X_n$ and $Y_1 \ldots Y_n$

$X_i$ and $Y_i$ are more similar to each other than to $X_j$ or $Y_j$

E.g., drug v. placebo on each of $n$ patients, two types of fertilizer on each of $n$ fields

We will discuss:

- weak pairing: null h. is $X_i \sim Y_i$ (but distribution of $X_i$ and $X_j$ not related)
- strong pairing: assume all samples have same distribution up to location, null h. is that $i$th pair has same location
Weak pairing

How nonparametric do we want to be? (I.e., invariant under which transformations?)

Completely nonparametric:

• Invariant to monotone transform of each pair separately
• Max invariant statistic is count of $X_i > Y_i$
• This is sign test — asymptotically $N\left(\frac{n}{2}, \frac{\sqrt{n}}{2}\right)$
Weak pairing, cont’d

“Mostly” nonparametric:

• Invariant to monotone transform of all data simultaneously
• Max invariant stat is combined rank vector
• Can compute scores as before
• Condition on observed score pairing
• Permutation distribution: $i$th score equally likely to come from $X_i$ or $Y_i$

\[
\sum_i (\xi_i - \xi_i') \sim N(0, \sum_i (\xi_i - \xi_i')^2)
\]
Strong pairing

\((X_i - Y_i - \theta_i) \sim g\) for some symmetric \(g\)

Split into \(\text{sign}(X_i - Y_i - \theta_i), |X_i - Y_i - \theta_i|\)

Invariant to monotone transform of \(|X_i - Y_i - \theta_i|\)

Max invariant stat: signs, ranks for \(|X_i - Y_i - \theta_i|\)
(under \(H_0\), ranks for \(|X_i - Y_i|\))

Compute scores as before, except we now want expected abs values of scores — examples:

- double-exponential: sign test
- logistic: signed ranks (paired Wilcoxon)
- normal: signed normal scores

Permutation distribution: \(\sum_i s_i \xi_i \sim N(0, \sum_i \xi_i^2)\)