Applied Nonparametrics
STA 4502/5507

Yiyuan She

Department of Statistics, Florida State University

Fall 2009

- Project due: Friday, December 4 2009 by 4pm
  - 5-12 pages
  - a copy of the complete journal article from which you obtained the data
- Final
  - Wednesday, December 9, 12:30 - 2:30 pm in HCB 0207
  - Closed book, calculator permitted
  - You may bring TWO pages (letter-size) of notes which must be handwritten. (You do not have to prepare the tables.)

Course Flow

- Estimating success probabilities
- Single location: estimates, tests, intervals
- Two locations: testing, estimating differences between locations
- Scale comparisons
- Multiple locations and factors
- Independence
- ‘Nonparametric regression’: Theil’s test
- Nonparametric regression: kernels, splines

Fundamental Statistics

- Some basic concepts: Type-I error, Type-II error, Power; point estimate, confidence interval
- Critical value method for hypothesis testing (use the quantile function in `R`)
- *p*-value: smallest significance level at which we would reject the null based on the given data (use the distribution function in `R`)
- *p*-value methods is more informative (why?)

Binomial Test

- $X_1, X_2, \cdots, X_n i.i.d. \sim \text{Bernoulli}(p)$
- Test statistic: $B = \sum X_i$
- Exact Binomial Test: $B \sim Bin(n, p_0)$
- Large Sample Test: standardization, CLT
- Estimation: point, confidence interval

Wilcoxon Signed Rank Test

- Paired replicates data
- What are the assumptions? (pairs, continuity, symmetry, independence)
- Null: $H_0 : \theta = 0$
- No difference before and after

- Test statistic is $T^+ = \sum_{i=1}^{n} R_i \psi_i$
- Reject when T^+ big
- Null distribution: range, symmetry, moments, large sample approximation
- `R: wilcox.test`
- Use `(y, x)` or `y-x`
- Set `paired` to be `TRUE`

- Point estimate: $\hat{\theta} =$ median $\left\{ \frac{Z_i + Z_j}{2}, i \leq j = 1, 2, \ldots, n \right\}$
- Confidence Interval: Tukey's idea

Fisher Sign Test

- Assumptions: Paired observations again, independence, common median θ : $F_i(\theta) = 1 - F_i(\theta)$
- Not necessarily symmetric (weaker assumption)
- Use signs, instead of ranks (comparison)
- Test statistic is $B = \sum_{i=1}^{n} \psi_i$
- Null distribution: $B \sim \text{Bin}(n, 0.5)$
- Large Sample Approximation
- Estimation: $\hat{\theta} = \text{ median} \{Z_i, i = 1, 2, \ldots, n\}$
- CI

Wilcoxon Rank Sum (Mann-Whitney) Test

- Assumptions (continuous, iid, location shift model)
- Null: $H_0 : F(t) = G(t)$ for all t
- Location shift: $Y \overset{d}{=} X + \Delta$ (or $G(t) = F(t - \Delta)$ for all t), Δ: location shift or treatment effect
- $W = \sum_{j=1}^{n} S_j = \sum_{j=1}^{n} rank(Y_j)$
- $W = U + \frac{n(n+1)}{2}$
- Null distribution

- R: `wilcox.test`
- The `R` example in class!

Robust Rank Test (Fligner-Policello)

- Assumptions: distribution of X (Y) is symmetric about median θ_x (θ_y)
- Compare with Wilcoxon rank-sum
- $H_0 : \theta_x = \theta_y$ (not $F = G$)
- Statistic: (compare the procedure to two sample t-test)

Ansari-Bradley Test for Two-Sample Scale Comparison

- Assumptions: iid, continuity, location-shift model, and common median
- Location-scale model assumption: $\frac{X-\theta_1}{\eta_1} \stackrel{d}{=} \frac{Y-\theta_2}{\eta_2} \sim H(\cdot)$, where H is a continuous distribution with median 0
- Common median: $\theta_1 = \theta_2$
- Parameter of interest: $\gamma^2 = \eta_1^2/\eta_2^2$
- $C = \sum_{j=1}^{n} R_j$ is the test statistic (symmetric ranking)

Jackknife & Bootstrap

- **Resampling** methods
- When to use them?
- Difference

(Two-Sample) Kolmogorov-Smirnov Test

- Test for differences in two populations
- Not location, not scale specific
- Assume X and Y independent (within and between samples)
- $H_0 : F(t) = G(t)$ vs. H_1: any difference, $F(t) \neq G(t)$ for at least one t
- Goodness of fit test
- The K Statistic, EDF

Kruskal-Wallis test

- Assumptions: independent + continuous + $F_j(t) = F(t - \tau_j)$, $t \in (-\infty, \infty)$, $j = 1, 2, \ldots, k$ where F is a continuous distribution function with *unknown* median θ
- $H_0 : \tau_1 = \tau_2 = \cdots = \tau_k$ vs. H_1: $\tau_1, \cdots, \tau_k$ not all equal
- Explain the parameters! (τ_j, $R_{.j}$, etc)
- Test statistic:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(R_{.j} - \frac{N+1}{2} \right)^2$$

or,

$$H = \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} \right) - 3(N+1)$$

- Null distribution: not symmetric
- Large sample approximation, χ^2_{k-1}

- Ties, H'
- `kruskal.test`
- Compare KW and Wilcoxon

Fligner-Wolfe test

- Assume one of the treatments is a control ($j = 1$)

$$H_0 : \tau_i = \tau_1, \quad i = 2, 3, \ldots, k$$

- Test statistic

$$FW = \sum_{j=2}^{k} \sum_{i=1}^{n_j} r_{i,j}$$

- Wilcoxon test!

Multiple Comparisons

- Suppose the null was rejected in Kruskal - Wallis test.
- Which treatments show differences?
- Pair-wise comparisons $(k(k-1)/2)$
- What is FWER? Why not use the canonical $\alpha = 0.05$ for each test?

D-S-C-F Test

- Test statistics:

$$W^*_{i,j} = \frac{W_{i,j} - \frac{n_j(n_i+n_j+1)}{2}}{\sqrt{\frac{n_i n_j(n_i+n_j+1)}{24}}}$$

- $\sqrt{2}\times$ standardarized $W_{i,j}$
- α is the experiment-wise rate (familywise error rate, FWER)
- Large sample approximation

Some General Procedures for Multiple Testings

- Bonferroni, Holm, BH
- What are these procedures?
- FDR vs. FWER

Kendall Test

- Assumptions: continuous + paired
- $H_0 : \tau = 0$
- Kendall's τ $\tau = 2P((Y_2 - Y_1)(X_2 - X_1) > 0) - 1$
- Explain the parameters!
- The test statistic $K = K' - K''$,
- based on signs (and ranks)
- Null distribution, large-sample approximation
- How to estimate τ?

- `cor.test`

Spearman Test

- Test statistic: r_s
- Large sample approximation
- `cor.test`!

Theil Test

- Simplest case: linear
- $Y_i = \alpha + \beta x_i + e_i, \quad i = 1, 2, \ldots, n$
- x known, α and β unknown
- e_i are continuous random variables with median 0
- $x_1 < x_2 < \cdots < x_n$
- Null $H_0 : \beta = \beta_0$

- Test statistic: C, based on Kendall test
- How to estimate the slope parameter and the intercept?

Nonparametric Regression

Kernel methods

- bandwidth
- What is a kernel function?
- Compare Theil's test with nonparametric regression