Applied Nonparametrics
STA 4502/5507

Yiyuan She

Department of Statistics, Florida State University

Fall 2009

Midterm Review

- Estimating success probabilities
- Single location: estimates, tests, intervals
- Two locations: testing, estimating differences between locations
- Scale comparisons and others

Problems

- One sample location (& location difference for paired replicates data)
- Two sample location difference
- Two sample scale difference
- Distribution comparison (goodness of fit)

Methods

- Rank/sign tests: ranks, signs
  - Fisher sign test, Wilcoxon signed rank test
  - Wilcoxon rank sum test, robust rank test
  - Ansari-Bradley test
- Jackknife
- Goodness of fit test: Kolmogorov-Smirnov

Binomial Test

- 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?)

Exact Binomial Test

- $X_1, X_2, \cdots, X_n \sim \text{Bernoulli}(p)$
- Consider $H_0 : p = p_0$
- Test statistic: $B = \sum X_i$
- Null distribution $B \sim Bin(n, p_0)$
- Rejection region

Large Sample Test

- B is asymptotically normal (CLT)
- This test will be approximate (OK for large samples)
- Under H_0:
  - $E(B) = np_0$
  - $\text{var}(B) = np_0(1 - p_0)$
- Standardize B

$$B^* = \frac{B - np_0}{\sqrt{np_0(1 - p_0)}}$$

- B^* approximately normal(0, 1)
- Critical values are now z_α

Estimation

- Point estimate for p: $\hat{p} = B/n$
- Confidence interval: $p_L(\alpha) = \hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$, $p_U(\alpha) = \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$
- Asymptotic method, based on

$$P\left(-z_{\alpha/2} < \frac{p-\hat{p}}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) = 1-\alpha$$

- `binom.test`

Wilcoxon Signed Rank Test

- Paired replicates data
- Distribution-free (X, Y)
- What are the assumptions? (pairs, continuity, symmetry, independence)
- Null:

$$H_0 : \theta = 0$$

- No difference before and after

Wilcoxon Test

- Set $\psi_i = \begin{cases} 1, & Z_i > 0, \\ 0, & Z_i < 0. \end{cases}$
- Get ranks R_i of $|Z_i|$
- Test statistic is $T^+ = \sum_{i=1}^n R_i \psi_i$
- Intuition: reject when T^+ big

Null distribution

- $0 \leq T^+ \leq \frac{n(n+1)}{2}$
- $E(T^+) = \frac{n(n+1)}{4}$
- $var(T^+) = \frac{n(n+1)(2n+1)}{24}$
- Symmetry: T^+ is symmetric about ET^+
- $T^+ \stackrel{d}{=} \sum_1^n ib_i$, where $b_i \stackrel{i.i.d.}{\sim}$ Bernoulli$(1/2)$
- Exact distribution available
- Asymptotically, $T^* = \frac{T^+ - E(T^+)}{\sqrt{\text{var}(T^+)}} \dot{\sim} N(0,1)$
- Some variants: ties, $H_0 : \theta = \theta_0$

- R: `wilcox.test`
- Use `(y, x)` or `y-x`
- Set `paired` to be `TRUE`

Estimation

- Point estimate: $\hat{\theta} =$ median $\left\{ \frac{Z_i + Z_j}{2}, i \leq j = 1, 2, \ldots, n \right\}$
- $n(n+1)/2$ of these Walsh averages
- Closely related to the signed rank test
- Hodges-Lehmann method
- Robust to outliers (compare to $\sum Z_i/n$)
- Interval estimate?

Confidence Interval

- Tukey's idea to get the interval estimate (θ_L, θ_U)
  - $W^{(i)}$ are the ordered pairwise averages of the Z_j: $W^{(1)} \leq W^{(2)} \leq \cdots \leq W^{(M)}$
  - Count in C_α from each end: $[W^{(i_1)}, W^{(i_2)}]$ $(i_1 + i_2 = M + 1 = \frac{n(n+1)}{2} + 1)$
- Calculate C_α $C_\alpha = \frac{n(n+1)}{2} + 1 - t_{\alpha/2}$ where $t_{\alpha/2}$ is the upper $\alpha/2$th percentile of the null distribution of T^+ (A.4).
- $\theta_U = W^{(t_{\alpha/2})}$, $\theta_L = W^{(C_\alpha)} = W^{(M+1-t_{\alpha/2})}$ with confidence level $1 - \alpha$

Fisher Sign Test

- Assumptions: Paired observations again, independence, common median θ : $F_i(\theta) = 1 - F_i(\theta)$
- Not necessarily symmetric (weaker assumption)
- Set $\psi_i = \begin{cases} 1, & Z_i > 0, \\ 0, & Z_i < 0. \end{cases}$
- Test statistic is $B = \sum_{i=1}^{n} \psi_i$
- Null distribution: ψ_i i.i.d. $\sim$ Bernoulli(0.5) $\Rightarrow B \sim \text{Bin}(n, 0.5)$ which is symmetric
- Toss zeros, reduce n; no ranks

Large Sample Approximation

- Standardize

$$B^* = \frac{B - E(B)}{\sqrt{\text{var}(B)}} \sim N(0,1)$$

approximately under null, where $E(B) = np = n/2$ and $\text{var}(B) = np(1-p) = n/4$

Estimation

- Point estimate: $\hat{\theta} =$ median $\{Z_i, i = 1, 2, \ldots, n\}$
- A symmetric two-sided interval (θ_L, θ_U): $\theta_L = Z^{(C_\alpha)} = Z^{(n+1-b_{\alpha/2,n,1/2})}$, $\theta_U = Z^{(b_{\alpha/2,n,1/2})}$ with confidence level $1 - \alpha$

Signed Rank vs. Sign

- Robustness
- Efficiency
- Computational feasibility
- Both apply to one sample location problem as well

Wilcoxon Rank Sum (Mann-Whitney) Test

- Distribution-free
- Assumptions (continuous, iid, location shift model)
- Null: $H_0 : F(t) = G(t)$ for all t
- Location shift
  - $Y \stackrel{d}{=} X + \Delta$ (or $G(t) = F(t - \Delta)$ for all t), Δ: location shift or treatment effect
  - $E(X)$ and $E(Y)$ may not exist
  - $H_0 : \Delta = 0$ vs. $H_1 : \Delta >, <, \neq 0$

Wilcoxon Rank Sum Statistic

- Rank the $N(= m + n)$ combined samples
- Denote the ranks of Y within this ranking as S_i
- $W = \sum_{j=1}^{n} S_j = \sum_{j=1}^{n} rank(Y_j)$
- $U = \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(X_i, Y_j)$, where $\phi(X_i, Y_j) = 1_{X_i < Y_j}$
- $W = U + \frac{n(n+1)}{2}$

Null Distribution

- $n(n+1)/2 \leq W \leq n(2m+n+1)/2$
- The null distribution of W is symmetric about its mean $n(N+1)/2$, namely, $P(W \leq x) = P(W \geq n(N+1) - x)$
- Large sample approximation: $W^* = \frac{W - E(W)}{\sqrt{\text{var}(W)}} \dot\sim N(0,1)$ under null, where $E(W) = \frac{n(m+n+1)}{2}$, $\text{var}(W) = \frac{mn(m+n+1)}{12}$
- Variants: ties, $H_0 : \Delta = \Delta_0$

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

Estimation

- Point estimate based on Hodges-Lehmann method: $\hat{\Delta} = \text{ median}\{(Y_j - X_i), i = 1, 2, \ldots, m, j = 1, 2, \ldots, n\}$
- Interval estimate with confidence level $1 - \alpha$: $\Delta_L = U^{(C_\alpha)}$, $\Delta_U = U^{(mn+1-C_\alpha)}$

- Wilcoxon rank-sum test only assumes location difference
- No dispersion or shape differences
- No dependency
- Analogue: two-sample t-test with *equal* variances
- What if the variances are not equal? Behrens-Fisher problem
- Robust rank test: Welch's t-test (two-sample t-test with unequal variances)

Robust Rank Test (Fligner-Policello)

- Assumptions: distribution of X (Y) is symmetric about median θ_x (θ_y)
- $H_0 : \theta_x = \theta_y$ (not $F = G$)
- Statistic: (compare the procedure to two sample t-test)
  - P_i = number of sample Y observations less than X_i
  - Q_j = number of sample X observations less than Y_j
  - $\overline{P}$ = average X sample placement
  - $\overline{Q}$ = average Y sample placement
  - $V_1 = \sum_{i=1}^{m}(P_i - \overline{P})^2$
  - $V_2 = \sum_{j=1}^{n}(Q_j - \overline{Q})^2$
  - $\hat{U} = \frac{\sum_{j=1}^{n} Q_j - \sum_{i=1}^{m} P_i}{2(V_1 + V_2 + \overline{P}\,\overline{Q})^{1/2}}$
  - Asymptotically, $\hat{U} \sim N(0, 1)$ under null

Ansari-Bradley Test for Two-Sample Scale Comparison

- Assumptions: iid, continuity, location-shift model, and common median
- Location-scale model assumption: $\frac{X-\theta_1}{\eta_1} \overset{d}{=} \frac{Y-\theta_2}{\eta_2} \sim H(\cdot)$, 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 Statistic

- Ranks again
- Order the N combined sample values
- Assign 1 to smallest and largest
- Assign 2 to next smallest and next largest
- Continue ...
- R_j = score assigned to Y_j
- $C = \sum_{j=1}^{n} R_j$ is the test statistic

Null Distribution

- Not necessarily symmetric
- Exact distribution available though
- Large sample approximation: $C^* = \frac{C-E(C)}{\sqrt{\text{var}(C)}} \sim$ standard normal under null; use different formulas for E and Var
- Variants: ties, $H_0 : \gamma^2 = \gamma_0^2$
- R: `ansari.test`

Miller Jackknife Test

- Assumptions: Medians not equal (or known)
- Previous location-scale model assumption holds (Ansari - Bradley)
- Also assume: $E(V^4) < \infty$ where $V \sim H$ (and thus γ^2 is ratio of variances)
- Jackknife is a **resampling** method, but we usually set $k = 1$ — "leave one out" (no randomness)

General Jackknife Procedure

- Let $\hat{\theta}$ be the estimate using all the data
- For the i-th sample (without using piece i), calculate the estimate $\hat{\theta}_{(i)}$ in the same way, $i = 1, \ldots, n$
- Form $\tilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)}$, $i = 1, \ldots, n$
- The jackknife estimate $\hat{\theta}_J$ is the mean of $\tilde{\theta}_i$, namely, $\sum \tilde{\theta}_i / n$.
- The standard error of this estimate is estimated by $\sqrt{\frac{\sum_{i=1}^{n}(\tilde{\theta}_i - \hat{\theta}_J)^2}{(n-1)n}}$

Miller Jackknife Test

- Construct S_i, T_j as estimates for $\ln \eta_1^2$ and $\ln \eta_2^2$ respectively
- Set $A_i = mS_0 - (m-1)S_i$, $i = 1, 2, \ldots, m$, and $B_j = nT_0 - (n-1)T_j$, $j = 1, 2, \ldots, n$
- Then use $\bar{A}$ and $\bar{B}$ to estimate η_1^2 and η_2^2 respectively, with the variances given by V_1 and V_2. Therefore $\tilde{\gamma}^2 = e^{\{\overline{A}-\overline{B}\}}$ is an estimate of variance ratio
- The Q statistic: $Q = \frac{\overline{A}-\overline{B}}{\sqrt{V_1+V_2}}$
- Asymptotically, $Q \sim N(0,1)$ under null (independent of H) – asymptotically distribution free
- Not a rank test

(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

- Maximum distance (scaled) between the two empirical distribution functions
- EDF: $F_m(t) = \sum_{i=1}^{m} 1_{X_i \le t}/m$ and $G_n(t) = \sum_{i=1}^{n} 1_{Y_j \le t}/n$ which are non-decreasing, step functions
- Define the distance: $D_{m,n} = \max_{-\infty < t < \infty} \{|F_m(t) - G_n(t)|\}$
- $K_{m,n} = \sqrt{\frac{mn}{m+n}} D_{m,n}$
- Its (exact/asymptotic) null distribution does not depend on F or G!

- `R: ks.test`
- Can test X and Y, or,
- One-sample KS test: Test X against a particular distribution
- `pnorm(mean, sd)`, `pexp(mean)`, etc.