Applied Nonparametrics
STA 4502/5507

Yiyuan She

Department of Statistics, Florida State University

Fall 2009

Possible 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
- Other topics ...

Multiple Location

- Compared two population centers via locations (medians) in chapter 4
- Now, compare multiple locations
- Parametric equivalent is analysis of variance (ANOVA)
- ANOVA assumes means exist, variances exist, data follows particular distribution

Problem Setup

- k populations (treatments)
- $k - 1$ treatments and 1 control
- In first case, looking for differences in locations
- Second case compares treatments to control
- Also want to compare all k treatments to each other

Assumptions

- $N = \sum_{j=1}^{k} n_j$, n_j observations from jth treatment
- All N observations are independent
- $X_{1,j}, X_{2,j}, \ldots, X_{n_j,j}$ are data from treatment j, following continuous distribution F_j
- $F_j(t) = F(t - \tau_j), \quad t \in (-\infty, \infty), \quad j = 1, 2, \ldots, k$ where F is a continuous distribution function with *unknown* median θ
- τ_j is the treatment effect for population j

$$
\begin{array}{ccccc}
X_{1,1} & X_{1,2} & X_{1,3} & \cdots & X_{1,k} \\
X_{2,1} & X_{2,2} & X_{2,3} & \cdots & X_{2,k} \\
X_{3,1} & X_{3,2} & X_{3,3} & \cdots & X_{3,k} \\
 & X_{4,2} & X_{4,3} & \cdots & X_{4,k} \\
 & X_{5,2} & & \cdots & X_{5,k} \\
 & X_{6,2} & & \cdots & X_{6,k} \\
 & \vdots & & & \vdots \\
 & X_{n_2,2} & & & \vdots \\
 & & & & X_{n_k,k}
\end{array}
$$

Assumptions

- This corresponds to the parametric one-way ANOVA

$$X_{i,j} = \theta + \tau_j + \varepsilon_{i,j}, \quad i = 1, 2, \ldots, n_j, \quad j = 1, 2, \ldots, k$$

- θ is overall median
- τ_j is the treatment j effect
- Errors $\varepsilon_{i,j}$ are iid with median 0 from continuous distribution
- If errors are normally distributed, then medians = means = 0, constant variance

Kruskal - Wallis Test

- No difference among treatment effects τ_j

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_k$$

or,

$$H_0 : F_1 = F_2 = \cdots = F_k = F$$

- Only difference is in medians, all have same scale

Kruskal - Wallis Test

- Alternative

$$H_1 : \text{ Not } H_0$$

- At least one treatment effect is different, $\tau_1, \cdots, \tau_k$ not all equal

Kruskal - Wallis Test

- Order the N combined sample values $X_{i,j}$
- Get ranks $r_{i,j}$
- For each j, set

$$R_j = \sum_{i=1}^{n_j} r_{i,j}$$

and

$$R_{.j} = R_j / n_j$$

- $R_{.j}$ is the average rank for sample from treatment j

Kruskal - Wallis Test

- Test statistic:

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

or,

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

- Second slightly easier if doing this by hand
- Reject H_0 if $H \geq h_\alpha$
- Exact null distribution available. See Table A.12 ($k = 3, 4, 5$)

Kruskal - Wallis Test

- Why $\frac{N+1}{2}$?
- Motivation: under H_0, the average rank of the jth population should be close to $(N+1)/2$ for any j
- Note: $k = 2$ is equivalent to Wilcoxon rank-sum test

Kruskal - Wallis Test

- Large sample approximation
- $\min n_j \to \infty$
- $H \sim \chi^2_{k-1}$
- Reject H_0 if $H \geq \chi^2_{k-1,\alpha}$
- Chart A.2 (from 1944)

- Use R when possible

$$\texttt{pchisq(q, df, ncp=0, lower.tail=T)} = P(\chi^2_{df} \leq q)$$

$$\texttt{pchisq(1, 4, ncp=0, lower.tail=F)} = P(\chi^2_4 > 1)$$

$$\texttt{qchisq(p, df, ncp=0, lower.tail=T)} = q$$
$$\Rightarrow P(\chi^2_{df} \leq q) = p$$

$$\texttt{qchisq(.05, 4, lower.tail=F)=0.710723}$$
$$\Rightarrow P(\chi^2_4 > 0.710723) = 0.05$$

Kruskal - Wallis Test

- Continuous $\Rightarrow$ No ties, strictly increasing ranks
- Ties will occur in practice
- Give each group in tie the average of the scores

Kruskal - Wallis Test

- Modify H

$$H' = \frac{H}{1 - \left(\sum_{j=1}^{g} \frac{t_j^3 - t_j}{N^3 - N} \right)}$$

- H is computed as before with average ranks
- g is the number of tied groups
- t_j is the size of tie group j
- Untied observation is a tie group of size $t_j = 1$
- If no ties, $g = N$, $t_j^3 = t_j = 1$, and $H' = H$
- Approximately a level-α test

Kruskal - Wallis Test

- R
- `kruskal.test(x, g)`
- `x` is the vector of all the data, each treatment sample laid end-to-end
- `g` is the vector specifying which treatment the data belongs to

Kruskal - Wallis Test

- Imagine 3 treatments with $n_1 = 3, n_2 = 4, n_3 = 4$
- Treatment 1 data: 3, 4, 4
- Treatment 2 data: 4, 4, 5, 4
- Treatment 3 data: 1, 2, 2, 4
- Then set `x=c(3, 4, 4, 4, 4, 5, 4, 1, 2, 2, 4)`
- The corresponding g is `g=c(1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3)`
- Or, `g=c(rep(1, 3), rep(2, 4), rep(3, 4))`

Kruskal - Wallis Test

- Use R
- Understand output
- The limitations of the assumptions

Other Alternatives

- Ordered alternative
  - Jonckheere-Terpstra trend test

$$H_1 : \tau_1 \leq \tau_2 \leq \cdots \leq \tau_k$$

  - `JT.test` in R package `SAGx`,
  - Has large sample approximation (normal)
  - 6.2
- Umbrella alternatives
  - $H_1 : \tau_1 \leq \tau_2 \leq \cdots \leq \tau_{p-1} \leq \tau_p \geq \tau_{p+1} \geq \cdots \geq \tau_k$
  - p (peak) fixed or unspecified
  - 6.3
- Skip

Control vs. Treatment

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

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

- Same null as before
- Test statistic

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

- Fligner - Wolfe

Control vs. Treatment

- We're adding the ranks of treatments $2, 3, \ldots, k$ with respect to all k treatments
- Two sample rank-sum test!
- X is the n_1 samples from control
- Y is the $N^* = N - n_1$ samples from other treatments

Control vs. Treatment

- One-sided test

$$H_1 : \tau_i \geq \tau_1, \quad i = 2, 3, \cdots, k$$

- Reject H_0 when $FW \geq f_\alpha$
- To use A.6 (which assumes $n \leq m$)
  - If $n_1 \geq N^*$ then $f_\alpha = w_\alpha$ $(m = n_1, n = N^*)$(A.6)
  - If $n_1 < N^*$ then

$$f_\alpha = w'_\alpha + \frac{(N - 2n_1)(N + 1)}{2}$$

    where w'_α comes from A.6 with $m = N^*$ and $n = n_1$
  - Why?

Control vs. Treatment

- One-sided test

$$H_1 : \tau_i \leq \tau_1, \quad i = 2, 3, \cdots, k$$

- Reject H_0 when $FW \leq N^*(N+1) - f_\alpha$

Control vs. Treatment

- Large sample approximation
- $E(FW) = \frac{N^*(N+1)}{2}$
- $\text{var}(FW) = \frac{n_1 N^*(N+1)}{12}$
- $FW^* = \frac{FW - EFW}{\sqrt{\text{var}(FW)}} \sim N(0,1)$

Control vs. Treatment

- Ties ...
- Average ranks
- Approximately level α
- Modify variance in large sample approximation

Multiple Comparisons

- Suppose the null was rejected in Kruskal - Wallis test
- Which treatments show differences?
- Look at all pair-wise comparisons
- Many of them:

$$\binom{k}{2} = \frac{k(k-1)}{2}$$

- Two-sided

Multiple Comparisons

- For each pair (i, j),

$$W_{i,j} = \sum_{b=1}^{n_j} R_{i,b}, \quad 1 \leq i < j \leq k$$

- $R_{i,b}$ are the ranks of the sample from treatment j with respect to the combined sample of treatments i and j
- $W_{i,j}$: Wilcoxon rank sum of the jth sample in the joint two-sample ranking of the i-th and j-th sample observations
- $\frac{k(k-1)}{2}$ of these $W_{i,j}$
- Problem: control the familywise error rate

Multiple Comparisons

- Goal: control the familywise error rate (overall Type-I error)
- Why? Consider $m = 100$ independent tests, each with significance level $\alpha = 0.05$
  - $P(\text{at least one false positive}|H_0) = 1 - (1 - \alpha)^m = 0.994 \gg 0.05$
  - $E(\#\text{ of false positives}|H_0) = m\alpha = 5$ due to chance!

When m is large (gene microarray analysis $\sim 10k, 100k$), choosing $\alpha = 0.05$ does not have a good control of the (familywise) Type-I error, and results in too many false positives.

Multiple Comparisons

- Set

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

- $\sqrt{2}\times$ standardarized $W_{i,j}$
- α is the experiment-wise rate (familywise error rate, FWER)
- P(at least 1 type I error among all pair-wise comparisons) $= \alpha$
- P(making all correct decisions$|H_0$ is true$) = 1 - \alpha$

Multiple Comparisons

- For each pair of treatments τ_u, τ_v

$$\tau_u \neq \tau_v \text{ if } |W^*_{u,v}| \geq w^*_\alpha$$

- $P(|W^*_{u,v}| < w^*_\alpha \text{ for all pairs } u, v | H_0) = 1 - \alpha$
- Table A.16 (exact but small)
- Ties ... average ranks ... modify variance ...
- Distribution free due to the use of Wilcoxon test

Limitations of the D-S-C-F Multiple Comparison Procedure

- Table A.16 limited use ($k \leq 3$ and $n_i \leq 7$)
- Depend on the distribution of Wilcoxon rank-sum statistic (though independent of the data distribution), not general for multiple comparisons
- FWER may be very conservative. Hard to tabulate for other criteria (such as FDR)

Multiple Comparisons

- Large sample approximation
- min $n_j \to \infty$
- Asymptotically, $\max_{u<v} |W^*_{u,v}| \stackrel{d}{=}$ range of k independent standard normal random variables, i.e., $\max_{1\le u<v\le k} |Z_u - Z_v|$, where Z_i are iid $\sim N(0,1)$
- $R_{u,v} = \{\tau_u \neq \tau_v \text{ if } |W^*_{u,v}| \geq q_\alpha\}$
- q_α: Table A.17

More General Procedures for Multiple Comparisons

- Bonferroni
- Holm
- Benjamini-Hochberg

The error control will be approximate, but you are free to use any test for each comparison

Bonferroni Procedure

$$P(\text{at least one false prositive}) \leq P(\text{reject 1}|H_0) + P(\text{reject 2}|H_0) + \cdots + P(\text{reject m}|H_0)$$

- FWER control: $P(\text{at least one false prositive}) \leq \alpha$
- $P(\text{reject } j|H_0) \leq \alpha/m \ \forall j$ suffices
- For example, if you have 50 hypothesis tests, and want $FWER \leq 0.05$, Bonferroni's choice is $0.05/50 = 1e-3$ as the significance level for each test
- Extremely conservative for large m
- May fail to identify true positives (Type-II error large)
- Other alternatives possible?

Holm Procedure

Given the FWER bound α (say 0.05)

- Order the p-values. WLOG, assume $p_1 \leq p_2 \leq \cdots \leq p_m$
- Compare p_1 to $\frac{\alpha}{m}$, p_2 to $\frac{\alpha}{m-1}$, $\cdots$, p_j to $\frac{\alpha}{m+1-j}$, $\cdots$, p_m to α
- Let j_0 be the first index such that $p_j > \frac{\alpha}{m+1-j}$, $1 \leq j \leq m$.
- Reject $1, 2, \cdots, j_0 - 1$. Accept $j_0, \cdots, m$.
- Stepwise (or sequential)
- Always dominates Bonferroni
- Still not powerful enough to detect all positives (due to the FWER)

False Discovery Rate

- A better control: $\frac{\#\text{ of false rejected}}{\#\text{ of rejected in total}}$

$$\begin{aligned} FWER &= P(\#\text{ of false rejected} > 0) \\ &= E\left[1_{\#\text{ of false rejected}>0}\right] \end{aligned}$$

$$FDR = E\left[\frac{\#\text{ of false rejected}}{\#\text{ of rejected in total}}\right]$$

False Discovery Rate

	Accepted	Rejected	
H_0 true	U	V	m_0
H_1 true	S	T	m_1
	W	R	

$FWER = P(V > 0) = E[1_{V>0}]$ while $FDR = E\left[\frac{V}{R}\right]$.
Obviously, $FDR \leq FWER$. Less stringent and more meaningful.

Benjamini-Hochberg Procedure
To have FDR control $\leq \alpha$

- Order the p-values. WLOG, assume $p_1 \leq p_2 \leq \cdots \leq p_m$
- Compare p_1 to $\frac{1}{m}\alpha$, p_2 to $\frac{2}{m}\alpha$, $\cdots$, p_j to $\frac{j}{m}\alpha$, $\cdots$, p_m to α
- Let j_0 be the first index such that $p_j > \frac{j}{m}\alpha$, $1 \leq j \leq m$.
- Reject $1, 2, \cdots, j_0 - 1$. Accept $j_0, \cdots, m$.
- For dependent tests, replace α by $\frac{\alpha}{\sum_1^m \frac{1}{i}}$

FDR is more appropriate than FWER in multiple hypothesis testing (comparison). FDR procedures (BH) are more powerful.

Others

- q-value (Storey 2003)
- Empirical null, Local FDR (Efron 2004)
- k-FWER (Romano 2005)
- ...

A lot of R packages

Multiple Comparisons

- Can do multiple comparison for ordered alternatives (6.6)
- Can do multiple comparison for treatments vs. control (6.7)
- Can estimate linear combinations of treatments (contrasts) and obtain confidence intervals for these contrasts (6.8, 6.9)
- Skip

Chapter 7

- Chapter 6 is one-way model
- One effect - τ_j
- Chapter 7 examines two-way models
- Two effects: treatment and block