Possible Course Flow

- Estimating success probabilities
- Single location: estimates, tests, intervals
- Two locations: testing, estimating differences between locations
- **Scale comparisons and others**
- Multiple locations and factors
- Independence
- Nonparametric regression
- Other topics ...

Two-Sample Problem

- Compare two population centers via locations (medians)
- Now, compare scale parameters
- Perhaps same location, perhaps not
- Even more generally, compare two distributions in all respects

Assumptions

- $X_i, i = 1, 2, \ldots, m$ iid
- $Y_i, i = 1, 2, \ldots, n$ iid
- $N = m + n$ observations
- X_i's and Y_i's are independent
- Continuous populations
- F is distribution of X, population 1
- G is distribution of Y, population 2

Ansari - Bradley Test

- Distribution-free
- Ranks again
- Null:

$$H_0 : F(t) = G(t) \text{ for all } t$$

Same distribution (but no specified)

- Assume same median $\theta_1 = \theta_2$

Ansari - Bradley Test

- Interested in knowing if one distribution has different variability than the other
- Suppose
  - $F(t) = H\left(\frac{t-\theta_1}{\eta_1}\right)$
  - $G(t) = H\left(\frac{t-\theta_2}{\eta_2}\right)$
  - Equivalently: $X \stackrel{d}{=} \eta_1 Z + \theta_1$, $Y \stackrel{d}{=} \eta_2 Z + \theta_2$, with $Z \sim H$
- H is continuous with median 0 $\Rightarrow F(\theta_1) = G(\theta_2) = 1/2$
- Further assumption: $\theta_1 = \theta_2$ (common median)
- In summary, $\frac{X-\theta}{\eta_1} \stackrel{d}{=} \frac{Y-\theta}{\eta_2}$
- If $\theta_1 \neq \theta_2$, but both are **known**, shift each sample: $X_i' = X_i - \theta_1, \quad Y_i' = Y_i - \theta_2$. Now have common median 0

Ansari - Bradley Test

- Look at ratio of scales: $\gamma = \eta_1/\eta_2$
- If variances exist for X and Y, then

$$\gamma^2 = \frac{\text{var}(X)}{\text{var}(Y)}$$

- Write null as

$$H_0 : \gamma^2 = 1$$

Ansari - Bradley Test

- 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

Ansari - Bradley Test

- One-tail alternative

$$H_1 : \gamma^2 > 1$$

- Reject H_0 if $C \geq c_\alpha$
- Table A.8
- Assumes Y is the smaller sample size ($n \leq m$)

Ansari - Bradley Test

- One-tail alternative

$$H_1 : \gamma^2 < 1$$

- Reject H_0 if $C \leq c_{1-\alpha} - 1$
- Two-tail alternative

$$H_1 : \gamma^2 \neq 1$$

- Reject H_0 if $C \geq c_{\alpha_1}$ or $C \leq c_{1-\alpha_2} - 1$
- Typically set $\alpha_1 = \alpha_2 = \alpha/2$ (valid for even N due to symmetry)

Ansari - Bradley Test

- Large sample approximation
- When N even:
  - $E(C) = \frac{n(N+2)}{4}$
  - $\text{var}(C) = \frac{mn(N+2)(N-2)}{48(N-1)}$

  Null distribution is symmetric $\Rightarrow c_{1-\alpha} - 1 = \frac{n(N+2)}{2} - c_\alpha$
- Otherwise:
  - $E(C) = \frac{n(N+1)^2}{4N}$
  - $\text{var}(C) = \frac{mn(N+1)(N^2+3)}{48N^2}$

Ansari - Bradley Test

- CLT, standardize

$$C^* = \frac{C - E(C)}{\sqrt{\text{var}(C)}} \sim \text{ standard normal}$$

- Need smaller of n and m large
- Use $z_\alpha, z_{\alpha/2}$

Ansari - Bradley Test

- Continuous $\Rightarrow$ No ties, strictly increasing ranks
- Ties will occur in practice
- Give each group in tie the average of the scores
- Approximately a level-α test
- In large sample approximation, have different value for variance
- If N even,

$$var(C) = \frac{mn\left[16\sum_{j=1}^{g} t_j r_j^2 - N(N+2)^2\right]}{16N(N-1)}$$

- g is number of groups, t_j is size of group, r_j is average in group

Assumptions

- $E(X)$ $E(Y)$ may not exist
- Only scale difference
- Common median assumption is essential

Ansari - Bradley Test

- Test $H_0 : \gamma^2 = \gamma_0$ with common median θ_0 (*known*)
- Use $X_i' = (X_i - \theta_0)/\gamma_0$ and $Y_i' = (Y_i - \theta_0)$
- Perform test with Y_i' and X_i'

Ansari - Bradley Test

- `R`
- `ansari.test(x, y, exact, conf.int, conf.level)`
- Confidence interval (Bauer, Comment 12)
- Estimates the ratio of the scales
- `R` uses different method when ties cross the center point

Miller Jackknife Test

- Medians not equal (or known)
- Previous location-scale model assumption holds (Ansari - Bradley)
- Also assume: $E(V^4) < \infty$ where $V \sim H$
- This assumption implies that γ^2 is ratio of variances
- Uses jackknife method
- More applicable

Miller Jackknife Test

- Jackknife is a **resampling** method
- Sample the data repeatedly (*without* replacement), form estimates for each sample
- Combine these estimates; solutions are functions of the estimates
- In jackknife, each sample leaves out one particular piece of data
- If there are n pieces of data, then there are n jackknife samples
- Sometimes referred to as "leave one out" method

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 given by $\sqrt{\frac{\sum_{i=1}^{n}(\tilde{\theta}_i - \hat{\theta}_J)^2}{(n-1)n}}$
- Why there's an extra factor n?

Miller Jackknife Test

- Get $\overline{X}_i$, $\overline{Y}_j$ for each jackknife sample
- $\overline{X}_i$ is the sample mean without using data piece i
- Also get sample variances for X and Y, leaving out one data piece each time
- The Miller test will also use full sample mean and sample variance to construct test statistic

Miller Jackknife Test

$$\overline{X}_i = \frac{1}{m-1} \sum_{s \neq i}^{m} X_s$$

$$D_i^2 = \frac{1}{m-2} \sum_{s \neq i}^{m} (X_s - \overline{X}_i)^2$$

Get $\overline{Y}_j$ and E_j^2 for Y

Miller Jackknife Test

Set

$$S_i = \ln D_i^2, \qquad i = 1, 2, \ldots, m$$

$$T_j = \ln E_j^2, \qquad j = 1, 2, \ldots, n$$

Also get S_0 and T_0 using all data

ln: stable the variance, make the statistic more normal

Miller Jackknife Test

Set

$$A_i = mS_0 - (m-1)S_i, \qquad i = 1, 2, \ldots, m$$

$$B_j = nT_0 - (n-1)T_j, \qquad j = 1, 2, \ldots, n$$

$$V_1 = \sum_{i=1}^{m} \frac{(A_i - \overline{A})^2}{m(m-1)} (\text{estimate var}\bar{A}), \qquad V_2 = \sum_{i=1}^{m} \frac{(B_j - \overline{B})^2}{n(n-1)} (\text{estimate var}$$

$$Q = \frac{\overline{A} - \overline{B}}{\sqrt{V_1 + V_2}}$$

Miller Jackknife Test

- Write null as

$$H_0 : \gamma^2 = 1$$

- One-tail alternative

$$H_1 : \gamma^2 > 1$$

- Reject H_0 if $Q \geq z_\alpha$
- Table A.1 or `qnorm`

Miller Jackknife Test

- One-tail alternative

$$H_1 : \gamma^2 < 1$$

- Reject H_0 if $Q \leq -z_\alpha$
- One-tail alternative

$$H_1 : \gamma^2 \neq 1$$

- Reject H_0 if $Q \leq -z_{\alpha/2}$ or $Q \geq z_{\alpha/2}$

Miller Jackknife Test

- Asymptotically distribution free
    - F-test: extremely nonrobust
- These are approximately α significance tests
- Asymptotic tests - good as sample size $\rightarrow \infty$
- Not a rank test
- No ties

Miller Jackknife Test

- Estimate of ratio:

$$\tilde{\gamma}^2 = e^{\{\overline{A} - \overline{B}\}}$$

- $(1 - \alpha)$ confidence intervals (approximately)
- Two-sided

$$\gamma_U^2 = e^{\{\overline{A} - \overline{B} + z_{\alpha/2}\sqrt{V_1 + V_2}\}}$$
$$\gamma_L^2 = e^{\{\overline{A} - \overline{B} - z_{\alpha/2}\sqrt{V_1 + V_2}\}}$$

- One-sided (lower)

$$\gamma_L^2 = e^{\{\overline{A} - \overline{B} - z_{\alpha}\sqrt{V_1 + V_2}\}}$$

- One-sided (upper)

$$\gamma_U^2 = e^{\{\overline{A} - \overline{B} + z_{\alpha}\sqrt{V_1 + V_2}\}}$$

Lepage Test

- Test for either scale or location differences
- Null:

$$H_0 : F(t) = G(t) \text{ for all } t$$

- Alternative:

$$H_1 : \theta_1 \neq \theta_2 \text{ and/or } \eta_1 \neq \eta_2$$

- Rank test
- Large sample approximation
- Section 5.3 (skip)

Kolmogorov - Smirnov Test

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

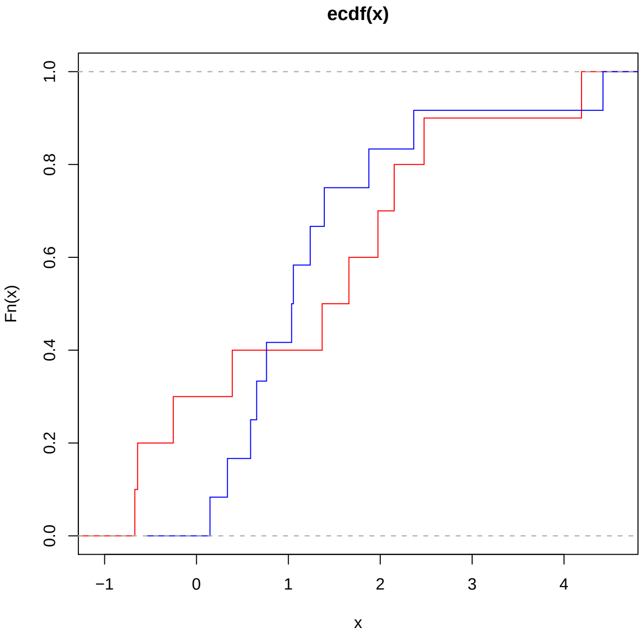
Kolmogorov - Smirnov Test

Set

$$F_m(t) = \frac{\text{number of sample } X\text{'s } \leq t}{m}$$

$$G_n(t) = \frac{\text{number of sample } Y\text{'s } \leq t}{n}$$

$F_m(t)$ and $G_n(t)$ are empirical distribution functions which are non-decreasing, step functions



ecdf(x)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
−1
0
1
2
3
4
x

Kolmogorov - Smirnov Test

$$J = \frac{mn}{d} \max_{-\infty<t<\infty} \{|F_m(t) - G_n(t)|\}$$

where $d =$ greatest common divisor of n and m. Or

$$K_{m,n}(J^*) = \sqrt{\frac{mn}{m+n}} D_{m,n} = \sqrt{\frac{mn}{m+n}} \max_{-\infty<t<\infty} |F_m(t) - G_n(t)|$$

- Order the $m + n$ values as $Z_{(1)}, Z_{(2)}, \ldots, Z_{(m+n)}$
- Sufficient to consider these (finitely many) differences

$$D_{m,n} = \max_i |F_m(Z_{(i)}) - G_n(Z_{(i)})|$$

Kolmogorov - Smirnov Test

- Null:

$$H_0 : F(t) = G(t) \text{ for all } t$$

- Alternative:

$$H_1 : F(t) \neq G(t) \text{ for at least one } t$$

- Reject H_0 if $J \geq j_\alpha$
- Table A.10 (X smaller)

Kolmogorov - Smirnov Test

- Large sample approximation $J^* = \frac{Jd}{\sqrt{mnN}}$, or in fact

$$K_{m,n} = \sqrt{\frac{mn}{m+n}} D_{m,n}$$

- Reject H_0 if $J^* \geq q^*_\alpha$
- $P(J^* \geq q^*_\alpha) = \alpha$
- Limiting distribution: *Kolmogorov distribution* (cf. (5.74) P179)
- Table A.11
- Not normal
- No ties

Kolmogorov - Smirnov Test

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

(Skip 5.5)