ANOVA

Christof Seiler

Stanford University, Spring 2016, STATS 205

Overview

Before:

- One-Sample, Two-Sample Problems

Today:

- ANOVA (two or more samples)

One-Way ANOVA

- Determine the effect of a single factor A on a response over a specific population
- Assume A consists of k levels or treatmens
- In a completely randomize design, n subjects are randomly selected from the reference population
- n_j randomly assigned to treatment $j = 1, \ldots, k$
- Let ith be the response in the jth treatment denoted by Y_{ij}, $i = 1, \ldots, n_j$
- **Assumptions:**
  - Responses are independent of another
  - Distribution among levels differ by at most shifts in location

One-Way ANOVA

- **Data**:

Treatment			
1	2	. . .	k
Y_{11}	Y_{12}	. . .	Y_{1k}
Y_{21}	Y_{22}	. . .	Y_{2k}
$\vdots$	$\vdots$		$\vdots$
Y_{n_j1}	Y_{n_j2}	. . .	Y_{n_jk}

- **Model**

$$Y_{ij} = \theta + \mu_i + e_{ij}, \qquad i = 1, \ldots, n_j, \qquad j = 1, \ldots, k$$

with

- overall median θ
- μ_i is the treatment effect
- e_{ij} samples from continuous distribution with median 0

One-Way ANOVA

- The null hypothesis

  $$H_0 : \mu_1 = \cdots = \mu_k$$

  underlying distributions $F_1, \ldots, F_k$ are connected through the relationship

  $$F_j(t) = F(t - \mu_j), -\infty < t < \infty$$

- The alternative is that at least two of the treatment are not equal

  $$H_A : \mu_1, \ldots, \mu_k \text{ not all equal}$$

Kruskal-Wallis Test

- Total sample size $n = \sum_{j=1}^{k} n_j$
- Rank R_{ij} of response Y_{ij} among all n observations; ranking done without knowledge of treatment
- Let $R_{.j}$ denotes the average of the ranks for sample j
- The **Kruskal-Wallis test statistic**

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

- Asymptotically χ^2 distributed with $k-1$ degrees of freedom

Kruskal-Wallis Test

- Motivation for the test
- The **Kruskal-Wallis test statistic**

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

- The average rank sample $j = 1, \ldots, k$ is

$$\mathrm{E}_{\mathrm{H}_0}(R_{.j}) = \mathrm{E}_{\mathrm{H}_0} \left(\frac{1}{n_j} \sum_{i=1}^{n_j} R_{ij} \right) = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathrm{E}_{\mathrm{H}_0}(R_{ij}) = \frac{n+1}{2}$$

Kruskal-Wallis Test (Example)

- Efficiency self-clearing mechanism of respiratory tract that conducts air into the lungs from the rate of dust in the three groups:
  - normal subjects,
  - subjects with obstructive airway disease, and
  - subjects with asbestosis
- Responses are the clearance half-times of the subjects
- Sample sizes: $n_1 = n_3 = 5$ and $n_2 = 4$

Kruskal-Wallis Test (Example)

```r
normal = c(2.9,3.0,2.5,2.6,3.2)
obstruct = c(3.8,2.7,4.0,2.4)
asbestosis = c(2.8,3.4,3.7,2.2,2.0)
x = c(normal,obstruct,asbestosis)
g = c(rep(1,5),rep(2,4),rep(3,5))
test = kruskal.test(x,g)
test$statistic
```

```
## Kruskal-Wallis chi-squared
##                  0.7714286
```

```r
test$p.value
```

```
## [1] 0.6799648
```

Two-Way ANOVA

- Same as before but now we have blocks:

	Treatment			
Blocks	1	2	…	k
1	Y_{111}	Y_{121}	…	Y_{1k1}
	$\vdots$	$\vdots$		$\vdots$
	$Y_{11c_{11}}$	$Y_{12c_{12}}$	…	$Y_{1kc_{1k}}$
2	Y_{211}	Y_{221}	…	Y_{2k1}
	$\vdots$	$\vdots$		$\vdots$
	$Y_{21c_{21}}$	$Y_{22_{c_{22}}}$	…	$Y_{2kc_{2k}}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$

- The **Friedman test** is analog to the Kruskal-Wallis test

Median Polish

- For special case of no repetitions (one observation per block/treatment cell)
- This may be the actual data we observe or someone may have summarized all the entries in each cell with a single number
- **Data**:

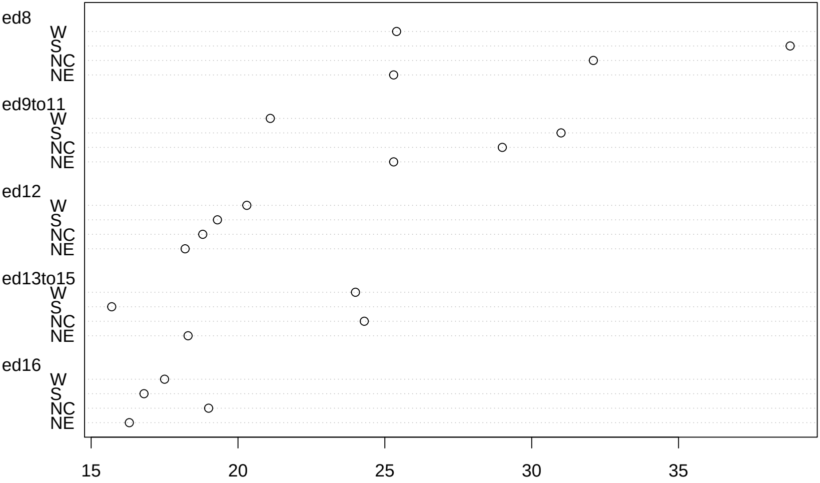
		j	
i	1	$\dots$	J
1	Y_{11}	$\dots$	Y_{1J}
$\vdots$	$\vdots$		$\vdots$
I	Y_{I1}	$\dots$	Y_{IJ}

Median Polish

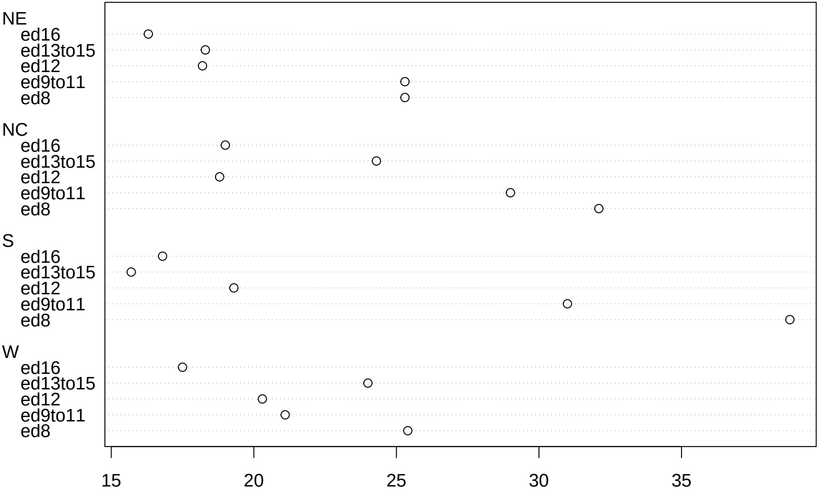
- Infant mortality rates in the United States 1964-1966 by region and father's eduction
- Cell entires are number of deaths (under one year old) per 1000 live births

```
##      ed8 ed9to11 ed12 ed13to15 ed16
## NE 25.3    25.3 18.2     18.3 16.3
## NC 32.1    29.0 18.8     24.3 19.0
## S  38.8    31.0 19.3     15.7 16.8
## W  25.4    21.1 20.3     24.0 17.5
```

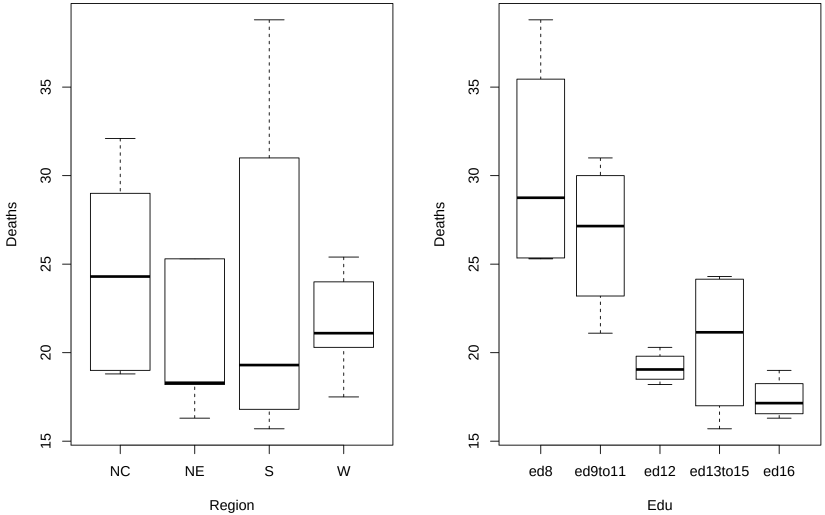
Median Polish – Cleveland Dot Plot



Median Polish – Cleveland Dot Plot



Median Polish



Median Polish

- **Additive model**:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

- Overall typical value μ
- Row effect α_i
- Column effect β_j
- Random fluctuation ϵ_{ij}

Median Polish

- Compute overall typical value μ

```
mu = median(as.matrix(df)); mu; df
```

```
## [1] 20.7
```

```
##      ed8 ed9to11 ed12 ed13to15 ed16
## NE 25.3    25.3 18.2     18.3 16.3
## NC 32.1    29.0 18.8     24.3 19.0
## S  38.8    31.0 19.3     15.7 16.8
## W  25.4    21.1 20.3     24.0 17.5
```

Median Polish

- Compute the row medians

```
df = as.matrix(df) - mu
rowMedian = apply(df,1,median)
df = df - rowMedian; df
```

```
##     ed8 ed9to11 ed12 ed13to15 ed16
## NE  7.0     7.0 -0.1      0.0 -2.0
## NC  7.8     4.7 -5.5      0.0 -5.3
## S  19.5    11.7  0.0     -3.6 -2.5
## W   4.3     0.0 -0.8      2.9 -3.6
```

Median Polish

- Add row median to residual table

```
df = cbind(roweff=c(rowMedian),df); df
```

```
##      roweff  ed8 ed9to11 ed12 ed13to15 ed16
## NE     -2.4  7.0     7.0 -0.1      0.0 -2.0
## NC      3.6  7.8     4.7 -5.5      0.0 -5.3
## S      -1.4 19.5    11.7  0.0     -3.6 -2.5
## W       0.4  4.3     0.0 -0.8      2.9 -3.6
```

Median Polish

- Add and overall value to residual table

```
df = rbind(coleff=rep(0,6),df); df[1,1] = mu; df
```

```
##          roweff  ed8 ed9to11 ed12 ed13to15 ed16
## coleff     20.7  0.0     0.0  0.0      0.0  0.0
## NE         -2.4  7.0     7.0 -0.1      0.0 -2.0
## NC          3.6  7.8     4.7 -5.5      0.0 -5.3
## S          -1.4 19.5    11.7  0.0     -3.6 -2.5
## W           0.4  4.3     0.0 -0.8      2.9 -3.6
```

Median Polish

- Compute column median

```
colMedian = apply(df[2:5,],2,median); colMedian
```

```
##   roweff      ed8  ed9to11     ed12 ed13to15     ed16
##    -0.50     7.40     5.85    -0.45     0.00    -3.05
```

Median Polish

- Create new residual table from column medians

```
df[1,] = df[1,]+colMedian
df[2:5,] = sweep(df[2:5,],2,colMedian); df
```

```
##        roweff  ed8 ed9to11  ed12 ed13to15  ed16
## coleff   20.2  7.4    5.85 -0.45      0.0 -3.05
## NE       -1.9 -0.4    1.15  0.35      0.0  1.05
## NC        4.1  0.4   -1.15 -5.05      0.0 -2.25
## S        -0.9 12.1    5.85  0.45     -3.6  0.55
## W         0.9 -3.1   -5.85 -0.35      2.9 -0.55
```

Median Polish

- Second iteration: Add row effects to left margin and subtract from residuals

```
rowMedian = apply(df[,2:6],1,median); rowMedian
```

```
## coleff      NE      NC       S       W
##   0.00    0.35   -1.15    0.55   -0.55
```

```
df[,1] = df[,1]+rowMedian
df[,2:6] = sweep(df[,2:6],1,rowMedian); df
```

```
##          roweff   ed8 ed9to11  ed12 ed13to15  ed16
## coleff    20.20  7.40    5.85 -0.45     0.00 -3.05
## NE        -1.55 -0.75    0.80  0.00    -0.35  0.70
## NC         2.95  1.55    0.00 -3.90     1.15 -1.10
## S         -0.35 11.55    5.30 -0.10    -4.15  0.00
## W          0.35 -2.55   -5.30  0.20     3.45  0.00
```

Median Polish

- Second iteration: Add column effects to top margin and subtract from residuals

```
colMedian = apply(df[2:5,],2,median); colMedian
```

```
##   roweff      ed8   ed9to11      ed12 ed13to15      ed16
##     0.00     0.40      0.40     -0.05     0.40      0.00
```

```
df[1,] = df[1,]+colMedian
df[2:5,] = sweep(df[2:5,],2,colMedian); df
```

```
##        roweff   ed8 ed9to11  ed12 ed13to15  ed16
## coleff  20.20  7.80    6.25 -0.50     0.40 -3.05
## NE      -1.55 -1.15    0.40  0.05    -0.75  0.70
## NC       2.95  1.15   -0.40 -3.85     0.75 -1.10
## S       -0.35 11.15    4.90 -0.05    -4.55  0.00
## W        0.35 -2.95   -5.70  0.25     3.05  0.00
```

Median Polish

```
##          roweff   ed8 ed9to11   ed12 ed13to15   ed16
## coleff  20.20  7.80    6.25 -0.50     0.40 -3.05
## NE      -1.55 -1.15    0.40  0.05    -0.75  0.70
## NC       2.95  1.15   -0.40 -3.85     0.75 -1.10
## S       -0.35 11.15    4.90 -0.05    -4.55  0.00
## W        0.35 -2.95   -5.70  0.25     3.05  0.00
```

- Infant mortality rates are highest in North Central region and lowest in Northeast
- The education of the father is a stronger factor in distinguishing among these rates than geography
- In particular, completion of high school appears to exert the greatest single influence in reducing the mortality rates among infant offspring

Median Polish

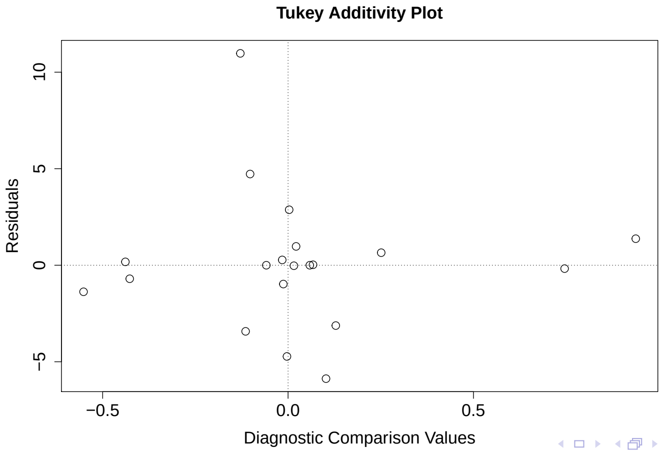
```
##         roweff   ed8 ed9to11  ed12 ed13to15  ed16
## coleff  20.20  7.80    6.25 -0.50     0.40 -3.05
## NE      -1.55 -1.15    0.40  0.05    -0.75  0.70
## NC       2.95  1.15   -0.40 -3.85     0.75 -1.10
## S       -0.35 11.15    4.90 -0.05    -4.55  0.00
## W        0.35 -2.95   -5.70  0.25     3.05  0.00
```

- The residual of 11.15 for the least educated fathers in the South calls for a closer look

Tukey Additivity Plot

- Comparison value $\alpha_i \beta_j / \mu$

```
## 1: 47.1
## 2: 42.9
## 3: 42.45
```



Tukey Additivity Plot

- Why $\alpha_i\beta_j/\mu$ against ϵ_{ij}?
- To show why this makes sense, we start by asking:
- Can we find a power transformation of the data so that model

  $$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

  will adequately summarize the transformed data?
- If so, then equation

  $$y_{ij}^p = m + a_i + b_j + r_{ij}$$

  will hold for some value p
- If fit exact, then

  $$y_{ij}^p = m + a_i + b_j$$

  or

  $$y_{ij} = (m + a_i + b_j)^{1/p}$$

Tukey Additivity Plot

- Compare

$$y_{ij} = (m + a_i + b_j)^{1/p}$$

  to simple additive model, we use a second-order approximation
- Rewrite

$$y_{ij} = m^{1/p}\left(1 + \frac{a_i}{m} + \frac{b_j}{m}\right)^{1/p}$$

- Taylor expand second factor $(1+t)^{1/p}$

$$y_{ij} \approx m^{1/p}\left(1 + \frac{1}{p}\left(\frac{a_i}{m} + \frac{b_j}{m}\right) + \frac{1-p}{2p^2}\left(\frac{a_i}{m} + \frac{b_j}{m}\right)^2\right)$$

Tukey Additivity Plot

- Arrange terms in this expansion into four groups, terms that depend
  - on neither i nor j
  - only on i
  - only on j
  - both i and j
- In simplified notation:

$$y_{ij} \approx D\left(1 + \frac{A_i}{D} + \frac{B_j}{D} + \frac{C_{ij}}{D}\right) \qquad y_{ij} \approx D + A_i + B_j + C_{ij}$$

$$D = m^{1/p} \qquad \frac{A_i}{D} = \frac{1}{p}\frac{a_i}{m} + \frac{1-p}{2p^2}\frac{a_i^2}{m^2} \qquad \frac{B_j}{D} = \frac{1}{p}\frac{b_j}{m} + \frac{1-p}{2p^2}\frac{b_j^2}{m^2}$$

$$\frac{C_{ij}}{D} = \frac{1-p}{2p^2}\frac{2a_i b_j}{m^2} = \frac{1-p}{p^2}\frac{a_i}{m}\frac{b_j}{m}$$

Tukey Additivity Plot

- Through Taylor expansion, we obtained

  $$y_{ij} \approx D + A_i + B_j + C_{ij}$$

  which is now a function of p
- Examine term when a_i/m and b_j/m are close to 0 (which means that row and column effects are much smaller than common value)
- With this assumption expressions a_i^2/m^2, b_j^2/m^2, and $a_i b_j/m^2$ can be ignored

  $$\frac{A_i}{D}\frac{B_j}{D} \approx \frac{1}{p^2}\frac{a_i}{m}\frac{b_j}{m}$$

- Using this

  $$\frac{C_{ij}}{D} \approx (1-p)\frac{A_i}{D}\frac{B_j}{D}$$

- Using this

  $$y_{ij} \approx D\left(1 + \frac{A_i}{D} + \frac{B_j}{D} + (1-p)\frac{A_i}{D}\frac{B_j}{D}\right)$$

Tukey Additivity Plot

- Rewrite
$$y_{ij} \approx D + A_i + B_j + (1-p)\frac{A_i B_j}{D}$$
- And we conclude that if y_{ij}^p is approximated by an additive model, then, to a second-order approximation, y_{ij} is given by the above approximation
- For the diagnostic plot
$$y_{ij} - D - A_i - B_j \approx (1-p)\frac{A_i B_j}{D}$$
- If $R_{ij} = y_{ij} - D - A_i - B_j$ are residuals, then
$$R_{ij} \approx (1-p)\frac{A_i B_j}{D}$$

References

- Hoaglin, Mosteller, and Tukey (1983). Understanding Robust and Exploratory Data Analysis
- Manuel Gimond Course Notes: `http://mgimond.github.io/ES218/Week11a.html`