# ANOVA Pairwise Comparisons

Stat 120

May 18 2022

# Post-ANOVA

- Inference **AFTER** doing **ANOVA** to compare means for several groups:
  - Confidence interval for a single mean
  - Confidence interval for a difference in two means
  - Pairwise t-test for a difference in two means

# ANOVA for Difference in Means

- **Data:** Random samples of size $n_1, n_2, \cdots, n_k$ from each of k populations (or groups)
- **Summary statistics:**
  - Sample mean for each group
  - Std. dev. for each group
  - Mean and std. dev. for all values

# ANOVA for Difference in Means

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_a : \text{at least one } \mu_i \text{ is different}$$

- **Conditions:** Similar variability AND either sample sizes in each group are large (each $n_i \geq 30$) OR the data are relatively normally distributed

# Cuckoo Birds

- Cuckoo birds lay their eggs in the nests of other birds
- When the cuckoo baby hatches, it kicks out all the original eggs/babies
- If the cuckoo is lucky, the mother will raise the cuckoo as if it were her own
    - Do cuckoo bird eggs found in nests of different species differ in size?

Cuckoo bird in nest

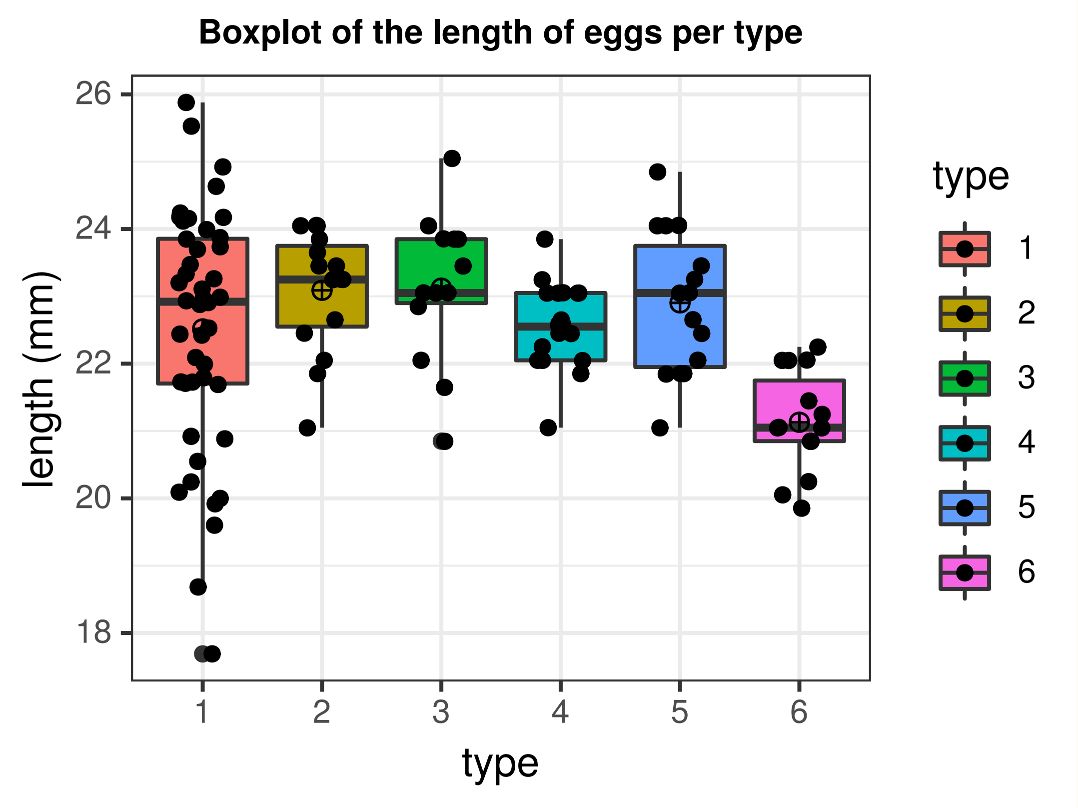
# Cuckoo Dataset

- **cuckoo** dataset contains information on 120 Cuckoo eggs, obtained from randomly selected "foster" nests.
- researchers have measured the `length` (in mm) and established the `type` (species) of foster parent.

- `type=1`: Meadow pipit
- `type=2`: Tree pipit
- `type=3`: Dunnock
- `type=4`: European robin
- `type=5`: White wagtail
- `type=6`: Eurasian wren

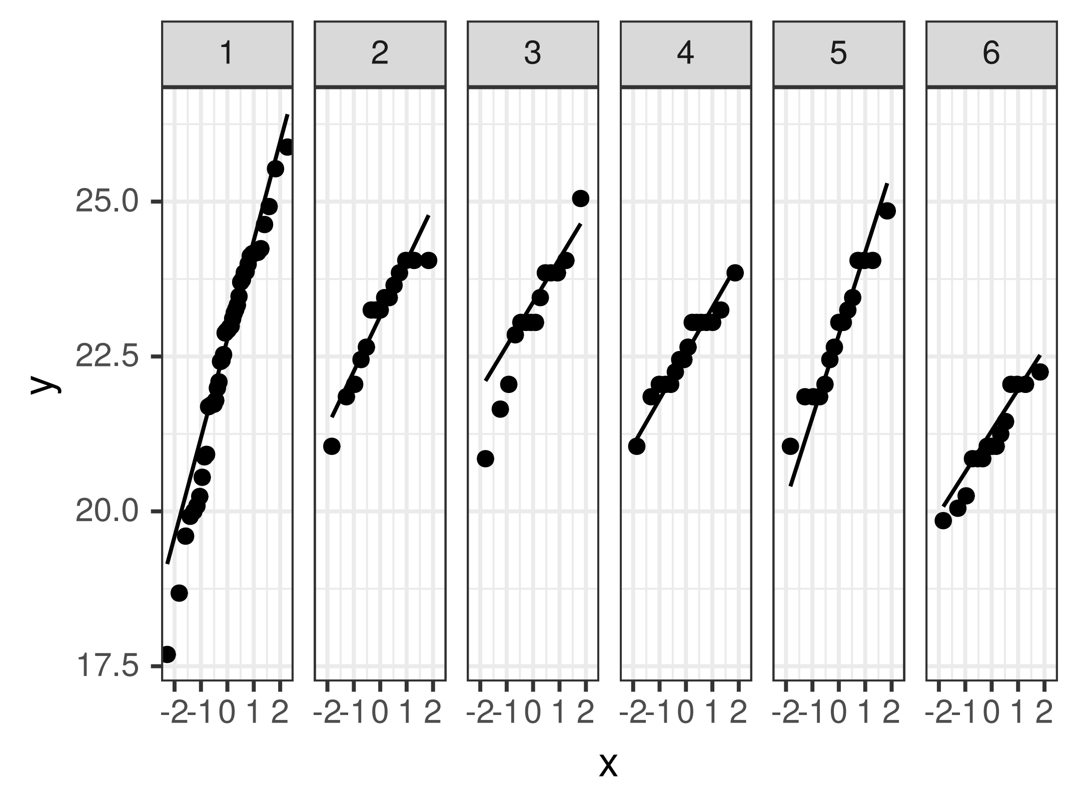
# Side-by-side Boxplot (1a)

```
cuckoo %>%
  ggplot(aes(x=type,y=length,fill=type)) +
  theme_bw() +
  geom_boxplot() +
  geom_jitter(width = 0.2) +
  labs(title ="Boxplot of the length of eggs per type",
       y = "length (mm)",
       x = "type") +
  stat_summary(fun=mean, geom="point", shape=10,
               size=2, color="black", fill="black") +
  theme(plot.title = element_text(hjust=0.5, size=9,
                                  face='bold'))
```



# Approximate normality in groups (1b)

```
cuckoo %>%
  ggplot(aes(sample=length)) +
  geom_qq() +
  geom_qq_line() +
  facet_grid(~type) +
  theme_bw()
```



# Fitting ANOVA (1c)

```
fit_anova <- aov(length~type, cuckoo)
summary(fit_anova)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)    
type          5  41.12   8.223   4.713 0.000602 ***
Residuals   113 197.18   1.745                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the **p-value** is very small, at the significance level of $5\%$, we have sufficient evidence to conclude that the mean egg length for at least one bird type is **different** from the mean egg length in at least one other bird type.

But which of the species are different?

# Inference after ANOVA

- Compute a CI for any $\mu_i$
- Compute a CI for $\mu_i - \mu_j$
- Pairwise t-test for difference in means

$$H_0 : \mu_i = \mu_j \text{ vs. } H_a : \mu_i \neq \mu_j$$

- Use the usual procedures except:
  - Estimate any $\sigma$ with the pooled standard deviation: $\sqrt{MSE}$
  - Use the error degrees of freedom for any t-values

# Inference after ANOVA

Compute a CI for any $\mu_i$

- The usual procedure:

$$\bar{x}_i \pm t^* \frac{s_i}{\sqrt{n_i}}$$

- **BUT** after **ANOVA**, estimate any $\sigma$ with the pooled standard deviation:

$$\bar{x}_i \pm t^* \frac{\sqrt{MSE}}{\sqrt{n_i}}$$

- the corresponding `df=n-k`

# Cuckoo Eggs (1d)

Find a 95% confidence interval for the mean cuckoo egg length in `European robin` nests (Type = 4).

```
MSE <- 1.745
summary(fit_anova)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)    
type          5  41.12   8.223   4.713 0.000602 ***
Residuals   113 197.18   1.745                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
```

| Type | mean | sd | n |
|---|---|---|---|
| 1 | 22.51636 | 1.8058292 | 44 |
| 2 | 23.09000 | 0.9014274 | 15 |
| 3 | 23.12143 | 1.0687365 | 14 |
| 4 | 22.57500 | 0.6845923 | 16 |
| 5 | 22.90333 | 1.0676186 | 15 |
| 6 | 21.13000 | 0.7437357 | 15 |

$$\bar{x}_i \pm t^* \frac{\sqrt{MSE}}{\sqrt{n_i}}, \text{ df} = \text{n-k}$$

# Inferences after ANOVA

## Compute a CI for any $\mu_i - \mu_j$

- The usual procedure:

$$(\bar{x}_i - \bar{x}_j) \pm t^* \sqrt{\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}}$$

- BUT after ANOVA, estimate any $\sigma$ with the pooled standard deviation:

$$(\bar{x}_i - \bar{x}_j) \pm t^* \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

- the corresponding `df=n-k`

# Cuckoo Eggs (1e)

Find a 95% CI for the difference in mean egg length between `European robin`(type = 4) and `Eurasian wren`(type = 6) nests.

```
              Df Sum Sq Mean Sq F value   Pr(>F)    
type           5  41.12   8.223   4.713 0.000602 ***
Residuals    113 197.18   1.745                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

| Type | mean | sd | n |
|---|---|---|---|
| 1 | 22.51636 | 1.8058292 | 44 |
| 2 | 23.09000 | 0.9014274 | 15 |
| 3 | 23.12143 | 1.0687365 | 14 |
| 4 | 22.57500 | 0.6845923 | 16 |
| 5 | 22.90333 | 1.0676186 | 15 |
| 6 | 21.13000 | 0.7437357 | 15 |

$$(22.575 - 21.130) \pm 1.981 \cdot \sqrt{1.745\left(\frac{1}{16} + \frac{1}{15}\right)} = (0.50, 2.39)$$

```
(stat[4,2] - stat[6,2]) + c(-1,1)* (qt(1-0.05/2, df=113))* sqrt(MSE*(1/stat[4,4] + 1/stat[6,4]))
[1] 0.5044174 2.3855826
```

Why is it important that the interval contains only positive values?

# Cuckoo Eggs (1f)

Find a 95% CI for the difference in mean egg length between `Dunnock` (type = 3) and `European robin`(type = 4) nests.

```
              Df Sum Sq Mean Sq F value   Pr(>F)    
type           5  41.12   8.223   4.713 0.000602 ***
Residuals    113 197.18   1.745                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

| Type | mean | sd | n |
|---|---|---|---|
| 1 | 22.51636 | 1.8058292 | 44 |
| 2 | 23.09000 | 0.9014274 | 15 |
| 3 | 23.12143 | 1.0687365 | 14 |
| 4 | 22.57500 | 0.6845923 | 16 |
| 5 | 22.90333 | 1.0676186 | 15 |
| 6 | 21.13000 | 0.7437357 | 15 |

```
(stat[3,2] - stat[4,2]) + c(-1,1)* (qt(1-0.05/2, df=113))*sqrt(MSE*(1/stat[3,4] + 1/stat[4,4]))
[1] -0.4113351  1.5041922
```

$$(23.121 - 22.575) \pm 1.981 \cdot \sqrt{1.188\left(\frac{1}{14} + \frac{1}{16}\right)} = (-0.41, 1.50)$$

What does it mean if the interval contains 0?

# Mutiple Comparisons

- Often, doing pairwise comparisons after ANOVA involves many tests
  - e.g. $k$ groups/categories,then we have $\frac{k(k-1)}{2}$ comparisons
  - $k = 6$ bird species then 15 pairwise tests.
- If each test has an $\alpha$ chance of a Type I error (finding a difference between a pair that aren't different), the overall Type I error rate can be much higher.
  - Use a smaller $\alpha$ for each pairwise test (Bonferroni)
  - $\alpha^* = \frac{\alpha}{k}$
  - e.g $\alpha = 0.05$ and $k = 6$, then $\alpha^* = 0.05/6 = 0.0083$

# Cuckoo Eggs (1g)

- Which means are "different" at a $5\%$ significance level?

```
pairwise.t.test(cuckoo$length, cuckoo$type, p.adjust.method =  "bonferroni")
```

```


    Pairwise comparisons using t tests with pooled SD

data:  cuckoo$length and cuckoo$type

  1      2      3      4      5
2 1.0000 -      -      -      -
3 1.0000 1.0000 -      -      -
4 1.0000 1.0000 1.0000 -      -
5 1.0000 1.0000 1.0000 1.0000 -
6 0.0097 0.0013 0.0014 0.0436 0.0055

P value adjustment method: bonferroni
```

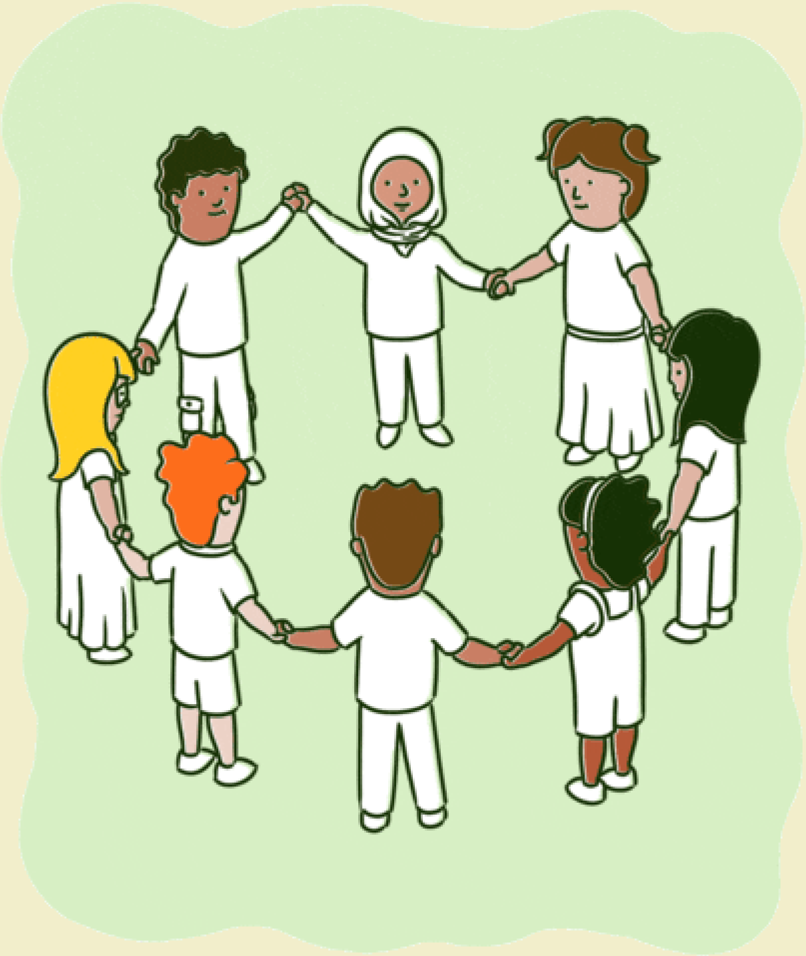
# Multiple Comparisons (1h)

```r
library(multcomp, quietly = TRUE)
fit <- lm(length~type, cuckoo)
mcp <- glht(fit, linfct = mcp(type =
summary(mcp)
```

```

	 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = length ~ type, data = cuckoo)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0  0.57364    0.39495   1.452  0.68900
3 - 1 == 0  0.60506    0.40534   1.493  0.66348
4 - 1 == 0  0.05864    0.38564   0.152  0.99999
5 - 1 == 0  0.38697    0.39495   0.980  0.92128
6 - 1 == 0 -1.38636    0.39495  -3.510  0.00813 **
3 - 2 == 0  0.03143    0.49089   0.064  1.00000
4 - 2 == 0 -0.51500    0.47475  -1.085  0.88349
5 - 2 == 0 -0.18667    0.48235  -0.387  0.99880
6 - 2 == 0 -1.96000    0.48235  -4.063  0.00124 **
4 - 3 == 0 -0.54643    0.48342  -1.130  0.86433
5 - 3 == 0 -0.21810    0.49089  -0.444  0.99768
6 - 3 == 0 -1.99143    0.49089  -4.057  0.00122 **
5 - 4 == 0  0.32833    0.47475   0.692  0.98196
6 - 4 == 0 -1.44500    0.47475  -3.044  0.03291 *
6 - 5 == 0 -1.77333    0.48235  -3.676  0.00469 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

# Your Turn 1

05:00



- Go over to the in class activity file
- Complete the remaining activity