Analysis of Variance (ANOVA)

Sumanta Basu

Comparing more than two means with ANOVA
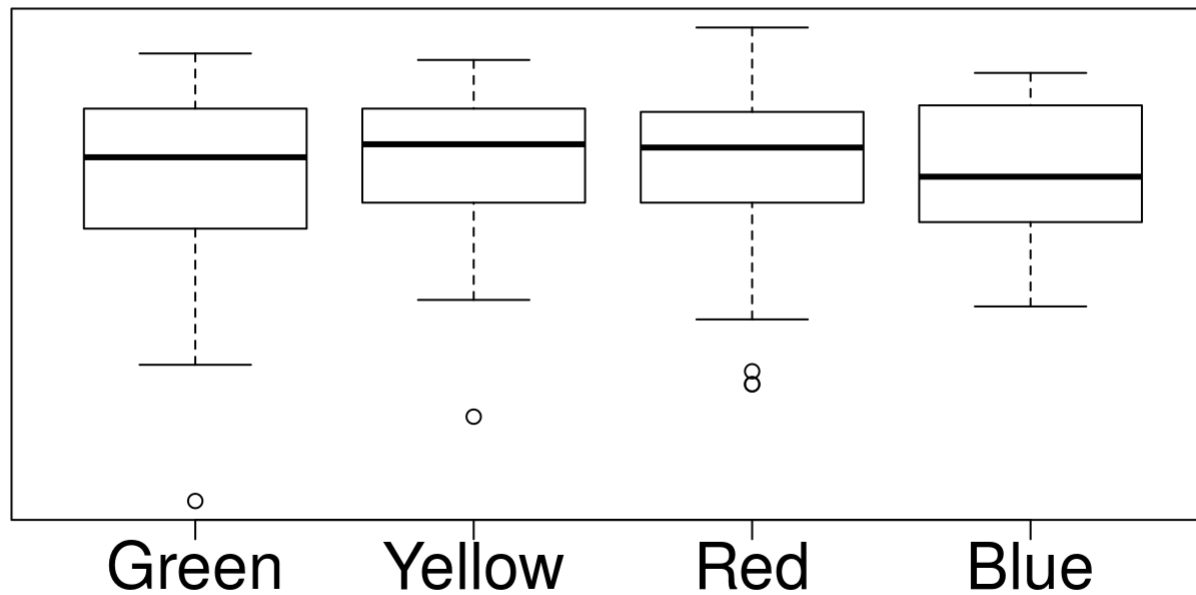
Outline 1

> *How about we compare all possible pairs?*

- *Pro: Can simply use t-tests many times*

- *Con: Type-I error inflates quickly*

- *Correct for multiple hypothesis testing, at the expense of losing power*

Is there a difference between labs?

Prelims score by lab (not this year)

Green Yellow Red Blue

Should we just compare all pairs?

There are $\binom{4}{2} = 6$ pairs: Could do a two-sample test for each:

- $H_0 : \mu_1 = \mu_2$

- $H_0 : \mu_1 = \mu_3$

- $H_0 : \mu_1 = \mu_4$

- $H_0 : \mu_2 = \mu_3$

- $H_0 : \mu_2 = \mu_4$

- $H_0 : \mu_3 = \mu_4$

Not a good idea.

Multiple testing problem

Suppose we do M independent tests at significance level α. Suppose all nulls are true.

$$P(\text{Fail to reject all } H_0 \mid \text{all } H_0\text{'s are true}) = (1 - \alpha)^M$$

Suppose we write up a paper on our findings if we rejected any of the nulls. **What's the probability of writing a paper when we shouldn't?**

$$P(\text{Reject at least 1 } H_0 \mid \text{all } H_0\text{'s are true}) = 1 - (1 - \alpha)^M$$

- E.g. if $\alpha = 0.05$ and $M = 6$, this is 26%.

- if $M = 13$, it's about 50%

- if $M = 30$, it's nearly 80%

Data snooping / fishing

> If you look at the data and then decide which hypothesis to test, your p-value is invalid and you are not actually getting the stated Type I error rate.

It's as if you did some large number of tests M (informally)... so if you chose the "most promising looking" test out of $M = 6$ possibilities, you'd have a 25% of making a Type I error! Even though you used $\alpha = 0.05$.
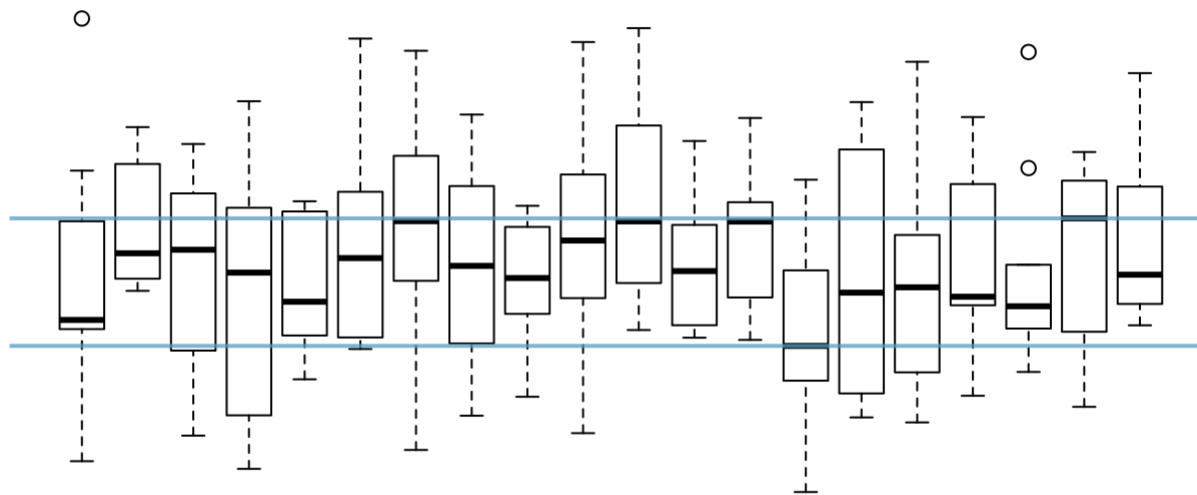
In pictures

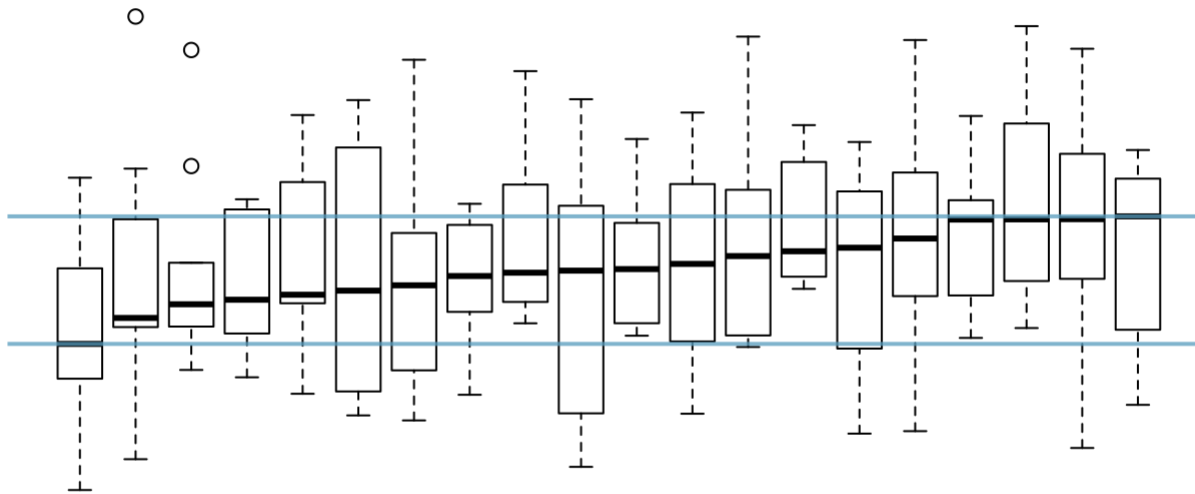Generate 20 groups of size 10 from $N(0, 1)$.
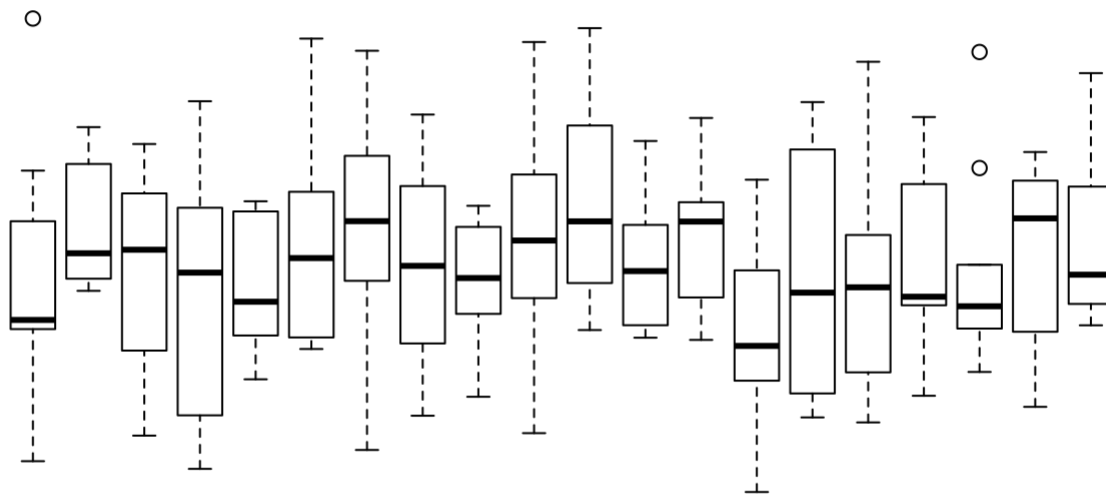
In pictures

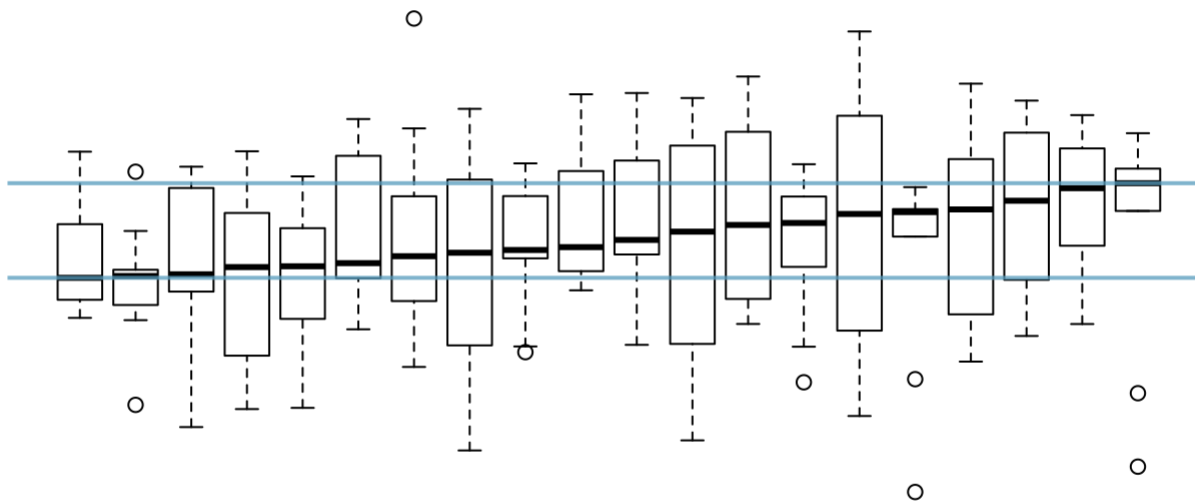A lot of variation in medians… though all from from $N(0, 1)$.

In pictures

Sort these. Lowest dependably far from highest…
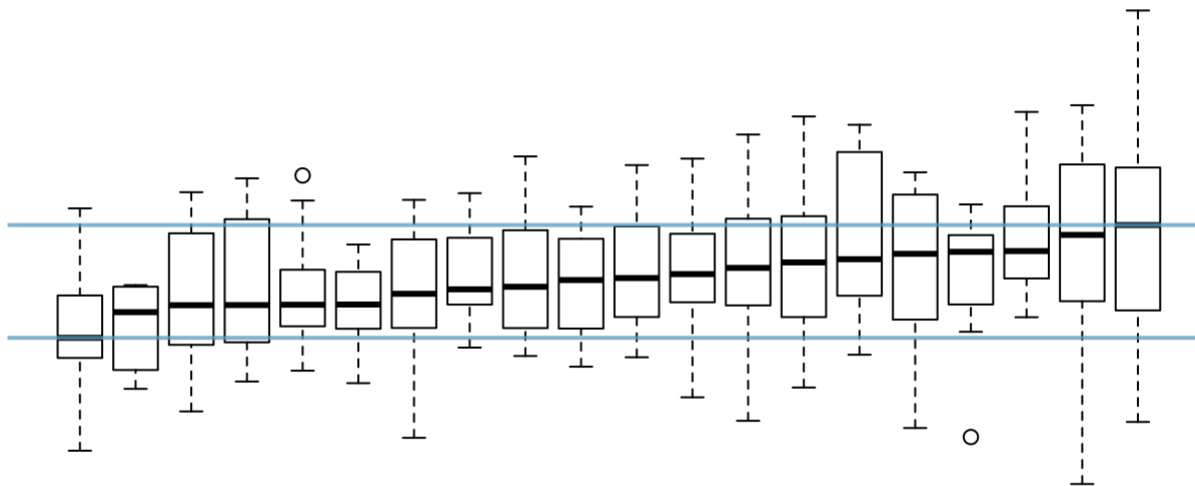
In pictures (repeat)

Sort these. Lowest dependably far from highest…

In pictures (repeat)

Sort these. Lowest dependably far from highest…

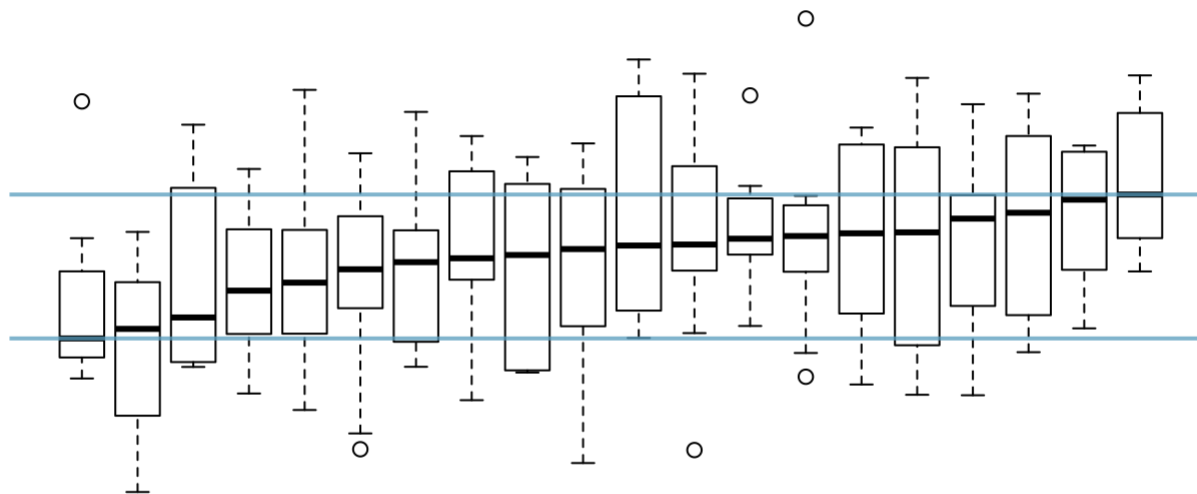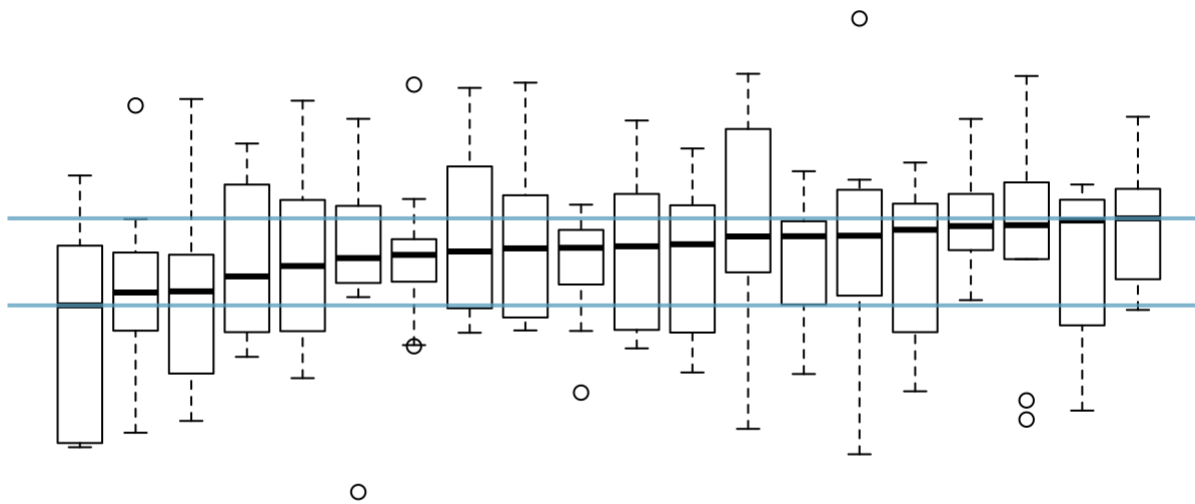# In pictures (repeat)

Sort these. Lowest dependably far from highest…

In pictures (repeat)

Sort these. Lowest dependably far from highest…

In pictures (repeat)

Sort these. Lowest dependably far from highest…
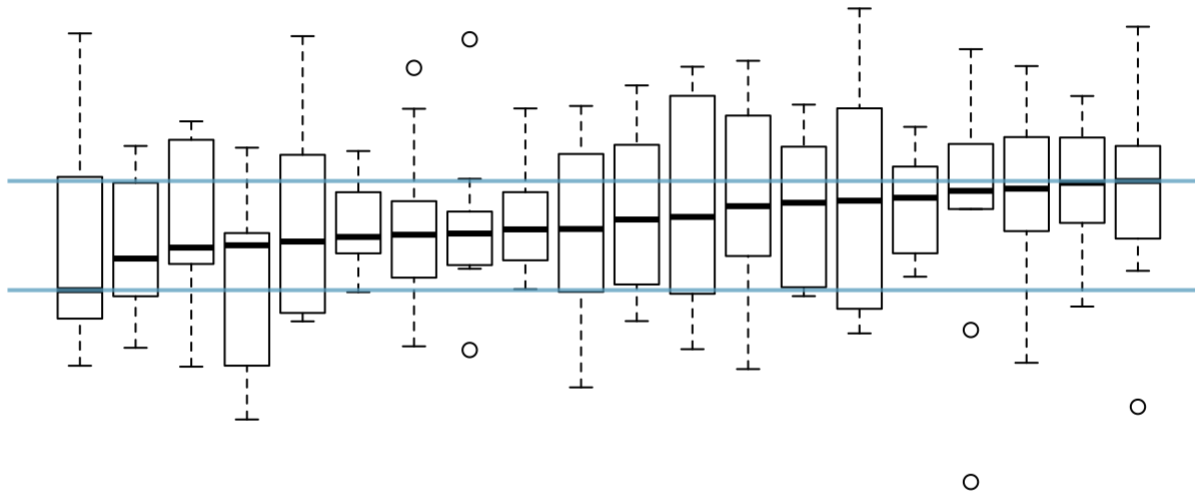
A simple (but suboptimal) fix

> **Bonferroni correction** If you're doing M tests and you want an overall Type I error rate of α, then do your individual tests at level $\alpha^* = \frac{\alpha}{M}$.

- works even if tests are dependent

- for $M = 6$, we should do tests with $\alpha^* = 0.05/6$.

- not a great fix since hard to reject a null - i.e., we **lose power**

- Bonferroni is simple but overly conservative

- What if you are doing $M = 100,000$ as in genomics??

Test all pairs of labs at 0.05/6 level

Compute p-values: e.g., `t.test(red,blue)$p.value`

```
## [1] 0.9529006
```

```
## [1] 0.4620212
```

```
## [1] 0.9529006
```

```
## [1] 0.3863428
```

```
## [1] 0.9583997
```

```
## [1] 0.4285551
```

- none are even close to $0.05/6 \approx 0.008$

So what should we do?

- after correction for multiple testing, doing all $\binom{k}{2}$ tests can have low power

- sometimes we just want to say that **not all means are equal** without saying which pair it is.
- e.g., lab assignment is *associated* with prelim grades
- not trying to say something specific like lab "Red" is different from lab "Blue"

- We wish to test

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_A : \text{at least one mean is different}$$

This is the goal of **ANOVA** (**AN**alysis **O**f **VA**riance)

One-way ANOVA

Outline 2

Assume: independence, normality and equal variance

Model each observation as ``group mean + noise''

Compare variability between groups to variability within groups

How?

- *Assuming no difference in group means, measure how between/within group variability should look*

- *See if between/within variability is substantially higher in your data*

General guideline: be careful about making causal claims from observational data

Contexts for using ANOVA

Observational study

- draw independent samples of size

$$n_1, \ldots, n_k$$

  from $k \geq 2$ populations.

- Goal: Are the means of the k populations all the same?

- allows us to make inferences about **associations** (as in correlations)

Controlled experiments

- draw independent samples then randomly assign subjects to one of k treatment groups

- Goal: Are the means of the k populations all the same?

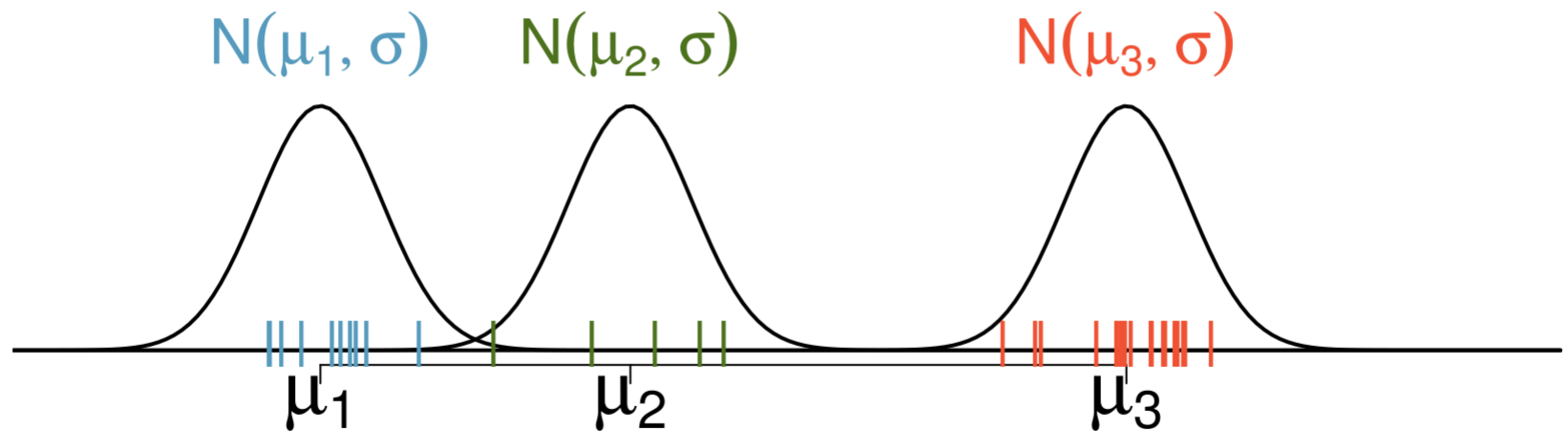- allows us to make **causal** inferences

Question

Example: Prelim scores in labs

Suppose we reject the null of $\mu_{\text{Green}} = \mu_{\text{Red}}$ in favor of $\mu_{\text{Green}} > \mu_{\text{Red}}$, what can we conclude?

(Confounders?)

0. Assumption and ANOVA model

Assumptions to use ANOVA

- independence (both between groups and within)

- normality

- **equal variances**

$$N(\mu_1, \sigma) \quad N(\mu_2, \sigma) \quad N(\mu_3, \sigma)$$

(groups can only differ by mean shifts)

Basic ANOVA: model

Let Y_{ij} = response for unit j from group i (e.g., score of student j in lab i)

All assumptions incorportated by following linear model:

$$Y_{1j} = \mu_1 + \epsilon_{1j} \text{ for } j = 1, \ldots, n_1$$
$$Y_{2j} = \mu_2 + \epsilon_{2j} \text{ for } j = 1, \ldots, n_2$$
$$\vdots$$
$$Y_{kj} = \mu_k + \epsilon_{kj} \text{ for } j = 1, \ldots, n_k$$

where $\epsilon_{ij} \sim N(0, \sigma)$ are all independent.

> Response = Mean of group + Noise

Basic ANOVA: model in words

- each group (potentially) has its own mean.

- response is simply the mean corrupted by noise

- noise is independent, $N(0, \sigma)$ with same σ across all groups

Basic ANOVA: model in pictures

Means

Basic ANOVA: model in pictures

Response = Mean + Noise

1. Null and Alternative Hypotheses

ANOVA

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

What null looks like

Basic ANOVA: model (alternate notation)

$$Y_{1j} = \mu + \alpha_1 + \epsilon_{1j} \text{ for } j = 1, \ldots, n_1$$

$$Y_{2j} = \mu + \alpha_2 + \epsilon_{2j} \text{ for } j = 1, \ldots, n_2$$

$$\vdots$$

$$Y_{kj} = \mu + \alpha_k + \epsilon_{kj} \text{ for } j = 1, \ldots, n_k$$

In this case, we test whether

$$H_0 : \alpha_1 = \cdots = \alpha_k = 0$$

Think of α_i as how group i's mean differs from μ, i.e. $\alpha_i = \mu_i - \mu$

(We require $\sum_{i=1}^{k} \alpha_i = 0$ so μ represents the "overall mean")

One-way fixed effects ANOVA

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \text{ for } j = 1, \ldots, n_i$$

Notation:

- n_i is number of units in group i

- Y_{ij} is response of unit j in group i

- μ is overall mean response value

- α_i is **effect** on mean response due to group i

- ϵ_{ij} random error for unit j of group i

# Fixed effects vs. random effects

Fixed effects ANOVA (our focus) - use when we care about the specific k groups present in sample - e.g., we care about these particular four labs

Random effects ANOVA (in BTRY 6020) - use when these specific k groups are not themselves of intrinsic interest, but rather have been chosen to represent a larger population of potential groups - e.g., company wants to know if employee interviews influenced by who's interviewing. SRS of 5 interviewers (out of, say, 100s) and has each interview 10 applicants
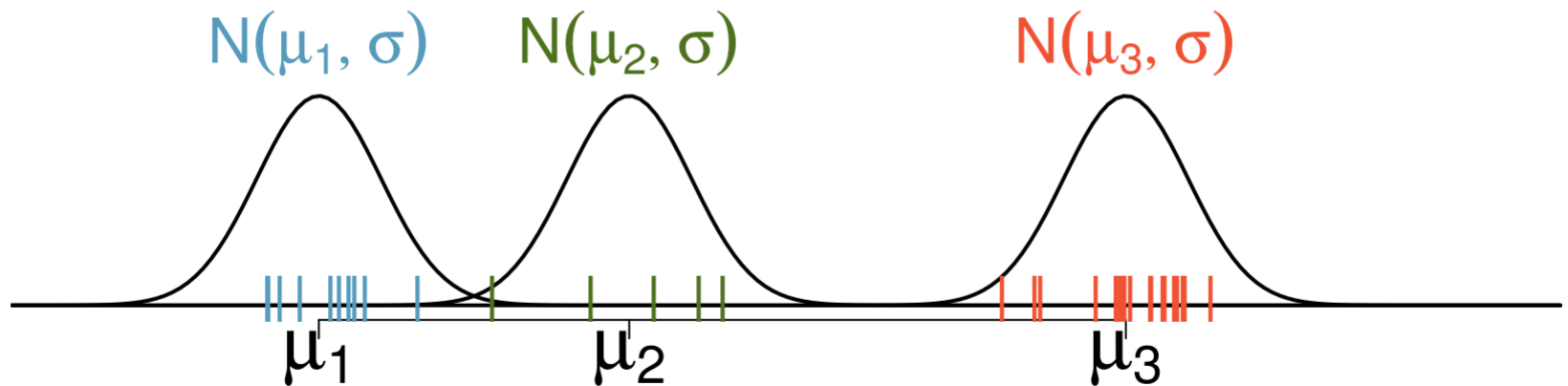
2. Measure within and between group variability
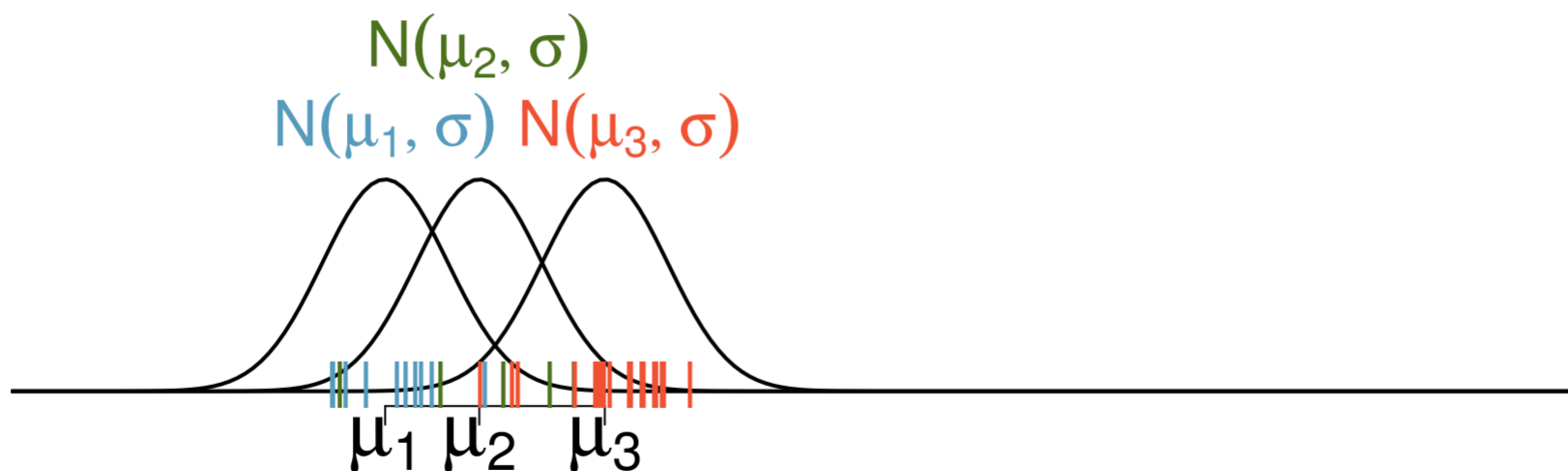
Principal idea behind ANOVA

Compare **variability between** the k estimated group-specific means to a pooled estimate of **within-group variability**
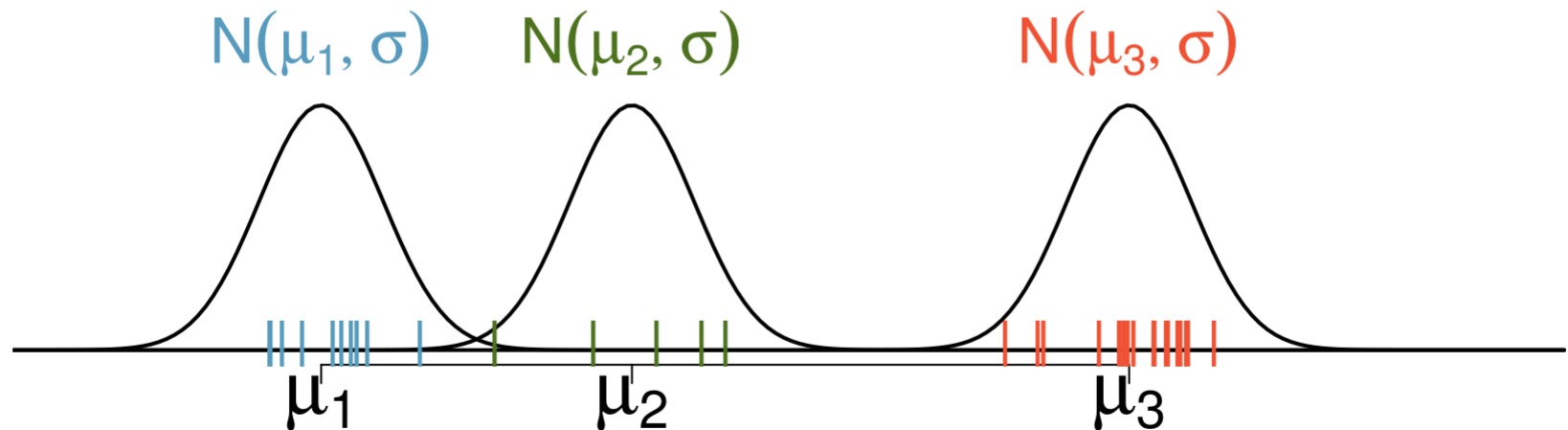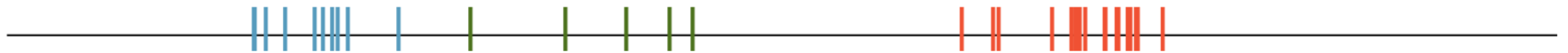
Within group variance smaller than between group variance

$N(\mu_1, \sigma)$ $N(\mu_2, \sigma)$ $N(\mu_3, \sigma)$
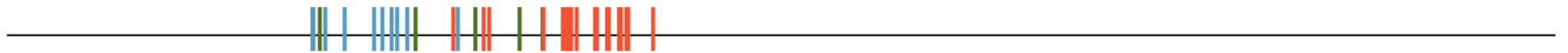
μ_1 μ_2 μ_3

Within group variance bigger than between group variance

Within group variance smaller than between group variance

Within group variance bigger than between group variance

Estimation in fixed effects model

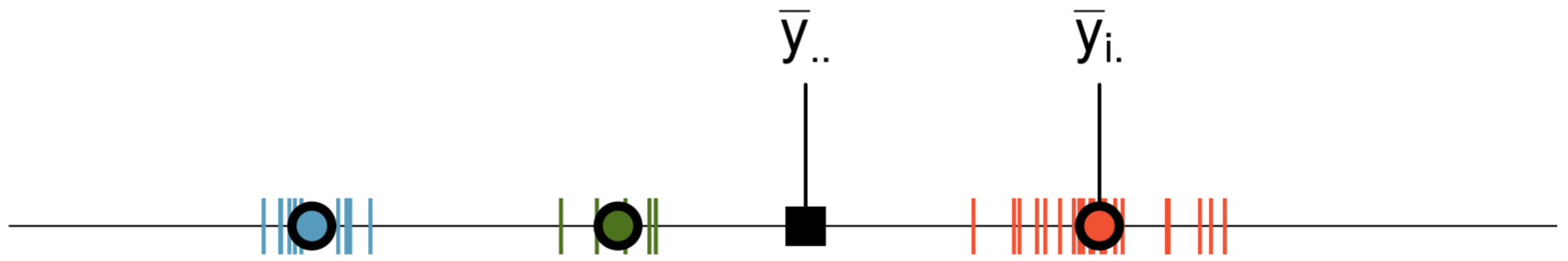**Model:** $\quad Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$

Observed data

$$y_{ij} \quad \text{for } j = 1, \ldots, n_i \text{ (and } i = 1, \ldots, k)$$

Notation:

- group-specific sample mean: $\quad \bar{y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$

- overall sample mean: $\quad \bar{y}_{\cdot\cdot} = \frac{1}{n_{tot}} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$

(here $n_{tot} = \sum_{i=1}^{k} n_i$)
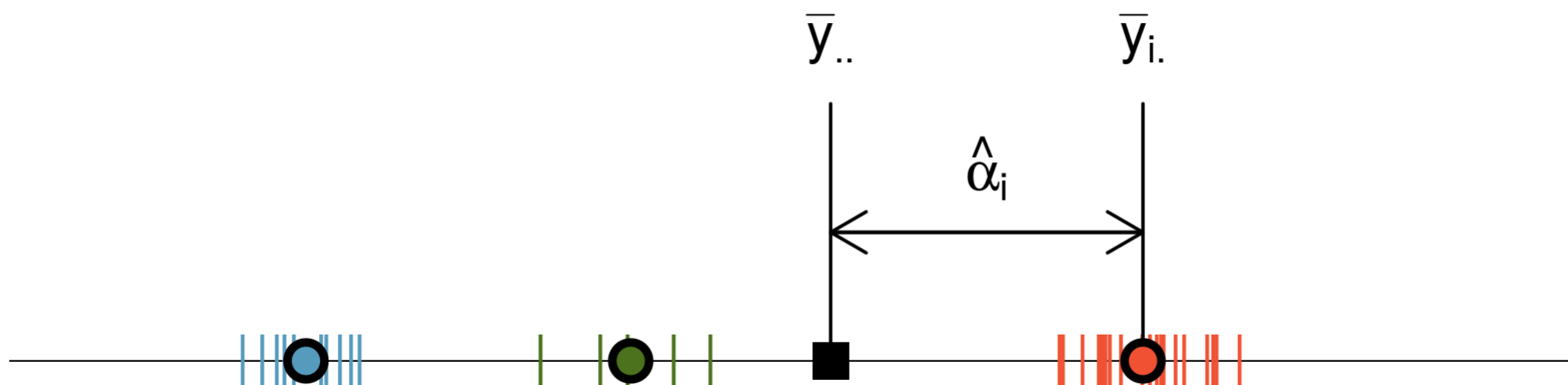
$\overline{y}_{..}$ $\overline{y}_{i.}$

Estimation in fixed effects model

Model: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$
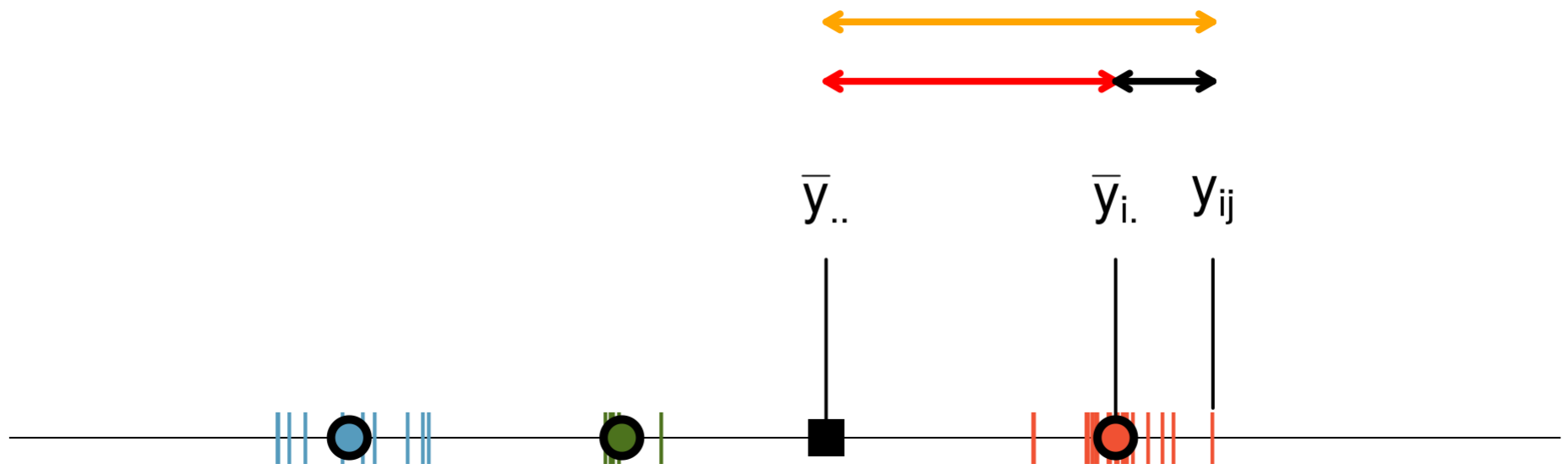
Obvious estimates:

$$\hat{\mu} = \bar{y}_{..} \qquad\qquad \hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{..}$$

ANOVA decomposition

$$y_{ij} - \bar{y}_{..} = (y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..})$$

$$\overline{y}_{..} \qquad \overline{y}_{i.} \qquad y_{ij}$$

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^{k} n_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

ANOVA decomposition

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2 + \sum_{i=1}^{k} n_i(\bar{y}_{i.}-\bar{y}_{..})^2$$

$$\mathrm{tss} = \mathrm{sse} + \mathrm{ssb}$$

Total sum of squares = sum of squares within group + sum of squares between groups

- **tss:** variability about the overall mean
- recognize $tss/(n_{tot}-1)$?

2a. Within group variability

sse (sum of squares of errors)

mse (mean squares of errors)

sse (as pooled error variance)

Regardless of whether H_0 or H_A holds, an unbiased estimate for σ^2 given by

$$\frac{1}{n_{tot} - k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 = \frac{sse}{n_{tot} - k}$$
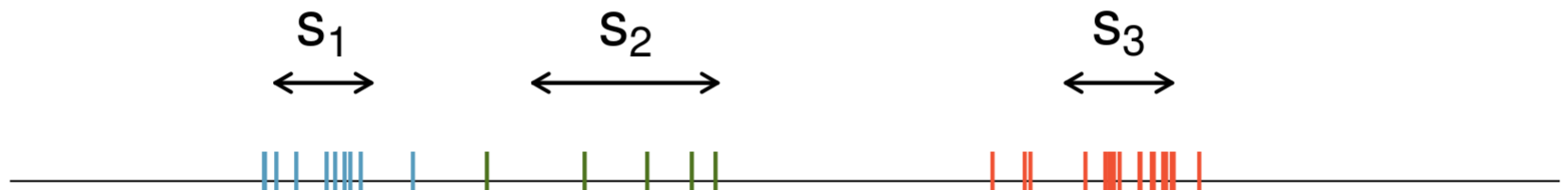
Recall **sample variance for group** i:

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

$$" \; mse \; " = \frac{sse}{n_{tot} - k} = \frac{1}{n_{tot} - k} \sum_{i=1}^{k} (n_i - 1) s_i^2$$

mse (as pooled error variance)

$$mse = \frac{sse}{n_{tot} - k} = \frac{1}{n_{tot} - k} \sum_{i=1}^{k} (n_i - 1)s_i^2$$

- weighted average of groups' sample variances

- bigger groups get more weight

2b. Between group variability

ssb (sum of squares between groups)

msb (mean squares between groups)

ssb: Between groups variability

ANOVA decomposition suggests that

$$ssb = \sum_{i=1}^{k} n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

is a way to measure variability **between groups**.

How much *on average* does a group vary from grand mean?

$$msb = \frac{ssb}{k-1}$$

Under H_0:

- should be small (purely due to random fluctuations of noise)

- in particular, would expect $msb \approx mse$

3. Compare between to within

- Test Statistic: MSB/MSE

- Under H_0, follows **F distribution** with appropriate df

- Reject H_0 if test statistic **is large**

F statistic

Under H_0, we expect $msb \approx mse$. More precisely...

Let MSB and MSE be the random variables (of which msb and mse are realizations)
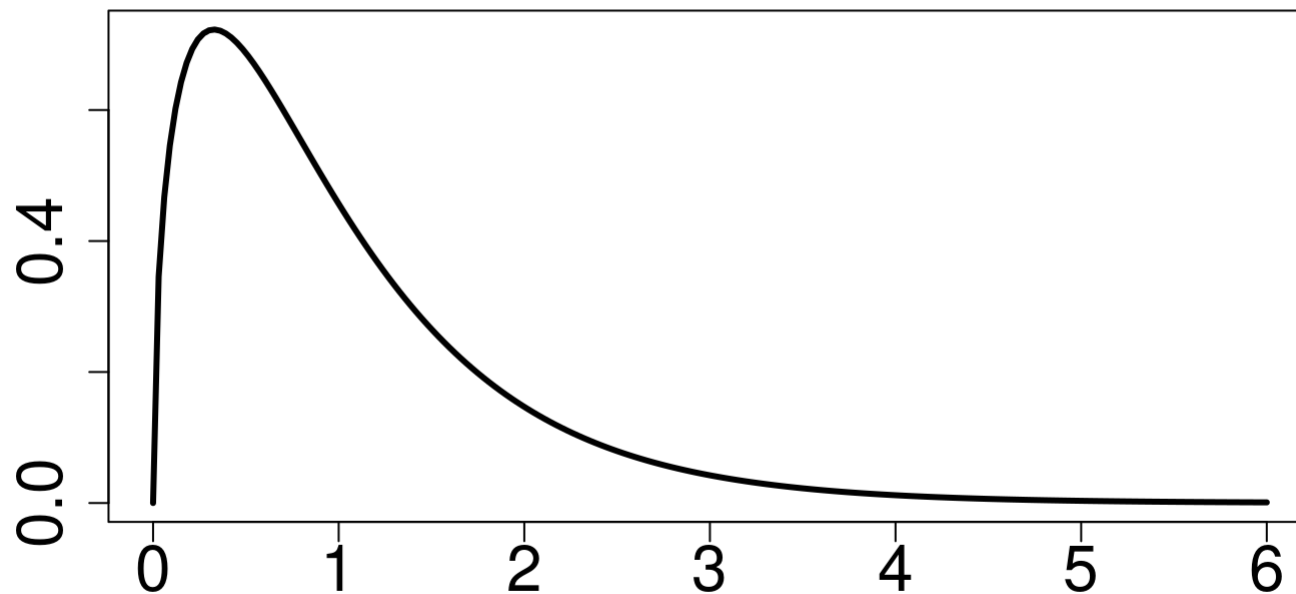
Under H_0, the statistic

$$\frac{MSB}{MSE} \sim F_{k-1, n_{tot}-k},$$

that is, it has an F distribution with $k-1$ and $n_{tot} - k$ degrees of freedom.

F distribution

F stands for Fisher

Density of F with 3 and 120 dfs

Rejecting null

Under H_0, the statistic

$$\frac{MSB}{MSE} \sim F_{k-1, n_{tot}-k},$$

If $\frac{msb}{mse}$ is "larger than expected under null", we reject H_0 in favor of H_A.

The ANOVA p-value is given by

$$P\left(F_{k-1, n_{tot}-k} > \frac{msb}{mse}\right)$$

The ANOVA Table

| Source of variability	Degrees of freedom	Sum of squares	Mean square	F statistic	p-value
Between	$k-1$	ssb	$msb = \frac{ssb}{k-1}$	$\frac{msb}{mse}$	$P\left(F_{k-1,n_{tot}-k} > \frac{msb}{mse}\right)$
Error (within)	$n_{tot} - k$	sse	$mse = \frac{sse}{n_{tot}-k}$		
Total	$n_{tot} - 1$	tss			

4. Check Assumptions

Independence: revisit sampling strategy used in data collection
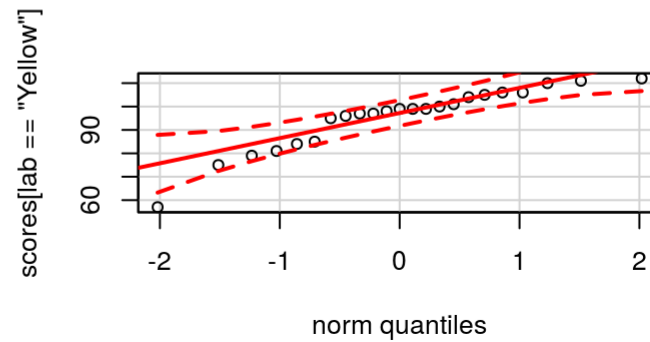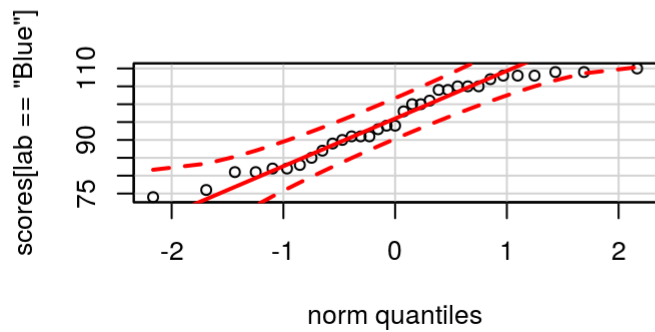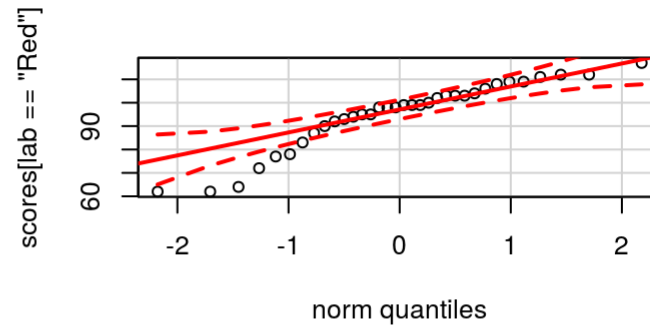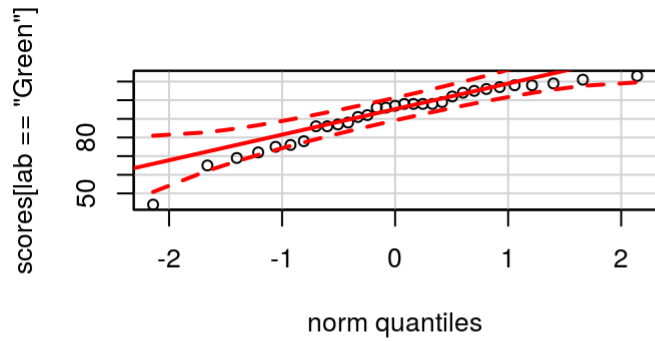
Normality: Q-Q plots, histograms

Equal variance: side-by-side boxplots, Levene's Test (to be covered in lab this week)

Back to the labs question

Prelims score by lab

Green   Yellow   Red   Blue
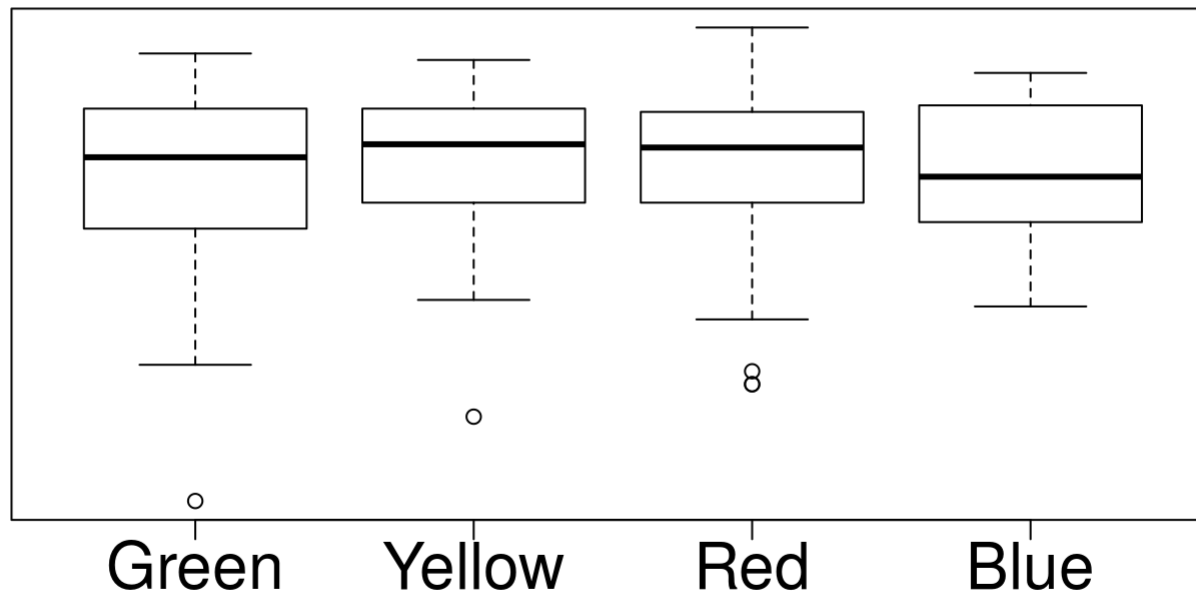
Assumptions met?

A look at Red lab

Histogram of scores[lab == "Red"]
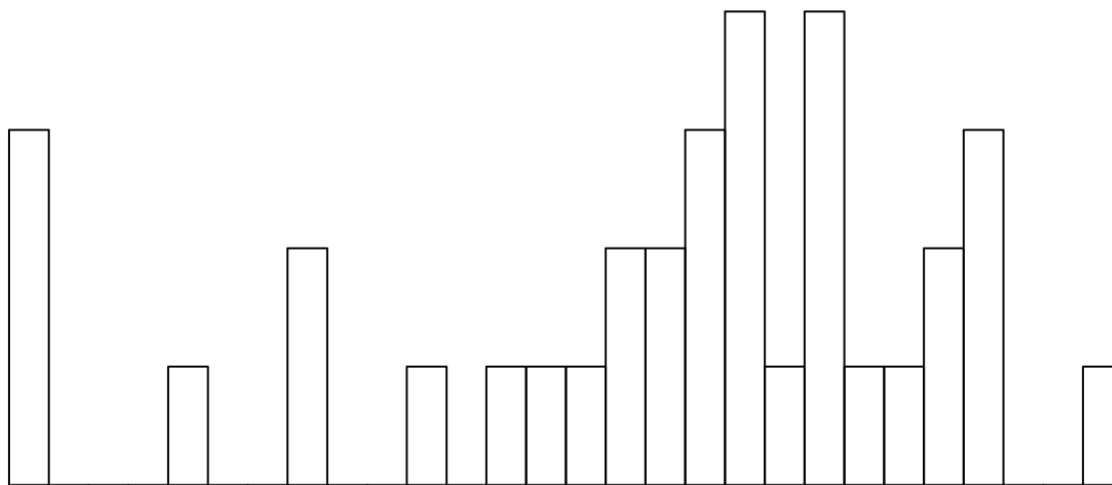
Importance of assumptions for ANOVA

- **independence** is very important

- **normality** is less important especially if sample sizes are close to equal (and not too small)

- **equal** σ_i assumption is crucial

- watch out for **outliers** (can repeat analysis with removed and see if there's a difference)

Let's proceed...

```
fit = aov(scores ~ lab)
summary(fit)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## lab          3    203   67.73   0.359  0.783
## Residuals  117  22059  188.54
```

Rerun after removing outliers

```
fit2 = aov(scores[-ii]~lab[-ii])
summary(fit2)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lab[-ii]       3    334   111.4   0.853  0.468
## Residuals    112  14631   130.6
```

Still no evidence against null.

5. R commands

aov

```
fit = aov(scores~lab)
summary(fit)
```

lm

```
scores.lm = lm(scores ~ lab)
anova(scores.lm)
```

Finishing up ANOVA

Basic ANOVA: model in pictures

Response = Mean + Noise

ANOVA

Want to test

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

versus H_A : means are not all the same

Basic ANOVA: model

$$Y_{1j} = \mu + \alpha_1 + \epsilon_{1j} \text{ for } j = 1, \ldots, n_1$$
$$Y_{2j} = \mu + \alpha_2 + \epsilon_{2j} \text{ for } j = 1, \ldots, n_2$$
$$\vdots$$
$$Y_{kj} = \mu + \alpha_k + \epsilon_{kj} \text{ for } j = 1, \ldots, n_k$$

In this case, we test whether $H_0 : \alpha_1 = \cdots = \alpha_k = 0$

Think of α_i as how group i's mean differs from μ, i.e. $\alpha_i = \mu_i - \mu$

(We require $\sum_{i=1}^{k} \alpha_i = 0$ so μ represents the "overall mean")
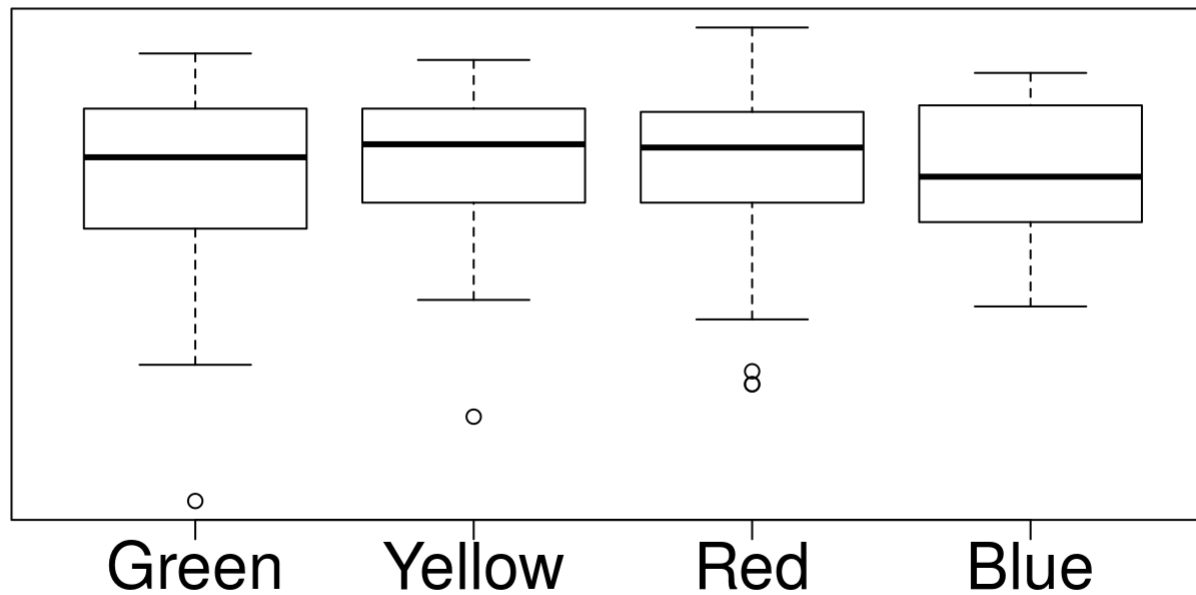
One-way fixed effects ANOVA

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \text{ for } j = 1, \ldots, n_i$$

Notation:

- n_i is number of units in group i

- Y_{ij} is response of unit j in group i

- μ is overall mean response value

- α_i is **effect** on mean response due to group i

- $\epsilon_{ij} \sim N(0, \sigma)$ random error for unit j of group i

Back to the labs question

Prelims score by lab

Green   Yellow   Red   Blue

In R...

```
scores.lm = lm(scores ~ lab)
anova(scores.lm)
```

```
## Analysis of Variance Table
##
## Response: scores
##             Df  Sum Sq Mean Sq F value Pr(>F)
## lab          3   203.2  67.729  0.3592 0.7826
## Residuals  117 22059.0 188.539
```

Rerun after removing outliers

```
scores2.lm = lm(scores[-ii] ~ lab[-ii])
anova(scores2.lm)
```

```
## Analysis of Variance Table
##
## Response: scores[-ii]
##              Df  Sum Sq Mean Sq F value Pr(>F)
## lab[-ii]      3   334.1  111.37  0.8526 0.4681
## Residuals   112 14630.6  130.63
```
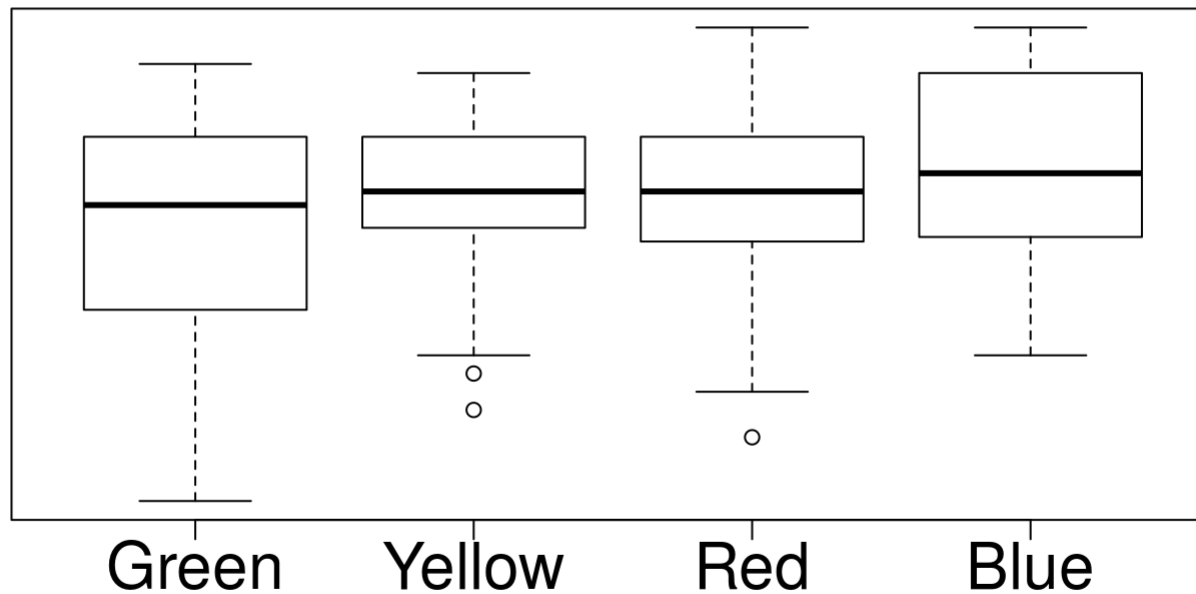
Still no evidence against null.

6. If H_0 rejected, which pairs differ in mean?

· Can go back to pairwise t-tests, **with multiple testing correction (using, say, Bonferroni)**

· Can use **Tukey's Honest Significance Difference (HSD) test**

Suppose the data looked like this

(Fake) prelims score by lab

Green   Yellow   Red   Blue

ANOVA on fake data

```
scores3.lm = lm(scores~lab, data = dat2)
anova(scores3.lm)
```

```
## Analysis of Variance Table
##
## Response: scores
##              Df  Sum Sq Mean Sq F value  Pr(>F)
## lab           3  1121.3  373.77  2.8613 0.04007 *
## Residuals   112 14630.6  130.63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- At $\alpha = 0.05$ level, we reject H_0 in favor of H_A.

Which pair(s) are different?

- can do pairwise tests as long as we correct for multiple testing
- in this context, there's something better than Bonferroni...
- **Tukey Honest Significant Differences** (Tukey HSD)
- takes into account that we are looking at $\binom{k}{2}$ comparisons

In R

```
library(multcomp)
hsd <- glht(scores3.lm, linfct = mcp(lab = "Tukey"))
summary(hsd)
```

[Output on next slide.]

In R

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = scores ~ lab, data = dat2)
##
## Linear Hypotheses:
##                     Estimate Std. Error t value Pr(>|t|)
## Yellow - Green == 0    3.294      3.208   1.027   0.7336
## Red - Green == 0       4.325      2.927   1.477   0.4537
## Blue - Green == 0      8.370      2.883   2.903   0.0225 *
## Red - Yellow == 0      1.031      3.186   0.324   0.9882
## Blue - Yellow == 0     5.076      3.146   1.613   0.3746
## Blue - Red == 0        4.045      2.859   1.415   0.4920
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Final note

- can sometimes reject $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ and yet no pairwise test would reject!

- ANOVA pools information across all groups and tries to answer a more general question