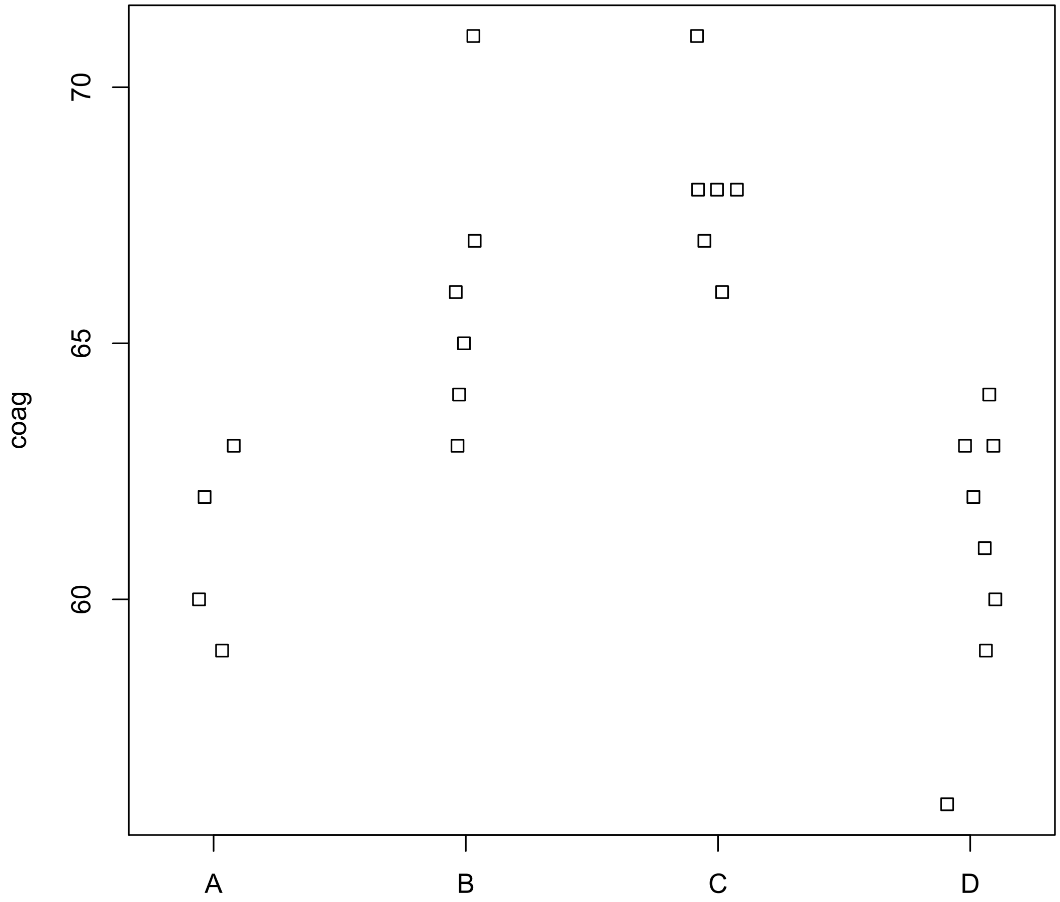
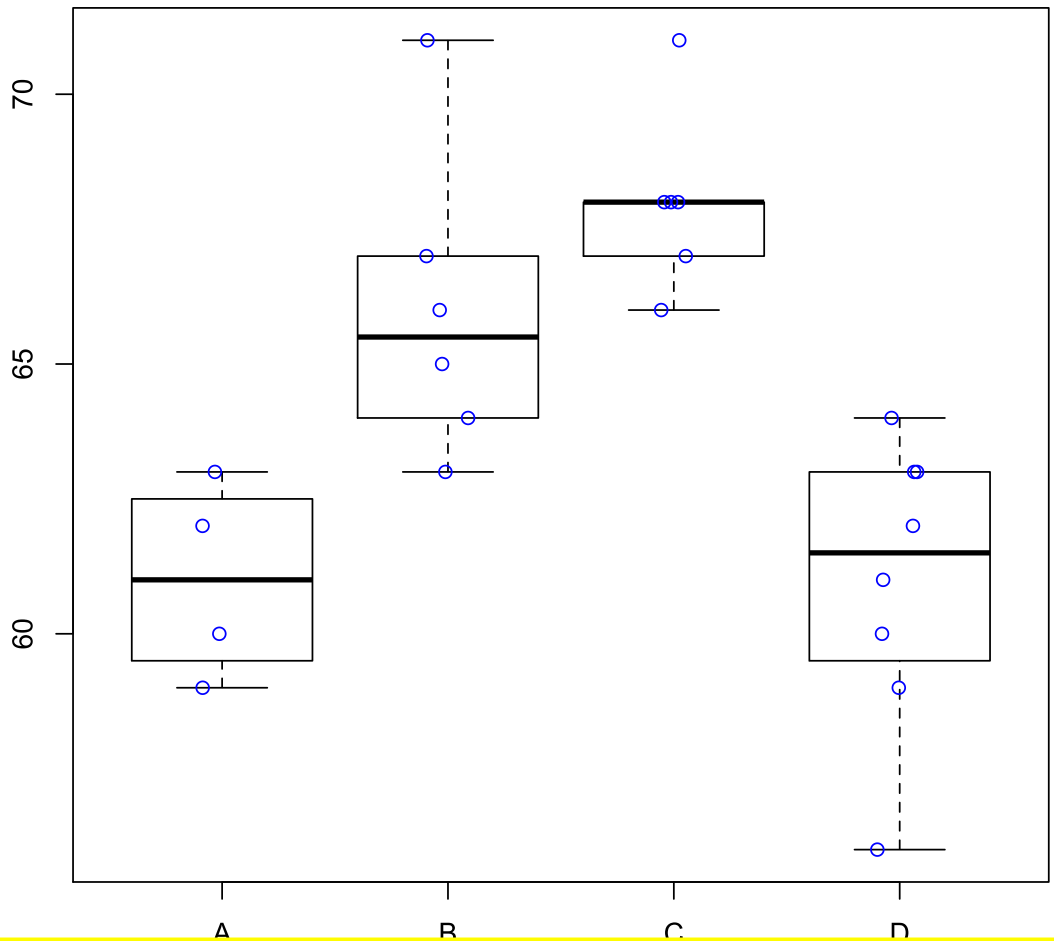


```
stripchart(coag ~ diet, vertical=TRUE,
method="jitter")
```



```
> boxplot(coag ~ diet, outline=FALSE)
> stripchart(coag ~ diet, vertical=TRUE,
add=TRUE, col="blue",
 pch=1, method="jitter")
```

Diagnostics

- Q-Q plot for residuals.
- Check outliers.
- Test for equal variance.

  Levene's test: run regression abs(residual) $\sim X$, i.e., use abs(residuals) as the response in a new one-way ANOVA. If the p-value for the F-test is less than 1% level, then we conclude that there is no evidence of a non-constant variance.

```
> g = lm(coag ~ diet)
> summary(lm(abs(g$res) ~ diet))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5000     0.7159   2.095   0.0491 *
dietB         0.5000     0.9242   0.541   0.5945
dietC        -0.5000     0.9242  -0.541   0.5945
dietD         0.5000     0.8768   0.570   0.5748

Residual standard error: 1.432 on 20 df
Multiple R-squared: 0.09559,
Adjusted R-squared: -0.04007
F-statistic: 0.7046 on 3 and 20 DF,  p-value: 0.5604
```

Detecting the Difference among Groups

- Consider the one-way ANOVA model [a]

$$y_{ij} = \alpha_i + e_{ij}, \quad e_{ij} \text{ iid } \sim \mathsf{N}(0, \sigma^2).$$

- After detecting some difference among the groups using the F-test, interest centers on which groups or combinations of them are different.

---

[a] Here we use the parameterization which sets $\mu = 0$.

There are two cases:

- Pairwise difference: $\alpha_i - \alpha_j$
- Contrasts: $\sum_{i=1}^{g} c_i \alpha_i, \quad \sum_i c_i = 0$. (Of course, the pairwise diff is a special case of contrasts).

We'll focus on CIs for differences, which also tells us the corresponding testing result due to the duality between statistical tests and CIs.

Pairwise Comparisons

- α_i: unknown group mean

  Estimate $\hat{\alpha}_i = \bar{y}_{i\cdot}$ with s.e. $\hat{\sigma}\sqrt{1/n_i}$.

- $\alpha_i - \alpha_j$: unknown group difference

  Estimate $\hat{\alpha}_i - \hat{\alpha}_j = \bar{y}_{i\cdot} - \bar{y}_{j\cdot}$ with s.e. $\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$.

- $(1 - \alpha)$ CI for $\alpha_i - \alpha_j$

$$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm t_{n-g}^{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

- The ordinary t-based CI (on the previous slide) is the CI for just one comparison.

- Recall its interpretation (assume $\alpha = 5\%$):

  The (random) CI covers the true parameter $\alpha_i - \alpha_j$ with prob 95%. In other words, the chance of making an error (i.e., not covering the true difference) is controlled to be 5%.

- In practice we need to construct CIs for multiple pairwise differences, e.g., for the `coagulation` data, there are totally 6 pairwise comparisons
- If we construct 95% CI for each pairwise difference, then the chance of making an error is 5% for each CI. However, the chance that at least one of the CI does not cover the true difference (i.e., the family wise error rate) will be much bigger than 5%.
- We need to adjust for multiple comparisons. How?

Bonferroni Correction

- Suppose there are m pairwise comparisons. To control the family wise error rate to be α, we need to reduce the error rate for each individual comparison to be α/m.
- That is, we need to incase the significant level from $(1-\alpha)$ to be $(1-\alpha/m)$. For example, if $m = 10$ and $\alpha = 5\%$, then we need to set the significant level for each individual comparison to be as high as 99.5%.
- Not applicable when m is large, since the CIs would be too wide (of little practical interest) due to the increase of the significant level.

Tukey's Honest Significant Difference (HSD)

- The Tukey's CIs for $\alpha_i - \alpha_j$ are

$$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm q^{\alpha}_{g,n-g} \frac{\hat{\sigma}}{\sqrt{2}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

- Let $X_1, \ldots, X_m$ be iid N$(0, \sigma^2)$ and the following random variable

$$\frac{\max_i X_i - \min_j X_j}{\hat{\sigma}} \sim q_{m,v}$$

aka the studentized range distribution, where v is the df used in estimating σ.

- $q^{\alpha}_{g,n-g}$ is the $(1-\alpha)$ quantile of $q_{g,n-g}$.

Recall how we derive the t-based CI.

$$\frac{\bar{y}_{i\cdot} - \bar{y}_{j\cdot} - (\alpha_i - \alpha_j)}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim t\text{-dist}\ (\text{df} = n - g)$$

$$\mathbb{P}\left(\frac{|\bar{y}_{i\cdot} - \bar{y}_{j\cdot} - (\alpha_i - \alpha_j)|}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \leq t_{n-g}^{\alpha/2} \right) = 1 - \alpha$$

$$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm t_{n-g}^{\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

$$\mathbb{P}\left(\underset{\underset{\displaystyle\max_{i,j=1,\ldots,g}}{\Big\uparrow}}{\frac{|\bar{y}_{i\cdot} - \bar{y}_{j\cdot} - (\alpha_i - \alpha_j)|}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}} \le \underset{\underset{\displaystyle C}{\Big\uparrow}}{t_{n-g}^{\alpha/2}} \right) = 1 - \alpha$$

Suppose $n_i = n_j = n_0$.

$$\frac{|\bar{y}_{i\cdot} - \bar{y}_{j\cdot} - (\alpha_i - \alpha_j)|}{\hat{\sigma}\sqrt{\frac{1}{n_0} + \frac{1}{n_0}}}$$

$$= \frac{1}{\hat{\sigma}\sqrt{2}} \left| \frac{\bar{y}_{i\cdot} - \alpha_i}{\sqrt{\frac{1}{n_0}}} - \frac{\bar{y}_{j\cdot} - \alpha_j}{\sqrt{\frac{1}{n_0}}} \right|$$

$$= \frac{1}{\sqrt{2}} \frac{|X_i - X_j|}{\hat{\sigma}}$$

$$X_1, \ldots, X_g \text{ iid } \sim \mathsf{N}(0, \sigma^2)$$

$$\mathbb{P}\left(\frac{|\bar{y}_{i\cdot} - \bar{y}_{j\cdot} - (\alpha_i - \alpha_j)|}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \leq t_{n-g}^{\alpha/2}\right) = 1 - \alpha$$

$$\max_{i,j=1,\ldots,g}$$

$$\mathbb{P}\left(\frac{1}{\sqrt{2}} \max_{i,j} \frac{|X_i - X_j|}{\hat{\sigma}} \leq C\right) = 1 - \alpha$$

```
> TukeyHSD(aov(coag~diet, coagulation))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = coag ~ diet, data =
coagulation)

$diet
    diff         lwr       upr     p adj
B-A    5   0.7245544  9.275446 0.0183283
C-A    7   2.7245544 11.275446 0.0009577
D-A    0  -4.0560438  4.056044 1.0000000
C-B    2  -1.8240748  5.824075 0.4766005
D-B   -5  -8.5770944 -1.422906 0.0044114
D-C   -7 -10.5770944 -3.422906 0.0001268
```

Scheffé's Method for Contrasts

- A linear combination of the group means $\sum_{i=1}^{g} c_i \alpha_i$ is called a contrast if $\sum_i c_i = 0$.
  - $\alpha_1 - \alpha_2$: $c_1 = 1$, $c_2 = -1$, and other c_i's $= 0$.
  - $(\alpha_1 + \alpha_2)/2 - \alpha_3$: $c_1 = c_2 = 1/2$, $c_3 = -1$, and other c_i's $= 0$.
- The estimate of $\sum_{i=1}^{g} c_i \alpha_i$ is $\sum_{i=1}^{g} c_i \bar{y}_{i\cdot}$ with s.e. $\hat{\sigma}\sqrt{\sum_i c_i^2/n_i}$.
- The Scheffé's CIs are

$$\sum_i c_i \bar{y}_{i\cdot} \ \pm \ \sqrt{(g-1)F^{\alpha}_{g-1,n-g}}\hat{\sigma}\sqrt{\sum_i \frac{c_i^2}{n_i}}.$$

$$\frac{\left(\sum_i c_i \bar{y}_{i\cdot} - \sum_i c_i \alpha_i\right)^2}{\hat{\sigma}^2 \left(\sum_i c_i^2 / n_i \right)}$$

$$= \frac{\left[\sum_i c_i (\bar{y}_{i\cdot} - \alpha_i) \right]^2 / (\sum_i c_i^2 / n_i)}{\hat{\sigma}^2}$$

$$\leq \frac{\chi^2_{g-1}}{\chi^2_{n-g}/(n-g)} = (g-1) F_{g-1,n-g}$$

$$\mathbb{P}\left(\frac{\left|\sum_i c_i\bar{y}_{i\cdot} - \sum_i c_i\alpha_i\right|}{\hat{\sigma}\sqrt{\sum_i c_i^2/n_i}} \le t_{n-g}^{\alpha/2}\right) = 1-\alpha$$

$$\max_{c_1,\ldots,c_g}$$

C

A Summary

- One pairwise/contrast: The ordinary t-based CI
- A small number of comparisons: Bonferroni CIs
- A large number of pairwise diffs: Tukey's CIs (adjusted for all possible pairwise comparisons)
- A large number of contrasts: Scheffé's CIs (adjusted for all possible contrasts)

How to decided between Bonferroni and Tukey's (or Scheffé's)? Just pick the approach giving your CIs of (overall) shorter length.