Again we deal with circuit boards but this time investigate 3 types of fluxes: X, Y, Z.

We have 18 circuit boards, randomly assign each flux to 6 boards. This gives us the randomization reference distribution and a logical basis for a test of the hypothesis $H_0$ : no flux differences.

Randomize soldering/cleaning, coating, and humidity chamber slots. These randomizations avoid unintended biases from hidden factors.

There are $\binom{18}{6} \times \binom{12}{6} = 18,564 \times 924 = 17,153,136$ flux allocations.

Note the growth in the number of splits when dividing 18 into 3 groups of 6.

The full randomization reference distribution may be pushing the limits $\Rightarrow$ simulated reference distribution.
### The Flux3 Data

#### Applied Statistics and Experimental Design

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.9</td>
<td>10.7</td>
<td>10.9</td>
</tr>
<tr>
<td>9.6</td>
<td>10.4</td>
<td>11.0</td>
</tr>
<tr>
<td>9.6</td>
<td>9.5</td>
<td>9.5</td>
</tr>
<tr>
<td>9.7</td>
<td>9.6</td>
<td>10.0</td>
</tr>
<tr>
<td>9.5</td>
<td>9.8</td>
<td>11.7</td>
</tr>
<tr>
<td>10.0</td>
<td>9.9</td>
<td>10.2</td>
</tr>
</tbody>
</table>

**Units**: $\log_{10}$(Ohm)

![Box plot](image)
To answer whether the fluxes are different in their effects we could perform all 3 possible two-sample tests for the following respective testing problems:

\[
H_{0,XY} : \mu_X = \mu_Y \text{ vs. } H_{1,XY} : \mu_X \neq \mu_Y \\
H_{0,XZ} : \mu_X = \mu_Z \text{ vs. } H_{1,XZ} : \mu_X \neq \mu_Z \\
H_{0,YZ} : \mu_Y = \mu_Z \text{ vs. } H_{1,YZ} : \mu_Y \neq \mu_Z \\
\]

If we do each such test at level $\alpha$ what is our chance of getting a rejection by at least one of these tests when in fact the 3 fluxes show no difference?

If we assume that these 3 tests are independent of each other we would have

\[
P_0(\text{Type I error}) = 1 - P_0(\text{accept } H_{0,XY} \cap \text{accept } H_{0,XZ} \cap \text{accept } H_{0,YZ}) \\
= 1 - (1 - \alpha)^3 = 0.142625 \quad \text{for } \alpha = .05 .
\]

$P_0$ indicates that all 3 fluxes are the same.
The Multiple Comparison Issue

If you expose yourself to multiple opportunities of making a wrong decision, the chance of making a wrong decision at least once is much higher than planned for in the individual tests. This is called the multiple comparison issue.

How much higher is it? The calculation based on independence is not quite correct. The same sample is involved in any two such comparisons \( \implies \) dependence.

An upper bound on the overall type I error probability by Boole’s inequality:

\[
P_0(\text{overall type I error}) = P_0(\text{reject } H_{0,12} \cup \text{reject } H_{0,13} \cup \text{reject } H_{0,23})
\leq P_0(\text{reject } H_{0,12}) + P(\text{reject } H_{0,13}) + P(\text{reject } H_{0,23})
= 3\alpha = .15 \quad \text{when } \alpha = .05.
\]

How much smaller than this upper bound is the true \( P_0(\text{overall type I error}) \)?
Get the randomization reference distribution of $\bar{X} - \bar{Y}$ for splits of the 18 SIR values into 3 groups of 6 and taking the difference of averages for the first two groups.

Do this by simulation: $N_{sim0} = 10000$ times.

Get the .95-quantile $t_{crit}$ of this simulated $|\bar{X} - \bar{Y}|$ reference distribution.

Then simulate another $N_{sim1} = 10000$ such splits, computing $|\bar{X} - \bar{Y}|$, $|\bar{X} - \bar{Z}|$, and $|\bar{Y} - \bar{Z}|$ each time, and tallying the proportions of each exceeding $t_{crit}$ and the proportion of at least one of them exceeding $t_{crit}$.

The resulting proportions are: 0.0451 0.0460 0.0491 for the individual tests and 0.1186 for the overall type I error rate.

The code for running this, `typeIerror.rateRand`, is posted on web.
Rather than doing all possible pairwise tests we will address this in a global way.

In the context of a 3 population model we will test the hypothesis

\[ H_0 : \mu_1 = \mu_2 = \mu_3 \text{ (common value unspecified)} \]

against the alternative

\[ H_1 : \mu_i \neq \mu_j \text{ for some } i \neq j. \]

More generally we may have \( t \) treatments

and \( n_i \) observations \( Y_{i,1}, \ldots, Y_{i,n_i} \) for the \( i^{th} \) treatment, \( i = 1, \ldots, t. \)

Test \( H_0 : \mu_1 = \ldots = \mu_t \) against \( H_1 : \mu_i \neq \mu_j \) for some \( i \neq j. \)

For the Flux3 data we have: \( t = 3 \) and \( n_1 = n_2 = n_3 = 6 \), a balanced design.

When the \( n_i \) are not all the same we have an unbalanced design.
We have measurements $Y_{ij}$, the $j^{th}$ response under the $i^{th}$ treatment, $i = 1, \ldots, t$ and $j = 1, \ldots, n_t$. A total of $N = n_1 + \ldots + n_t$ measurements.

**Treatment Means Model:** $Y_{ij} = \mu_i + \varepsilon_{ij}$ with $E(\varepsilon_{ij}) = 0$ and $\text{var}(\varepsilon_{ij}) = \sigma^2$.

View $\varepsilon_{ij}$ (i.i.d.) as response variation/error/noise that occurs within treatment or after the treatment mean $\mu_i$ is subtracted from the response $Y_{ij}$.

**Treatment Effects Model:** $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ with $E(\varepsilon_{ij}) = 0$ and $\text{var}(\varepsilon_{ij}) = \sigma^2$.

$\mu = \bar{\mu} = \sum_{i,j} \mu_i / N = \sum_i n_i \mu_i / N$ is the grand mean (weighted average)

$\tau_i = \mu_i - \mu = \mu_i - \bar{\mu}$ is the $i^{th}$ treatment effect and $\varepsilon_{ij}$ (i.i.d.) is the within treatment variation with $E(\varepsilon_{ij}) = 0$ and $\text{var}(\varepsilon_{ij}) = \sigma^2$.

Note that the $\tau_i$ satisfy the constraint: $\sum_{ij} \tau_i = \sum_i n_i \tau_i = 0$. 

In the reduced model we assume

\[ Y_{ij} = \mu + \varepsilon_{ij} \quad \text{with} \quad E(\varepsilon_{ij}) = 0 \quad \text{with} \quad \text{var}(\varepsilon_{ij}) = \sigma^2, \]

i.e., there is no variation due to treatments.

This corresponds to our previously stated hypothesis

\[ H_0 : \mu_1 = \ldots = \mu_t \quad \text{or} \quad H_0 : \tau_1 = \ldots = \tau_t = 0 \]

which is a special case of our previous more general model.

We will test this hypothesis by fitting models to the full model and the reduced model and compare the quality of fits relative to each other.
The method of Least Squares originated with Gauss and Legendre.

Minimize the sum of squares criterion
\[ SS(\mu) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 \]
over \( \mu \).

Via calculus solving the equations
\[ \frac{\partial SS(\mu)}{\partial \mu_i} = 0, \quad i = 1, \ldots, t, \]
or directly using the notation
\[ \bar{Y}_i = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} \]
and the fact
\[ \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i) = 0, \]
we get
\[
SS(\mu) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i + \bar{Y}_i - \mu_i)^2 = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{t} \sum_{j=1}^{n_i} (\bar{Y}_i - \mu_i)^2
+ 2 \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)(\bar{Y}_i - \mu_i)
\]
\[
= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{t} \sum_{j=1}^{n_i} (\bar{Y}_i - \mu_i)^2 \geq \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 = SS(\hat{\mu})
\]

\( \Rightarrow \) the least squares estimates (LSE)
\[ \hat{\mu}_i = \bar{Y}_i \]
minimize \( SS(\mu) \).
Minimize the sum of squares criterion $SS(\mu) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \mu)^2$

With $\bar{Y}_{..} = \sum_{i} \sum_{j} Y_{ij} / \sum_{i} n_i = \sum_{i} \sum_{j} Y_{ij} / N = \sum_{i} (n_i / N) \bar{Y}_i.$ and $\sum_{i} \sum_{j} (Y_{ij} - \bar{Y}_{..}) = 0$

$\implies SS(\mu) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \mu)^2 = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..} + \bar{Y}_{..} - \mu)^2$

$= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 + \sum_{i=1}^{t} \sum_{j=1}^{n_i} (\bar{Y}_{..} - \mu)^2 - 2 \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})(\bar{Y}_{..} - \mu)$

$= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 + \sum_{i=1}^{t} \sum_{j=1}^{n_i} (\bar{Y}_{..} - \mu)^2 \geq \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 = SS(\hat{\mu})$

$\implies$ the least squares estimate (LSE) $\hat{\mu} = \bar{Y}_{..}$ minimizes $SS(\mu)$
Using $\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.}) = 0$ we have the following sum of squares decomposition

$$SS_T = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.} + \bar{Y}_{i.} - \bar{Y}_{..})^2$$

$$= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 + \sum_{i=1}^{t} \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2$$

$$+ 2 \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.}) (\bar{Y}_{i.} - \bar{Y}_{..})$$

$$= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 + \sum_{i=1}^{t} \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2 = SS_E + SS_{Treat}$$

This is the fundamental ANOVA identity: $SS_T = SS_E + SS_{Treat} = SS_W + SS_B$.

$SS$ of total variation = error variation + treatment variation

or $SS$ of total variation = variation within samples + variation between samples.
How should we compare the two model fits

\[ \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \quad \text{and} \quad \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 \]

The previous slide showed

\[ \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 \geq \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \]

Thus we should makes allowances for the fact that we had more freedom in the former minimization than in the latter.
Assume that $X_1, \ldots, X_n$ are i.i.d. with mean $\mu$ and variance $\sigma^2$.

If in addition we assume a normal distribution for the $X_i$ we have

$$E \left( \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right) = E(s^2) = \sigma^2 \quad \Rightarrow s^2 \text{ is an unbiased estimate of } \sigma^2.$$

The normality assumption is not essential. Using $E(Y^2) = \text{var}(Y) + [E(Y)]^2$

$$\implies E((n-1)s^2) = E \left( \sum_{i=1}^{n} (X_i - \bar{X})^2 \right) = E \left( \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right)$$

$$= n(\sigma^2 + \mu^2) - n(\text{var}(\bar{X}) + [E(\bar{X})]^2)$$

$$= n(\sigma^2 + \mu^2) - n(\sigma^2/n + \mu^2) = (n-1)\sigma^2.$$

i.e., $E(s^2) = \sigma^2$ \quad Q.E.D.
\[ E(\text{MS}_E) = \sigma^2 \]

With

\[ s_i^2 = \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 / (n_i - 1) \quad \sum_{i=1}^{t} (n_i - 1)s_i^2 = SS_E \]

the result from the previous slide shows

\[ E\left( \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \right) = E\left( \sum_{i=1}^{t} (n_i - 1)s_i^2 \right) = \sum_{i=1}^{t} (n_i - 1)\sigma^2 = (N - t)\sigma^2 \]

or the Mean Square for Error

\[ \text{MS}_E = \frac{SS_E}{N - t} = \frac{\sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}{N - t} \]

is an unbiased estimate for \( \sigma^2 \)

This is true whether \( H_0 : \mu_1 = \ldots = \mu_t \) holds or not.
\( E(\text{MS}_{\text{Treat}}) = \sigma^2 + ? \)

Note that \( E(\bar{Y}_{..}) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \mu_i/N = \sum_{i=1}^{t} n_i \mu_i/N = \bar{\mu} \) (weighted average)

and \( SS_{\text{Treat}} = \sum_{i=1}^{t} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2 = \sum_{i=1}^{t} n_i\bar{Y}_{i.}^2 - N\bar{Y}_{..}^2 \) (exercise)

\[ E(SS_{\text{Treat}}) = \sum_{i=1}^{t} n_i E(\bar{Y}_{i.}^2) - N E(\bar{Y}_{..}^2) \]

\[ = \sum_{i=1}^{t} n_i(\text{var}(\bar{Y}_{i.}) + [E(\bar{Y}_{i.})]^2) - N(\text{var}(\bar{Y}_{..}) + [E(\bar{Y}_{..})]^2) \]

\[ = \sum_{i=1}^{t} n_i(\sigma^2/n_i + \mu_i^2) - N(\sigma^2/N + \bar{\mu}^2) = (t-1)\sigma^2 + \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu})^2 \]

since \( \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu})^2 = \sum_{i=1}^{t} n_i\mu_i^2 + \sum_{i=1}^{t} n_i\bar{\mu}^2 - 2 \sum_{i=1}^{t} n_i\mu_i\bar{\mu} = \sum_{i=1}^{t} n_i\mu_i^2 - N\bar{\mu}^2 \)

\[ E(\text{MS}_{\text{Treat}}) = E(SS_{\text{Treat}}/(t-1)) = \sigma^2 + \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu})^2/(t-1) = \sigma^2 + \sum_{i=1}^{t} n_i\tau_i^2/(t-1). \]
When $H_0$ is true then both $MS_{Treat}$ and $MS_E$ are unbiased estimates of $\sigma^2$

$H_0$ is false $\Rightarrow \sum_{i=1}^{t} n_i (\mu_i - \bar{\mu})^2 / (t - 1) > 0 \Rightarrow E(MS_{Treat}) > E(MS_E)$

and $MS_{Treat}$ will generally be somewhat larger than $MS_E$

and more so when the $\mu_i$ are more dispersed. The $n_i$ act as magnifiers!

This suggests $F = MS_{Treat} / MS_E$ as a plausible test statistic.

Looking at the ratio makes more sense than looking at the difference, since any such difference should be viewed relative to the magnitude of $MS_E$.

By transferral we will use this test statistic in our randomization test, even though we are not quite in an i.i.d. situation there.
Equivalent Form for the $F$-Statistic under Randomization

First note that in the $SS$ decomposition $SS_T = SS_{Treat} + SS_E$ the sum $SS_T$ stays constant over all partitions of the full data set into $t$ groups of sizes $n_1, \ldots, n_t$.

In $SS_{Treat} = \sum_{i=1}^{t} n_i \bar{Y}_i^2 - N \bar{Y}^2_{..} = F_{equiv} - N \bar{Y}^2_{..}$ with $F_{equiv} = \sum_{i=1}^{t} n_i \bar{Y}_i^2$ the term $\bar{Y}^2_{..}$ stays constant over all such partitions.

Thus

$$F = \frac{N - t}{t - 1} \frac{SS_{Treat}}{SS_E} = \frac{N - t}{t - 1} \frac{SS_{Treat}}{SS_T - SS_{Treat}} = \frac{N - t}{t - 1} \frac{F_{equiv} - N \bar{Y}^2_{..}}{SS_T - (F_{equiv} - N \bar{Y}^2_{..})} \uparrow$$

in $F_{equiv}$

Thus the randomization distribution of $F$ is in 1-1 correspondence with the randomization distribution of $F_{equiv}$ which we can then take as an alternate and more easily calculable test statistic for computing p-values under $H_0$. 

17
Simulated Randomization Distribution

F–equivalent test statistic 1832.3
p–value = 0.04296
based on 1e+05 simulations
R Code for Randomization Distribution

```r
code
Ftest.rand = function (y=SIR,n=c(6,6,6),Nsim=10000){#try Nsim=10000 first for speed
  F.eq=NULL
  for(i in 1:Nsim){
    ind=sample(1:18)
    F.eq=c(F.eq,n[1]*mean(y[ind[1:n[1]])^2+
  }
  out=hist(F.eq,nclass=100,main="Simulated Randomization Distribution",
    xlab="F-equivalent Test Statistic",col=c("blue","orange"))
  abline(v=F.obs,col="red",lwd=2)
  pval=mean(F.eq>=F.obs)
  text(F.obs+.2,.24*max(out$counts),
    paste("F-equivalent test statistic ",format(signif(F.obs,5)),adj=0)
  text(F.obs+.2,.2*max(out$counts),paste("p-value =",format(signif(pval,4)),adj=0)
  text(F.obs+.2,.16*max(out$counts),paste("based on ",Nsim," simulations"),adj=0)
  c(F.obs,pval)
}
```

This would need to be adapted to other ANOVA data situations!

19
As in the case of the 2-sample problem one finds that the \( F_{t-1,N-t} \) distribution provides a good approximation to the randomization distribution of \( F \).

The randomization distribution of \( F \) is obtained from that of \( F_{\text{equiv}} \) via

\[
F = \frac{N - t}{t - 1} \frac{F_{\text{equiv}} - N\bar{Y}^2}{SS_T - (F_{\text{equiv}} - N\bar{Y}^2)}
\]

The next slide shows the quality of this approximation for the Flux3 data set.
Simulated Randomization Distribution

F-statistic 3.6452
p-value = 0.04296 & p-value = 0.05126 from F-distribution based on 1e+05 simulations

superimposed F-density
We will now assume in addition that the $Y_{ij}$ have normal distributions with previously indicated model parameters, and are independent.

First we consider $H_0: \mu_1 = \ldots = \mu_t$ to get the null distribution of $F$

$$SS_E = \sum_{i=1}^{t} (n_i - 1)s_i^2 \sim \sigma^2 \chi_{n_1-1}^2 + \ldots + \sigma^2 \chi_{n_t-1}^2 \sim \sigma^2 \chi_{N-t}^2$$ for $H_0$ true or not

$SS_E$ is independent of $\bar{Y}_1, \ldots, \bar{Y}_t$.

Is $SS_{Treat} = \sum_{i=1}^{t} n_i \bar{Y}_{i.}^2 - N\bar{Y}_{..}^2$ distributed like $\sigma^2 \chi^2$? What degrees of freedom $f$?

Under $H_0$ we would expect $f = t - 1$ since $E(MS_{Treat}) = E(SS_{Treat}/(t - 1)) = \sigma^2$. 

22
Distribution of $SS_{Treat}$ ($H_0$ true or not)

\[ \bar{Y}_i \sim \mathcal{N}(\mu_i, \sigma^2/n_i) \implies \sqrt{n_i} \bar{Y}_i \sim \mathcal{N}(\sqrt{n_i} \mu_i, \sigma^2) \implies \sqrt{n_i} \bar{Y}_i = \sqrt{n_i} \mu_i + \sigma Z_i \]

with $Z_1, \ldots, Z_t$ being i.i.d. standard normal random variables.

With a new orthonormal basis $g_1, \ldots, g_t$ with $g'_1 = (\sqrt{n_1/N}, \ldots, \sqrt{n_t/N})$

\[
(\sqrt{n_1} \bar{Y}_1, \ldots, \sqrt{n_t} \bar{Y}_t)' = \sqrt{n_1} \bar{Y}_1 \cdot e_1 + \ldots + \sqrt{n_t} \bar{Y}_t \cdot e_t = V_1 g_1 + \ldots + V_t g_t
\]

$V_1 = (\sqrt{n_1} \bar{Y}_1, \ldots, \sqrt{n_t} \bar{Y}_t)g_1 = \sum_{i=1}^t n_i \bar{Y}_i / \sqrt{N} = \sqrt{N} \bar{Y}_. \quad \text{and}$

$\sum_{i=2}^t V_i^2 = \sum_{i=1}^t n_i \bar{Y}_i^2 - V_1^2 = \sum_{i=1}^t n_i \bar{Y}_i^2 - N \bar{Y}_2^2 = SS_{Treat}$

$V_i = (\sqrt{n_1} \bar{Y}_1, \ldots, \sqrt{n_t} \bar{Y}_t)g_i = (\sqrt{n_1} \mu_1, \ldots, \sqrt{n_t} \mu_t)g_i + \sigma(Z_1, \ldots, Z_t)g_i$

$= v_i + \sigma U_i \quad \text{with } U_1, \ldots, U_t \text{ being again i.i.d. } \mathcal{N}(0, 1)$

$v_1 = \sum_{i=1}^t n_i \mu_i / \sqrt{N} = \sqrt{N} \bar{\mu} \text{ and } \sum_{i=2}^t v_i^2 = \sum_{i=1}^t n_i \mu_i^2 - v_1^2 = \sum_{i=1}^t n_i(\mu_i - \bar{\mu})^2$
\[ \mu_1 = \ldots = \mu_t \implies \sum_{i=2}^{t} v_i^2 = 0 \implies v_2 = \ldots = v_t = 0 \]

\[ \implies \frac{V_i}{\sigma} = U_i, \quad i = 2, \ldots, t \implies \frac{SS_{\text{Treat}}}{\sigma^2} = \sum_{i=2}^{t} U_i^2 \sim \chi^2_{t-1} \]

\[ \mu_1, \ldots, \mu_t \text{ not all equal} \implies \frac{V_i}{\sigma} = \frac{v_i}{\sigma} + U_i \quad \text{for} \quad i = 2, \ldots, t \]

and

\[ \frac{SS_{\text{Treat}}}{\sigma^2} = \sum_{i=2}^{t} (U_i + \frac{v_i}{\sigma})^2 \quad \text{with} \quad \sum_{i=2}^{t} \frac{v_i^2}{\sigma^2} = \sum_{i=1}^{t} n_i (\mu_i - \bar{\mu})^2 / \sigma^2 \]

What is the distribution of \[ \sum_{i=2}^{t} (U_i + \frac{v_i}{\sigma})^2 \]?
Suppose $X_1 \sim \mathcal{N}(d_1, 1), \ldots, X_f \sim \mathcal{N}(d_f, 1)$ are independent. Then we say that
\[ X_1^2 + \ldots + X_f^2 \sim \chi^2_f, \lambda \]
has a noncentral $\chi^2$ distribution with $f$ degrees of freedom and noncentrality parameter $\lambda = \sum_{i=1}^f d_i^2$.

$(d_1, \ldots, d_f) = (0, \ldots, 0) \implies$ previously defined (central) $\chi^2_f$ distribution.

The way the noncentral $\chi^2$ distribution is defined it seems that the $d_i$ only affect the distribution through $\lambda$. This will be shown below.

Again we employ two sets of orthonormal basis vectors, the standard ones, $\mathbf{e}_1, \ldots, \mathbf{e}_f$, and another set, $\mathbf{h}_1, \ldots, \mathbf{h}_f$, where we only specify $\mathbf{h}'_1 = (d_1, \ldots, d_f)/\sqrt{\lambda} = \mathbf{d}'/\sqrt{\lambda}$.

Note $\mathbf{h}'_1 \mathbf{h}_1 = 1$ and assume $\mathbf{h}'_i \mathbf{h}_i = 1$ for $i > 1$ and $\mathbf{h}'_i \mathbf{h}_j = 0$ for $i < j$. 
We have $X' = Z' + d' = X_1 e_1 + \ldots + X_f e_f = Y_1 h_1 + \ldots + Y_f h_f$

where $Z' = (Z_1, \ldots, Z_f)$ consists of i.i.d. $\mathcal{N}(0, 1)$ components.

Note that as before we have $\sum_{i=1}^f X_i^2 = \sum_{i=1}^f Y_i^2$.

We argued before that $Z'h_1, \ldots, Z'h_f$ are again i.i.d. $\mathcal{N}(0, 1)$

$\implies (Y_1, \ldots, Y_f) = (X'h_1, \ldots, X'h_f) = (Z'h_1, \ldots, Z'h_f) + (d'h_1, \ldots, d'h_f)$

are independent normal random variables with common variance 1

and with respective means $d'h_1 = \lambda / \sqrt{\lambda} = \sqrt{\lambda}$, and $d'h_i = 0$ for $i > 1$.

Thus the distribution of $\sum_{i=1}^f X_i^2 = \sum_{i=1}^f Y_i^2$ depends on $d$ only through $\lambda$. 
Let \( C_1 \sim \chi_{f_1, \lambda} \) be a noncentral \( \chi^2 \) random variable and let \( C_2 \sim \chi^2_{f_2} \) be a (central) \( \chi^2 \) random variable which is independent of \( C_1 \), then we say that

\[
F = \frac{C_1/f_1}{C_2/f_2} \sim F_{f_1, f_2, \lambda}
\]

has a noncentral \( F \)-distribution with \( f_1 \) and \( f_2 \) degrees of freedom and with noncentrality parameter \( \lambda \).

What does R give us?

For the noncentral \( \chi^2 \) we have: \( \text{dchisq}(x, df, ncp=0) \), \( \text{pchisq}(q, df, ncp=0) \), \( \text{qchisq}(p, df, ncp=0) \), \( \text{rchisq}(n, df, ncp=0) \)

For the noncentral \( F \) we have: \( \text{pf}(q, df1, df2, ncp=0) \)
We have established the following: $SS_E$ and $SS_{Treat}$ are independent and

$$SS_E/\sigma^2 \sim \chi^2_{N-t} \quad \text{and} \quad SS_{Treat}/\sigma^2 \sim \chi^2_{t-1,\lambda} \quad \text{with} \quad \lambda = \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu})^2/\sigma^2$$

$$\implies F = \frac{SS_{Treat}/(t-1)}{SS_E/(N-t)} \sim F_{t-1,N-t,\lambda}$$

Under $H_0: \mu_1 = \ldots = \mu_t$ this becomes the $F_{t-1,N-t}$ distribution.

We reject $H_0$ whenever $F \geq F_{t-1,N-t}(1-\alpha) = F_{crit} = qf(1-\alpha, t-1, N-t)$

which denotes the $(1-\alpha)$-quantile of the $F_{t-1,N-t}$ distribution.

Power function: $\beta(\lambda) = P(F \geq F_{t-1,N-t}(1-\alpha)) = 1 - pf(F_{crit}, t-1, N-t, \lambda)$
R’s `anova` and `lm` Applied to Flux3

```r
> SIR = c(Flux3$X, Flux3$Y, Flux3$Z)
> SIR
[1]  9.9  9.6  9.6  9.7  9.5 10.0 10.7 10.4  9.5  9.6  9.8
[12]  9.9 10.9 11.0  9.5 10.0 11.7 10.2
> FLUX = c(rep("X", 6), rep("Y", 6), rep("Z", 6))
> FLUX
[1] "X" "X" "X" "X" "X" "X" "Y" "Y" "Y" "Y" "Y" "Y" "Z" "Z"
[15] "Z" "Z" "Z" "Z"
> anova(lm(SIR ~ as.factor(FLUX))) # see ?anova & ?lm
Analysis of Variance Table

Response: SIR

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>as.factor(FLUX)</td>
<td>2</td>
<td>2.1733</td>
<td>1.0867</td>
<td>3.6452</td>
<td>0.0512</td>
</tr>
<tr>
<td>Residuals</td>
<td>15</td>
<td>4.4717</td>
<td>0.2981</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Noncentral $F$ tail probabilities are increasing in the noncentrality parameter $\lambda$.

This follows from: $|a| \nearrow \implies P(|Z + a| \geq x) \nearrow$, where $Z \sim \mathcal{N}(0, 1)$.

$$P(|Z + a| \geq x) = P(Z + a \geq x) + P(Z + a \leq -x) = 1 - \Phi(x - a) + \Phi(-x - a)$$

$$= \Phi(a - x) + \Phi(-x - a) \quad \text{with derivative}$$

$$\frac{\partial P(|Z + a| \geq x)}{\partial a} = \varphi(a - x) - \varphi(-x - a)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{a^2 + x^2}{2}\right) \left[\exp(ax) - \exp(-ax)\right] \geq 0$$

$$\implies P(F_{t-1,N-t,\lambda} \geq F_{\text{crit}}) = P\left((Z_1 + \sqrt{\lambda})^2 \geq F_{\text{crit}} \sum_{j=1}^{N-t} \tilde{Z}_j^2 \frac{t - 1}{N - t} - \sum_{i=2}^{t-1} Z_i^2\right)$$

with $Z_i, \tilde{Z}_j \sim \mathcal{N}(0, 1)$ and the monotonicity in $\lambda$ follows from the above.
The power of the ANOVA $F$-test is a monotone function of
\[ \lambda = \sum_{i=1}^{t} n_i (\mu_i - \bar{\mu})^2 / \sigma^2 \]
Thus we consider the drivers in $\lambda$.

$\lambda$ increases as $\sigma$ decreases (provided the $\mu_i$ are not all the same).

The more difference there is between the treatment means $\mu_i$ the higher $\lambda$

Increasing the sample sizes will magnify $n_i (\mu_i - \bar{\mu})^2$ (provided $\bar{\mu}$ is stable).

The sample sizes we can plan for.

Later: we can reduce $\sigma$ by blocking units into more homogeneous groups.
We have \( N \) experimental units available for testing the effects of \( t \) treatments and suppose that \( N \) is a multiple of \( t \), say \( N = r \times t \) (\( r \) and \( t \) integer).

It would seem best to use samples of equal size \( r \) for each of the \( t \) treatments i.e., we would opt for a balanced design.

That way we would not emphasize one treatment over any of the others.

Is there some other optimality criterion that could be used as justification?

We may plan for a balanced design upfront, but then something goes wrong with a few observations and they have to be discarded from analysis. Be careful that the deletion of observations does not bias any conclusions.
We may be concerned with alternatives where all means but one are the same. Since we won’t know upfront which mean sticks out, we would want to maximize the minimum power against all such contingencies. Max-Min Strategy!

If \( \mu_1 = \mu + \Delta \) and \( \mu_2 = \ldots = \mu_t = \mu \) then \( \bar{\mu} = \mu + n_1 \Delta / N \) and (algebra)

\[
\lambda_1 = \sum_{i=1}^{t} n_i (\mu_i - \bar{\mu})^2 / \sigma^2 = \frac{N \Delta^2}{\sigma^2} \frac{n_1}{N} \left( 1 - \frac{n_1}{N} \right)
\]

and similarly

\[
\lambda_i = \frac{N \Delta^2}{\sigma^2} \frac{n_i}{N} \left( 1 - \frac{n_i}{N} \right)
\]

for the other cases. It is easy to see now that for fixed \( \sigma \)

\[
\max_{n_1, \ldots, n_t} \min_{1 \leq i \leq t} \left[ \frac{N \Delta^2}{\sigma^2} \frac{n_i}{N} \left( 1 - \frac{n_i}{N} \right) \right]
\]

is achieved when \( n_1 = \ldots = n_t \). That is because

\[
\frac{n_i}{N} \left( 1 - \frac{n_i}{N} \right)
\]

increases for \( n_i / N \leq 1/2 \).
Suppose we have \( t = 3 \) treatments and want to determine the sample size \( n \) per treatment to achieve power \( \beta(\lambda) = .9 \) for level \( \alpha = .05 \).

It is desired to do this for a \( \lambda = \lambda_i \) corresponding to the alternatives on the previous slide with \( \Delta/\sigma = .5 \), i.e., with \( N = t \times n \)

\[
\lambda_i = \frac{N\Delta^2}{\sigma^2} \times \frac{n}{N} \left(1 - \frac{n}{N}\right) = n \times \frac{\Delta^2}{\sigma^2} \times \frac{t - 1}{t} = n \times \lambda_0.
\]

\( \lambda_0 = (\Delta^2/\sigma^2) \times (t - 1)/t \) can be interpreted more generally as \( \sum(\mu_i - \bar{\mu})^2/\sigma^2 \).

\[
> \text{sample.sizeANOVA()}
\]
\[
> \text{sample.sizeANOVA(nrange=30:100)}
\]
\[
> \text{sample.sizeANOVA(nrange=70:100,power0=.9)}
\]

produced the next 3 slides \( \implies n = 77 \).
\[ \Delta / \sigma = 0.5, \lambda_0 = \sum (\mu_i - \bar{\mu})^2 / \sigma^2 = 0.1667, \alpha = 0.05 \]
\[ \Delta / \sigma = 0.5 , \lambda_0 = \sum (\mu_i - \bar{\mu})^2 / \sigma^2 = 0.1667 , \alpha = 0.05 \]
\[ \Delta/\sigma = 0.5, \ \lambda_0 = \sum (\mu_i - \bar{\mu})^2 / \sigma^2 = 0.1667, \ \alpha = 0.05 \]
\[
\begin{pmatrix}
X_1 \\
X_2 \\
\vdots \\
\vdots \\
X_n
\end{pmatrix}
= \begin{pmatrix}
\bar{X} \\
\bar{X} \\
\vdots \\
\vdots \\
\bar{X}
\end{pmatrix} + \begin{pmatrix}
X_1 - \bar{X} \\
X_2 - \bar{X} \\
\vdots \\
\vdots \\
X_n - \bar{X}
\end{pmatrix}
\]

\[
\perp \text{ because } (\bar{X}, \ldots, \bar{X}) \cdot \begin{pmatrix}
X_1 - \bar{X} \\
X_2 - \bar{X} \\
\vdots \\
\vdots \\
X_n - \bar{X}
\end{pmatrix} = \bar{X} \cdot \sum_{i=1}^{n} (X_i - \bar{X}) = \bar{X} \cdot \left( \sum_{i=1}^{n} X_i - n \cdot \bar{X} \right) = 0
\]

\((\bar{X}, \ldots, \bar{X})\) varies in just one dimension, along \(1' = (1, \ldots, 1)\), and the residual vector \((X_1 - \bar{X}, \ldots, X_n - \bar{X})\) varies in its \((n - 1)\)-dimensional orthogonal complement. The \(n\) residuals thus have \(n - 1\) degrees of freedom.
Orthogonal Decomposition of Sample Vector

\[ x = (x_1, x_2) \]

\[ \bar{x} = (\bar{x}, \bar{x}) \]

\[ x - \bar{x} = (x_1 - \bar{x}, x_2 - \bar{x}) \]

Pythagoras

\[ |x|^2 = |\bar{x}|^2 + |x - \bar{x}|^2 \]

\[ \sum_i x_i^2 = \sum_i \bar{x}^2 + \sum_i (x_i - \bar{x})^2 \]

\[ = n\bar{x}^2 + \sum_i (x_i - \bar{x})^2 \]

our previous

SS decomposition
Decomposition of total dimension $N = \sum n_i$ into subspace dimensions

$$N = 1 + N - 1 = 1 + \sum (n_i - 1) + t - 1$$

$$N - t$$

$$\begin{pmatrix}
Y_{11} \\
\vdots \\
Y_{1n_1} \\
\vdots \\
Y_{t1} \\
\vdots \\
Y_{tn_t}
\end{pmatrix} = \begin{pmatrix}
\bar{Y}_{..} \\
\vdots \\
\bar{Y}_{..} \\
\vdots \\
\bar{Y}_{..} \\
\vdots \\
\bar{Y}_{..}
\end{pmatrix} + \begin{pmatrix}
Y_{11} - \bar{Y}_{..} \\
\vdots \\
Y_{1n_1} - \bar{Y}_{..} \\
\vdots \\
Y_{t1} - \bar{Y}_{..} \\
\vdots \\
Y_{tn_t} - \bar{Y}_{..}
\end{pmatrix} + \begin{pmatrix}
\bar{Y}_{1.} - \bar{Y}_{1.} \\
\vdots \\
\bar{Y}_{1.} - \bar{Y}_{1.} \\
\vdots \\
\bar{Y}_{t.} - \bar{Y}_{t.} \\
\vdots \\
\bar{Y}_{t.} - \bar{Y}_{t.}
\end{pmatrix} + \begin{pmatrix}
\bar{Y}_{..} \\
\vdots \\
\bar{Y}_{..} \\
\vdots \\
\bar{Y}_{..} \\
\vdots \\
\bar{Y}_{..}
\end{pmatrix}$$

$$\sum \sum Y_{ij}^2 = \sum \sum \bar{Y}_{..}^2 + \sum \sum (Y_{ij} - \bar{Y}_{..})^2 = \sum \sum \bar{Y}_{..}^2 + \sum \sum (Y_{ij} - \bar{Y}_{i.})^2 + \sum \sum (\bar{Y}_{i.} - \bar{Y}_{..})^2$$
A Decomposition Detail to Previous Slide

\[
\begin{pmatrix}
Y_{11} - \bar{Y}.
& \vdots
& \vdots
& \vdots
Y_{1n_1} - \bar{Y}.
& \vdots
& \vdots
Y_{t_1} - \bar{Y}.
& \vdots
Y_{t_n} - \bar{Y}.
\end{pmatrix}
= \begin{pmatrix}
Y_{11} - \bar{Y}.
& \vdots
& \vdots
\vdots
Y_{1n_1} - \bar{Y}.
& \vdots
\vdots
Y_{t_1} - \bar{Y}.
& \vdots
Y_{t_n} - \bar{Y}.
\end{pmatrix}
+ \begin{pmatrix}
\bar{Y}_1.
& \vdots
& \vdots
\vdots
\bar{Y}_{1n_1}.
& \vdots
\vdots
\bar{Y}_t.
& \vdots
\bar{Y}_{t_n}.
\end{pmatrix}
\]

\[
= \begin{pmatrix}
\bar{Y}_1.
& \vdots
& \vdots
\vdots
\bar{Y}_{1n_1}.
& \vdots
\vdots
\bar{Y}_t.
& \vdots
\bar{Y}_{t_n}.
\end{pmatrix}
\]
\[
\sum_{i} \sum_{j} \bar{Y}_{..}(\bar{Y}_{i} - \bar{Y}_{..}) = \bar{Y}_{..} \sum_{i} n_{i}(\bar{Y}_{i} - \bar{Y}_{..}) = \bar{Y}_{..} \left( \sum_{i} \sum_{j} Y_{ij} - N\bar{Y}_{..} \right) = 0
\]

\[
\sum_{i} \sum_{j} \bar{Y}_{..}(Y_{ij} - \bar{Y}_{..}) = \bar{Y}_{..} \sum_{i} (n_{i}\bar{Y}_{i} - n_{i}\bar{Y}_{..}) = 0
\]

\[
\sum_{i} \sum_{j} (\bar{Y}_{i} - \bar{Y}_{..})(Y_{ij} - \bar{Y}_{..}) = \sum_{i} (\bar{Y}_{i} - \bar{Y}_{..}) \sum_{j} (Y_{ij} - \bar{Y}_{..}) = 0
\]
Let $1'_n = (1, 1, \ldots, 1)$ denote an $n$-vector filled with 1’s. With varying $Y_{ij}$ the vectors

$$
\begin{pmatrix}
\bar{Y}_{1.} - \bar{Y}. \\
\vdots \\
\bar{Y}_{1.} - \bar{Y}. \\
\vdots \\
\bar{Y}_{t.} - \bar{Y}. \\
\vdots \\
\bar{Y}_{t.} - \bar{Y}. \\
\end{pmatrix}
= (\bar{Y}_{1.} - \bar{Y}.)
\begin{pmatrix}
1_{n1} \\
0 \\
\vdots \\
0 \\
0 \\
\end{pmatrix} + \ldots + (\bar{Y}_{t.} - \bar{Y}.)
\begin{pmatrix}
0 \\
0 \\
\vdots \\
0 \\
1_{nt} \\
\end{pmatrix}
$$

span a $(t - 1)$-dimensional subspace of $R^N$ orthogonal to $1'_N = (1'_{n1}, \ldots, 1'_{nt})$.

Because of the previously shown orthogonality the vectors

$$(Y_{11} - \bar{Y}_{1.}, \ldots, Y_{1n1} - \bar{Y}_{1.}, \ldots, Y_{t1} - \bar{Y}_{t.}, \ldots, Y_{tn1} - \bar{Y}_{t.})$$

span an $(N - t)$-dimensional subspace of $R^N$. 
Orthogonal Decomposition of Sample Space

\[ |(Y_{11}, \ldots, Y_{tn})|^2 = |(\bar{Y}_., \ldots, \bar{Y}_.)|^2 + |(Y_{11} - \bar{Y}_1., \ldots, Y_{tn} - \bar{Y}_t.)|^2 + |(\bar{Y}_1. - \bar{Y}_., \ldots, \bar{Y}_t. - \bar{Y}_.)|^2 \]
In order to understand the blood coagulation behavior in relation to various diets, lab animals were given 4 different diets and their subsequent blood draws were then measured for their respective coagulation times in seconds.

The lab animals were assigned randomly to the various diets.

The results were as follows:

```r
> ctime
 [1]  59  60  62  63  63  64  65  66  67  71  66  67  68  68  68  71  56  59
 [19]  60  61  62  63  63  64
> diet
 [1] "A" "A" "A" "A" "B" "B" "B" "B" "B" "B" "C" "C" "C" "C"
 [14] "C" "C" "C" "D" "D" "D" "D" "D" "D" "D" "D" "D"
```
Plot for Coagulation Example

N_A = 4
N_B = 6
N_C = 6
N_D = 8

coagulation time (sec)

50 55 60 65 70 75 80

diet

A
B
C
D
Note that in the previous plot we used `jitter(ctime)` to plot `ctime` in the vertical direction and to plot its horizontal mean lines. This perturbs tied observations a small random amount to make tied observations more visible. For example, the mean lines for diet A and D would have been the same otherwise.

```r
> anova(lm(ctime~as.factor(diet)))
Analysis of Variance Table

Response: ctime

                      Df Sum Sq Mean Sq F value Pr(>F)  
as.factor(diet)      3  228.0   76.00 13.571  4.66e-05 ***
Residuals            20  112.0    5.60          
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
> out=lm(ctime~as.factor(diet))
> names(out)
[1] "coefficients" "residuals"  "effects"
[4] "rank"             "fitted.values" "assign"
[7] "qr"              "df.residual"  "contrasts"
[10] "xlevels"         "call"       "terms"
[13] "model"
> out$coefficients
(Intercept) as.factor(diet)B as.factor(diet)C
  6.100000e+01  5.000000e+00  7.000000e+00
as.factor(diet)D
  -1.095919e-14

Note that these are the estimates $\hat{\mu}_A, \hat{\mu}_B - \hat{\mu}_A, \hat{\mu}_C - \hat{\mu}_A, \hat{\mu}_D - \hat{\mu}_A$. 
Residuals from \texttt{lm} for Coagulation Example

\begin{verbatim}
> out$residuals
1             2             3             4
-2.000000e+00 -1.000000e+00  1.000000e+00  2.000000e+00
5             6             7             8
-3.000000e+00 -2.000000e+00 -1.000000e+00  1.111849e-16
9            10            11            12
  1.000000e+00  5.000000e+00 -2.000000e+00 -1.000000e+00
13           14            15            16
-5.534852e-17 -5.534852e-17 -5.534852e-17  3.000000e+00
17           18            19            20
-5.000000e+00 -2.000000e+00 -1.000000e+00 -1.663708e-16
21           22            23            24
  1.000000e+00  2.000000e+00  2.000000e+00  3.000000e+00
\end{verbatim}

Numbers such as \(-5.534852e-17\) should be treated as 0 (computing quirks).
> out$fitted.values
  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
 61  61  61  61  66  66  66  66  66  68  68  68  68  68  68  61  61  61
 20 21 22 23 24
 61  61  61  61  61
Randomization Test for Coagulation Example

Simulated Randomization Distribution

F-equivalent test statistic 98532
p-value = 7e-05
based on 1e+05 simulations
Simulated Randomization Distribution

F-statistic 13.571
p-value = 7e-05
p-value = 4.658e-05 from F-distribution based on 1e+05 simulations
When the hypothesis $H_0 : \mu_1 = \ldots = \mu_t$ is not rejected at level $\alpha$ then there is little purpose in looking closer at differences between the sample means $\bar{Y}_i$ for the various treatments.

Any such perceived differences could easily have come about by simple random variation, even when the hypothesis is true.

Why then read something into randomness? It is like reading tea leaves!

However, when the hypothesis is rejected it is quite natural to ask in which way the hypothesis was contradicted.

The best indicators for any analysis as to how the means $\mu_i$ may be different would be the sample or treatment means $\hat{\mu}_i = \bar{Y}_i$, $i = 1, \ldots, t$. 
A first step in understanding differences in the $\mu_i$ is to look at their estimates $\hat{\mu}_i = \bar{Y}_i$.

We should do this in the context of the sampling variability of $\hat{\mu}_i$.

In the past we addressed this via confidence intervals for $\mu_i$ based on $\hat{\mu}_i$.

In any such confidence interval we can now use the pooled variance $s^2$ from all $t$ samples and not just the variance $s^2_i$ from the $i^{th}$ sample, i.e. we get

$$\hat{\mu}_i \pm t_{1-\alpha/2, N-t} \times \frac{s}{\sqrt{n_i}}$$

as our $100(1 - \alpha)\%$ confidence interval for $\mu_i$.

This follows as before (exercise) from the independence of $\hat{\mu}_i$ and $s$, the fact that $(\hat{\mu}_i - \mu_i)/(\sigma/\sqrt{n_i}) \sim \mathcal{N}(0, 1)$, and from $s^2/\sigma^2 \sim \chi^2_{N-t}/(N-t)$.

The validity of this improvement ($N-t \gg n_i - 1$ when using $s^2$ instead of $s^2_i$) depends strongly on the assumption that the population variances $\sigma^2$ behind all $t$ samples are the same, or at least approximately so.
Suppose \( \hat{\theta} \) is an estimator for a parameter \( \theta \) of interest. We denote by 
\[
\sigma^2_{\hat{\theta}} = \text{var}(\hat{\theta}) = g(\theta, \psi)
\]
its sampling variance and by 
\[
\sigma_{\hat{\theta}} = \sqrt{g(\theta, \psi)}
\]
its sampling standard deviation.

The estimated sampling standard deviation of \( \hat{\theta} \), i.e., 
\[
\hat{\sigma}_{\theta} = \sqrt{g(\hat{\theta}, \hat{\psi})},
\]
is also called the standard error of \( \hat{\theta} \) and is denoted by \( SE(\hat{\theta}) \).

Example 1: \( \hat{\mu} = \bar{X} \) as estimate of \( \mu \) has variance 
\[
\text{var}(\hat{\mu}) = \sigma^2 / n \Rightarrow SE(\hat{\mu}) = s / \sqrt{n}.
\]
Example 2: \( s^2 \sim \sigma^2 \chi^2_{n-1} / (n - 1) \) as estimate of \( \sigma^2 \) has sampling variance

\[
\text{var}(s^2) = \frac{\sigma^4 2(n - 1)}{(n - 1)^2} = \frac{2\sigma^4}{n - 1} \implies SE(s^2) = s^2 \sqrt{\frac{2}{n - 1}}
\]

Note the different roles of \((\theta, \psi)\) in these two examples.

In Example 1: \( \theta = \mu \) and \( \psi = \sigma^2 \) and we only use \( \hat{\psi} \) in \( SE(\hat{\theta}) \).

In Example 2: \( \theta = \sigma^2 \) and there is no \( \psi \). We only use \( \hat{\theta} \) in \( SE(\hat{\theta}) \).
If \( \hat{\theta} \) has an approximate normal distribution with mean \( \theta \) and standard deviation \( \sigma_{\hat{\theta}} \), i.e.,

\[
\hat{\theta} \approx \mathcal{N}(\theta, \sigma^2_{\hat{\theta}}) \approx \mathcal{N}(\theta, SE^2(\hat{\theta}))
\]

\[\implies \hat{\theta} \pm 2 \times SE(\hat{\theta}) \text{ is an approximately 95\% confidence interval for } \theta\]

because \( z_{.975} = \text{qnorm}(.975) = 1.959964 \approx 2 \).

This works especially well for Student-\( t \) based intervals

\[
\bar{\mu}_i \pm t_{.975, f} \times \frac{s}{\sqrt{f}} = \bar{Y}_i \pm t_{.975, N-t} \times \frac{s}{\sqrt{N-t}}
\]

because \( t_{.975, f} \approx z_{.975} \) for large \( f \), see next slide.
$t_{0.975, f} \rightarrow z_{0.975} = 1.96 \approx 2$
Why should the rule of thumb work for $s^2$ as estimator of $\sigma^2$?

Recall: $s^2 \sim \sigma^2 \chi^2_{n-1}/(n-1)$. CLT $\implies$ approximate normality for $s^2$ since

$$\frac{(n-1)s^2}{\sigma^2} = \chi^2_{n-1} = \sum_{i=1}^{n-1} Z_i^2 \approx \mathcal{N}(n - 1, 2(n - 1)) \implies s^2 \approx \mathcal{N} \left( \sigma^2, 2\sigma^4/(n - 1) \right)$$

$$\implies s^2 \pm 2 \times SE(s^2) = s^2 \pm 2 \times s^2 \sqrt{\frac{2}{n-1}}$$

since $SE(s^2) = s^2 \sqrt{2/(n - 1)}$ is the estimate of $\sigma^2 \sqrt{2/(n - 1)}$,

the sampling standard deviation of $s^2$. 

58
Although for testing $H_0 : \mu_1 = \mu_2 = \mu_3$ in the case of the Flux3 data the p-value of $0.05126$ was not significant at level $\alpha = 0.05$ we illustrate the concepts of the different types of confidence intervals for the means.

<table>
<thead>
<tr>
<th>Flux</th>
<th>$\hat{\mu}_i$</th>
<th>$s_i$</th>
<th>$s$</th>
<th>95% intervals using $s_i$</th>
<th>95% intervals using $s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>9.717</td>
<td>0.194</td>
<td>0.546</td>
<td>[9.513, 9.920]</td>
<td>[9.242, 10.192]</td>
</tr>
<tr>
<td>Y</td>
<td>9.983</td>
<td>0.471</td>
<td>0.546</td>
<td>[9.489, 10.477]</td>
<td>[9.508, 10.458]</td>
</tr>
<tr>
<td>Z</td>
<td>10.550</td>
<td>0.797</td>
<td>0.546</td>
<td>[9.714, 11.386]</td>
<td>[10.075, 11.025]</td>
</tr>
</tbody>
</table>
Plots of Confidence Intervals for Flux3 Data

8.5 9.0 9.5 10.0 10.5 11.0 11.5 12.0

- Using pooled $s^2 = \sum_{i=1}^{t} s_i^2(n_i - 1)/(N - t)$
- Using individual $s_i^2$
For testing $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ in the case of the coagulation data the p-value of $4.7 \cdot 10^{-5}$ is highly significant. We again illustrate the concepts of the different types of confidence intervals for the means.

<table>
<thead>
<tr>
<th>Diet</th>
<th>$\hat{\mu}_i$</th>
<th>$s_i$</th>
<th>$s$</th>
<th>95% intervals using $s_i$</th>
<th>95% intervals using $s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>61</td>
<td>1.9</td>
<td>2.2</td>
<td>[57.9, 64.1]</td>
<td>[58.7, 63.3]</td>
</tr>
<tr>
<td>B</td>
<td>66</td>
<td>2.1</td>
<td>2.2</td>
<td>[63.8, 68.2]</td>
<td>[64.1, 67.9]</td>
</tr>
<tr>
<td>C</td>
<td>68</td>
<td>1.5</td>
<td>2.2</td>
<td>[66.4, 69.6]</td>
<td>[66.1, 69.9]</td>
</tr>
<tr>
<td>D</td>
<td>61</td>
<td>2.6</td>
<td>2.2</td>
<td>[58.8, 63.2]</td>
<td>[59.4, 59.4]</td>
</tr>
</tbody>
</table>
Plots of Confidence Intervals for Coagulation Data

Applied Statistics and Experimental Design

Fritz Scholz — Fall 2006

using pooled $s^2 = \sum_{i=1}^{t} s_i^2 (n_i - 1) / (N - t)$
using individual $s_i^2$
When constructing intervals of the type:

\[ \hat{\mu}_i \pm t_{1-\alpha/2} \frac{s}{\sqrt{n_i}} \quad \text{or} \quad \hat{\mu}_i \pm t_{1-\alpha/2} \frac{s_i}{\sqrt{n_i}} \quad \text{for } i = 1, \ldots, t \]

we should be aware that these intervals don’t simultaneously cover their respective targets \( \mu_i \) with probability \( 1 - \alpha \). They do so individually. For example

\[
\begin{align*}
P \left( \mu_i \in \hat{\mu}_i \pm t_{1-\alpha/2} \frac{s_i}{\sqrt{n_i}}, \ i = 1, \ldots, t \right) &= \prod_{i=1}^{t} P \left( \mu_i \in \hat{\mu}_i \pm t_{1-\alpha/2} \frac{s_i}{\sqrt{n_i}} \right) \\
&= (1 - \alpha)^t < 1 - \alpha.
\end{align*}
\]

Thus we should choose \( \alpha^* \) for individual intervals to get

\[
(1 - \alpha^*)^t = 1 - \alpha \quad \text{or} \quad \alpha^* = 1 - (1 - \alpha)^{1/t} \approx \frac{\alpha}{t} = \tilde{\alpha}_t.
\]

Some problem remains when using common pooled estimate \( s \). No independence!
\[ \alpha^* = 1 - (1 - \alpha)^{1/t} \approx \frac{\alpha}{t} \]

\[ \alpha_t^* = 1 - (1 - \alpha)^{(1/t)} \]

\[ \tilde{\alpha}_t = \frac{\alpha}{t} \]

\[ t = 3 \]

\[ t = 5 \]

\[ t = 6 \]

\[ t = 10 \]
When we use a common pooled estimate $s$ for the standard deviation $\sigma$
the previous confidence intervals are no longer independent.

However, it can be shown that

$$P\left(\mu_i \in \hat{\mu}_i \pm t_{1-\alpha^*/2} \frac{s}{\sqrt{n_i}}, \ i = 1, \ldots, t\right) \geq \prod_{i=1}^{t} P\left(\mu_i \in \hat{\mu}_i \pm t_{1-\alpha^*/2} \frac{s}{\sqrt{n_i}}\right)$$

$$= \left(1 - \alpha^*\right)^t = 1 - \alpha$$

This comes from the positive dependence between confidence intervals through $s$,
i.e., if one interval is more (less) likely to cover its target $\mu_i$ due to $s$, so are the other
intervals more (less) likely to cover their targets $\mu_j$.

Using the same compensation as in the independence case would let us err on the
conservative side, i.e., give us higher confidence than the targeted $1 - \alpha$. 

65
Boole’s and Bonferroni’s Inequality

For any \( m \) events \( E_1, \ldots, E_m \) Boole’s inequality states

\[
P(E_1 \cup \ldots \cup E_m) \leq P(E_1) + \ldots + P(E_m)
\]

For any \( m \) events \( E_1, \ldots, E_m \) Bonferroni’s inequality states

\[
P(E_1 \cap \ldots \cap E_m) \geq 1 - \sum_{i=1}^{m} (1 - P(E_i))
\]

The statement are equivalent, since \( P(E_1 \cup \ldots \cup E_m) = 1 - P(E_1^c \cap \ldots \cap E_m^c) \).

If \( E_i \) denotes the \( i^{th} \) coverage event \( \left\{ \mu_i \in \hat{\mu}_i \pm t_1 - \bar{\alpha}/2 \frac{s}{\sqrt{n_i}} \right\} \) with \( P(E_i) = 1 - \bar{\alpha} \), then the simultaneous coverage probability is bounded from below as follows

\[
P \left( \bigcap_{i=1}^{t} E_i \right) \geq 1 - \sum_{i=1}^{t} (1 - P(E_i)) = 1 - t\bar{\alpha} = 1 - \alpha \quad \text{if} \quad \bar{\alpha} = \bar{\alpha}_t = \alpha / t,
\]

i.e., we can achieve at least \( 1 - \alpha \) probability coverage by choosing the individual coverage appropriately, namely \( 1 - \bar{\alpha} = 1 - \alpha / t \). Almost same adjustment.
Decomposing the Mean Vector $\mu$

Variation in the means $\mu_i$ is best understood through the familiar decomposition:

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_1 \\ \vdots \\ \mu_t \\ \vdots \\ \mu_t \end{pmatrix} = \bar{\mu} \cdot 1_N + \begin{pmatrix} \mu_1 - \bar{\mu} \\ \vdots \\ \mu_1 - \bar{\mu} \\ \vdots \\ \mu_t - \bar{\mu} \\ \vdots \\ \mu_t - \bar{\mu} \end{pmatrix}$$

The two vectors on the right are orthogonal to each other, with the first vector representing the projection of $\mu$ onto $1_N$ (with all components equal to $\bar{\mu}$) and the second representing the projection of $\mu$ onto $(t - 1)$-dimensional subspace of the $(N - 1)$-dimensional orthogonal complement to $1_N$.

It is this second vector that captures all aspects of variation in $\mu$. 

Any linear function of the distinct components \((\mu_1 - \bar{\mu}, \ldots, \mu_t - \bar{\mu})\) has to be of the form \(C = \sum_{i=1}^{t} c_i \mu_i\) with \(\sum_{i=1}^{t} c_i = 0\).

\[
\sum_{i=1}^{t} a_i (\mu_i - \bar{\mu}) = \sum_{i=1}^{t} a_i \mu_i - \sum_{i=1}^{t} a_i \sum_{j=1}^{t} \frac{n_j}{N} \mu_j = \sum_{i=1}^{t} a_i \mu_i - \sum_{j=1}^{t} a_j \sum_{i=1}^{t} \frac{n_i}{N} \mu_i
\]

\[
= \sum_{i=1}^{t} a_i \mu_i - \sum_{i=1}^{t} \frac{n_i}{N} \mu_i \sum_{j=1}^{t} a_j = \sum_{i=1}^{t} c_i \mu_i \quad \text{with} \quad c_i = a_i - \frac{n_i}{N} \sum_{j=1}^{t} a_j
\]

where \(\sum_{i=1}^{t} c_i = \sum_{i=1}^{t} a_i - \sum_{i=1}^{t} \frac{n_i}{N} \sum_{j=1}^{t} a_j = \sum_{i=1}^{t} a_i - \sum_{j=1}^{t} a_j = 0\).

Such a function \(C = \sum_{i=1}^{t} c_i \mu_i\) of the \(\mu_i\), with \(\sum_{i=1}^{t} c_i = 0\), is called a contrast.
Suppose we have 4 treatments with respective means $\mu_1, \ldots, \mu_4$.

We may be interested in contrasts of the following form $C_{12} = \mu_1 - \mu_2$
with $c' = (c_1, \ldots, c_4) = (1, -1, 0, 0)$. Similarly for the other differences
$C_{ij} = \mu_i - \mu_j$. There are $\binom{4}{2} = 6$ such contrasts.

Sometimes one of the treatments, say the first, is singled out as the control.

We may then be interested in just the 3 contrasts $C_{12}, C_{13}$ and $C_{14}$ or we may be interested in $C_{1.234} = \mu_1 - (\mu_2 + \mu_3 + \mu_4)/3$ with $c' = (1, -1/3, -1/3, -1/3)$.

Sometimes the first 2 treatment share something in common and so do the last 2.

One might then try: $C_{12.34} = (\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4)/2$ with $c = (1/2, 1/2, -1/2, -1/2)$.
A natural estimate of $C = \sum_{i=1}^{t} c_i \mu_i$ is $\hat{C} = \sum_{i=1}^{t} c_i \bar{Y}_i$.

We have
\[
E(\hat{C}) = E\left(\sum_{i=1}^{t} c_i \bar{Y}_i\right) = \sum_{i=1}^{t} c_i E(\bar{Y}_i) = \sum_{i=1}^{t} c_i \mu_i = C
\]

and
\[
\text{var}(\hat{C}) = \text{var}\left(\sum_{i=1}^{t} c_i \bar{Y}_i\right) = \sum_{i=1}^{t} c_i^2 \text{var}(\bar{Y}_i) = \sum_{i=1}^{t} c_i^2 \sigma^2 / n_i.
\]

Under the normality assumption for the $Y_{ij}$ we have
\[
\frac{\hat{C} - C}{s \sqrt{\sum_{i=1}^{t} c_i^2 / n_i}} \sim t_{N-t} \quad \text{where} \quad s^2 = \frac{\sum_{i=1}^{t}(n_i - 1)s_i^2}{N - t} = \frac{\sum_{ij}(Y_{ij} - \bar{Y}_i)^2}{N - t} = M_{SE}.
\]

\[
\implies \hat{C} \pm t_{N-t, 1-\alpha/2} \times s \times \sqrt{\sum_{i=1}^{t} c_i^2 / n_i} \quad \text{is a } 100(1 - \alpha)\% \text{ confidence interval for } C.
\]
Based on the duality of testing and confidence intervals we can test the hypothesis $H_0 : C = 0$ by rejecting it whenever the previous confidence interval does not contain $C = 0$.

Similarly, reject $H_0 : C = C_0$ by rejecting it whenever the previous confidence interval does not contain $C = C_0$.

Another notation for this interval is $\hat{C} \pm t_{N-t, 1- \alpha/2} \times SE(\hat{C})$ where

$$SE(\hat{C}) = s \times \sqrt{\sum_{i=1}^{t} \frac{c_i^2}{n_i}}.$$

$SE(\hat{C})$ is the standard error of $\hat{C}$, the estimate of the standard deviation of $\hat{C}$. 
After rejecting $H_0 : \mu_1 = \ldots = \mu_t$ one is often interested in looking at all $\binom{t}{2}$ pairwise contrasts $C_{ij} = \mu_i - \mu_j$. The following procedure is referred to as Fisher’s Protected Least Significant Difference (LSD) Method.

It consists of possibly two stages:

1) Perform $\alpha$ level $F$-test for testing $H_0$. If $H_0$ is not rejected, stop.

2) If $H_0$ is rejected, form all $\binom{t}{2} (1 - \alpha)$-level confidence intervals for $C_{ij} = \mu_i - \mu_j$:

$$\hat{I}_{ij} = \hat{\mu}_i - \hat{\mu}_j \pm t_{N-t,1-\alpha/2} \times s \times \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

and declare all $\mu_i - \mu_j \neq 0$ for which $\hat{I}_{ij}$ does not contain zero.
If $H_0$ is true, the chance of making any statements contradicting $H_0$ is at most $\alpha$. This is the protected aspect of this procedure.

However, when $H_0$ is not true there are many possible contingencies, some of which can give us a higher than desired chance of pronouncing a significant difference, when in fact there is none.

E.g., if all but one mean (say $\mu_1$) are equal and $\mu_1$ is far away from $\mu_2 = \ldots = \mu_t$ our chance of rejecting $H_0$ is almost 1.

However, among the intervals for $\mu_i - \mu_j$, $2 \leq i < j$ we may find a significantly higher than $\alpha$ proportion of cases with wrongly declared differences.

This is due to the multiple comparison issue.
The Tukey-Kramer method is based on the distribution of

\[ Q_{t,f} = \max_{1 \leq i < j \leq t} \left\{ \frac{|Z_i - Z_j|}{s} \right\} \]

where \( Z_1, \ldots, Z_t \sim \mathcal{N}(0, 1) \) and \( f \times s^2 \sim \chi^2_f \).

Its cdf and quantile function are given in R as `ptukey(q, nmeans, df)` and `qtukey(p, nmeans, df)`, \( n \text{means} = t \) is the number of means, \( df = f = N - t \) denotes the degrees of freedom in \( s \).

Applying this to \( Z_i = (\hat{\mu}_i - \mu_i) / (\sigma / \sqrt{n}) \) and assuming \( n_1 = \ldots = n_t = n \) we get

\[
\max_{i < j} \left\{ \frac{\sqrt{n} |\hat{\mu}_i - \hat{\mu}_j - (\mu_i - \mu_j)|}{s} \right\} = \max_{i < j} \left\{ \frac{|\hat{\mu}_i - \mu_i - \hat{\mu}_j + \mu_j|}{\sigma / \sqrt{n}} \right\} = Q_{t,f}
\]

\[
P(\mu_i - \mu_j \in \hat{\mu}_i - \hat{\mu}_j \pm q_{t,f,1-\alpha} \ s / \sqrt{n} \ \forall \ i < j) = 1 - \alpha
\]

simultaneous \((1 - \alpha)\)-coverage confidence intervals.

Here \( P(Q_{t,f} \leq q_{t,f,1-\alpha}) = 1 - \alpha \) or \( q_{t,f,1-\alpha} = \text{qtukey}(1 - \alpha, t, f) \).
The simultaneous intervals for all pairwise mean differences was due to Tukey, but it was hampered by the requirement of equal sample sizes.

This was addressed by Kramer in the following way. In the above confidence intervals replace $n$ in $1/\sqrt{n} = \sqrt{1/n}$ by $n^*_{ij}$, where $n^*_{ij}$ is the harmonic mean of $n_i$ and $n_j$, i.e., $1/n^*_{ij} = (1/n_i + 1/n_j)/2$. Different adjustment for each pair $(i, j)$!

It was possible to show

$$P \left( \mu_i - \mu_j \in \hat{\mu}_i - \hat{\mu}_j \pm Q_{t,f,1-\alpha} \frac{s}{\sqrt{n^*_{ij}}} \quad \forall \quad i < j \right) \geq 1 - \alpha$$

simultaneous confidence intervals with coverage $\geq 1 - \alpha$. 
Tukey-Kramer Method for Coagulation Data

coag.tukey = function (alpha=.05)
{
  diets=unique(diet)
  mu.vec=NULL
  nvec=NULL
  mean.vec=NULL
  for(i in 1:length(diets)){
    mu.vec=c(mu.vec,mean(ctime[diet==diets[i]]))
    nvec=c(nvec,length(ctime[diet==diets[i]]))
    mean.vec=c(mean.vec,rep(mu.vec[i],nvec[i]))
  }
  tr=length(nvec)
  N=sum(nvec)
  MSE=sum((ctime-mean.vec)^2/(N-tr))
s=sqrt(MSE)
intervals=NULL
for(i in 1:3){
    for(j in (i+1):4){
        nijstar=1/(.5*(1/nvec[i]+1/nvec[j]))
        qTK=qtukey(1-alpha,tr,N-tr)
        Diff=mu.vec[i]-mu.vec[j]
        lower=Diff - qTK*s/sqrt(nijstar)
        upper=Diff + qTK*s/sqrt(nijstar)
        intervals=rbind(intervals,c(lower,upper))
    }
}
intervals
Tukey-Kramer Results for Coagulation Data

> coag.tukey()

    [,1]      [,2]
[1,]  -9.275446 -0.7245544
[2,] -11.275446 -2.7245544
[3,]  -4.056044  4.0560438
[4,]  -5.824075  1.8240748
[5,]   1.422906  8.5770944
[6,]   3.422906 10.5770944

Declare significant differences in $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, $\mu_2 - \mu_4$, and $\mu_3 - \mu_4$. 
Scheffé’s Confidence Intervals for All Contrasts

Scheffé took the $F$-test for testing $H_0 : \mu_1 = \ldots = \mu_t$ and converted it into a simultaneous coverage statement about confidence intervals for all contrasts $c' = (c_1, \ldots, c_t)$:

$$P\left(\sum_{i=1}^{t} c_i \mu_i \in \sum_{i=1}^{t} c_i \hat{\mu}_i \pm \sqrt{(t-1) \cdot F_{t-1,N-t,1-\alpha} \times s \times \left(\sum_{i=1}^{t} \frac{c_i^2}{n_i}\right)^{1/2}} \quad \forall \ c\right) = 1 - \alpha$$

This is a coverage statement about an infinite number of contrasts, but can be applied conservatively to all pairwise contrasts. The resulting intervals tend to be quite conservative.

But it compares well with Bonferroni type intervals if applied to many contrasts.
Pairwise Comparison Intervals for Coagulation Data

<table>
<thead>
<tr>
<th>mean difference</th>
<th>Tukey-Kramer</th>
<th>Fisher’s protected LSD</th>
<th>Bonferroni inequality</th>
<th>Scheffé’s all contrasts method</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1 - \mu_2$</td>
<td>-9.28 -0.72</td>
<td>-8.19 -1.81</td>
<td>-9.47 -0.53</td>
<td>-9.66 -0.34</td>
</tr>
<tr>
<td>$\mu_1 - \mu_3$</td>
<td>-11.28 -2.72</td>
<td>-10.19 -3.81</td>
<td>-11.47 -2.53</td>
<td>-11.66 -2.34</td>
</tr>
<tr>
<td>$\mu_1 - \mu_4$</td>
<td>-4.06 4.06</td>
<td>-3.02 3.02</td>
<td>-4.24 4.24</td>
<td>-4.42 4.42</td>
</tr>
<tr>
<td>$\mu_2 - \mu_3$</td>
<td>-5.82 1.82</td>
<td>-4.85 0.85</td>
<td>-6.00 2.00</td>
<td>-6.17 2.17</td>
</tr>
<tr>
<td>$\mu_2 - \mu_4$</td>
<td>1.42 8.58</td>
<td>2.33 7.67</td>
<td>1.26 8.74</td>
<td>1.10 8.90</td>
</tr>
<tr>
<td>$\mu_3 - \mu_4$</td>
<td>3.42 10.58</td>
<td>4.33 9.67</td>
<td>3.26 10.74</td>
<td>3.10 10.90</td>
</tr>
</tbody>
</table>

Declare significant differences in $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, $\mu_2 - \mu_4$, and $\mu_3 - \mu_4$, using any of the four methods.
Simultaneous Paired Comparisons (95%)

Pairwise Comparisons of Means (Coagulation Data): $1 - \alpha = 0.95$

- Tukey–Kramer pairwise comparisons
- Fisher's protected LSD
- Bonferroni intervals
- Scheffe's intervals for all contrasts

Graph showing comparisons between means $\mu_1$, $\mu_2$, $\mu_3$, and $\mu_4$. The x-axis represents the differences $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, $\mu_1 - \mu_4$, $\mu_2 - \mu_3$, $\mu_2 - \mu_4$, and $\mu_3 - \mu_4$, and the y-axis represents the value differences ranging from -15 to 15.
Simultaneous Paired Comparisons (99%) 

Pairwise Comparisons of Means (Coagulation Data): $1 - \alpha = 0.99$

- Tukey–Kramer pairwise comparisons
- Fisher's protected LSD
- Bonferroni intervals
- Scheffe's intervals for all contrasts
All $\binom{t}{2}$ pairwise comparisons for $\mu_i - \mu_j$ could by very many and simultaneous intervals would become quite conservative.

Since all these contrasts span a $(t - 1)$-dimensional space one should be able to capture all differences with just $t - 1$ orthogonal contrasts.

$$C_1 = \sum_{i=1}^{t} c_1i\mu_i \quad \perp \quad C_2 = \sum_{i=1}^{t} c_2i\mu_i \iff \sum_{i=1}^{t} c_1i c_2i/n_i = 0 \quad (\text{wrong def. in Montgomery p.91})$$

$$C_1 \perp C_2 \implies \text{cov}(\hat{C}_1, \hat{C}_2) = \sum_{i=1}^{t} \sum_{j=1}^{t} c_1i c_2j \text{cov}(\hat{\mu}_i, \hat{\mu}_j) = \sum_{i=1}^{t} c_1i c_2i\sigma^2/n_i = 0 ,$$

i.e., $\hat{C}_1$ and $\hat{C}_2$ are independent and simultaneous statements for $C_1, C_2, \ldots$ are easier to handle, just as before when making simultaneous intervals for $\mu_1, \ldots, \mu_t$ based on independent $\hat{\mu}_1, \ldots, \hat{\mu}_t$.

The trick is to have meaningful or interpretable orthogonal contrast.
Suppose we have $t = 3$ treatments of which the third is a control, i.e., we are familiar with its performance.

Assume further that we have a balanced design, i.e., $n_1 = n_2 = n_3$.

We could try the following $t - 1 = 2$ orthogonal contrasts:

$c'_1 = (.5, .5, -1)$ and $c'_2 = (1, -1, 0)$.

Note that $C_1 = (\mu_1 + \mu_2)/2 - \mu_3$ and $C_2 = \mu_1 - \mu_2$, of which the first assesses how much the average mean of the two new treatments differs from the control mean and the second assesses the difference between the two new treatments. These are seemingly “orthogonal” issues.
We have an unbalanced design, i.e., \( n_1, n_2, n_3 \) may be different.

Then the following \( t - 1 = 2 \) vectors:

\[
c'_1 = \left( \frac{n_1}{n_1 + n_2}, \frac{n_2}{n_1 + n_2}, -1 \right) \quad \text{and} \quad c'_2 = (1, -1, 0)
\]

are indeed contrast vectors:

\[
\frac{n_1}{n_1 + n_2} + \frac{n_2}{n_1 + n_2} - 1 = 0 \quad \text{and} \quad 1 - 1 + 0 = 0
\]

and they are orthogonal:

\[
\frac{n_1}{[(n_1 + n_2)n_1]} - \frac{n_2}{[(n_1 + n_2)n_2]} - 1 \cdot 0/n_3 = 0.
\]

\[\Rightarrow C_1 = \frac{(n_1 \mu_1 + n_2 \mu_2)}{(n_1 + n_2)} - \mu_3 = \bar{\mu}_{12} - \mu_3 \quad \text{and} \quad C_2 = \mu_1 - \mu_2,
\]

of which the first assesses how much the weighted average mean of the two new treatments differs from the control mean and the second assesses the difference between the two new treatments.

These are seemingly “orthogonal” issues.
### Service Center Data

**Applied Statistics and Experimental Design**

Fritz Scholz — Fall 2006

#### Table:

<table>
<thead>
<tr>
<th># of persons on call</th>
<th># of calls processed per hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.7 2.7 2.5 1.9</td>
</tr>
<tr>
<td>3</td>
<td>4.5 3.5 4.7 5.4</td>
</tr>
<tr>
<td>4</td>
<td>4.7 4.8 5.6 5.1</td>
</tr>
<tr>
<td>5</td>
<td>6.3 5.2 6.6 4.9</td>
</tr>
<tr>
<td>7</td>
<td>6.3 5.7 6.1 6.1</td>
</tr>
</tbody>
</table>

#### Graph:

- **X-axis**: Number of people on call
- **Y-axis**: Number of calls processed per hour

The graph shows a scatter plot with data points representing the relationship between the number of people on call and the number of calls processed per hour. The data points are plotted on a Cartesian coordinate system.
Here we have a new type of treatment (number of persons on call), where the different treatment levels are scalar and not just qualitative.

In such situations the following orthogonal contrasts are of practical interest:

<table>
<thead>
<tr>
<th></th>
<th>$c_{i1}$</th>
<th>$c_{i2}$</th>
<th>$c_{i3}$</th>
<th>$c_{i4}$</th>
<th>$c_{i5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1 = \sum_{j=1}^{5} c_{1j} \mu_j$</td>
<td>-2</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>$C_2 = \sum_{j=1}^{5} c_{2j} \mu_j$</td>
<td>2</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>2</td>
</tr>
<tr>
<td>$C_3 = \sum_{j=1}^{5} c_{3j} \mu_j$</td>
<td>-1</td>
<td>2</td>
<td>0</td>
<td>-2</td>
<td>1</td>
</tr>
<tr>
<td>$C_4 = \sum_{j=1}^{5} c_{4j} \mu_j$</td>
<td>1</td>
<td>-4</td>
<td>6</td>
<td>-4</td>
<td>1</td>
</tr>
</tbody>
</table>

For what kind of mean patterns in $\mu_1, \ldots, \mu_5$ would $|C_i|$ and consequently $|\hat{C}_i|$ be large?
$C_i = \sum_{j=1}^{5} c_{i,j} \times j$ using $\mu_j = j$
The previous plot suggests that a pattern in the means \( \mu_j \) in relation to \( j = 1, \ldots, 5 \) that correlates most strongly with the corresponding pattern in the plot should yield a high value for the corresponding absolute contrast \( |C_i| \).

Thus a large value \( |C_1| \) indicates a strong linear component in the mean pattern.

A large value \( |C_2| \) indicates a strong quadratic component in the mean pattern.

A large value \( |C_3| \) indicates a strong cubic component in the mean pattern.

A large value \( |C_4| \) indicates a strong quartic component in the mean pattern.

Typically, one hopes to rule out some (if not all) of the latter possibilities.
Simultaneous Contrast Intervals

<table>
<thead>
<tr>
<th></th>
<th>95%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1$</td>
<td>[6.27, 11.58]</td>
<td>[5.53, 12.32]</td>
</tr>
<tr>
<td>$C_2$</td>
<td>[-7.02, -0.73]</td>
<td>[-7.89, 0.14]</td>
</tr>
<tr>
<td>$C_3$</td>
<td>[-1.26, 4.06]</td>
<td>[-1.99, 4.79]</td>
</tr>
<tr>
<td>$C_4$</td>
<td>[-9.58, 4.48]</td>
<td>[-11.53, 6.43]</td>
</tr>
</tbody>
</table>

From these intervals one sees that $C_1$ and $C_2$ are significantly different from zero with 95% confidence, but $C_2$ not quite with 99% confidence. Hence there appears to be a sufficiently strong linear and mildly quadratic component.

The original data plot suggested this and its strength is now assessed statistically.
The previous orthogonal contrasts for linear, quadratic, cubic, quartic behavior were tailored to five treatments.

How do we get similar contrast vectors when we have \( t \) treatments?

R has a function `contr.poly(t)` that gives you orthogonal vectors representing the various polynomial components: linear, quadratic, …

\[
\begin{align*}
> \text{round(contr.poly(7),3)} \\
&\begin{array}{cccccc}
.L & .Q & .C & ^4 & ^5 & ^6 \\
[1,] & -0.567 & 0.546 & -0.408 & 0.242 & -0.109 & 0.033 \\
[2,] & -0.378 & 0.000 & 0.408 & -0.564 & 0.436 & -0.197 \\
[3,] & -0.189 & -0.327 & 0.408 & 0.081 & -0.546 & 0.493 \\
[4,] & 0.000 & -0.436 & 0.000 & 0.483 & 0.000 & -0.658 \\
[5,] & 0.189 & -0.327 & -0.408 & 0.081 & 0.546 & 0.493 \\
[6,] & 0.378 & 0.000 & -0.408 & -0.564 & -0.436 & -0.197 \\
[7,] & 0.567 & 0.546 & 0.408 & 0.242 & 0.109 & 0.033 \\
\end{array}
\end{align*}
\]
Orthogonal Polynomial Contrasts from `contr.poly(7)`

Applied Statistics and Experimental Design
Fritz Scholz — Fall 2006
Model: $Y_{ij} = \mu_i + \epsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, t,$ with the following assumptions.

A1: $\{\epsilon_{ij}\}$ are independent;
A2: $\text{var}(\epsilon_{ij}) = \text{var}(Y_{ij}) = \sigma^2$ for all $i, j$ (homoscedastic);
A3: $\{\epsilon_{ij}\}$ are normally distributed.

These assumption allow us to perform the $F$-test for homogeneity of means, do power calculations, plan sample sizes to achieve a desired power, and obtain simultaneous confidence intervals for contrasts.

We will examine A2 and A3 and deal with A1 when we exploit blocking.
Here we would like to check normality of $\epsilon_{ij} = Y_{ij} - \mu_i$, $j = 1, \ldots, n_i$, $i = 1, \ldots, t$.

Not knowing $\mu_i$ we estimate the error term $\epsilon_{ij}$ via $\hat{\epsilon}_{ij} = Y_{ij} - \hat{\mu}_i = Y_{ij} - \bar{Y}_i$.

If normality holds then a QQ-plot of all these $N = n_1 + \ldots + n_t$ estimated error terms (also called residuals) should look roughly linear with intercept near zero.

Slope $\approx \sigma$.

We have done this before in the single sample situation and won’t show repeats.

It is also possible to perform the formal EDF-based tests of fit (KS, CvM, and AD), but they would require minor modifications in the package nortest, not available right now.
Hermit Crab counts were obtained at 6 different coastline sites.

For each site counts were obtained at 25 randomly selected transects.

Download the data file **crab.csv** from the web into your work directory.
Import it into R via `crab=read.csv("crab.csv")`.

Since these are count data one should not expect good normality behavior.

```r
> names(crab)
[1] "count" "site"
> plot(crab$site,crab$count,xlab="site",ylab="count",col="blue")
```

produced the plot on the next slide.
> out.lm=lm(crab$count~as.factor(crab$site))
> anova(out.lm)
Analysis of Variance Table

Response: crab$count

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>76695</td>
<td>15339</td>
<td>2.9669</td>
<td>0.01401 *</td>
</tr>
<tr>
<td>Residuals</td>
<td>144</td>
<td>744493</td>
<td>5170</td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> qqnorm(out.lm$residuals)
> qqline(out.lm$residuals)

produced the (not so) normal QQ-plot for the ANOVA residuals on the next slide.
Normal QQ-Plot of Hermit Crab Count ANOVA Residuals

Theoretical Quantiles
Sample Quantiles

Theoretical Quantiles
Sample Quantiles

-2 -1 0 1 2
0 100 200 300 400

Normal QQ–Plot

Applied Statistics and Experimental Design
Fritz Scholz — Fall 2006
The appropriate indicators for checking constant variance over all $t$ treatment groups would seem to be $s_1^2, \ldots, s_t^2$.

Rule of thumb:
If $\min(s_1^2, \ldots, s_t^2) / \max(s_1^2, \ldots, s_t^2) > 1/3$ we should be OK in working with the constant variance assumption.
If $\min(s_1^2, \ldots, s_t^2) / \max(s_1^2, \ldots, s_t^2) < 1/7$ we should deal with it.

One first diagnostic is to plot the residuals $Y_{ij} - \bar{Y}_i$ versus the corresponding fitted values $\bar{Y}_i$ for $j = 1, \ldots, n_i, \ i = 1, \ldots, t$.

Compare the difference in information displayed in the next two plots.

Often there is a relationship between variability and the mean and there are ways to deal with this (variance stabilizing transforms of the $Y_{ij}$).
plot(crab$site, out.lm$residuals, col="blue", xlab="site", ylab="residuals")
```r
plot(out.lm$fitted.values, out.lm$residuals, col="blue",
xlab="fitted values", ylab="residuals")
```
The modified Levene test looks at the absolute deviations $X_{ij} = |Y_{ij} - \tilde{Y}_i|$ where $\tilde{Y}_i$ denotes the median of the $i^{th}$ treatment sample.

Originally this was proposed with using $\bar{Y}_i$ in place of $\tilde{Y}_i$, whence “modified.”

The idea is as follows:

If the standard deviations in the $t$ samples $Y_{i1}, \ldots, Y_{in_i}$, $i = 1, \ldots, t$ are the same, then one would expect to have roughly equal means for the $X_{ij}$.

One can check this by performing an ANOVA $F$-test on the $X_{ij}$ values.
crab.levene = function (){
d=NULL
for(i in 1:6){
    m=median(crab$count[crab$site==i])
    d=c(d,abs(crab$count[crab$site==i]-m))
}
anova(lm(d~as.factor(crab$site)))
}
> crab.levene()
Analysis of Variance Table

Response: d

                     Df Sum Sq Mean Sq  F value    Pr(>F)
as.factor(crab$site) 5  71146  14229 2.92783 0.01508 *
Residuals            144 699845   4860
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
We saw for the crab count data that the variability in counts seemed proportional to the averages of the counts and the variability did not show much normality.

Some random phenomena are not so much driven by additive accumulation of random contributions but more so by multiplicative accumulations. A crab colony could have started with a starting group of size $X_0$ that somehow found each other. This group produced a random number $X_0 \times X_1$ of new crabs, where $X_1$ represents the reproduction rate per crab. This rate is variable or random. The next generation would have $X_0 \times X_1 \times X_2$ crabs, and so on.

This motivates the following variation model: $Y = \mu \times \varepsilon = \mu \times (X_1 \times X_2 \times \ldots)$, where the random term $\varepsilon$ has mean $\mu_\varepsilon$ and standard deviation $\sigma_\varepsilon$.

$\Rightarrow \text{var}(Y) = \mu^2 \times \text{var}(\varepsilon)$ or $\sigma_Y = \mu \times \sigma_\varepsilon$ and $\mu_Y = E(Y) = \mu \times E(\varepsilon)$ and thus $\sigma_Y$ is proportional to $\mu_Y$ since both are proportional to $\mu$. 

104
Multiplicative error model $\implies \sigma \propto \mu$. However, using $\log(Y) = \log(\mu) + \log(\varepsilon)$

$\implies E(\log(Y)) = \log(\mu) + E(\log(\varepsilon))$ and $\var(\log(Y)) = \var(\log(\varepsilon))$

breaks the link, i.e., $\mu$ affects the mean but no longer the variance of $\log(Y)$, an example of variance stabilization!

There is further benefit in viewing the multiplicative error term $\varepsilon$ as a product of several random contributors. By taking the transform $\log(Y)$:

$$V = \log(Y) = \log(\mu) + \log(\varepsilon) = \log(\mu) + \log(X_1) + \log(X_2) + \ldots$$

we can appeal to the CLT, applied to the sum of the $\log(X_i)$ terms, to justify a normal additive error model for $V$, i.e., $V = \tilde{\mu} + \tilde{\varepsilon}$ with $\tilde{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$.

Applying this to all our count data we would have the following familiar model:

$$V_{ij} = \log(Y_{ij}) = \tilde{\mu}_i + \tilde{\varepsilon}_{ij} \text{ with } \tilde{\varepsilon}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$
Since some of the observed counts are zero there would be the problem of $\log(0)$.

We look at two ways of dealing with it.

1. Adding a small fraction, say $1/6$, to all counts. ($1/6 > 0$ is somewhat arbitrary)
   This is a technical solution, keeping all the data.

2. Eliminate all zero counts.
   This may be justified if a zero count just means that there were no crabs in that transect to begin with. It is not a matter of not seeing them because the population size is small. This reduces the count data to $150 - 33 = 117$ counts.
Box Plots for \( \text{count and } \log(\text{count+1/6}) \)
Analysis of Variance Table

Response: log(count + 1/6)

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>as.factor(site)</td>
<td>5</td>
<td>54.73</td>
<td>10.95</td>
<td>2.3226</td>
</tr>
<tr>
<td>Residuals</td>
<td>144</td>
<td>678.60</td>
<td>4.71</td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Box Plots for \( \text{count} \) and \( \log(\text{count}[\text{count}>0]) \)
Normal QQ-Plots of 117 Residuals
Analysis of Variance Table

Response: log(count[count > 0])

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>as.factor(site[count &gt; 0])</td>
<td>5</td>
<td>47.905</td>
<td>9.581</td>
<td>4.3866</td>
</tr>
<tr>
<td>Residuals</td>
<td>111</td>
<td>242.440</td>
<td>2.184</td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Levene Test for $\log(\text{count}+1/6)$ and $\log(\text{count}[\text{count}>0])$

> log.crab.levene16()
Analysis of Variance Table

Response: d

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>as.factor(site)</td>
<td>5</td>
<td>7.193</td>
<td>1.439</td>
<td>0.7513</td>
<td>0.5864</td>
</tr>
<tr>
<td>Residuals</td>
<td>144</td>
<td>275.748</td>
<td>1.915</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

> log.crab.levene0()
Analysis of Variance Table

Response: d

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>as.factor(site)</td>
<td>5</td>
<td>6.168</td>
<td>1.234</td>
<td>1.4711</td>
<td>0.205</td>
</tr>
<tr>
<td>Residuals</td>
<td>111</td>
<td>93.077</td>
<td>0.839</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The log(count[<count>0]) analysis appears to show stronger evidence of site differences, as indicated by the p-values: \( .0011 < .046 \).

The qqnorm plots for the residuals seem to show no gross violation of normality, when compared to qqnorm plots of true normal random samples of same size.

The qqnorm plot for the log(count+1/6) residual analysis shows the effect of the retained zeros strongly.

The boxplots for the log(count[<count>0]) analysis seem better regularized than in the case of the log(count+1/6) analysis (the box for site 6 is distorted by 9 zeros).

The Levene test shows no significant differences in \( \sigma \) across sites for either case.
For data with a multiplicative error model for $Y_{ij}$ we showed $\sigma_i \propto \mu_i$ or $\sigma_\mu \propto \mu$

and the beneficial variance stabilizing effect of the log-transform.

Suppose $\sigma_\mu = k \times \mu^\alpha$, a power relationship, somewhat more general than $\sigma_\mu \propto \mu$.

Can we find a transform $V = f(Y)$ for which the variance no longer depends on $\mu$?

A 1-term Taylor series expansion of $f$ around $\mu$, assumed to be the mean of $Y$,

$\Rightarrow f(Y) \approx f(\mu) + (Y - \mu)f'(\mu) \Rightarrow E(f(Y)) \approx f(\mu)$  and  $\text{var}(f(Y)) \approx \sigma_\mu^2 [f'(\mu)]^2$

To get $\text{var}(f(Y))$ independent of $\mu$ we need $\sigma_\mu^2 [f'(\mu)]^2 = k^2 \mu^{2\alpha} [f'(\mu)]^2 = c$, i.e.,

$f'(\mu) = \frac{\tilde{c}}{\mu^\alpha}$  or  $f(\mu) = \tilde{c} \frac{\mu^{1-\alpha}}{1-\alpha} + c^*$  with $\alpha = 1 \Rightarrow f(\mu) = \log(\mu)$ as special case.
According to the previous slide: If $\sigma_\mu = k \mu^\alpha$ we should analyze the transformed data $\tilde{Y} = f(Y) = Y^{1-\alpha}$ if $\alpha \neq 1$ and $\tilde{Y} = \log(Y)$ when $\alpha = 1$.

But what is the correct $\alpha$? Let the data speak for themselves.

$$\sigma_\mu \propto \mu^\alpha \iff \sigma_\mu = c \times \mu^\alpha \iff \log(\sigma_\mu) = k + \alpha \times \log(\mu)$$

Thus look for a linear relationship between $\log(s_i)$ and $\log(\hat{\mu}_i) = \log(\bar{Y}_i^*).$

Its slope $\hat{\alpha}$ is our estimate of $\alpha$.

$$\hat{\alpha} = \text{lm}(\log(s_i) \sim \log(\bar{Y}_i^*))\text{coef}[2]$$

Then perform the ANOVA for $\tilde{Y}_{ij} = Y_{ij}^{1-\hat{\alpha}} = Y_{ij}^{\hat{\lambda}}$. 
<table>
<thead>
<tr>
<th>Relation</th>
<th>$\sigma_Y$</th>
<th>$\alpha$</th>
<th>$\lambda$</th>
<th>Transform</th>
<th>$\tilde{Y}_{ij}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_Y \sim \mu_Y$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>no transform!</td>
<td>$Y_{ij}$</td>
</tr>
<tr>
<td>$\sigma_Y \sim \mu_Y^{1/2}$</td>
<td>1/2</td>
<td>1/2</td>
<td>1/2</td>
<td>square root</td>
<td>$Y_{ij}^{1/2} = \sqrt{Y_{ij}}$</td>
</tr>
<tr>
<td>$\sigma_Y \sim \mu_Y$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>log</td>
<td>$\log(Y_{ij})$</td>
</tr>
<tr>
<td>$\sigma_Y \sim \mu_Y^{3/2}$</td>
<td>3/2</td>
<td>-1/2</td>
<td>-1/2</td>
<td>reciproc. of sqrt</td>
<td>$Y_{ij}^{-1/2} = 1/\sqrt{Y_{ij}}$</td>
</tr>
<tr>
<td>$\sigma_Y \sim \mu_Y^2$</td>
<td>2</td>
<td>-1</td>
<td>-1</td>
<td>reciprocal</td>
<td>$1/Y_{ij}$</td>
</tr>
</tbody>
</table>
All the above transformations can be captured in the following unified format known as the *Box-Cox transformations*

\[
y(\lambda) = \frac{y^\lambda - 1}{\lambda} \quad \text{with} \quad y^{(0)} = \lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \log(y) \quad \text{by L’Hospital’s rule}.
\]

For any given \( \lambda \neq 0 \) the results of an ANOVA on \( \tilde{Y}_{ij} \) or an ANOVA on \( Y_{ij}^{(\lambda)} = (Y_{ij}^\lambda - 1)/\lambda = a \times Y_{ij}^\lambda + b = a \times \tilde{Y}_{ij} + b \) will be the same.
Don’t transform if \( \frac{\min(s_1^2, \ldots, s_t^2)}{\max(s_1^2, \ldots, s_t^2)} > \frac{1}{3} \)

Make sure the linear relationship between \( \log(s_i) \) and \( \log(\bar{Y}_i) \) is strong.

Use simple exponents \( \lambda \) in the transformations, i.e., use \( \lambda = 1/2 \) rather than \( \lambda = 1 - \alpha = 0.473 \), as possibly calculated from slope of the linear fit of \( \log(s_i) \approx \alpha \times \log(\bar{Y}_i) + b \).

Try to see whether the transform can be explained rationally, as with the multiplicative model motivating the log-transform.

When presenting the analysis, make sure to point out the transformation issue and show the transformed and untransformed data in graphical form.
<table>
<thead>
<tr>
<th>site</th>
<th>$s_i$</th>
<th>$\hat{\mu}_i$</th>
<th>log($s_i$)</th>
<th>log($\hat{\mu}_i$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>17.39</td>
<td>9.24</td>
<td>2.86</td>
<td>2.22</td>
</tr>
<tr>
<td>5</td>
<td>19.84</td>
<td>10.00</td>
<td>2.99</td>
<td>2.30</td>
</tr>
<tr>
<td>6</td>
<td>23.01</td>
<td>12.64</td>
<td>3.14</td>
<td>2.54</td>
</tr>
<tr>
<td>1</td>
<td>50.39</td>
<td>33.80</td>
<td>3.92</td>
<td>3.52</td>
</tr>
<tr>
<td>3</td>
<td>107.44</td>
<td>50.64</td>
<td>4.68</td>
<td>3.92</td>
</tr>
<tr>
<td>2</td>
<td>125.35</td>
<td>68.72</td>
<td>4.83</td>
<td>4.23</td>
</tr>
</tbody>
</table>
$\log(s_i) \text{ vs } \log(\hat{\mu}_i)$ Plot for Crab Data

$\alpha = 0.9858 \Rightarrow \lambda = 0$ (log-transform)

$\min s_i^2 / \max s_i^2 = 0.019 < 1/3$

$\hat{\alpha} = 0.9858$
Let \( Y_{11}, \ldots, Y_{1n_1} \overset{\text{i.i.d.}}{\sim} F_1, \; Y_{21}, \ldots, Y_{2n_2} \overset{\text{i.i.d.}}{\sim} F_2, \; \ldots, \; Y_{k1}, \ldots, Y_{kn_k} \overset{\text{i.i.d.}}{\sim} F_k \)

Test the hypothesis \( H_0 : F_1 = \ldots = F_k \) where the common \( F \) is not specified.

Since the problem stays invariant under the same strictly monotone transformation of the \( Y_{ij} \) values only their relative position to each other matters, i.e., their ranks.

Denote by \( R_{ij} \) the rank of observation \( Y_{ij} \) among all \( N \) observations \( Y_{11}, \ldots, Y_{kn_k} \), i.e., the smallest of the \( Y_{ij} \) gets rank 1, the second smallest gets rank 2, \ldots, and the largest of the \( Y_{ij} \) gets rank \( N \).

In the case of ties assign the same average rank to all these tied observations.
Let \( \bar{R}_i = \sum_{j=1}^{n_i} R_{ij} / n_i \) denote the average rank for the \( i \)th sample.

Note that the average \( \bar{R}_.. \) of all \( N \) ranks, \( R_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, k \), is just the midpoint between 1 and \( N \), i.e., \( \bar{R}_.. = (N + 1)/2 \).

If the distributions of these samples are the same, one would expect that the sets of ranks for the \( k \) samples are well intermeshed, i.e., their variability around their means should compare well with the variability of all \( N \) ranks around \((N + 1)/2\).

\[
H = \frac{SS_{Treat}}{SS_T/(N - 1)} = \frac{\sum_{i=1}^{k} n_i(\bar{R}_i^2 - \bar{R}_..)^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (R_{ij} - \bar{R}_..)^2/(N - 1)} = \frac{\sum_{i=1}^{k} n_i\bar{R}_i^2 - N\bar{R}_..^2}{[\sum_{i=1}^{k} \sum_{j=1}^{n_i} R_{ij}^2 - N\bar{R}_..^2] / (N - 1)}
\]

suggests itself as a reasonable test statistic.
Kruskal-Wallis $k$-Sample Test (continued)

\[ \sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6} \implies \]

\[ \sum_{i=1}^{k} \sum_{j=1}^{n_i} (R_{ij} - \bar{R}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} R_{ij}^2 - N \left( \frac{N+1}{2} \right)^2 = \frac{N(N+1)(2N+1)}{6} - N \left( \frac{N+1}{2} \right)^2 = \frac{N(N+1)(N-1)}{12} \implies SS_T = \frac{N(N+1)}{12} \]

Kruskal and Wallis showed that under $H_0$ (all rankings are equally likely)

\[ H = \left\{ \sum_{i=1}^{k} n_i \bar{R}_i^2 - N \left( \frac{N-1}{2} \right)^2 \right\} / [N(N+1)/12] = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \bar{R}_i^2 - 3(N+1) \]

has an approximate $\chi^2_{k-1}$ distribution.

We reject $H_0$ whenever $H \geq \chi^2_{k-1,1-\alpha} = qchisq(1-\alpha,k-1)$. 

124
> kruskal.test(list(Flux3$X, Flux3$Y, Flux3$Z))

Kruskal-Wallis rank sum test

data: list(Flux3$X, Flux3$Y, Flux3$Z)
Kruskal-Wallis chi-squared = 4.2633, df = 2, p-value = 0.1186

The $p$-value is not as small as in the normal ANOVA or randomization tests. Compared to the former test we no longer assume normality and compared to the latter we used $R_{ij}$ in place of the more informative $Y_{ij}$.

The K-W test in ineffective for changes in scale while locations are matched.

Look at the documentation of `kruskal.test` on how to use it.
Estimate $F_i(x)$ by the $i^{th}$ sample distribution function, i.e., by its EDF $\hat{F}_i(x)$ and estimate the common cdf $F(x)$ (under $H_0$) by the EDF $\hat{F}(x)$ of all samples combined.

Under $H_0$ we expect that the $\hat{F}_i(x)$ should not differ much from $\hat{F}(x)$.

We assess the difference between the $\hat{F}_i(x)$ and $\hat{F}(x)$ by the Anderson-Darling discrepancy metric (a function of only the ranks, why?)

$$AD_k = \sum_{i=1}^{k} n_i \int_B \frac{[\hat{F}_i(x) - \hat{F}(x)]^2}{\hat{F}(x)(1 - \hat{F}(x))} \, d\hat{F}(x) = \sum_{i=1}^{k} \frac{n_i}{N} \sum_{j=1}^{N-1} \frac{[\hat{F}_i(Z_j) - \hat{F}(Z_j)]^2}{\hat{F}(Z_j)(1 - \hat{F}(Z_j))}$$

where $B$ denotes the set of all $x$ for which $\hat{F}(x) < 1$ and $Z_1 < \ldots < Z_N$ denote the ordered combined sample values.

Reject $H_0$ for large $AD_k$. Approximate null distribution is available.
> AD.test(Flux3$X, Flux3$Y, Flux3$Z)

Anderson-Darling k-sample test.

Number of samples = 3
Sample sizes: 6 6 6
Total number of points: 18
Number of ties (identical points): 6
Mean of Anderson Darling Criterion: 2
Standard deviation of Anderson Darling Criterion: 0.9441525
T.kN = (Anderson Darling Criterion - mean)/sigma
Null Hypothesis: All samples come from a common population.

<table>
<thead>
<tr>
<th>T.kN</th>
<th>P-value extrapolation</th>
</tr>
</thead>
<tbody>
<tr>
<td>not adj. for ties</td>
<td>1.224926   0.1091177  0</td>
</tr>
<tr>
<td>adj. for ties</td>
<td>1.125149    0.1215943  0</td>
</tr>
</tbody>
</table>
For Flux3 the p-values were comparable.

The AD-test is effective against any alternatives of $H_0$, it is an omnibus test.
This is not the case for the KW-test (as mentioned w.r.t. variability changes).

The AD-test may have less power than a test geared against a specific alternative.
Similarly for the KW-test.

In large samples the AD-test rejects with probability $\to 1$ for any alternative to $H_0$.
Not always true for the KW-test.

It is advised to restrict use of the AD-test to $n_i \geq 5$, $i = 1, \ldots, k$.
Similar restriction may be appropriate to make $\chi^2_{k-1}$ approximation reasonable.

The AD-test is often used to justify pooling data when $H_0$ is not rejected.
It pays special attention to behavior in the sample tails.