University of Washington

STATISTICS

Applied Statistics and Experimental Design

One-Factor ANOVA

Fritz Scholz

Fall Quarter 2008

One-Factor ANOVA

ANOVA is an acronym for Analysis of Variance.

The primary focus is the difference in means of several populations or the difference in mean response under several treatments

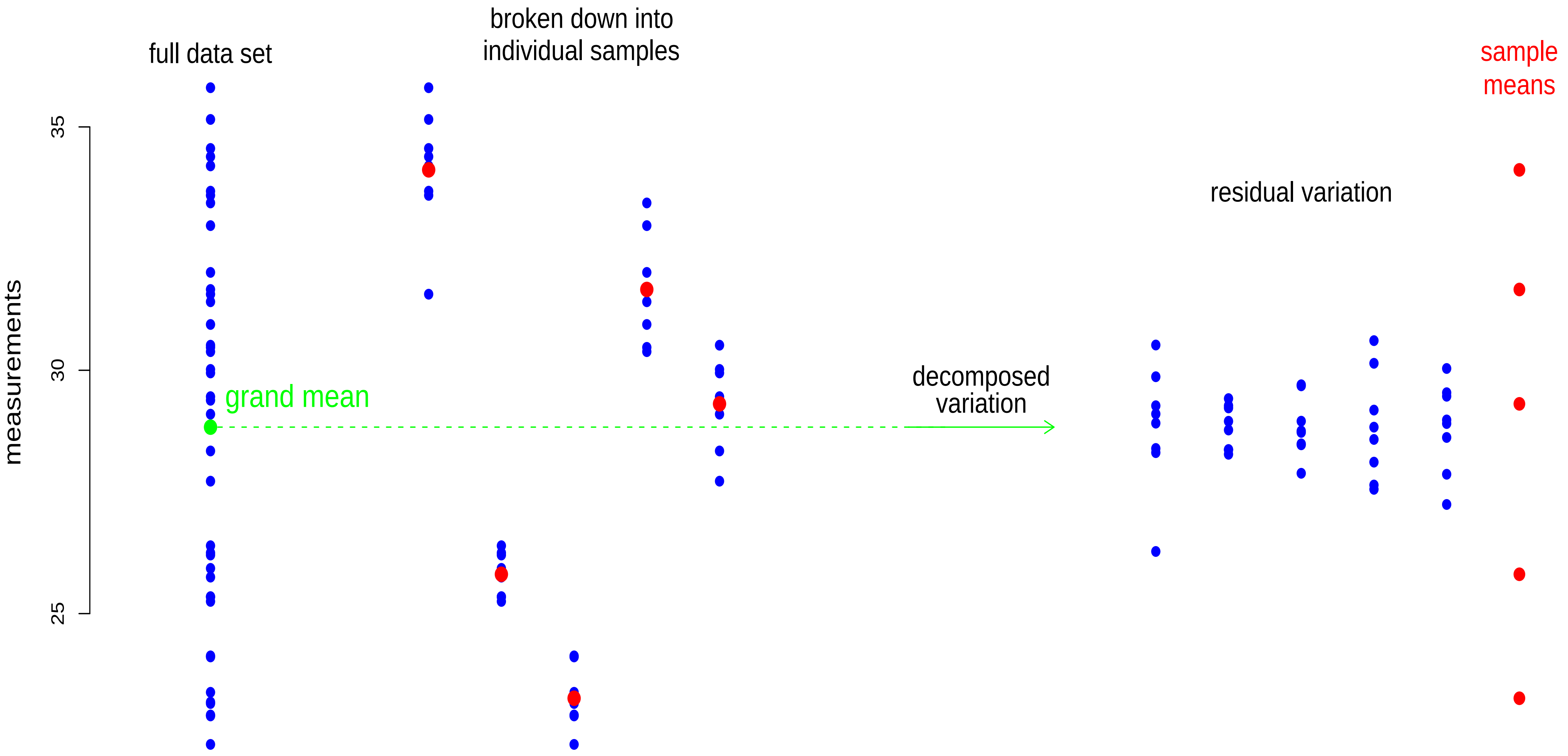
The reference to variance in ANOVA alludes to the analysis technique.

It is the overall data variation that is decomposed into several components.

How much of that variation is due to changing the sampled population or changing the treatment?

How much variation cannot not be attributed to such systematic changes?

ANOVA Illustrated



The Notion of Factor in One-Factor ANOVA

It is difficult to explain the notion of
2-dimensional space to someone who has lived only in 1-dimensional space,
or 3-dimensional space to someone who lives in flatland
or 4-dimensional space to us in the “real” 3-dimensional world.

The term Factor similarly alludes to different possible directions/dimensions in which changes can take place in populations or in treatments.

Example: In soldering circuit boards we could have several types of flux (say 3) and also several methods of cleaning the boards (say 4).

Combining each with each, we thus could have $3 \times 4 = 12$ distinct treatments.

However, it is more enlightening to view the effects of flux and cleaning method separately. Each would be called a factor, the flux factor and the cleaning factor.

We can then ask which factor is responsible for changes in the mean response.

More Than 2 Treatments or Populations

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

We have 18 circuit boards, randomly assign each flux to 6 boards.

In principle, this gives us the randomization reference distribution and

thus a logical basis for a test of the hypothesis H_0 : no flux differences.

Randomize the order of soldering/cleaning, coating, and humidity chamber slots.

These randomizations avoid unintended biases from hidden factors (dimensions).

There are $\binom{18}{6} \times \binom{12}{6} \times \binom{6}{6} = 18,564 \times 924 \times 1 = 17,153,136$ 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 computing limits

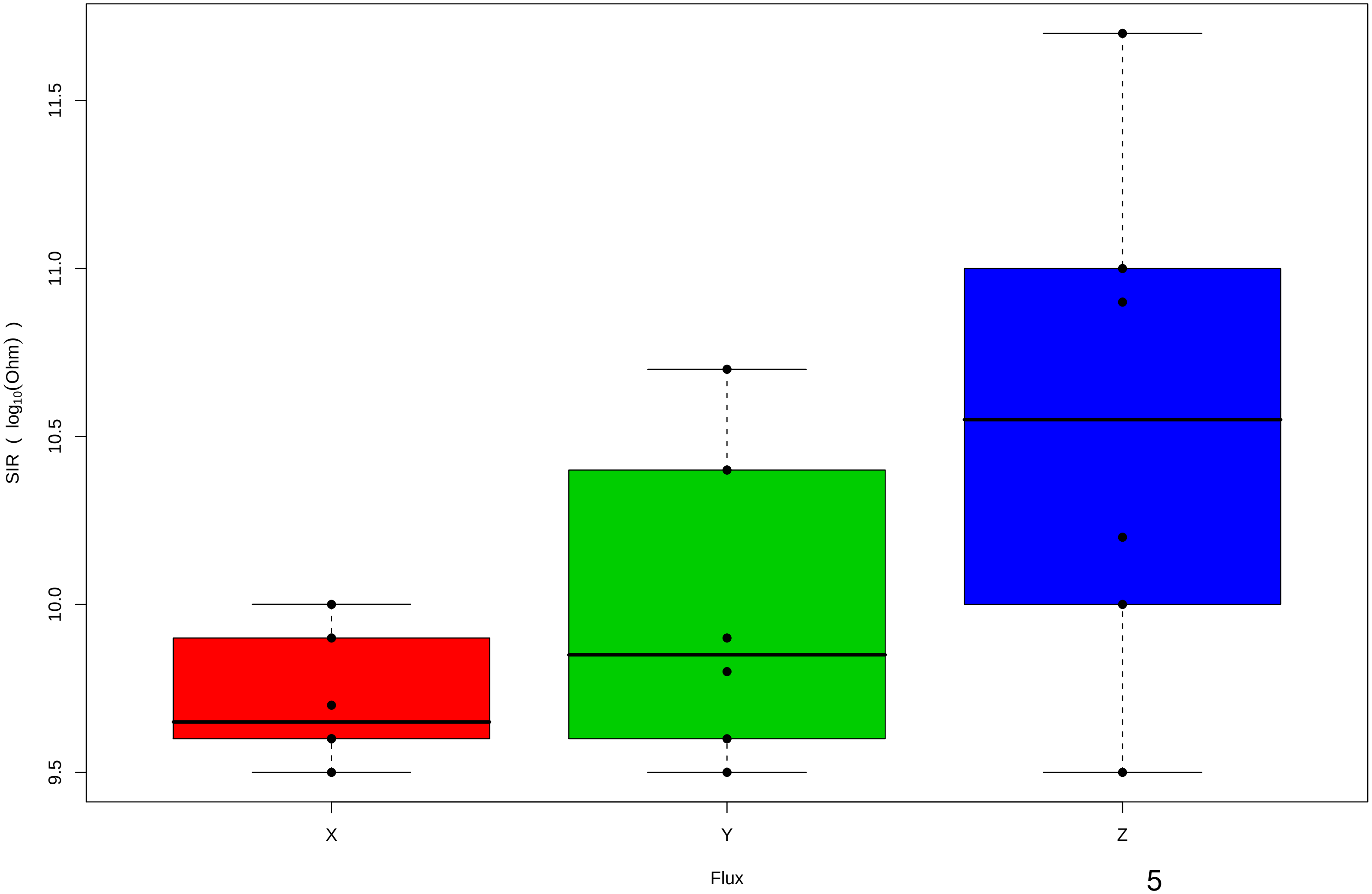
$\Longrightarrow$ simulated reference distribution.

The Flux3 Data

SIR Responses

X	Y	Z
9.9	10.7	10.9
9.6	10.4	11.0
9.6	9.5	9.5
9.7	9.6	10.0
9.5	9.8	11.7
10.0	9.9	10.2

units $\log_{10}(\mathrm{Ohm})$



Differences in the Fluxes?

To examine whether the fluxes are in some way different in their effects we could again focus on differences between the means of the SIR responses.

We denote these means by $\mu_1 = \mu_X$, $\mu_2 = \mu_Y$, and $\mu_3 = \mu_Z$.

Mathematically, $X \equiv Y$ and $Y \equiv Z \implies X \equiv Z$.

It would seem that testing $H_{0,XY} : X \equiv Y$ and $H_{0,YZ} : Y \equiv Z$ might suffice.

Statistically, $X \approx Y$ and $Y \approx Z$ allows for the possibility that X and Z are sufficiently different.

To guard against this we could perform all 3 possible two-sample tests for the following respective hypothesis testing problems:

$H_{0,XY} : X \equiv Y$ vs. $H_{1,XY} : \mu_X \neq \mu_Y$, $H_{0,YZ} : Y \equiv Z$ vs. $H_{1,YZ} : \mu_Y \neq \mu_Z$
$H_{0,XZ} : X \equiv Z$ vs. $H_{1,XZ} : \mu_X \neq \mu_Z$

Probability of Overall Type I Error?

If we do each such test at level α, what is our chance of getting a rejection by at least one of these tests when in fact all 3 fluxes are equivalent?

(2 versus 4 engines on aircraft, controversy between Boeing and Airbus)

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

$$\begin{aligned} P_0(\text{Overall Type I Error}) &= P_0(\text{reject at least one of the hypotheses}) \\ &= 1 - P_0(\text{accept all of the hypotheses}) \\ &= 1 - P_0(\text{ accept } H_{0,XY} \cap \text{ accept } H_{0,XZ} \cap \text{ accept } H_{0,YZ}) \\ \text{by independence} &= 1 - (1-\alpha)^3 = 0.142625 \quad \text{for } \alpha = .05 \text{ .} \end{aligned}$$

P_0 indicates that all 3 fluxes are the same and that we are dealing with the null or randomization reference distribution.

Engine Failure

If $p_F =$ probability of shutdown for a given engine ($p_F \approx 1$ in 10000 flights)

the chance of at least one shutdown on a flight with k engines is

$$P(\text{at least one shutdown}) = 1 - P(\text{no shutdown}) = 1 - (1 - p_F)^k \approx k \times p_F \ .$$

k	$1-(1-p_F)^k$	$k \times p_F$
2	.00019999	.0002
4	.00039994	.0004

$(1 - p_F)^k$ assumes that engine (non-)shutdowns are independent events.

This independence is the goal of ETOPS

(Extended-range Twin-engine Operational Performance Standards)

`http://en.wikipedia.org/wiki/ETOPS`

For example, different engines are serviced by different mechanics.

The Multiple Comparison Issue

If you expose yourself to multiple rare opportunities of making a wrong decision, the chance of making a wrong decision at least once (the overall type I error) is much higher than planned for in the individual tests.

This problem is referred to as 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 $\Longrightarrow$ dependence.

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

$$\begin{aligned} P_0(\text{Overall Type I Error }) &= P_0(\text{reject } H_{0,XY} \cup \text{ reject } H_{0,XZ} \cup \text{ reject } H_{0,YZ}) \\ &\leq P_0(\text{reject } H_{0,XY}) + P_0(\text{reject } H_{0,XZ}) + P_0(\text{reject } H_{0,YZ}) \\ &= 3\alpha = .15 \quad \text{when} \quad \alpha = .05 \; . \end{aligned}$$

How much smaller than this upper bound is the true $P_0(\text{Overall Type I Error })$?

Overall Type I Error Probability

We will evaluate it based on the randomization reference distribution.

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: $\texttt{Nsim0} = 10000$ times.

For $\alpha = .05$ get the .95-quantile `tcrit` of this simulated $|\bar{X}-\bar{Y}|$ reference distribution. It serves equally well for tests based on $|\bar{X}-\bar{Z}|$ or $|\bar{Y}-\bar{Z}|$. Why?

Then simulate another $\texttt{Nsim1} = 10000$ such splits, computing $|\bar{X}-\bar{Y}|$, $|\bar{X}-\bar{Z}|$, and $|\bar{Y}-\bar{Z}|$ each time, and tally the proportions of each individually exceeding `tcrit` and the proportion of at least one of them exceeding `tcrit`.

The resulting proportions are: `0.0451 0.0460 0.0491` for the individual tests ($\approx$ the targeted $\alpha = .05$) and `0.1186` for the overall type I error rate.
The code for running this, `typeIerror.rateRand`, is posted on web.

A Global Testing View

Rather than doing all 3 possible pairwise tests based on separate discrepancy statistics $|\bar{X}-\bar{Y}|$, $|\bar{X}-\bar{Z}|$, and $|\bar{Y}-\bar{Z}|$, we will address this in a global way, using a single discrepancy statistic. For now we will focus on the population view.

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

$H_0 : \mu_1 = \mu_2 = \mu_3$ (common value unspecified $\Longrightarrow$ composite hypothesis)

against the alternative $H_1 : \mu_i \neq \mu_j$ 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.

Useful Models for Treatment Variation

We have measurements Y_{ij}, the j^{th} response under the i^{th} treatment,

$j = 1, \ldots, n_i$ and $i = 1, \ldots, t$. 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 $\mathrm{var}(\varepsilon_{ij}) = \sigma^2$.

View ε_{ij} (i.i.d.) as response variation/error/noise that occurs within treatment

or after the treatment mean μ_i is subtracted from the response Y_{ij}.

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

$\mu = \bar{\mu} = \sum_{ij} \mu_i / N = \sum_i n_i \mu_i / N =$ grand mean (or n_i/N-weighted average of the μ_i)

The grand mean is the average of the means for all the observations.

$\tau_i = \mu_i - \mu = \mu_i - \bar{\mu}$ is the i^{th} treatment effect and

ε_{ij} (i.i.d.) is the within treatment variation with $E(\varepsilon_{ij}) = 0$ and $\mathrm{var}(\varepsilon_{ij}) = \sigma^2$.

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

The Reduced Model

In contrast to the full model with varying treatment means, as discussed on the previous slide, we assume in the reduced model a single mean for all observations:

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

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

The reduced model corresponds to our previously stated hypothesis

$$H_0 : \mu_1 = \ldots = \mu_t \qquad \text{or equivalently} \qquad H_0 : \tau_1 = \ldots = \tau_t = 0$$

which is a special case of our previous full population model.

Test this hypothesis by fitting the full model and the reduced model to the data and compare the quality of fits relative to each other via some discrepancy metric.

Full Model Fitting by Least Squares

The method of Least Squares originated with Gauss and Legendre.

Minimize the Sum of Squares criterion

$$SS(\mu_1,\ldots,\mu_t) = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij}-\mu_i)^2 \quad \text{over} \quad \mu = (\mu_1,\ldots,\mu_t)\ .$$

Using the notation $\bar{Y}_{i\bullet} = \sum_{j=1}^{n_i} Y_{ij}/n_i$ and the fact $\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet}) = 0$:

$$SS(\mu_1,\ldots,\mu_t) = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet}+\bar{Y}_{i\bullet}-\mu_i)^2 \qquad (a+b)^2 = a^2+b^2+2ab$$

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

$$= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2 + \sum_{i=1}^{t}\sum_{j=1}^{n_i}(\bar{Y}_{i\bullet}-\mu_i)^2 \geq \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2 = SS(\hat{\mu}_1,\ldots,\hat{\mu}_t)$$

$\Longrightarrow$ the least squares estimates (LSE) $\hat{\mu}_i = \bar{Y}_{i\bullet}$ minimize $SS(\mu_1,\ldots,\mu_t)$.

The Dot Notation

If $a_1, \ldots, a_n$ are n numbers then

$$a_{\bullet} = \sum_{i=1}^{n} a_i \qquad \text{and} \qquad \bar{a}_{\bullet} = \sum_{i=1}^{n} a_i/n\,.$$

For an array of numbers $a_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n,$ we write

$$a_{\bullet j} = \sum_{i=1}^{m} a_{ij} \qquad \bar{a}_{\bullet j} = \sum_{i=1}^{m} a_{ij}/m \qquad a_{i\bullet} = \sum_{j=1}^{m} a_{ij} \qquad \bar{a}_{i\bullet} = \sum_{j=1}^{m} a_{ij}/n$$

$$a_{\bullet\bullet} = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} \qquad \text{and} \qquad \bar{a}_{\bullet\bullet} = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}/(mn)$$

Similarly for higher dimensional arrays $a_{ijk},\ i = 1, \ldots, m,\ j = 1, \ldots, n,\ k = 1, \ldots, \ell$

$$a_{ij\bullet} = \sum_{k=1}^{\ell} a_{ijk} \qquad \text{and} \qquad \bar{a}_{ij\bullet} = \sum_{k=1}^{\ell} a_{ijk}/\ell \qquad \text{and so on.}$$

Reduced Model Fitting by Least Squares

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

With $\bar{Y}_{\bullet\bullet} = \sum_i \sum_j Y_{ij} / \sum_i n_i = \sum_i \sum_j Y_{ij}/N = \sum_i (n_i/N) \bar{Y}_{i\bullet}$ and $\sum_i \sum_j (Y_{ij} - \bar{Y}_{\bullet\bullet}) = 0$

$$\begin{aligned}
\Longrightarrow SS(\mu) &= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \mu)^2 = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\bullet\bullet} + \bar{Y}_{\bullet\bullet} - \mu)^2 \\
&= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 + \sum_{i=1}^{t} \sum_{j=1}^{n_i} (\bar{Y}_{\bullet\bullet} - \mu)^2 + 2 \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\bullet\bullet})(\bar{Y}_{\bullet\bullet} - \mu) \\
&= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 + \sum_{i=1}^{t} \sum_{j=1}^{n_i} (\bar{Y}_{\bullet\bullet} - \mu)^2 \geq \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 = SS(\hat{\mu})
\end{aligned}$$

$\Longrightarrow$ the least squares estimate (LSE) $\hat{\mu} = \bar{Y}_{\bullet\bullet}$ minimizes $SS(\mu)$

Means and Variances of Least Squares Estimates

$$
\begin{aligned}
E(\bar{Y}_{i\bullet}) &= E\left(\frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}\right) = \frac{1}{n_i}\sum_{j=1}^{n_i} E(Y_{ij}) = \frac{1}{n_i}\sum_{j=1}^{n_i} \mu_i = \mu_i \\
\operatorname{var}(\bar{Y}_{i\bullet}) &= \operatorname{var}\left(\frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}\right) = \frac{1}{n_i^2}\sum_{j=1}^{n_i} \operatorname{var}(Y_{ij}) = \frac{1}{n_i^2}\sum_{j=1}^{n_i} \sigma^2 = \frac{\sigma^2}{n_i} \\
E(\bar{Y}_{\bullet\bullet}) &= E\left(\frac{1}{N}\sum_{i=1}^{t}\sum_{j=1}^{n_i} Y_{ij}\right) = E\left(\sum_{i=1}^{t}\frac{n_i}{N}\bar{Y}_{i\bullet}\right) = \sum_{i=1}^{t}\frac{n_i}{N}E(\bar{Y}_{i\bullet}) = \sum_{i=1}^{t}\frac{n_i}{N}\mu_i = \bar{\mu} \\
\operatorname{var}(\bar{Y}_{\bullet\bullet}) &= \operatorname{var}\left(\sum_{i=1}^{t}\frac{n_i}{N}\bar{Y}_{i\bullet}\right) = \sum_{i=1}^{t}\left(\frac{n_i}{N}\right)^2 \operatorname{var}(\bar{Y}_{i\bullet}) = \sum_{i=1}^{t}\left(\frac{n_i}{N}\right)^2 \frac{\sigma^2}{n_i} = \sum_{i=1}^{t}\frac{n_i}{N^2}\sigma^2 = \frac{\sigma^2}{N} \\
\operatorname{var}(\bar{Y}_{\bullet\bullet}) &= \operatorname{var}\left(\frac{1}{N}\sum_{i=1}^{t}\sum_{j=1}^{n_i} Y_{ij}\right) = \frac{1}{N^2}\sum_{i=1}^{t}\sum_{j=1}^{n_i} \operatorname{var}(Y_{ij}) = \frac{1}{N^2}\sum_{i=1}^{t}\sum_{j=1}^{n_i} \sigma^2 = \frac{\sigma^2}{N}\,.
\end{aligned}
$$

Sum of Squares (SS) Decomposition

Using $\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})=0$ we have the following sum of squares decomposition

$$\begin{aligned} SS_{\mathrm{T}}=\sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{\bullet\bullet})^2 &= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet}+\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2 \\ &= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2+\sum_{i=1}^{t}\sum_{j=1}^{n_i}(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2 \\ &\qquad +2\sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet}) \\ &= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2+\sum_{i=1}^{t}\sum_{j=1}^{n_i}(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2 = SS_{\mathrm{E}}+SS_{\mathrm{Treat}} \end{aligned}$$

This is the fundamental ANOVA identity: $SS_{\mathrm{T}}=SS_{\mathrm{E}}+SS_{\mathrm{Treat}}=SS_{\mathrm{W}}+SS_{\mathrm{B}}$.

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

or SS of total variation $=$ variation within samples $+$ variation between samples.

How to Compare the Model Fits?

How should we compare the two model fits

$$SS_E = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\bullet})^2 \qquad \text{and} \qquad SS_T = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\bullet\bullet})^2 \qquad ?$$

Under H_0 (reduced model) both fits should be somewhat comparable, except that the full model fit gave us more freedom in minimizing the sum of squares.

The previous slide showed

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

$$\text{with} \qquad SS_T - SS_E = SS_{\text{Treat}} = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 \,.$$

To make a fair comparison we should make allowances for this extra freedom.

We need to understand $E(SS_T)$ and $E(SS_E)$ when H_0 is true or false.

Unbiasedness of s^2: $E(s^2) = \sigma^2$

Assume that $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2.

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

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

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

$$\begin{aligned}
\Longrightarrow \qquad E((n-1)s^2) &= E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = E\left(\sum_{i=1}^{n}\left(X_i^2 - 2X_i\bar{X} + \bar{X}^2\right)\right) \\
&= E\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = n(\sigma^2 + \mu^2) - n(\mathrm{var}(\bar{X}) + [E(\bar{X})]^2) \\
&= n(\sigma^2 + \mu^2) - n(\sigma^2/n + \mu^2) = (n-1)\sigma^2 \ \Rightarrow E(s^2) = \sigma^2.
\end{aligned}$$

$E(MS_\mathrm{E}) = \sigma^2$

With

$$s_i^2 = \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2/(n_i - 1) \qquad \text{we have} \qquad \sum_{i=1}^{t} (n_i - 1)s_i^2 = SS_\mathrm{E}$$

and the result from the previous slide shows

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

or the Mean Square for Error

$$MS_\mathrm{E} = \frac{SS_\mathrm{E}}{N-t} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\bullet})^2}{N-t} \qquad \text{is an unbiased estimate for } \sigma^2$$

This is true whether $H_0 : \mu_1 = \ldots = \mu_t$ holds or not (also without normality).

$E(MS_{\text{Treat}}) = \sigma^2 + ?$

$$SS_{\text{Treat}} = \sum_{i=1}^{t} n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{t} n_i(\bar{Y}_{i\bullet}^2 - 2\bar{Y}_{i\bullet}\bar{Y}_{\bullet\bullet} + \bar{Y}_{\bullet\bullet}^2) = \sum_{i=1}^{t} n_i\bar{Y}_{i\bullet}^2 - N\bar{Y}_{\bullet\bullet}^2$$

$$\Longrightarrow E\left(SS_{\text{Treat}}\right) = \sum_{i=1}^{t} n_i E(\bar{Y}_{i\bullet}^2) - N\,E(\bar{Y}_{\bullet\bullet}^2)$$ (with or without normality)

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

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

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

$$E(MS_{\text{Treat}}) = E\left(SS_{\text{Treat}}/(t-1)\right) = \sigma^2 + \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu})^2/(t-1) = \sigma^2 + \sum_{i=1}^{t} n_i\tau_i^2/(t-1)\,.$$

A Test Statistic for H_0

When H_0 is true then both MS_{Treat} and MS_{E} are unbiased estimates of σ^2

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

and MS_{Treat} will generally be somewhat larger than MS_{E}

and more so when the μ_i are more dispersed. The n_i act as magnifiers!

This suggests $F = MS_{\text{Treat}}/MS_{\text{E}}$ 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 $SS_{\rm T} = SS_{\rm Treat} + SS_{\rm E}$ the sum $SS_{\rm T}$ stays constant over all partitions of the full data set into t groups of sizes $n_1, \ldots, n_t$.

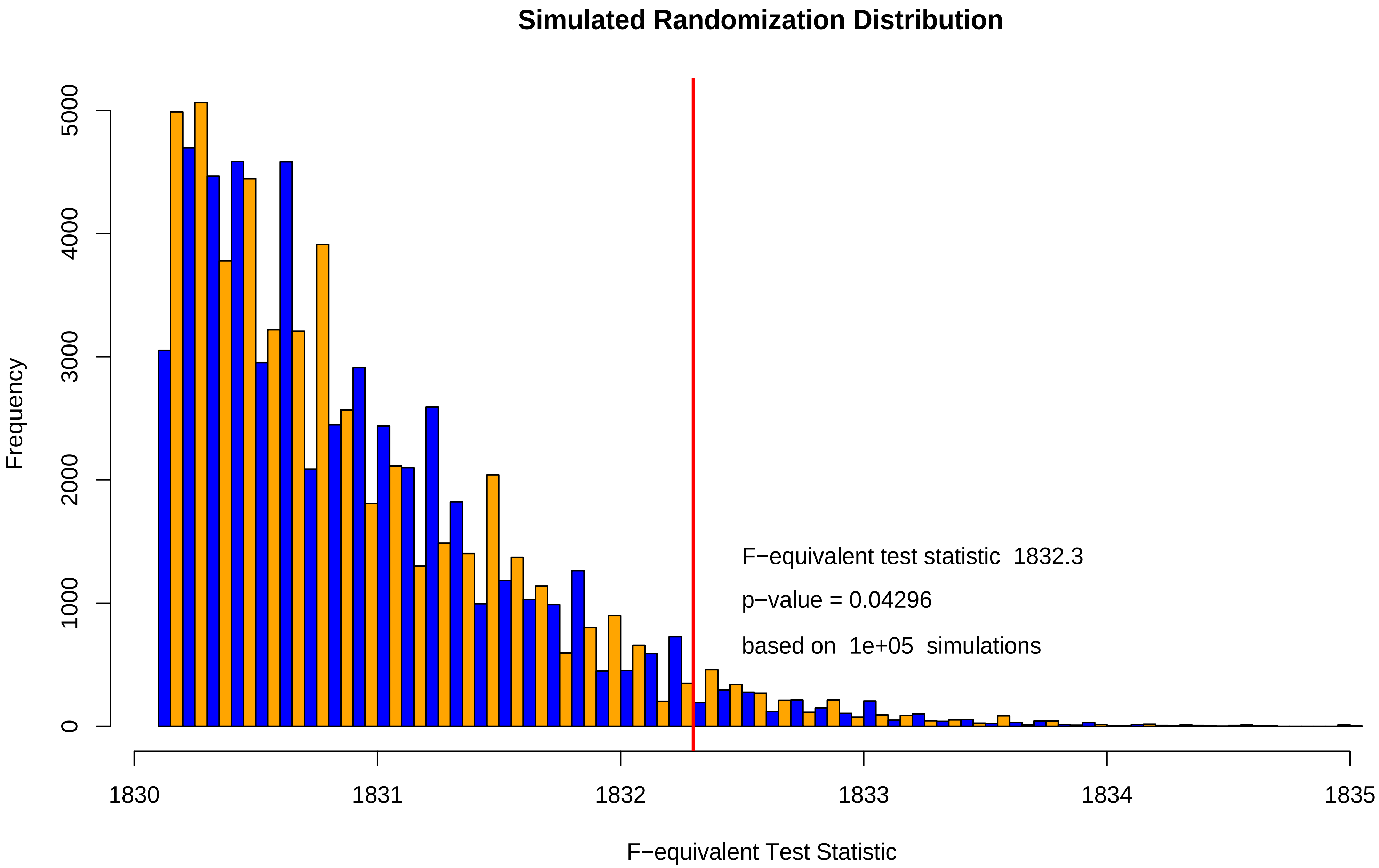
In $SS_{\rm Treat} = \sum_{i=1}^{t} n_i \bar{Y}_{i\bullet}^2 - N\bar{Y}_{\bullet\bullet}^2 = F_{\rm equiv} - N\bar{Y}_{\bullet\bullet}^2$ with $F_{\rm equiv} = \sum_{i=1}^{t} n_i \bar{Y}_{i\bullet}^2$ the term $\bar{Y}_{\bullet\bullet}$ stays constant over all such partitions.

Thus

$$F = \frac{N-t}{t-1}\,\frac{SS_{\rm Treat}}{SS_{\rm E}} = \frac{N-t}{t-1}\,\frac{SS_{\rm Treat}}{SS_{\rm T} - SS_{\rm Treat}} = \frac{N-t}{t-1}\,\frac{F_{\rm equiv} - N\bar{Y}_{\bullet\bullet}^2}{SS_{\rm T} - (F_{\rm equiv} - N\bar{Y}_{\bullet\bullet}^2)} \nearrow \text{ in } F_{\rm equiv}$$

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

Randomization Distribution for Flux3



R Code for Randomization Distribution

```
Ftest.rand = function (y=SIR,n=c(6,6,6),Nsim=10000){#try Nsim=10000 first for speed
F.obs=n[1]*mean(y[1:n[1]])^2+n[2]*mean(y[n[1]+
    1:n[2]])^2+n[3]*mean(y[n[1]+n[2]+1:n[3]])^2
F.eq=rep(0,Nsim)
for(i in 1:Nsim){
ind=sample(1:18)
F.eq[i]=n[1]*mean(y[ind[1:n[1]]])^2+
    n[2]*mean(y[ind[n[1]+1:n[2]]])^2+n[3]*mean(y[ind[n[1]+n[2]+1:n[3]]])^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!

F-Distribution as Approximation to the Randomization Distribution

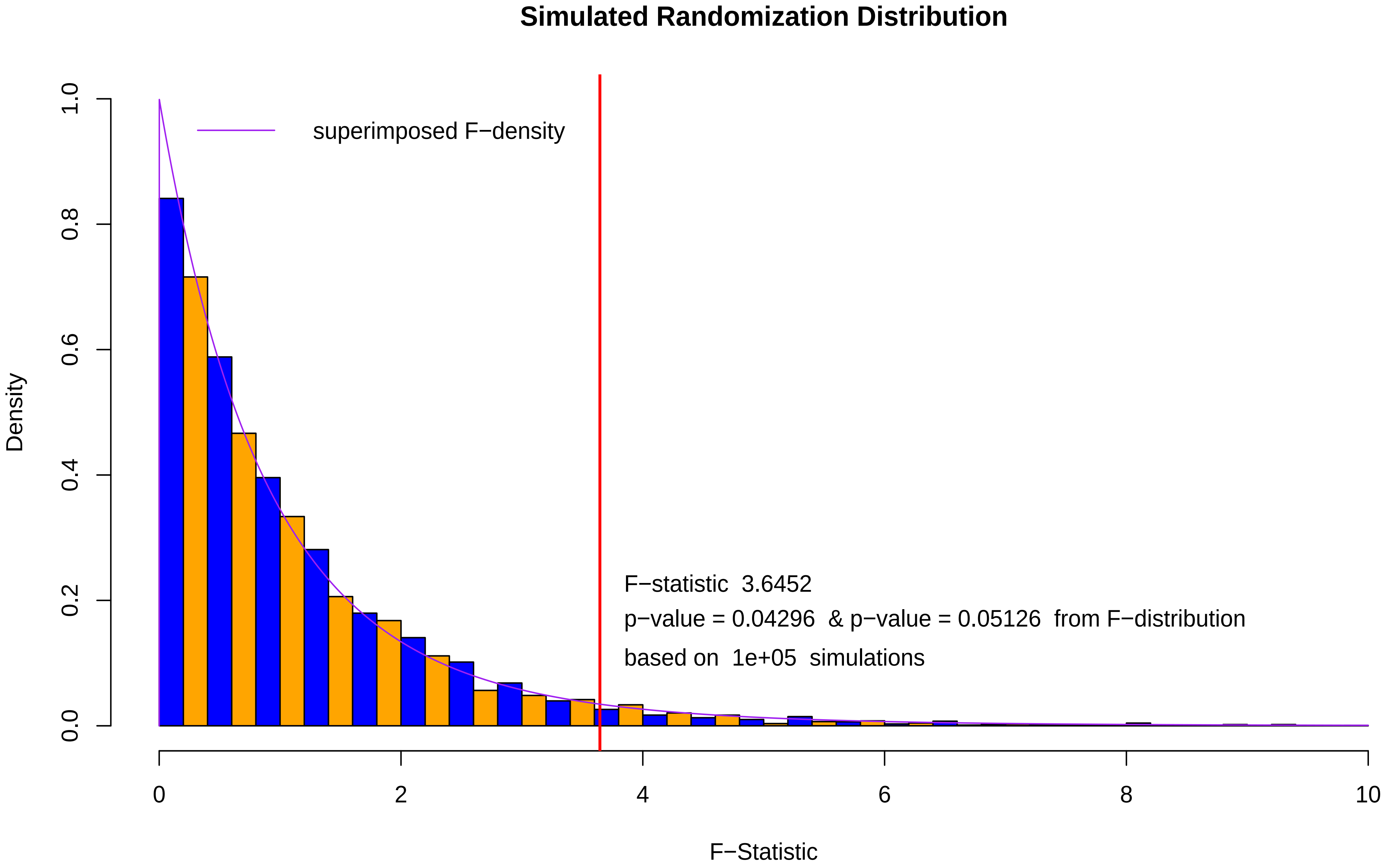
As in the case of the 2-sample problem one finds that the $F_{t-1,N-t}$ distribution often provides a good approximation to the randomization distribution of F.

The randomization distribution of F is obtained from that of F_{equiv} via

$$F = \frac{N-t}{t-1} \, \frac{F_{\text{equiv}} - N\bar{Y}_{\bullet\bullet}^2}{SS_{\text{T}} - (F_{\text{equiv}} - N\bar{Y}_{\bullet\bullet}^2)}$$

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

Randomization Distribution for Flux3



F−statistic 3.6452

p−value = 0.04296 & p−value = 0.05126 from F−distribution

based on 1e+05 simulations

Assuming Normality

In addition, we will now assume that the Y_{ij} are independent and have normal distributions with the previously indicated model parameters.

Whether $H_0 : \mu_1 = \ldots = \mu_t$ is true or not, we have $(n_i - 1)s_i^2 \sim \sigma^2\chi^2_{n_i-1}$.
Further, $s_1^2, \ldots, s_t^2$ are independent and thus

$$SS_{\text{E}} = \sum_{i=1}^{t}(n_i - 1)s_i^2 \sim \sigma^2\chi^2_{n_1-1} + \ldots + \sigma^2\chi^2_{n_t-1} \sim \sigma^2\chi^2_{N-t}$$

SS_{E} is independent of $\bar{Y}_{1\bullet}, \ldots, \bar{Y}_{t\bullet}$, since s_i^2 and $\bar{Y}_{i\bullet}$ are independent for all i and all pairs $(s_i^2, \bar{Y}_{i\bullet})$ are independent $\Longrightarrow$ SS_{E} and SS_{Treat} are independent.

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

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

The Distribution of F

The previous slide and Appendix A establish the following:

SS_E and SS_Treat are independent and

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

$$\Longrightarrow \qquad F = \frac{SS_\text{Treat}/(t-1)}{SS_\text{E}/(N-t)} \sim F_{t-1,N-t,\lambda}$$

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

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

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

Power function: $\beta(\lambda) = P(F \geq F_{t-1,N-t}(1-\alpha)) = \texttt{1} - \texttt{pf}(\texttt{Fcrit}, \texttt{t}-\texttt{1}, \texttt{N}-\texttt{t}, \lambda)$

R’s `anova` and `lm` Applied to Flux3

```
> 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
                Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(FLUX)  2 2.1733  1.0867  3.6452 0.05126 .
Residuals       15 4.4717  0.2981
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Discussion of Noncentrality Parameter λ

The power of the ANOVA F-test is a monotone function of $\lambda = \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu})^2/\sigma^2$

(See Appendix B)   Let us consider the drivers in λ.

λ increases as σ decreases (provided the μ_i are not all the same).

The more difference there is between the treatment means μ_i the higher λ

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

In fact: $\partial\lambda\sigma^2/\partial n_i = (\mu_i - \bar{\mu})^2 - \sum_j 2n_j(\mu_j - \bar{\mu})(\mu_i - \bar{\mu})/N = (\mu_i - \bar{\mu})^2 \geq 0$,
since $\partial\bar{\mu}/\partial n_i = \partial/\partial n_i \left(\sum_j n_j\mu_j/\sum_j n_j\right) = (\mu_i - \bar{\mu})/N$ i.e., increasing n_i never hurts.

The sample sizes we can plan for.

Later we address reducing σ by blocking units into more homogeneous groups.

Optimal Allocation of Sample Sizes?

We have N experimental units available for testing the effects of t treatments and suppose that N is a multiple of t, say $N = n \times t$ (n and t integer).

It would seem best to use samples of equal size n 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 optimality criterion that could be used as justification?

How many observations per treatment, i.e., how large should n be?

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.

Thus we need to be prepared for unbalanced designs.

A Sample Size Allocation Rationale

We may be concerned with alternatives where all means but one are the same. We want to achieve a given power β against such a mean, which deviates by Δ from the other means (which coincide).

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 $\bar{\mu} = \mu + n_1\Delta/N$.

With a bit of algebra we get

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

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.

The Max-Min Solution

It is easy to see now that for fixed σ

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

is achieved when $n_1 = \ldots = n_t$. That is because $R_i = (n_i/N)(1-n_i/N)$ increases for $n_i/N \le 1/2$. We can increase the smallest of these R_i only at the expense of lowering some of the other higher R_j, since $n_1+\ldots+n_t = N$ stays fixed. This increase can only happen when there is something left to lower.

Hence

$$\max_{n_1,\ldots,n_t} \min_{1\le i\le t} [\lambda_i] = \frac{N\Delta^2}{\sigma^2} \times \frac{n}{N}\left(1-\frac{n}{N}\right) = n \times \frac{\Delta^2}{\sigma^2} \times \left(1-\frac{n}{nt}\right) = n \times \frac{\Delta^2}{\sigma^2} \times \frac{t-1}{t} = n \times \lambda_0\,.$$

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

An Alternate Rationale

Dean and Voss discuss an alternate rationale for optimal sample size choice.

Find the optimal sample sizes $n_1, \ldots, n_t$ (with $\sum n_i = n \times t = N$), such that we have minimum power $\geq \beta$ when any two means differ by at least Δ, i.e., when

$$\max(\mu_1, \ldots, \mu_t) - \min(\mu_1, \ldots, \mu_t) \geq \Delta \, .$$

It can again be shown that equal sample size allocation, i.e., $n_1 = \ldots = n_t = n$, is the optimal (max-min) strategy.

A worst case mean scenario occurs when two means, say μ_1 and μ_t, differ by Δ while the other means coincide halfway between them, i.e.,

$$\mu_1 = \mu - \frac{\Delta}{2} \, , \quad \mu_t = \mu + \frac{\Delta}{2} \quad \text{and} \quad \mu_2 = \ldots = \mu_{t-1} = \mu \, .$$

Then $\lambda = n\Delta^2/(2\sigma^2) = n\lambda_1 \leq n(\Delta^2/\sigma^2) \times (t-1)/t$, with $=$ for $t = 2$.
Note that $\lambda_1 = \Delta^2/(2\sigma^2)$ does not depend on t.

Discussion of λ_0 and λ_1.

Δ is supposed to be the minimum mean difference to be detected with probability β under either rationale. We now make clear the difference between them.

Under the first rationale we basically assume that all but one treatment have no effect, and that the effect on the differing mean is at least $\pm\Delta$.

Under the second rationale we say that all treatments may have an effect, but that the maximum difference between some pair of means is at least Δ.
Among those scenarios the worst case is that one where $t-2$ treatments show no effect while the remaining two treatments have equal but opposite effects of size $\Delta/2$, relative to the unchanged means.

While the motivation seems acceptable, the worst case scenario appears contrived.

`sample.sizeANOVA` (see web page)

Just as in the case of planning appropriate sample sizes for the two-sample situation the F-test encounters the same difficulties in terms of the varying impacts of the common sample size n per treatment.

n affects the critical point of the level α F-test through

`tcrit=qf(1-alpha,t-1,N-t)=qf(alpha,t-1,n*t-t).`

n also enters the power function `1-pf(tcrit,t-1,n*t-t,lambda)` and n enters the power function through λ. Here $\lambda = n(\Delta/\sigma)^2(t-1)/t$ or $\lambda = n(\Delta/\sigma)^2/2$.

In either case we should know σ or have a reasonable upper bound σ_u, or express Δ not in absolute terms but in relation to the unknown σ by specifying Δ/σ.

To facilitate the choice of appropriate n per treatment the function `sample.sizeANOVA` is provided on the class web page.

Usage of `sample.sizeANOVA`

```
function (delta.per.sigma=.5,t.treat=3, nrange=2:30,alpha=.05,
   power0=NULL)
{
# delta.per.sigma is the ratio of delta over sigma for which
# one wants to detect a delta shift in one mean while all other
# means stay the same, or delta is the maximum difference
# between any two means to be detected. t.treat is the number of
# treatments. alpha is the desired significance level. nrange is a
# range of sample sizes over which the power will be calculated
# for that delta.per.sigma. power0 is on optional value for the
# target power that will be highlighted on the plot.
....
}
```

Example Usage of `sample.sizeANOVA`

The following three function calls invoke the default `t.treat=3` to produce the plots on the following three slides.

```
> sample.sizeANOVA()
> sample.sizeANOVA(nrange=30:100)
> sample.sizeANOVA(nrange=70:100,power0=.9)
```

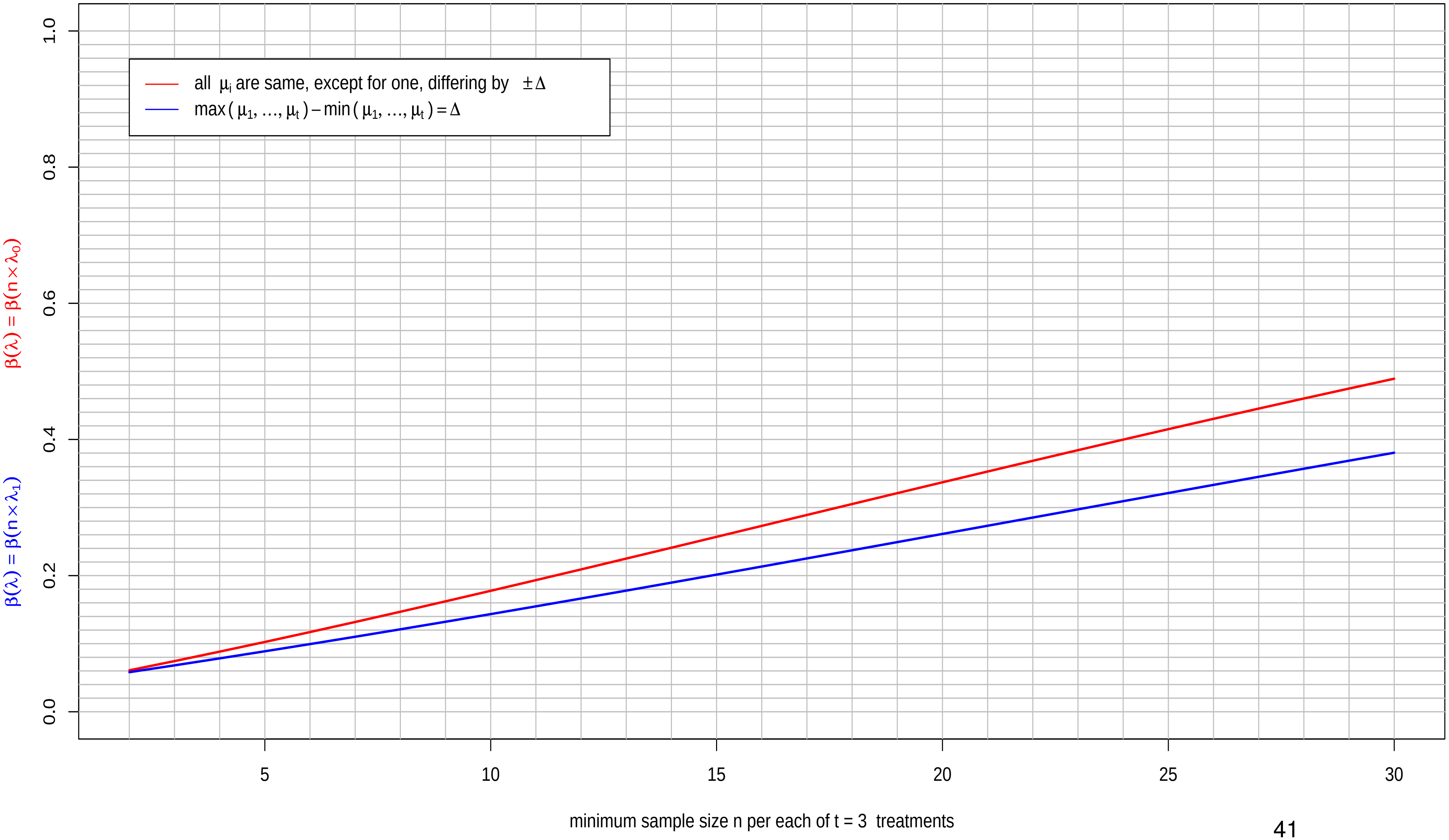
$\Longrightarrow \quad n = 77$ as the minimal sample size under the first rationale, w.r.t. λ_0

and

$\Longrightarrow \quad n = 103$ as the minimal sample size under the alternate rationale, w.r.t. λ_1

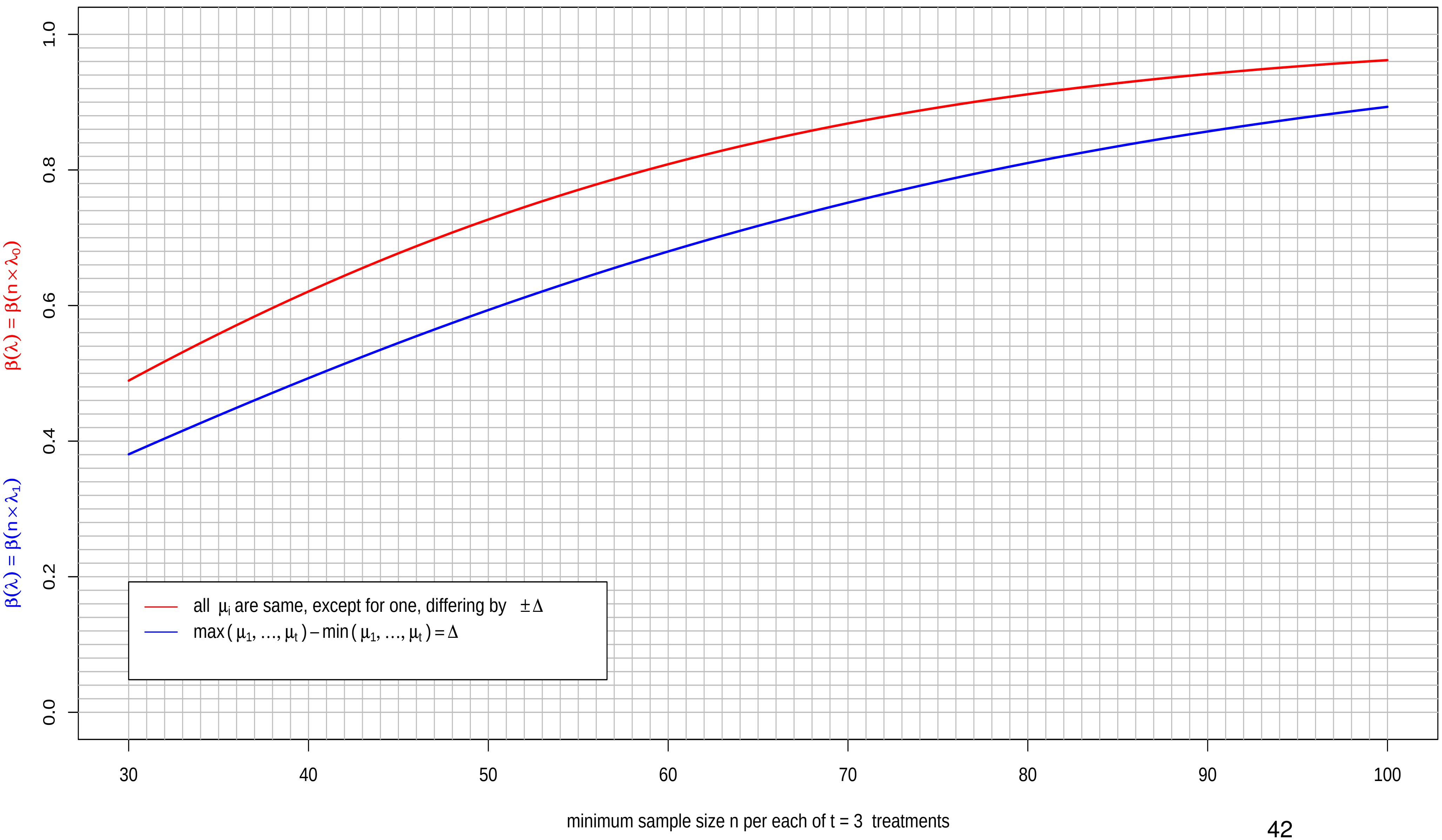
Sample Size Determination

$$\frac{\Delta}{\sigma} = 0.5 \,,\ \alpha = 0.05 \,, \quad \lambda_0 = \left(\frac{\Delta}{\sigma}\right)^2 \times \frac{t-1}{t} \,, \quad \lambda_1 = \frac{1}{2} \times \left(\frac{\Delta}{\sigma}\right)^2$$



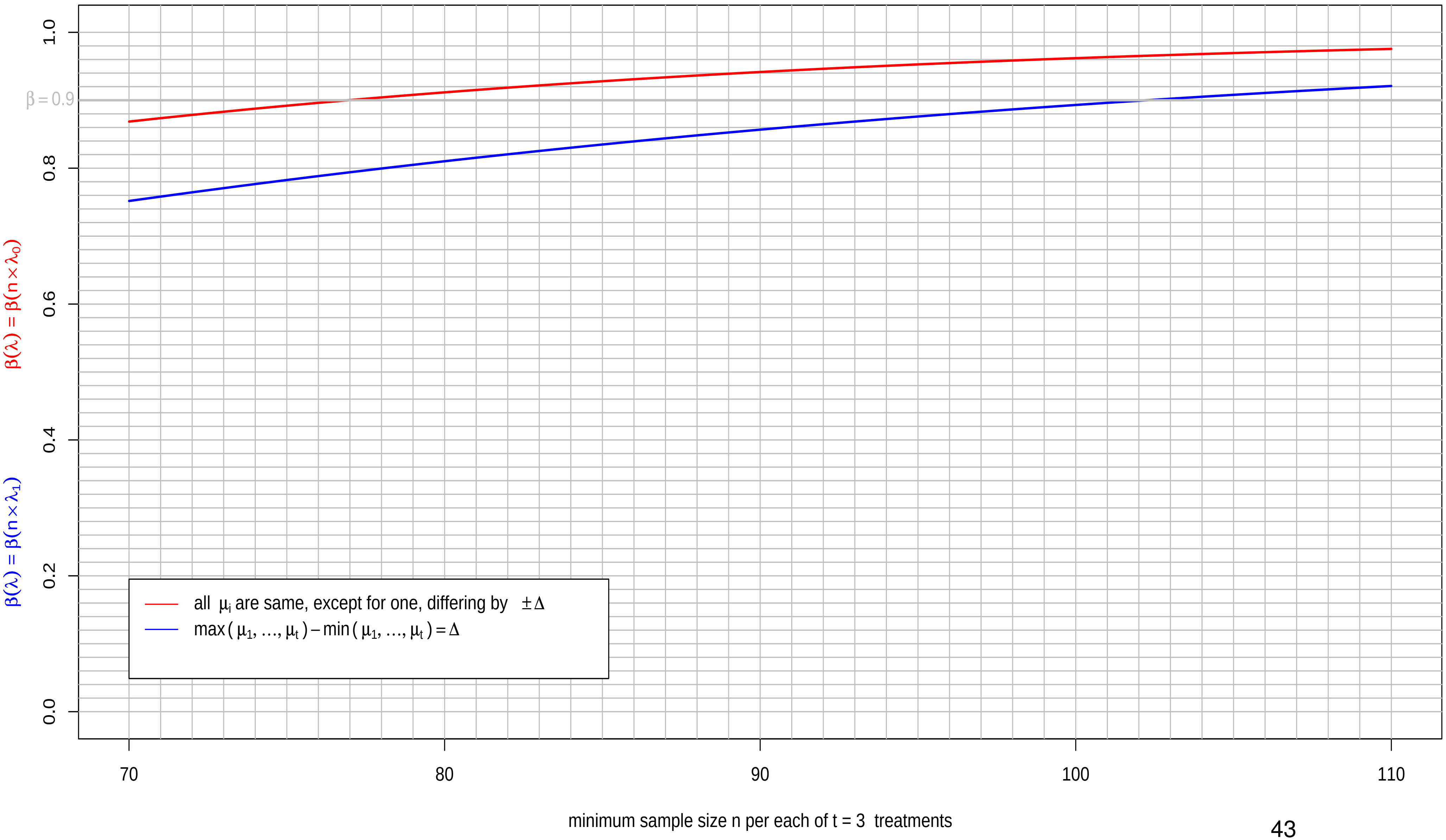
Sample Size Determination (increased *n*)

$$\frac{\Delta}{\sigma} = 0.5\ ,\ \alpha = 0.05\ ,\quad \lambda_0 = \left(\frac{\Delta}{\sigma}\right)^2 \times \frac{t-1}{t}\ ,\quad \lambda_1 = \frac{1}{2} \times \left(\frac{\Delta}{\sigma}\right)^2$$



Sample Size Determination (magnified)

$$\frac{\Delta}{\sigma} = 0.5\,,\ \alpha = 0.05\,,\quad \lambda_0 = \left(\frac{\Delta}{\sigma}\right)^2 \times \frac{t-1}{t}\,,\quad \lambda_1 = \frac{1}{2} \times \left(\frac{\Delta}{\sigma}\right)^2$$



The Effect of t

Even though the number of treatments does not affect λ_1 it affects the power function through the degrees of freedom

$\mathtt{tcrit} = \mathtt{qf}(1-\mathtt{alpha},\mathtt{t}-1,\mathtt{n}*\mathtt{t}-\mathtt{t})$ and $1-\mathtt{pf}(\mathtt{tcrit},\mathtt{t}-1,\mathtt{n}*\mathtt{t}-\mathtt{t},\mathtt{ncp})$

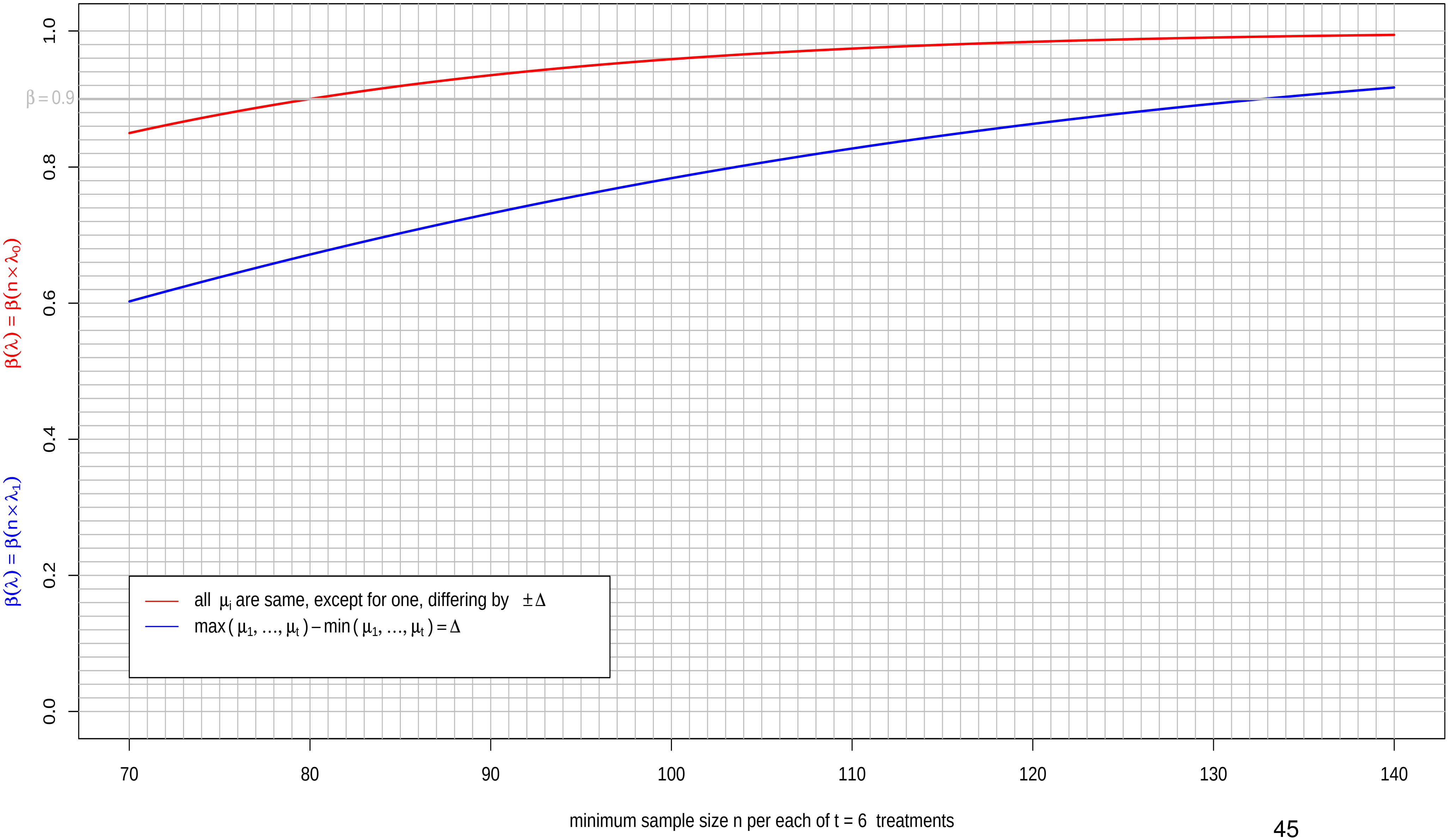
Thus the choice of n is very much affected, as can be seen in the following slide produced with $t = 6$

```
> sample.sizeANOVA(nrange=70:100,power0=.9,t.treat=6)
```

The minimum sample size per treatment is $n = 81$ under the first rationale (λ_0) and $n = 133$ under the alternate rationale (λ_1).

Sample Size Determination (magnified)

$$\frac{\Delta}{\sigma} = 0.5\ ,\ \alpha = 0.05\ ,\quad \lambda_0 = \left(\frac{\Delta}{\sigma}\right)^2 \times \frac{t-1}{t}\ ,\quad \lambda_1 = \frac{1}{2} \times \left(\frac{\Delta}{\sigma}\right)^2$$



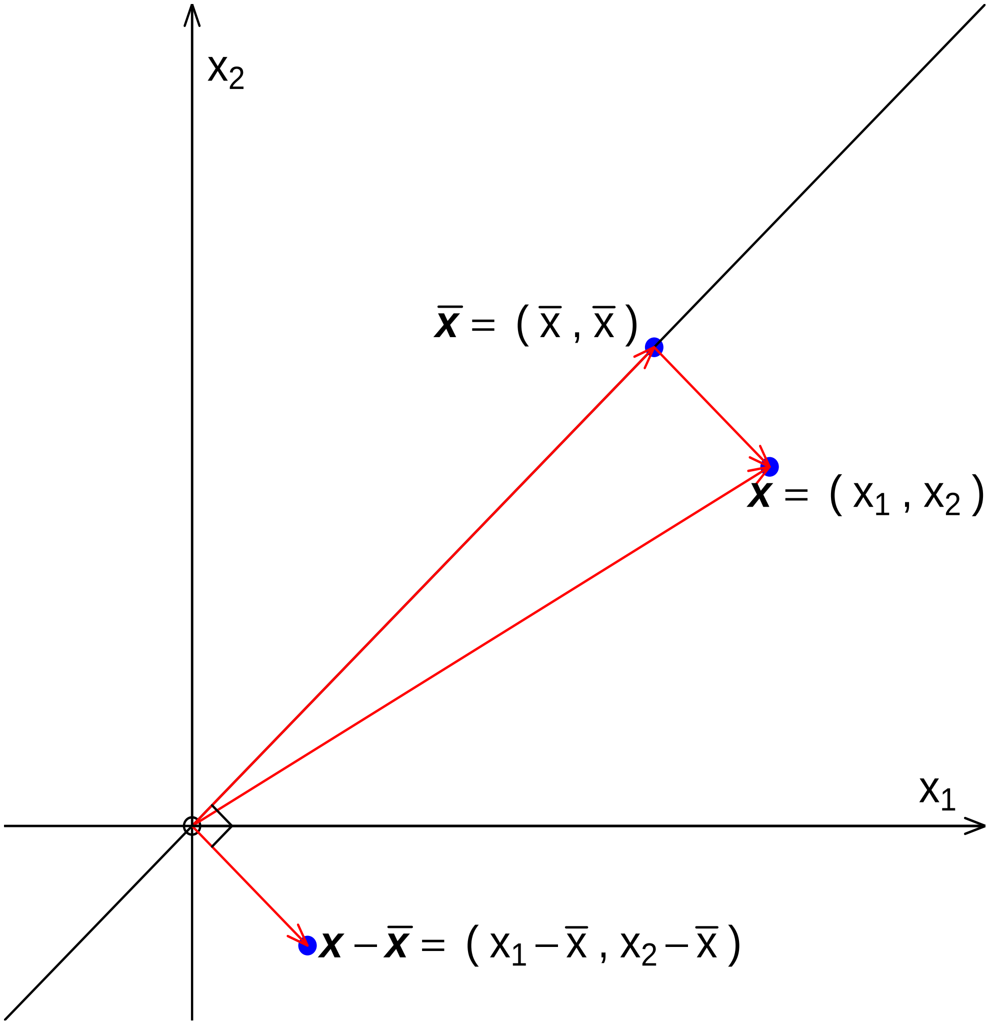
Degrees of Freedom and Geometry – Single Sample

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ \vdots \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} \bar{X} \\ \bar{X} \\ \vdots \\ \vdots \\ \vdots \\ \bar{X} \end{pmatrix} \quad \overset{\perp}{+} \quad \begin{pmatrix} X_1 - \bar{X} \\ X_2 - \bar{X} \\ \vdots \\ \vdots \\ \vdots \\ X_n - \bar{X} \end{pmatrix}$$

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

$(\bar{X},\ldots,\bar{X})$ varies in just one dimension, along $\mathbf{1}' = (1,\ldots,1)$, 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_2
x_1
$\bar{\boldsymbol{x}} = (\bar{x}, \bar{x})$
$\boldsymbol{x} = (x_1, x_2)$
$\boldsymbol{x} - \bar{\boldsymbol{x}} = (x_1 - \bar{x}, x_2 - \bar{x})$

Pythagoras

$$|\mathbf{x}|^2 = |\bar{\mathbf{x}}|^2 + |\mathbf{x} - \bar{\mathbf{x}}|^2$$

$$\begin{aligned} \textstyle\sum_i x_i^2 &= \textstyle\sum_i \bar{x}^2 + \sum_i (x_i - \bar{x})^2 \\ &= n\bar{x}^2 + \textstyle\sum_i (x_i - \bar{x})^2 \end{aligned}$$

our previous

SS decomposition

Degrees of Freedom and Geometry in t Samples

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

$$\begin{array}{ccccccccccc} N & = & 1 & + & N-1 & = & 1 & + & \sum(n_i - 1) & + & t-1 \\ & & & & & & & & N-t & & \end{array}$$

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

$$\sum_i \sum_j Y_{ij}^2 = \sum_i \sum_j \bar{Y}_{\bullet\bullet}^2 + \sum_i \sum_j (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 = \sum_i \sum_j \bar{Y}_{\bullet\bullet}^2 + \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\bullet})^2 + \sum_i \sum_j (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2$$

Orthogonalities

$$\sum_i \sum_j \bar{Y}_{\bullet\bullet}(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) = \bar{Y}_{\bullet\bullet} \sum_i n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) = \bar{Y}_{\bullet\bullet}(\sum_i \sum_j Y_{ij} - N\bar{Y}_{\bullet\bullet}) = 0$$

$$\sum_i \sum_j \bar{Y}_{\bullet\bullet}(Y_{ij} - \bar{Y}_{i\bullet}) = \bar{Y}_{\bullet\bullet} \sum_i (n_i\bar{Y}_{i\bullet} - n_i\bar{Y}_{i\bullet}) = 0$$

$$\sum_i \sum_j (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})(Y_{ij} - \bar{Y}_{i\bullet}) = \sum_i (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) \sum_j (Y_{ij} - \bar{Y}_{i\bullet}) = 0$$

Dimensions of Subspaces or Degrees of Freedom

Let $\mathbf{1}_n' = (1, 1, \ldots, 1)$ denote an n-vector filled with 1's. With varying Y_{ij}, the vectors

$$\begin{aligned}
\begin{pmatrix} \bar{Y}_{1\bullet} - \bar{Y}_{\bullet\bullet} \\ \vdots \\ \bar{Y}_{1\bullet} - \bar{Y}_{\bullet\bullet} \\ \vdots \\ \vdots \\ \bar{Y}_{t\bullet} - \bar{Y}_{\bullet\bullet} \\ \vdots \\ \bar{Y}_{t\bullet} - \bar{Y}_{\bullet\bullet} \end{pmatrix} &= (\bar{Y}_{1\bullet} - \bar{Y}_{\bullet\bullet}) \begin{pmatrix} \mathbf{1}_{n_1} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \ldots + (\bar{Y}_{t\bullet} - \bar{Y}_{\bullet\bullet}) \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \mathbf{1}_{n_t} \end{pmatrix} \\
&= (\bar{Y}_{1\bullet} - \bar{Y}_{\bullet\bullet})\mathbf{E}_1 + \ldots + (\bar{Y}_{t\bullet} - \bar{Y}_{\bullet\bullet})\mathbf{E}_t = \mathbf{D}
\end{aligned}$$

span a $(t-1)$-dimensional subspace of R^N, because the orthogonal vectors $\mathbf{E}_1, \ldots, \mathbf{E}_t$ span a t-dimensional subspace of R^N and $\mathbf{D}$ is always orthogonal to $\mathbf{1}_N' = (\mathbf{1}_{n_1}', \ldots, \mathbf{1}_{n_t}')$ $= \mathbf{E}_1' + \ldots + \mathbf{E}_t'$, since $\mathbf{1}_N'\mathbf{D} = (\mathbf{E}_1' + \ldots + \mathbf{E}_t')((\bar{Y}_{1\bullet} - \bar{Y}_{\bullet\bullet})\mathbf{E}_1 + \ldots + (\bar{Y}_{t\bullet} - \bar{Y}_{\bullet\bullet})\mathbf{E}_t)$ $= \sum_{i=1}^t n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) = 0$, because $\mathbf{E}_i'\mathbf{E}_i = n_i$ and $\mathbf{E}_i'\mathbf{E}_k = 0$ for $i \neq k$.

Note that $\sum_{i=1}^t a_i\mathbf{E}_i \perp \mathbf{1}_N = (\mathbf{E}_1 + \ldots + \mathbf{E}_t) \iff \sum_{i=1}^t n_i a_i = 0.$

More on Dimensions and Degrees of Freedom

Using the standard orthonormal basis vectors $\mathbf{e}_{ij}$ (with 1 in vector position (i,j) and 0 in all other positions) we have that

$$\mathbf{R} = \begin{pmatrix} Y_{11} - \bar{Y}_{1\bullet} \\ \vdots \\ Y_{1n_1} - \bar{Y}_{1\bullet} \\ \vdots \\ \vdots \\ Y_{t1} - \bar{Y}_{t\bullet} \\ \vdots \\ Y_{tn_t} - \bar{Y}_{t\bullet} \end{pmatrix} = \begin{array}{c} (Y_{11} - \bar{Y}_{1\bullet})\mathbf{e}_{11} + \ldots + (Y_{1n_1} - \bar{Y}_{1\bullet})\mathbf{e}_{1n_1} + \\ \cdots \\ \cdots \\ +(Y_{t1} - \bar{Y}_{t\bullet})\mathbf{e}_{t1} + \ldots + (Y_{tn_t} - \bar{Y}_{1\bullet})\mathbf{e}_{tn_t} \end{array} \qquad \perp \mathbf{E}_i \;\; \forall i$$

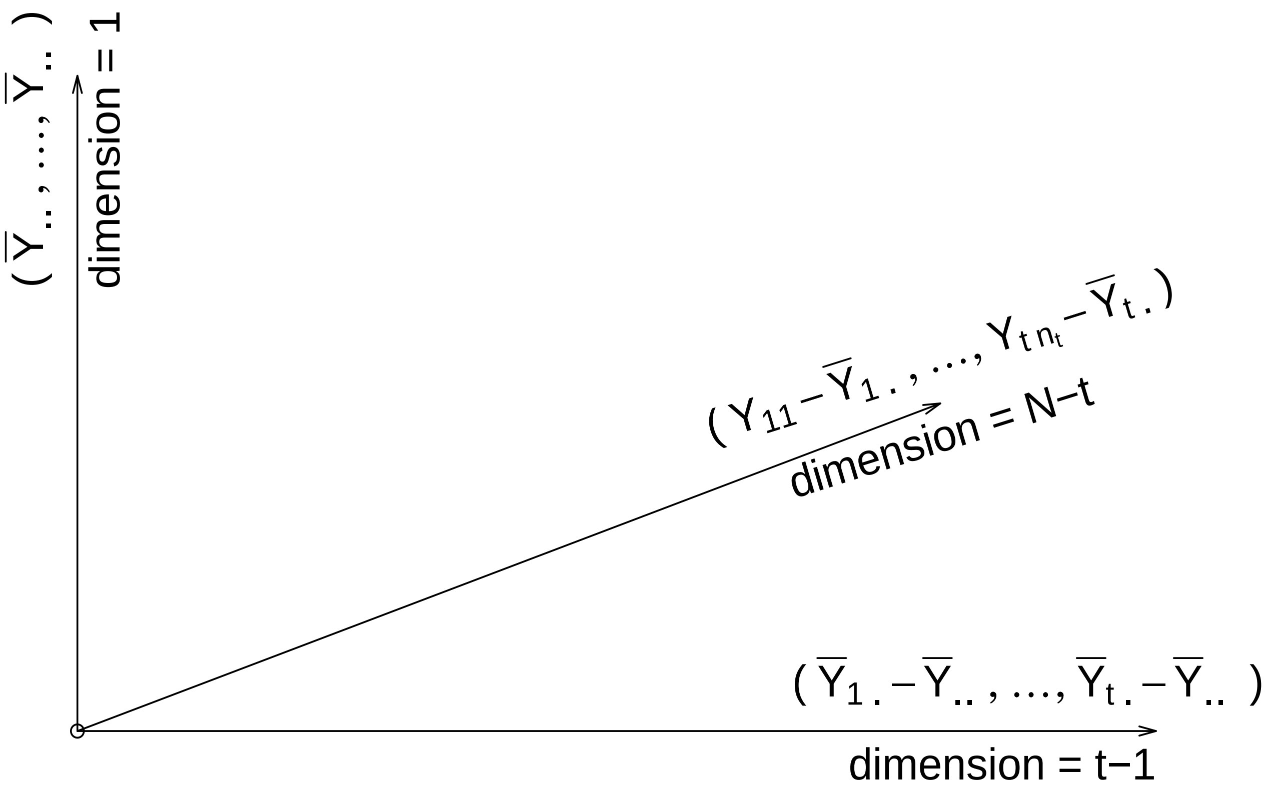
because $\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\bullet}) = 0$ for all i.

Thus $\mathbf{R}$ lives in the $N-t$ dimensional orthogonal complement M_{N-t} of $\mathbf{E}_1, \ldots, \mathbf{E}_t$.

Any vector $\mathbf{v}$ in M_{N-t} has to have the form

$\mathbf{v} = a_{11}\mathbf{e}_{11} + \ldots + a_{1n_1}\mathbf{e}_{1n_1} + \ldots + a_{t1}\mathbf{e}_{t1} + \ldots + a_{tn_1}\mathbf{e}_{tn_1}$ with $\sum_{j=1}^{n_i} a_{ij} = 0$ for $i = 1, \ldots, t$. Thus the $\mathbf{R}$ vectors span M_{N-t}.

Orthogonal Decomposition of Sample Space


(Ȳ.. , ..., Ȳ..)
dimension = 1
(Y₁₁ − Ȳ₁. , ..., Y_{t n_t} − Ȳ_t.)
dimension = N−t
(Ȳ₁. − Ȳ.. , ..., Ȳ_t. − Ȳ..)
dimension = t−1

$$|(Y_{11},\ldots,Y_{tn_t})|^2 = |(\bar{Y}_{\bullet\bullet},\ldots,\bar{Y}_{\bullet\bullet})|^2 + |(Y_{11}-\bar{Y}_{1\bullet},\ldots,Y_{tn_t}-\bar{Y}_{t\bullet})|^2 + |(\bar{Y}_{1\bullet}-\bar{Y}_{\bullet\bullet},\ldots,\bar{Y}_{t\bullet}-\bar{Y}_{\bullet\bullet})|^2$$

$$\sum_i\sum_j Y_{ij}^2 = \sum_i\sum_j \bar{Y}_{\bullet\bullet}^2 + \sum_i\sum_j (Y_{ij}-\bar{Y}_{i\bullet})^2 + \sum_i\sum_j (\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2$$

Coagulation Example

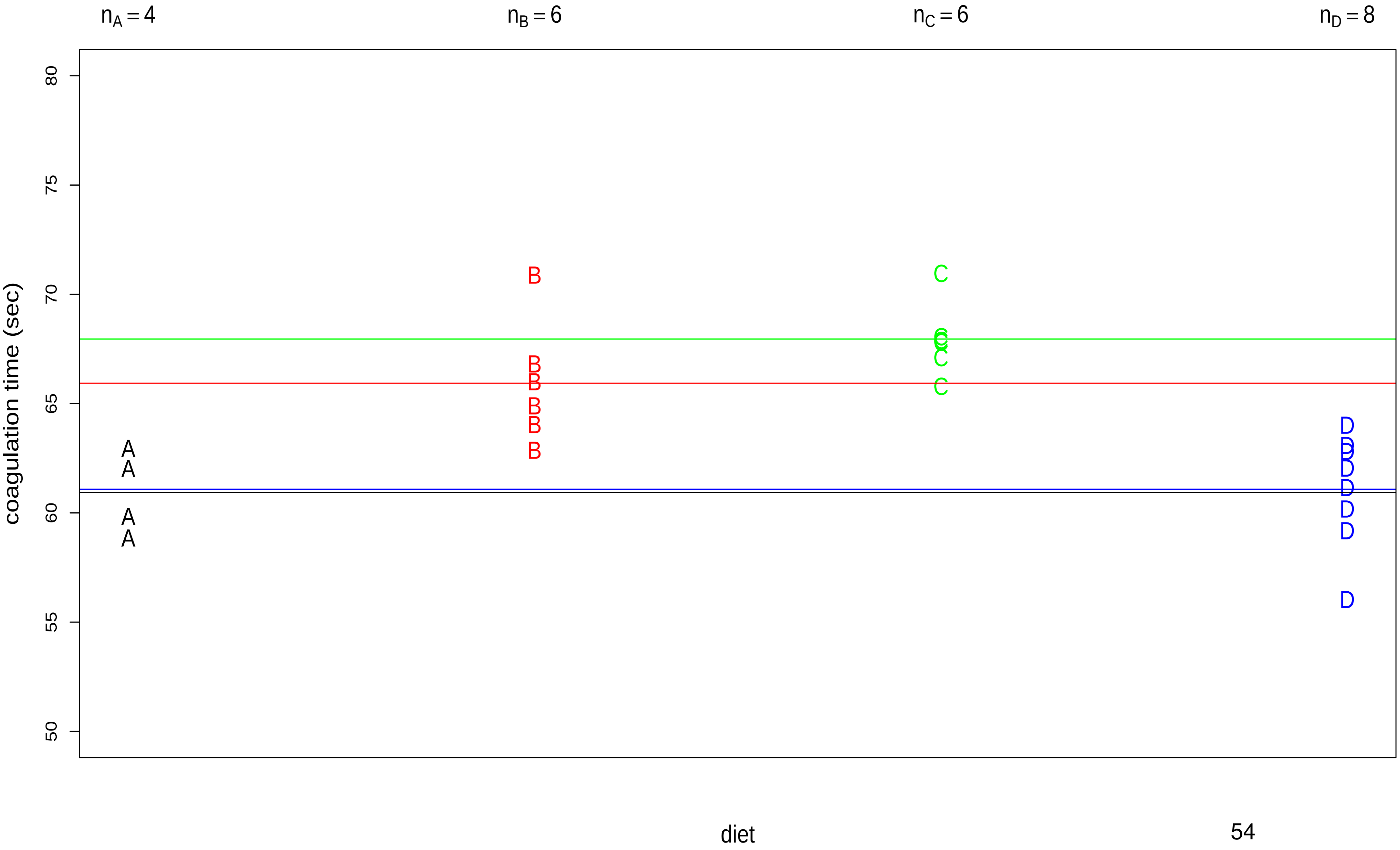
In order to understand the blood coagulation behavior in relation to various diets, lab animals were given 4 different diets and their subsequent blood draws were then measured for their respective coagulation times in seconds.

The lab animals were assigned randomly to the various diets.

The results were as follows:

```
> 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"
[14] "C" "C" "C" "D" "D" "D" "D" "D" "D" "D" "D"
```

Plot for Coagulation Example



ANOVA for Coagulation Example

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.

```
> anova(lm(ctime˜as.factor(diet))) # assumes ctime & diet in workspace
or > anova(lm(ctime˜as.factor(diet),data=coagulation.data))
# assumes coagulation.data is a list in the workspace
# with ctime & diet as components.
Analysis of Variance Table

Response: ctime
                Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(diet)  3  228.0    76.0  13.571 4.658e-05 ***
Residuals       20  112.0     5.6
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

`lm` for Coagulation Example

```
> out=lm(ctime˜as.factor(diet)) # this preserves all output from lm
> names(out)
 [1] "coefficients"  "residuals"      "effects"
 [4] "rank"          "fitted.values" "assign"
 [7] "qr"            "df.residual"    "contrasts"
[10] "xlevels"       "call"           "terms"
[13] "model"
> out$coefficients # or out$coef
      (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

$\hat{\mu}_A = 61$ `(Intercept)`, $\hat{\mu}_B - \hat{\mu}_A = 5$, $\hat{\mu}_C - \hat{\mu}_A = 7$, $\hat{\mu}_D - \hat{\mu}_A = 0$.

Residuals from `lm` for Coagulation Example

```
> 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
```

Numbers such as `-5.534852e-17` should be treated as 0 (computing quirks).

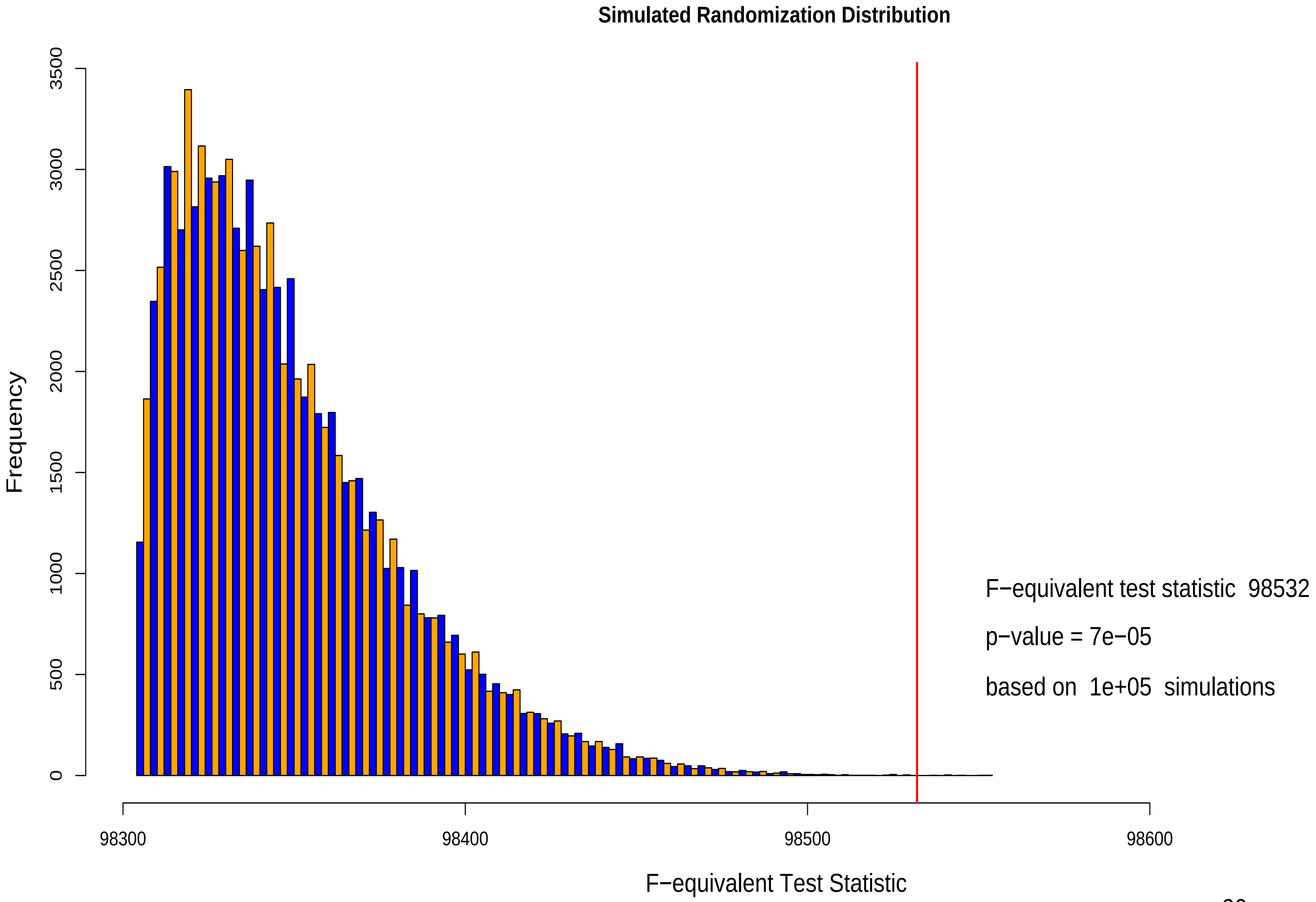
Rounded Residuals from `lm` for Coagulation Example

```
>  round(out$resid,4)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
-2 -1  1  2 -3 -2 -1  0  1  5 -2 -1  0  0  0  3 -5 -2 -1  0

21 22 23 24
 1  2  2  3
```

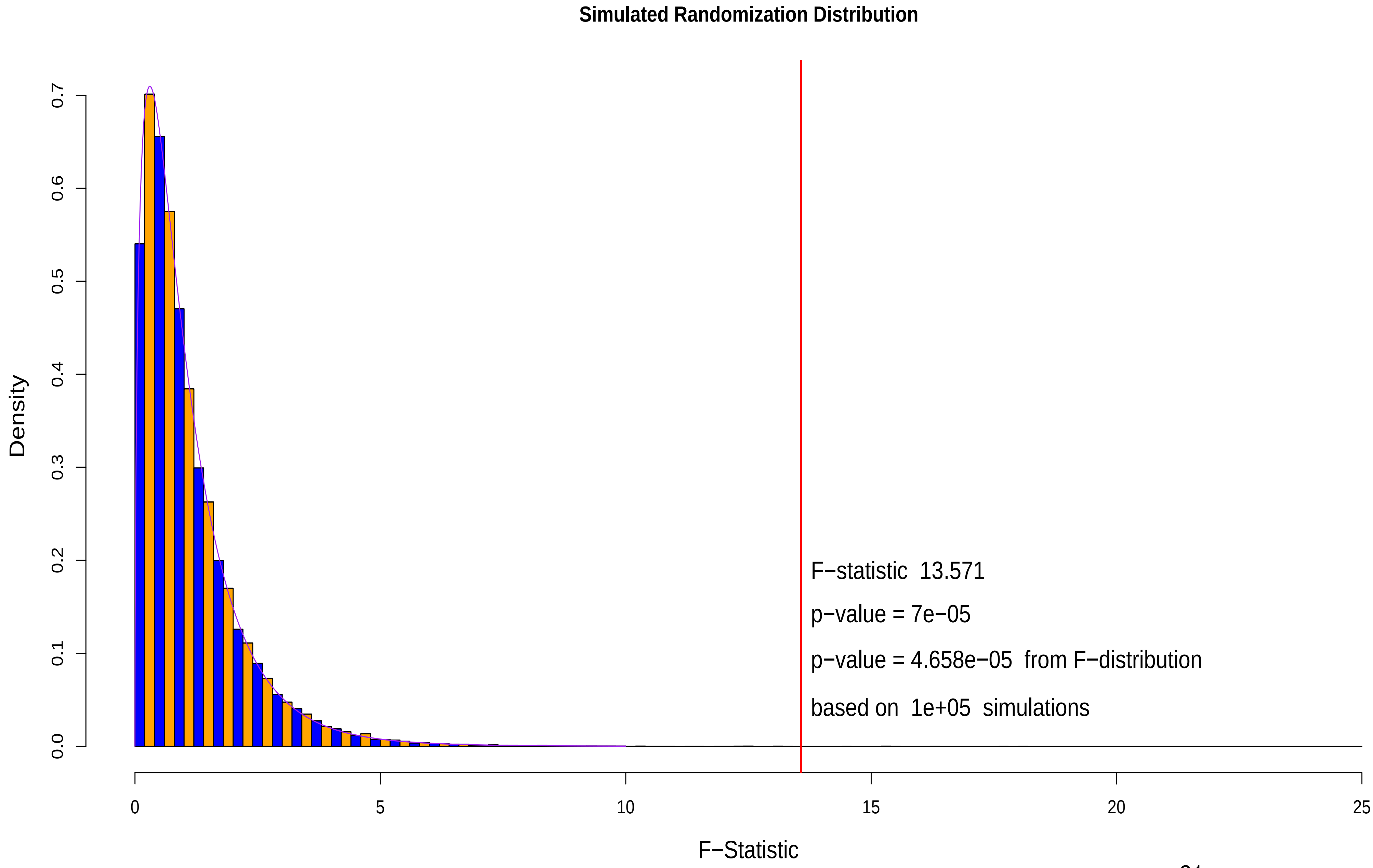
Fitted Values from `lm` for Coagulation Example

```
> 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 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



F-Approximation to Coagulation Randomization Test



Comparing Treatment Means $\bar{Y}_{i.}$

When the hypothesis $H_0 : \mu_1 = \ldots = \mu_t$ is not rejected at level α then there is little purpose in looking closer at differences between the sample means $\bar{Y}_{i\bullet}$ 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.

Confidence Intervals for μ_i

A first step in understanding differences in the μ_i is to look at their estimates $\hat{\mu}_i = \bar{Y}_{i\bullet}$.

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 μ_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_i^2 from the i^{th} sample, i.e. we get

$$\hat{\mu}_i \pm t_{N-t,1-\alpha/2} \times \frac{s}{\sqrt{n_i}} \qquad \text{as our } 100(1-\alpha)\% \text{ confidence interval for } \mu_i.$$

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 $s^2/\sigma^2 \sim \chi^2_{N-t}/(N-t)$ and combining this to

$$\frac{\hat{\mu}_i - \mu_i}{s/\sqrt{n_i}} = \frac{(\hat{\mu}_i - \mu_i)/(\sigma/\sqrt{n_i})}{s/\sigma} \sim t_{N-t}$$

Validity of Pooling?

Using s^2 instead of s_i^2 improves (narrows) the confidence intervals for μ_i.

This narrowing comes about because $t_{N-t,1-\alpha/2}$ then uses much higher degrees of freedom ($N-t \gg n_i - 1$) and thus shrinks, up to a point (see later plot).

The validity of this improvement depends strongly on the assumption that the population variances σ^2 behind all t samples are the same, or at least approximately so.

Recall our earlier discussion of this issue for the 2-sample t-test.

Standard Errors $SE(\hat{\theta})$

Suppose $\hat{\theta}$ is an estimator for a parameter θ of interest. We denote by $\sigma^2_{\hat{\theta}} = \mathrm{var}(\hat{\theta}) = g(\theta,\psi)$ 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., $\hat{\sigma}_{\hat{\theta}} = \sqrt{g(\hat{\theta},\hat{\psi})}$,

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 μ has variance $\mathrm{var}(\hat{\mu}) = \sigma^2/n \Rightarrow SE(\hat{\mu}) = s/\sqrt{n}$.

Example 2: $s^2 \sim \sigma^2\chi^2_{n-1}/(n-1)$ as estimate of σ^2 has sampling variance

$$\mathrm{var}(s^2) = \frac{\sigma^4\, 2(n-1)}{(n-1)^2} = \frac{2\sigma^4}{n-1} \quad\Longrightarrow\quad SE(s^2) = s^2\sqrt{\frac{2}{n-1}}$$

Note the different roles of (θ,ψ) 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 ψ. We only use $\hat{\theta}$ in $SE(\hat{\theta})$.

95%-Rule of Thumb Using SEs

If $\hat{\theta}$ has an approximate normal distribution with mean θ and standard deviation $\sigma_{\hat{\theta}}$,

i.e., $$\hat{\theta} \quad \approx \quad \mathcal{N}(\theta, \sigma^2_{\hat{\theta}}) \quad \approx \quad \mathcal{N}(\theta, SE^2(\hat{\theta}))\,,$$

as is often the case with many estimators

$\Longrightarrow \quad \hat{\theta} \pm 2 \times SE(\hat{\theta})$ is an approximately 95% confidence interval for θ

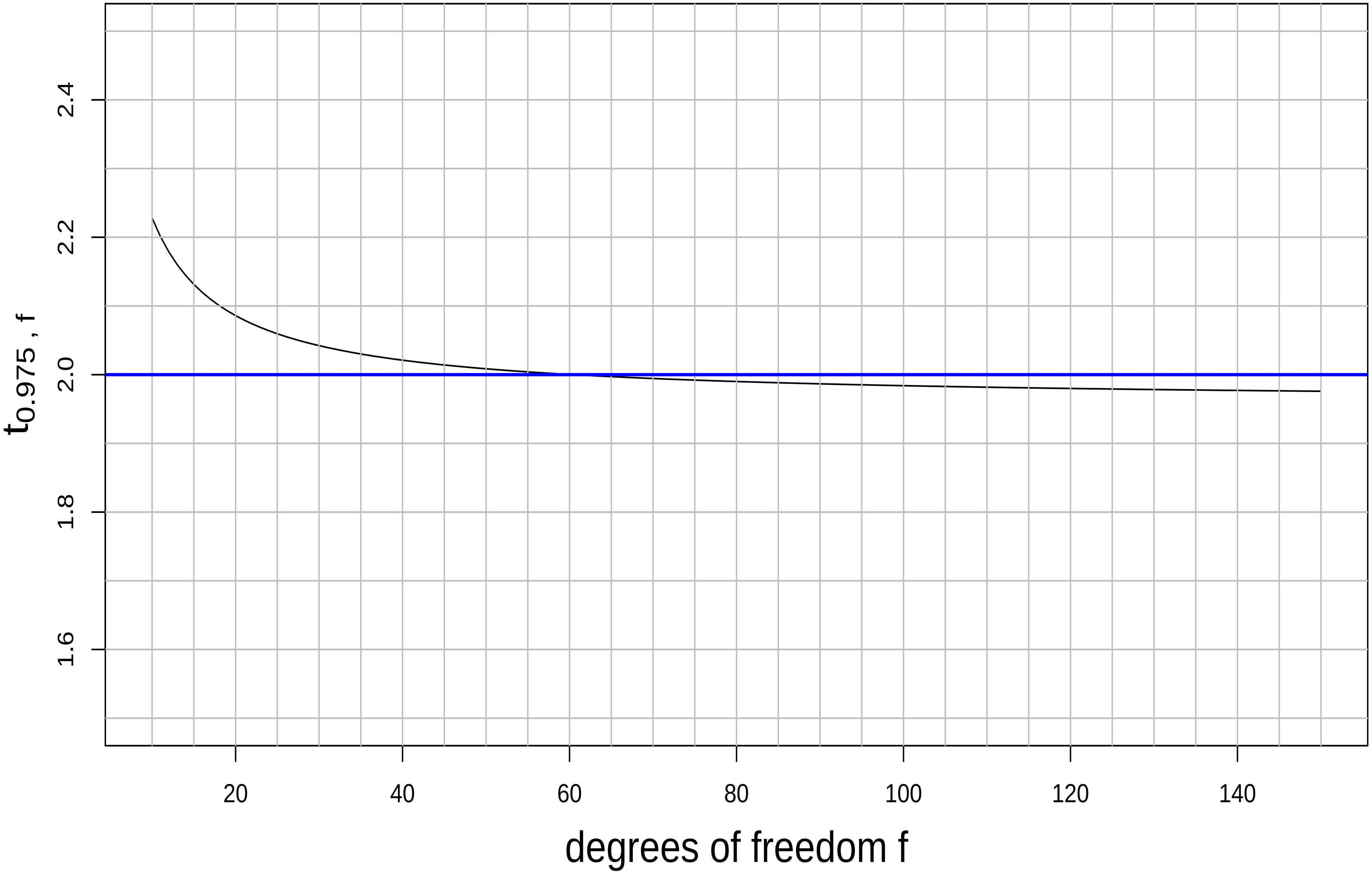
because $z_{.975} = \texttt{qnorm}(.975) = 1.959964 \approx 2$.

This works especially well for Student-t based intervals

$$\bar{\mu}_i \pm t_{f,.975} \times \frac{s}{\sqrt{n_i}} \quad = \quad \bar{Y}_{i\bullet} \pm t_{N-t,.975} \times \frac{s}{\sqrt{n_i}}$$

because $t_{f,.975} \approx z_{.975}$ for large f, see next slide.

$$t_{f,.975} \rightarrow z_{.975} = 1.96 \approx 2$$



Why Rule of Thumb Works for s^2

Why should the rule of thumb work for s^2 as estimator of σ^2?

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

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

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

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

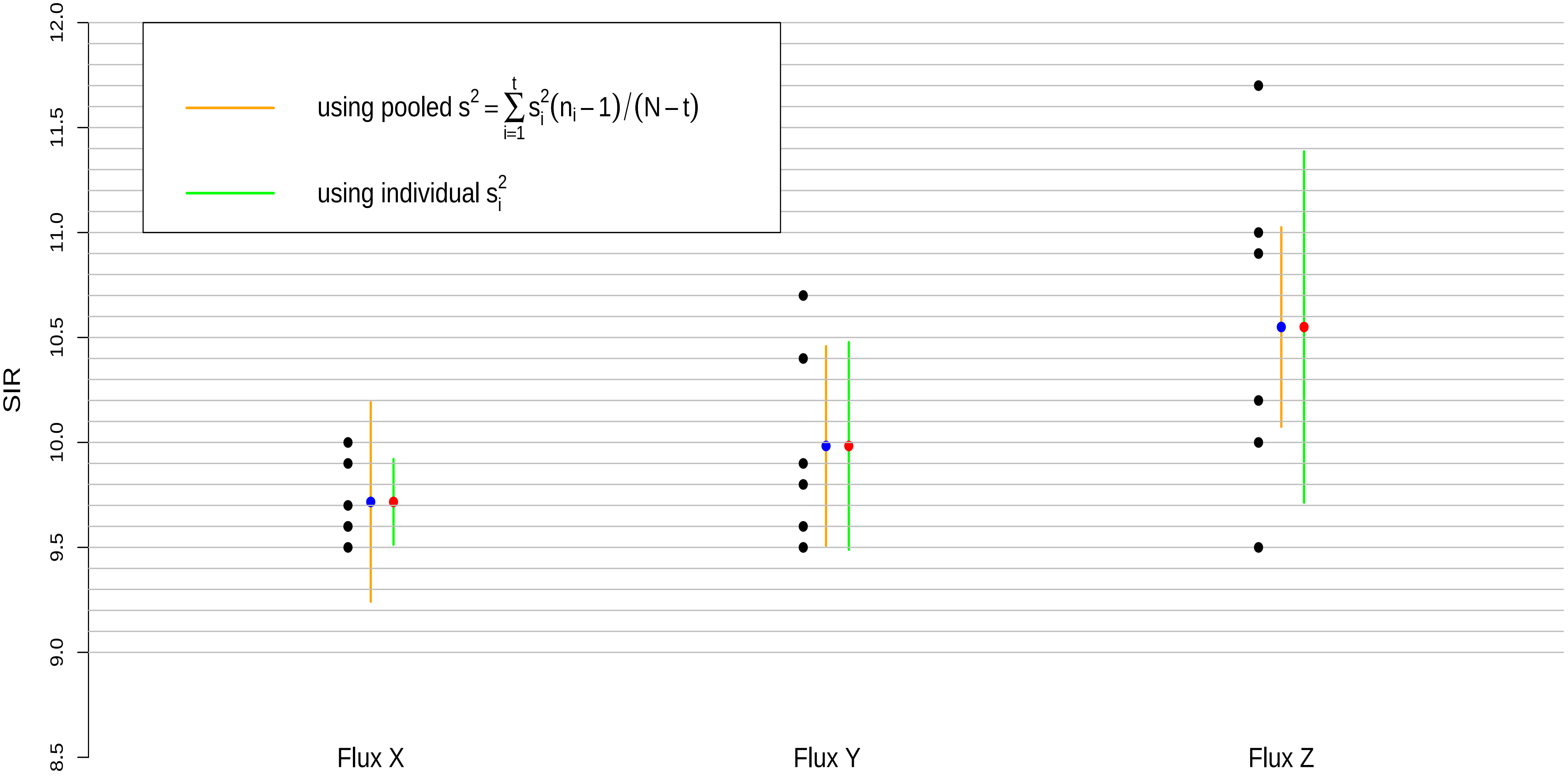
the sampling standard deviation of s^2.

Table of Confidence Intervals for Flux3 Data

Although for testing $H_0 : \mu_1 = \mu_2 = \mu_3$ in the case of the Flux3 data the p-value of $.05126$ was not significant at level $\alpha = .05$ we illustrate the concepts of the different types of confidence intervals for the means.

Flux	$\hat{\mu}_i$	s_i	s	95% intervals using s_i	95% intervals using s
X	9.717	0.194	0.546	$[9.513, \quad 9.920]$	$[\ 9.242, 10.192]$
Y	9.983	0.471	0.546	$[9.489, 10.477]$	$[\ 9.508, 10.458]$
Z	10.550	0.797	0.546	$[9.714, 11.386]$	$[10.075\ , 11.025]$

Plots of Confidence Intervals for Flux3 Data

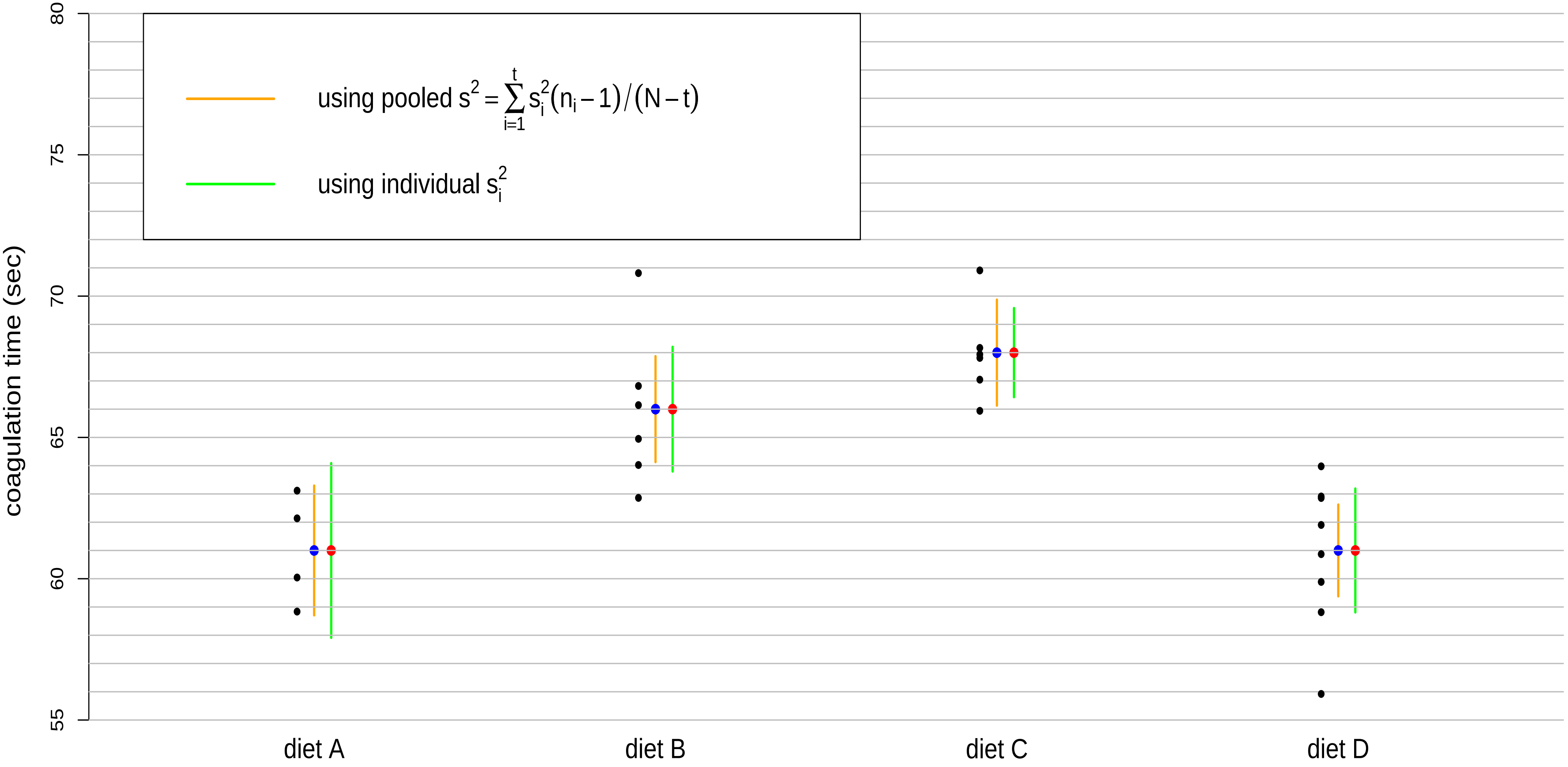


Tables of Confidence Intervals for the Coagulation Data

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.

Diet	$\hat{\mu}_i$	s_i	s	95% intervals using s_i	95% intervals using s
A	61	1.9	2.2	$[57.9, 64.1]$	$[58.7, 63.3]$
B	66	2.1	2.2	$[63.8, 68.2]$	$[64.1, 67.9]$
C	68	1.5	2.2	$[66.4, 69.6]$	$[66.1, 69.9]$
D	61	2.6	2.2	$[58.8, 63.2]$	$[59.4, 59.4]$

Plots of Confidence Intervals for Coagulation Data



Simultaneous Confidence Intervals

When constructing intervals of the type:

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

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

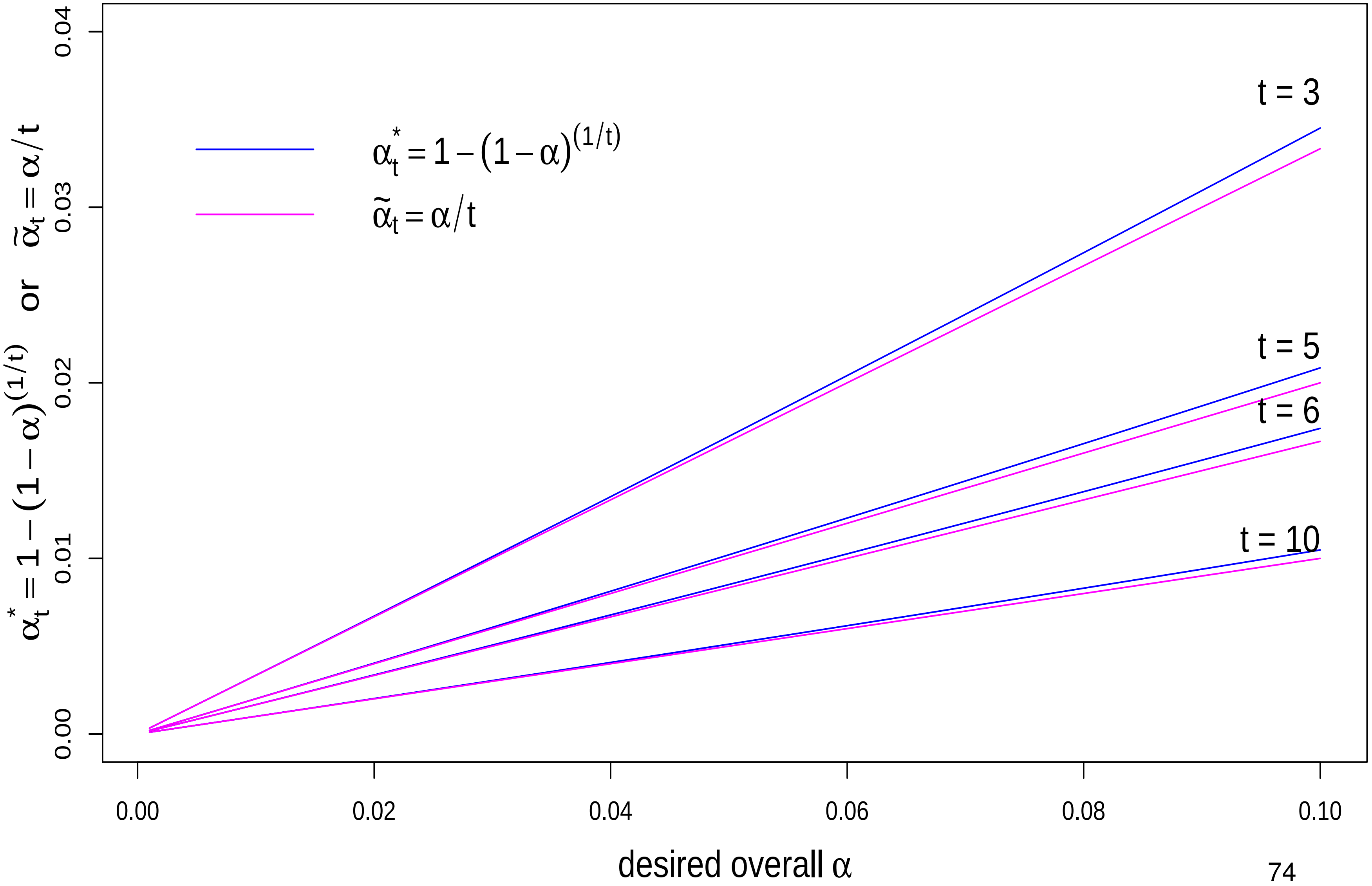
$$P\left(\mu_i \in \hat{\mu}_i \pm t_{n_i-1,1-\alpha/2}\frac{s_i}{\sqrt{n_i}},\ \ i = 1,\ldots,t\right) = \prod_{i=1}^{t} P\left(\mu_i \in \hat{\mu}_i \pm t_{n_i-1,1-\alpha/2}\frac{s_i}{\sqrt{n_i}}\right)$$
$$= (1-\alpha)^t < 1-\alpha.$$

To get simultaneous $1-\alpha$ coverage probability we should choose $1-\alpha^\star$ for individual interval coverage probability to get

$$(1-\alpha^\star)^t = 1-\alpha \qquad \text{or} \qquad \alpha^\star = 1-(1-\alpha)^{1/t} \approx \frac{\alpha}{t} = \tilde{\alpha}_t\ .$$

A problem remains when using a common pooled estimate s. No independence!

$$\alpha^{\star} = 1 - (1-\alpha)^{1/t} \approx \alpha/t$$



Dealing with Dependence from Using Pooled s

When we use a common pooled estimate s for the standard deviation σ

the previous confidence intervals are no longer independent.

However, it can be shown that

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

This comes from the positive dependence between confidence intervals through s,

i.e., if one interval is more (less) likely to cover its target μ_i due to s, so are the other intervals more (less) likely to cover their targets μ_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 $1-\alpha$.

Boole's and Bonferroni's Inequality

For any m events $E_1,\ldots,E_m$ Boole's inequality states

$$P(E_1\cup\ldots\cup E_m)\le P(E_1)+\ldots+P(E_m)$$

For any m events $E_1,\ldots,E_m$ Bonferroni's inequality states

$$P(E_1\cap\ldots\cap E_m)\ge 1-\sum_{i=1}^{m}(1-P(E_i))$$

The statements are equivalent, since $P(E_1^c\cup\ldots\cup E_m^c)\le P(E_1^c)+\ldots+P(E_m^c)\Longleftrightarrow$

$$P(E_1\cap\ldots\cap E_m)=1-P\left((E_1\cap\ldots\cap E_m)^c\right)=1-P(E_1^c\cup\ldots\cup E_m^c)\ge 1-\sum_{i=1}^{m}(1-P(E_i))$$

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

$$P\left(\bigcap_{i=1}^{t}E_i\right)\ge 1-\sum_{i=1}^{t}(1-P(E_i))=1-t\tilde{\alpha}=1-\alpha\quad\text{if}\quad\tilde{\alpha}=\tilde{\alpha}_t=\alpha/t\ ,$$

i.e., we can achieve at least $1-\alpha$ probability coverage by choosing the individual coverage appropriately, namely $1-\tilde{\alpha}=1-\alpha/t$.   Almost same adjustment.

Decomposing the Mean Vector μ

Variation in the means μ_i is best understood through the familiar decomposition:

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

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

It is this second vector that captures all aspects of variation in μ.

Why $(t-1)$-Dimensional Subspace V_{t-1}?

$$\begin{pmatrix}\mu_1-\bar{\mu}\\ \vdots\\ \mu_1-\bar{\mu}\\ \vdots\\ \vdots\\ 0\\ \vdots\\ 0\end{pmatrix} \overset{\perp}{+}\dots\overset{\perp}{+} \begin{pmatrix}0\\ \vdots\\ 0\\ \vdots\\ \vdots\\ \mu_t-\bar{\mu}\\ \vdots\\ \mu_t-\bar{\mu}\end{pmatrix} = \overset{\gamma_1}{n_1(\mu_1-\bar{\mu})} \begin{array}{c}\mathbf{a}_1\\ \begin{pmatrix}1/n_1\\ \vdots\\ 1/n_1\\ \vdots\\ \vdots\\ 0\\ \vdots\\ 0\end{pmatrix}\end{array} \overset{\perp}{+}\dots\overset{\perp}{+}\overset{\gamma_t}{n_t(\mu_t-\bar{\mu})} \begin{array}{c}\mathbf{a}_t\\ \begin{pmatrix}0\\ \vdots\\ 0\\ \vdots\\ \vdots\\ 1/n_t\\ \vdots\\ 1/n_t\end{pmatrix}\end{array}$$

$$= \gamma_1\mathbf{a}_1+\ldots+\gamma_t\mathbf{a}_t = \sum_{i=1}^{t-1}\gamma_i\mathbf{a}_i - \sum_{i=1}^{t-1}\gamma_i\mathbf{a}_t = \sum_{i=1}^{t-1}\gamma_i(\mathbf{a}_i-\mathbf{a}_t)$$

since

$$\sum_{i=1}^{t}\gamma_i = \sum_{i=1}^{t} n_i(\mu_i-\bar{\mu}) = 0 \quad \text{and thus} \quad \gamma_t = -\sum_{i=1}^{t-1}\gamma_i$$

and $\mathbf{a}_1-\mathbf{a}_t,\ldots,\mathbf{a}_{t-1}-\mathbf{a}_t$ are $t-1$ linearly independent vectors, spanning V_{t-1}.

$\sum_{i=1}^{t-1}x_i(\mathbf{a}_i-\mathbf{a}_t)=0 \quad\Longrightarrow\quad \sum_{i=1}^{t-1}x_i\mathbf{a}_i=\mathbf{a}_t\sum_{i=1}^{t-1}x_i \quad\Longrightarrow\quad x_1=\ldots=x_{t-1}=0.$

Motivating Contrasts

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 $\sum_{i=1}^{t} c_i = 0$.

$$\begin{aligned}
\sum_{i=1}^{t} a_i(\mu_i - \bar{\mu}) &= \sum_{i=1}^{t} a_i \mu_i - \sum_{i=1}^{t} a_i \sum_{j=1}^{t} \frac{n_j}{N} \mu_j \\
&= \sum_{i=1}^{t} a_i \mu_i - \sum_{i=1}^{t} \frac{n_i}{N} \mu_i \sum_{j=1}^{t} a_j = \sum_{i=1}^{t} c_i \mu_i \quad \text{with} \quad c_i = a_i - \frac{n_i}{N} \sum_{j=1}^{t} a_j \\
&\qquad \text{where} \quad \sum_{i=1}^{t} c_i = \sum_{i=1}^{t} a_i - \sum_{i=1}^{t} \frac{n_i}{N} \sum_{j=1}^{t} a_j = \sum_{i=1}^{t} a_i - \sum_{j=1}^{t} a_j = 0\,.
\end{aligned}$$

Such a function $C = \sum_{i=1}^{t} c_i \mu_i$ of the μ_i, with $\sum_{i=1}^{t} c_i = 0$, is called a contrast.

Examples of Contrasts

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 $\mathbf{c}'=(c_1,\ldots,c_4)=(1,-1,0,0)$. 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 $\mathbf{c}'=(1,-1/3,-1/3,-1/3)$.

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 $\mathbf{c}=(1/2,1/2,-1/2,-1/2)$

Estimates and Confidence Intervals for Contrasts

A natural estimate of $C = \sum_{i=1}^{t} c_i \mu_i$ is $\hat{C} = \sum_{i=1}^{t} c_i \hat{\mu}_i = \sum_{i=1}^{t} c_i \bar{Y}_{i\bullet}$.

We have $$E(\hat{C}) = E\left(\sum_{i=1}^{t} c_i \bar{Y}_{i\bullet}\right) = \sum_{i=1}^{t} c_i E\left(\bar{Y}_{i\bullet}\right) = \sum_{i=1}^{t} c_i \mu_i = C$$

and $$\mathrm{var}(\hat{C}) = \mathrm{var}\left(\sum_{i=1}^{t} c_i \bar{Y}_{i\bullet}\right) = \sum_{i=1}^{t} c_i^2 \mathrm{var}\left(\bar{Y}_{i\bullet}\right) = \sum_{i=1}^{t} c_i^2 \sigma^2 / n_i \, .$$

Under the normality assumption for the Y_{ij} we have

$$\frac{\hat{C} - C}{s\sqrt{\sum_{i=1}^{t} c_i^2 / n_i}} \sim t_{N-t} \quad \text{where} \quad s^2 = \frac{\sum_{i=1}^{t} (n_i - 1) s_i^2}{N - t} = \frac{\sum_{ij} (Y_{ij} - \bar{Y}_{i\bullet})^2}{N - t} = MS_{\mathrm{E}} \, .$$

$\Longrightarrow \hat{C} \pm t_{N-t,1-\alpha/2} \times s \times \sqrt{\sum_{i=1}^{t} c_i^2 / n_i}$ is a $100(1-\alpha)\%$ confidence interval for C.

Testing $H_0 : C = 0$

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

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

$SE(\hat{C})$ is the standard error of $\hat{C}$, the estimate of the standard deviation of $\hat{C}$.

Simultaneous Confidence Intervals for Contrasts

Just as with simultaneous confidence intervals for means we need to face the issue of simultaneous coverage probability in relation to the individual coverage probability for each such interval.

We will introduce/compare several such procedures, although there are still others.

The subject of such multiple comparisons is a very active research area.

Simultaneous Statistical Inference by Rupert Miller (1966)

Multiple Comparison Procedures by Yosef Hochberg and Ajit Tamhane (1987)

Multiple Comparisons: Theory and Methods by Jason Hsu (1996)

Multiple Comparisons and Multiple Tests by Peter Westfall (2000).

Paired Comparisons: Fisher's Protected LSD Method

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 α 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.

Comments on Fisher's Protected LSD Method

If H_0 is true, the chance of making any statements contradicting H_0 is at most α. 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 μ_1) are equal and μ_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 α proportion of cases with wrongly declared differences.

This is due to the multiple comparison issue.

Pairwise Comparisons: Tukey-Kramer Method

The Tukey-Kramer method is based on the distribution of

$$Q_{t,f} = \max_{1\le i<j\le t}\left\{\frac{|Z_i - Z_j|}{s}\right\} \qquad \text{where } Z_1,\ldots,Z_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \text{ and } f\times s^2 \sim \chi^2_f$$

Its cdf and quantile function are given in R as `ptukey(q,nmeans,df)` and `qtukey(p,nmeans,df)`, $\texttt{nmeans} = t$ is the number of means, $\texttt{df} = f = N - t$ 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

$$\max_{i<j}\left\{\frac{\sqrt{n}|\hat\mu_i - \hat\mu_j - (\mu_i - \mu_j)|}{s}\right\} = \max_{i<j}\left\{\frac{\left|\frac{\hat\mu_i - \mu_i}{\sigma/\sqrt{n}} - \frac{\hat\mu_j - \mu_j}{\sigma/\sqrt{n}}\right|}{s/\sigma}\right\} = Q_{t,f}$$

$$P\left(\mu_i - \mu_j \in \hat\mu_i - \hat\mu_j \pm q_{t,f,1-\alpha}\; s/\sqrt{n} \;\forall\; i<j\right) = 1-\alpha$$

simultaneous $(1-\alpha)$-coverage confidence intervals. $\forall =$ “for all.”

Here $P(Q_{t,f} \le q_{t,f,1-\alpha}) = 1-\alpha$ or $q_{t,f,1-\alpha} = \texttt{qtukey}(1-\alpha,\texttt{t},\texttt{f})$.

Tukey-Kramer Method: Unequal Sample Sizes

The simultaneous intervals for all pairwise mean differences was due to Tukey, but it was limited 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 $n^{\star}_{ij}$, where $n^{\star}_{ij}$ is the harmonic mean of n_i and n_j, i.e., $1/n^{\star}_{ij} = (1/n_i + 1/n_j)/2$. Different adjustment for each pair (i, j)!

It was possible to show

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

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))
```

Tukey-Kramer Method for Coagulation Data

```
  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 $\mu_3 - \mu_4$.

Bonferroni Confidence Intervals for Pairwise Contrasts

Applying Bonferroni's methods for simultaneous confidence statement we take $\tilde{\alpha} = \alpha / \binom{t}{2}$ for the individual confidence statements

$$\mu_i - \mu_j \in \hat{\mu}_i - \hat{\mu}_j \pm t_{N-t,1-\tilde{\alpha}/2} \times s \times \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

with $1 - \tilde{\alpha}$ individual coverage probability.

Then

$$\begin{aligned} P(\mu_i - \mu_j \in \hat{\mu}_i - \hat{\mu}_j \pm t_{N-t,1-\tilde{\alpha}/2} \times s \; \forall i < j) &\geq 1 - \binom{t}{2}(1 - (1 - \tilde{\alpha})) \\ &= 1 - \binom{t}{2}\tilde{\alpha} = 1 - \alpha \end{aligned}$$

i.e., the joint coverage probability for all pairwise contrasts is at least $1 - \alpha$.

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 $\mathbf{c}' = (c_1, \ldots, c_t)$:

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

$$= 1-\alpha$$

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.

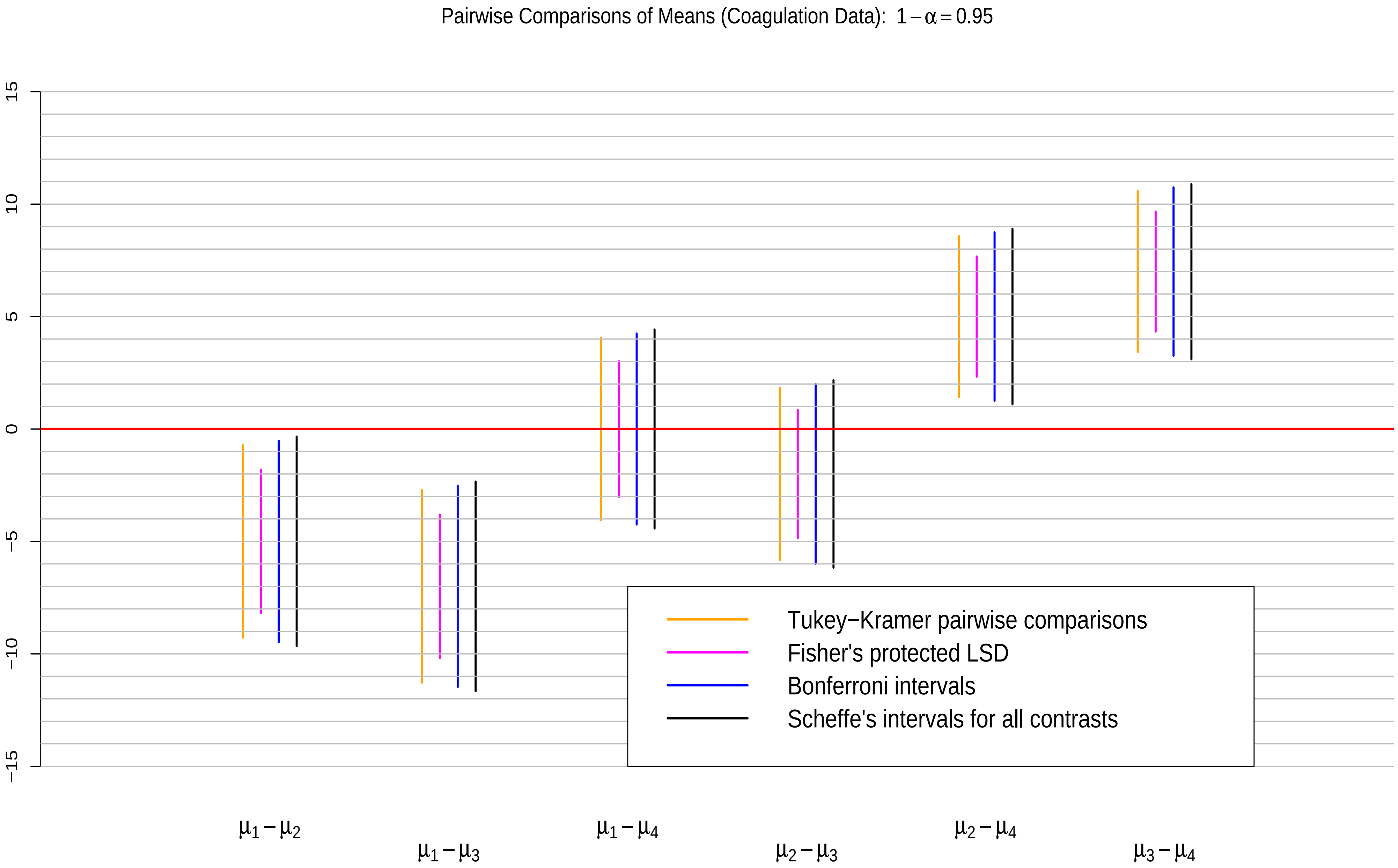
But it compares well with Bonferroni type intervals if applied to many contrasts.

Pairwise Comparison Intervals for Coagulation Data

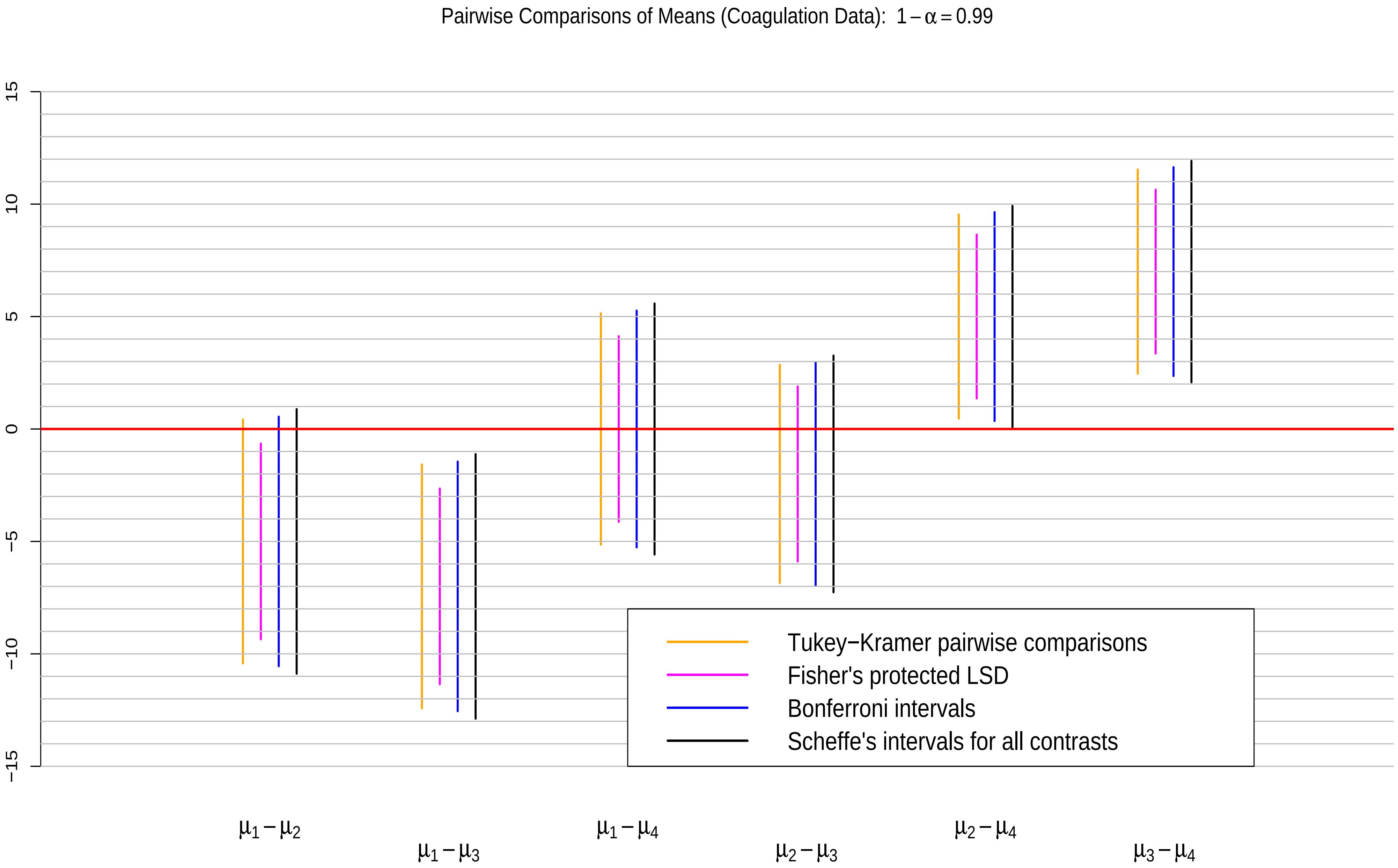
	(simultaneous) 95%-Intervals							
mean difference	Tukey-Kramer		Fisher's protected LSD		Bonferroni inequality		Scheffé's all contrasts method	
$\mu_1 - \mu_2$	-9.28	-0.72	-8.19	-1.81	-9.47	-0.53	-9.66	-0.34
$\mu_1 - \mu_3$	-11.28	-2.72	-10.19	-3.81	-11.47	-2.53	-11.66	-2.34
$\mu_1 - \mu_4$	-4.06	4.06	-3.02	3.02	-4.24	4.24	-4.42	4.42
$\mu_2 - \mu_3$	-5.82	1.82	-4.85	0.85	-6.00	2.00	-6.17	2.17
$\mu_2 - \mu_4$	1.42	8.58	2.33	7.67	1.26	8.74	1.10	8.90
$\mu_3 - \mu_4$	3.42	10.58	4.33	9.67	3.26	10.74	3.10	10.90

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%)



Simultaneous Paired Comparisons (99%)



Orthogonal Contrast

All $\binom{t}{2}$ pairwise comparisons for $\mu_i - \mu_j$ could be 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.

$$C_1 = \sum_{i=1}^{t} c_{1i}\mu_i \quad \perp \quad C_2 = \sum_{i=1}^{t} c_{2i}\mu_i \quad \Longleftrightarrow \quad \sum_{i=1}^{t} c_{1i}c_{2i}/n_i = 0$$

$$C_1 \perp C_2 \quad \Longrightarrow \quad \mathrm{cov}(\hat{C}_1,\hat{C}_2) = \sum_{i=1}^{t}\sum_{j=1}^{t} c_{1i}c_{2j}\mathrm{cov}(\hat{\mu}_i,\hat{\mu}_j) = \sum_{i=1}^{t} c_{1i}c_{2i}\sigma^2/n_i = 0\,,$$

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 independence of the contrast estimates motivates orthogonal contrasts.

An Orthogonal Contrast Example

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:

$\mathbf{c}_1' = (.5, .5, -1)$ and $\mathbf{c}_2' = (1, -1, 0)$.

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.

Unbalanced Case of Previous Example

We have an unbalanced design, i.e., $n_1,\ n_2,\ n_3$ may be different.

Then the following $t-1=2$ vectors:

$\mathbf{c}_1' = (n_1/(n_1+n_2), n_2/(n_1+n_2), -1)$ and $\mathbf{c}_2' = (1,-1,0)$ are indeed contrast vectors: $n_1/(n_1+n_2)+n_2/(n_1+n_2)-1=0$ and $1-1+0=0$

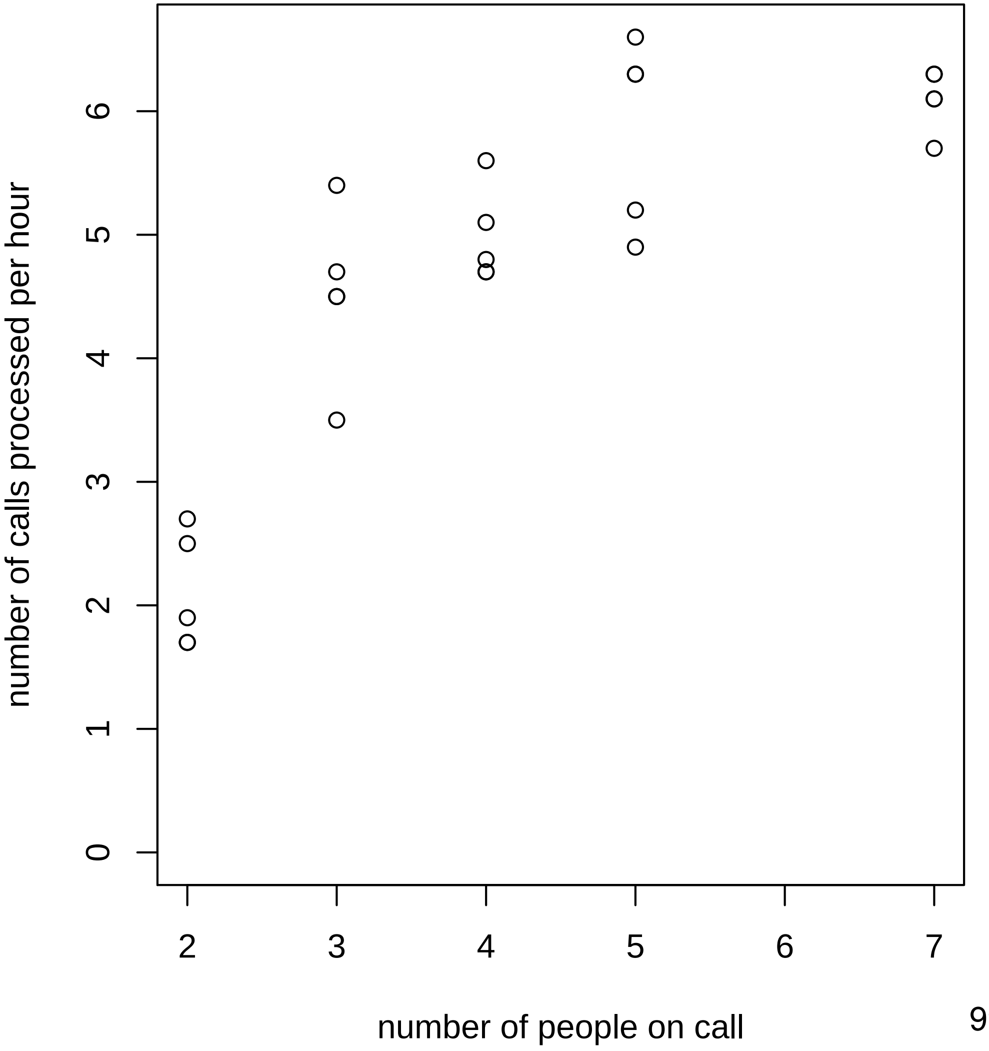
and they are orthogonal: $n_1/[(n_1+n_2)n_1]-n_2/[(n_1+n_2)n_2]-1\cdot 0/n_3=0$.

$\Longrightarrow C_1=(n_1\mu_1+n_2\mu_2)/(n_1+n_2)-\mu_3=\bar{\mu}_{12}-\mu_3$ and $C_2=\mu_1-\mu_2$, 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

# of persons on call	# of calls processed per hour			
2	1.7	2.7	2.5	1.9
3	4.5	3.5	4.7	5.4
4	4.7	4.8	5.6	5.1
5	6.3	5.2	6.6	4.9
7	6.3	5.7	6.1	6.1



Service Center Data

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:

	c_{i1}	c_{i2}	c_{i3}	c_{i4}	c_{i5}
$C_1 = \sum_{j=1}^{5} c_{1j}\,\mu_j$	-2	-1	0	1	2
$C_2 = \sum_{j=1}^{5} c_{2j}\,\mu_j$	2	-1	-2	-1	2
$C_3 = \sum_{j=1}^{5} c_{3j}\,\mu_j$	-1	2	0	-2	1
$C_4 = \sum_{j=1}^{5} c_{4j}\,\mu_j$	1	-4	6	-4	1

For what kind of mean patterns in $\mu_1,\ldots,\mu_5$ would $|C_i|$ and consequently $|\hat{C}_i|$ be large?

Correlations and Contrasts

For a contrast vector $\mathbf{c}$ let $C = \mathbf{c}'\mu = \sum_{j=1}^{t} c_j\mu_j$ be the corresponding contrast.

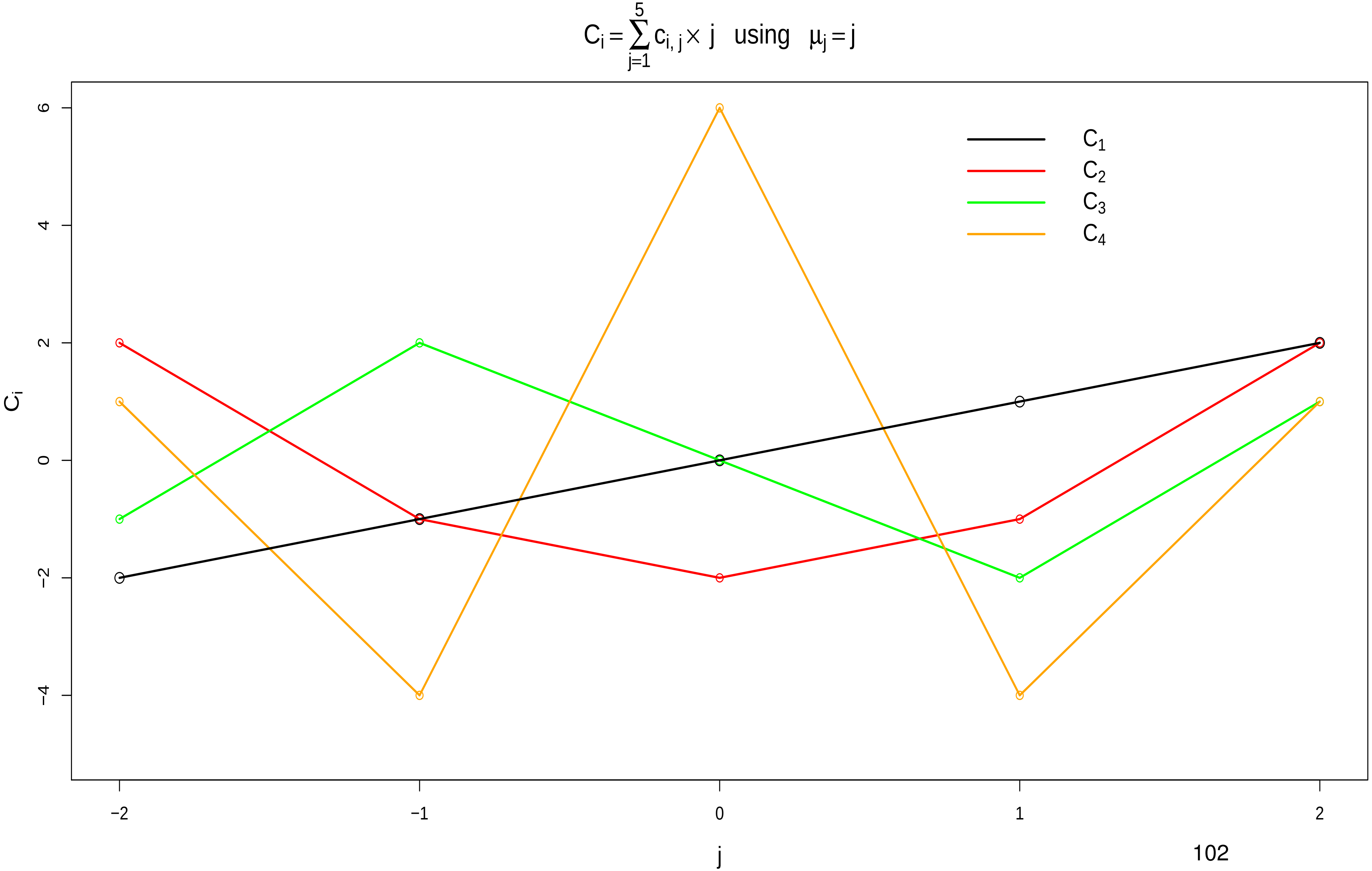
Then
$$\begin{aligned} C = \mathbf{c}'\mu &= \sum_{j=1}^{t} c_j\mu_j = \sum_{j=1}^{t} c_j(\mu_j - \bar{\mu}) = \sum_{j=1}^{t} (c_j - \bar{c})(\mu_j - \bar{\mu}) \\ &= \frac{\sum_{j=1}^{t}(c_j - \bar{c})(\mu_j - \bar{\mu})}{\sqrt{\sum_{j=1}^{t}(c_j - \bar{c})^2 \sum_{j=1}^{t}(\mu_j - \bar{\mu})^2}} \times \sqrt{\sum_{j=1}^{t}(c_j - \bar{c})^2 \sum_{j=1}^{t}(\mu_j - \bar{\mu})^2} \\ &= \rho(\mathbf{c},\mu) \times \sqrt{\sum_{j=1}^{t}(c_j - \bar{c})^2 \sum_{j=1}^{t}(\mu_j - \bar{\mu})^2} \end{aligned}$$

where the third and fourth $=$ come from $\sum_{j=1}^{t} c_j = 0$ and thus $\bar{c} = 0$.

Here $\rho(\mathbf{c},\mu)$ is the ordinary correlation coefficient of the vectors $\mathbf{c}$ and μ.

Aside from scaling $\mathbf{c}$ and μ, the absolute contrast $|C|$ becomes large when the absolute correlation $|\rho(\mathbf{c},\mu)|$ is large, i.e., when $\mathbf{c}$ and μ align reasonably well.

Orthogonal Contrast Plots



$$C_i = \sum_{j=1}^{5} c_{i,j} \times j \quad \text{using} \quad \mu_j = j$$

Interpretation of Orthogonal Contrast Plots

The previous plot suggests that a pattern in the means μ_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 Bonferroni Contrast Intervals

for Service Center Data

	95%	99%
C_1	[6.27, 11.58]	[5.53, 12.32]
C_2	[-7.02, -0.73]	[-7.89, 0.14]
C_3	[-1.26, 4.06]	[-1.99, 4.79]
C_4	[-9.58, 4.48]	[-11.53, 6.43]

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.

Orthogonal Polynomial Contrast Vectors

The previous orthogonal contrasts for linear, quadratic, cubic, quartic behavior were tailored to five treatments.

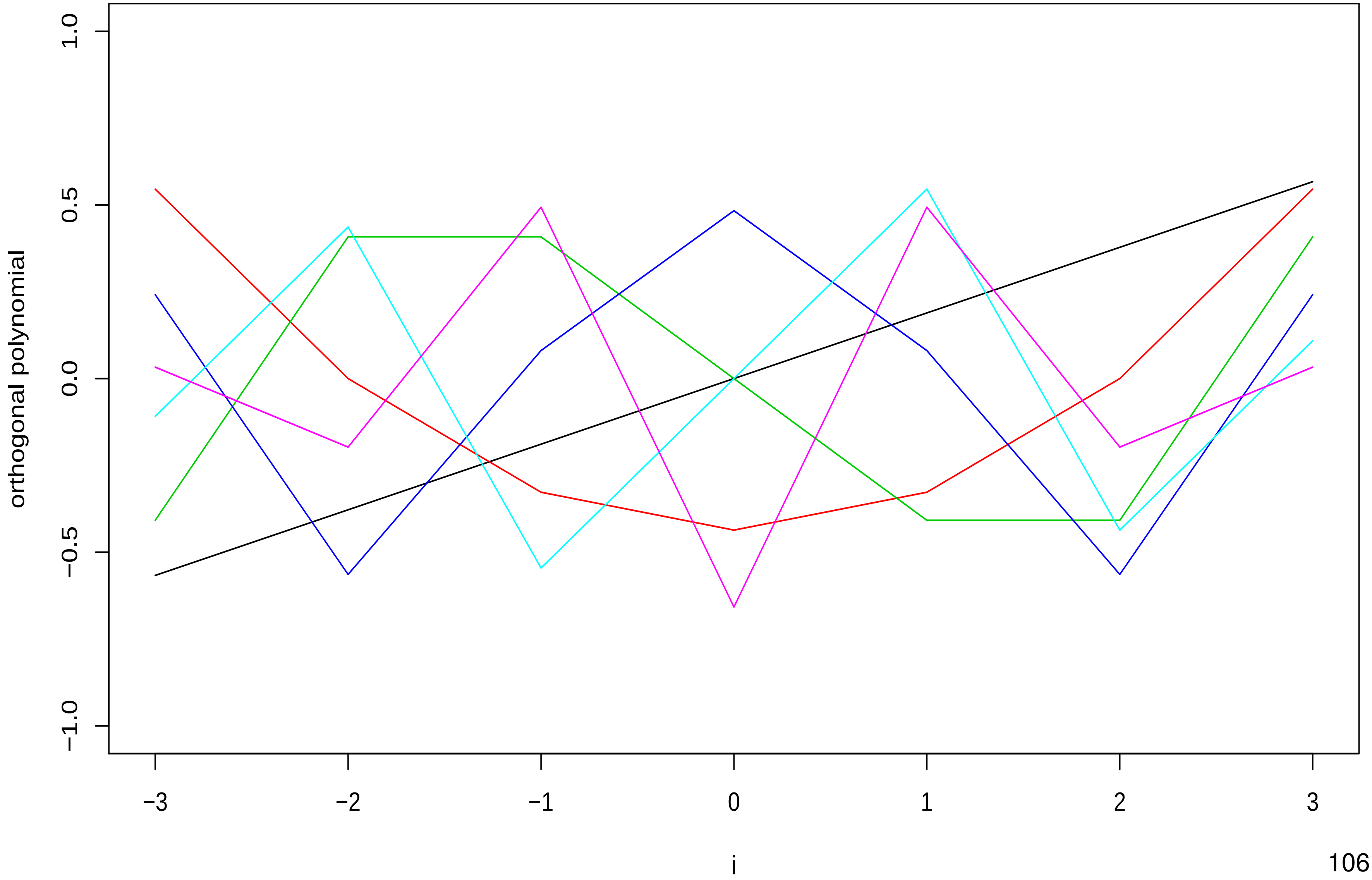
How do we get similar contrast vectors when we have t treatments?

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

```
> round(contr.poly(7),3)
         .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
```

More on this under Regression in Stat 423.

Orthogonal Polynomial Contrasts from `contr.poly(7)`



Model Diagnostics

Model: $Y_{ij} = \mu_i + \varepsilon_{ij}, \quad j = 1, \ldots, n_i\,,\ i = 1, \ldots, t,$ with the following assumptions:

A1: $\{\varepsilon_{ij}\}$ are independent;

A2: $\mathrm{var}(\varepsilon_{ij}) = \mathrm{var}(Y_{ij}) = \sigma^2$ for all i, j

(homogeneity of variances or homoscedasticity);

A3: $\{\varepsilon_{ij}\}$ 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.

Checking Normality

Here we would like to check normality of $\varepsilon_{ij} = Y_{ij} - \mu_i,\ j = 1,\ldots,n_i,\ i = 1,\ldots,t.$

Not knowing μ_i we estimate the error term ε_{ij} via $\hat{\varepsilon}_{ij} = Y_{ij} - \hat{\mu}_i = Y_{ij} - \bar{Y}_{i\bullet}.$

If normality holds then a normal 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. `qqnorm(residual.vector)` $\Longrightarrow$ normal QQ-plot. 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.

Checking Normality by Simulation

We can just adapt the KS, CvM, and AD EDF test of fit criteria and simulate their null distribution, in order to judge any significant non-normality in the residuals.

$$\begin{aligned}
D_{\mathrm{KS}} &= \max\left\{\max_i\left[\frac{i}{N}-U_{(i)}\right],\max_i\left[U_{(i)}-\frac{i-1}{N}\right]\right\}\\
D_{\mathrm{CvM}} &= \sum_{i=1}^{N}\left[U_{(i)}-\frac{2i-1}{2N}\right]^2+\frac{1}{12N}\\
D_{\mathrm{AD}} &= -N-\frac{1}{N}\sum_{i=1}^{N}(2i-1)\left[\log(U_{(i)})+\log(1-U_{(i)})\right]
\end{aligned}$$

where

$$U_{ij}=\Phi\left(\frac{Y_{ij}-\bar{Y}_{i\bullet}}{s}\right)$$

and $U_{(1)}\le\ldots\le U_{(N)}$ are the U_{ij} in increasing order.

The Simulation

The distribution of

$$U_{ij} = \Phi\left(\frac{Y_{ij} - \bar{Y}_{i\bullet}}{s}\right) = \Phi\left(\frac{(Y_{ij} - \mu_i)/\sigma - (\bar{Y}_{i\bullet} - \mu_i)/\sigma}{s/\sigma}\right)$$

does not depend on any unknown parameters.

Thus we may as well simulate the $Y_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, compute $\bar{Y}_{i\bullet}, i = 1, \ldots, t$ and s and then U_{ij}, sort these values and compute the respective EDF criteria.

Repeat this over and over, say $N_{\text{sim}} = 10000$ times, and compare the EDF criteria for the actual data set against these simulated null distributions to obtain estimated p-values. View this as potential homework.

It may be advantageous to modify the above EDF criteria if sample sizes are quite different (uncharted territory).

Hermit Crab Count Data

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.

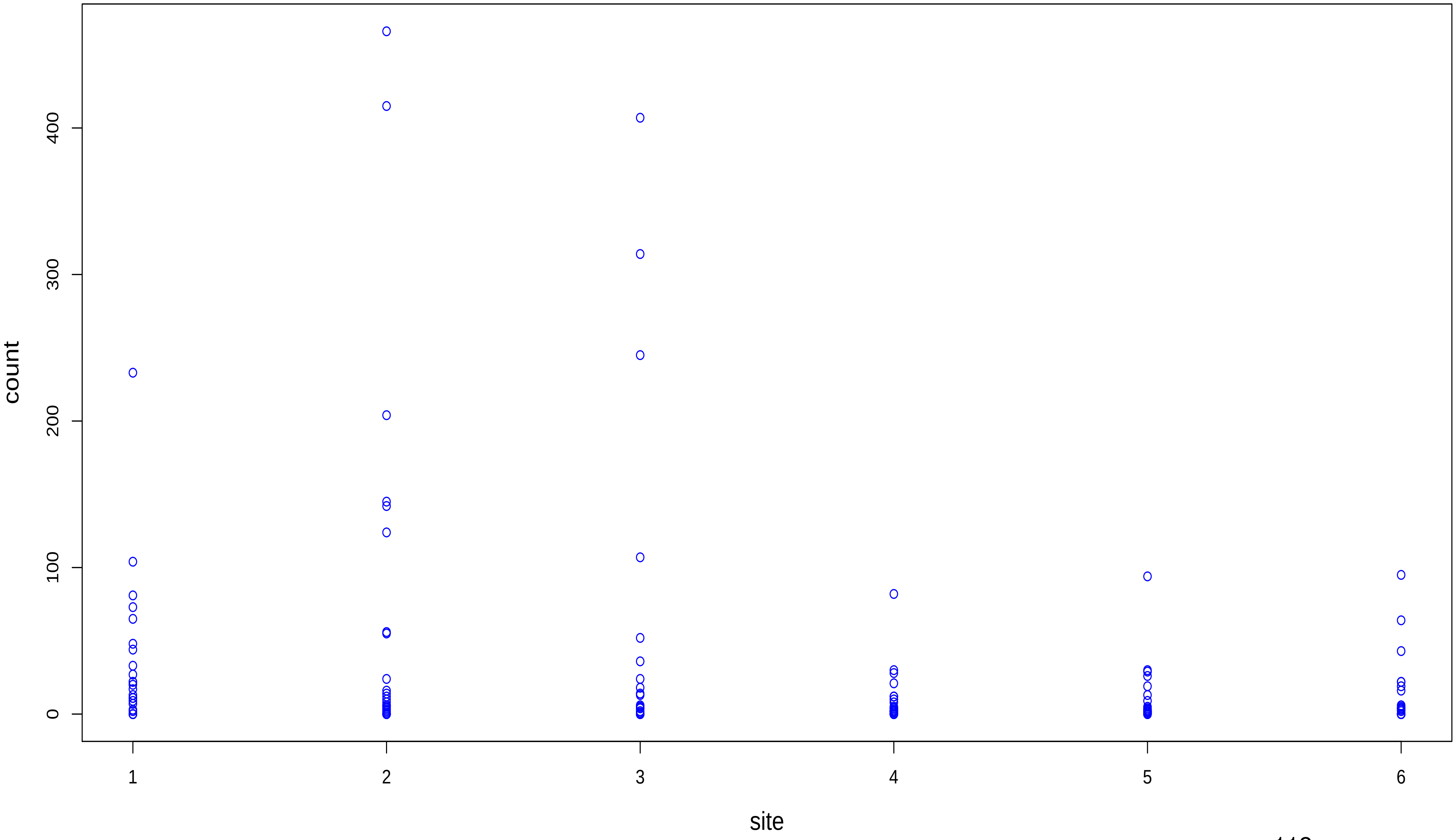
Import it into R via `crab=read.csv("crab.csv")`.

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

```
> names(crab)
[1] "count" "site"
> plot(crab$site,crab$count,xlab="site",ylab="count",
+ col="blue",cex.lab=1.3)
```

produced the plot on the next slide.

Plot of Hermit Crab Counts

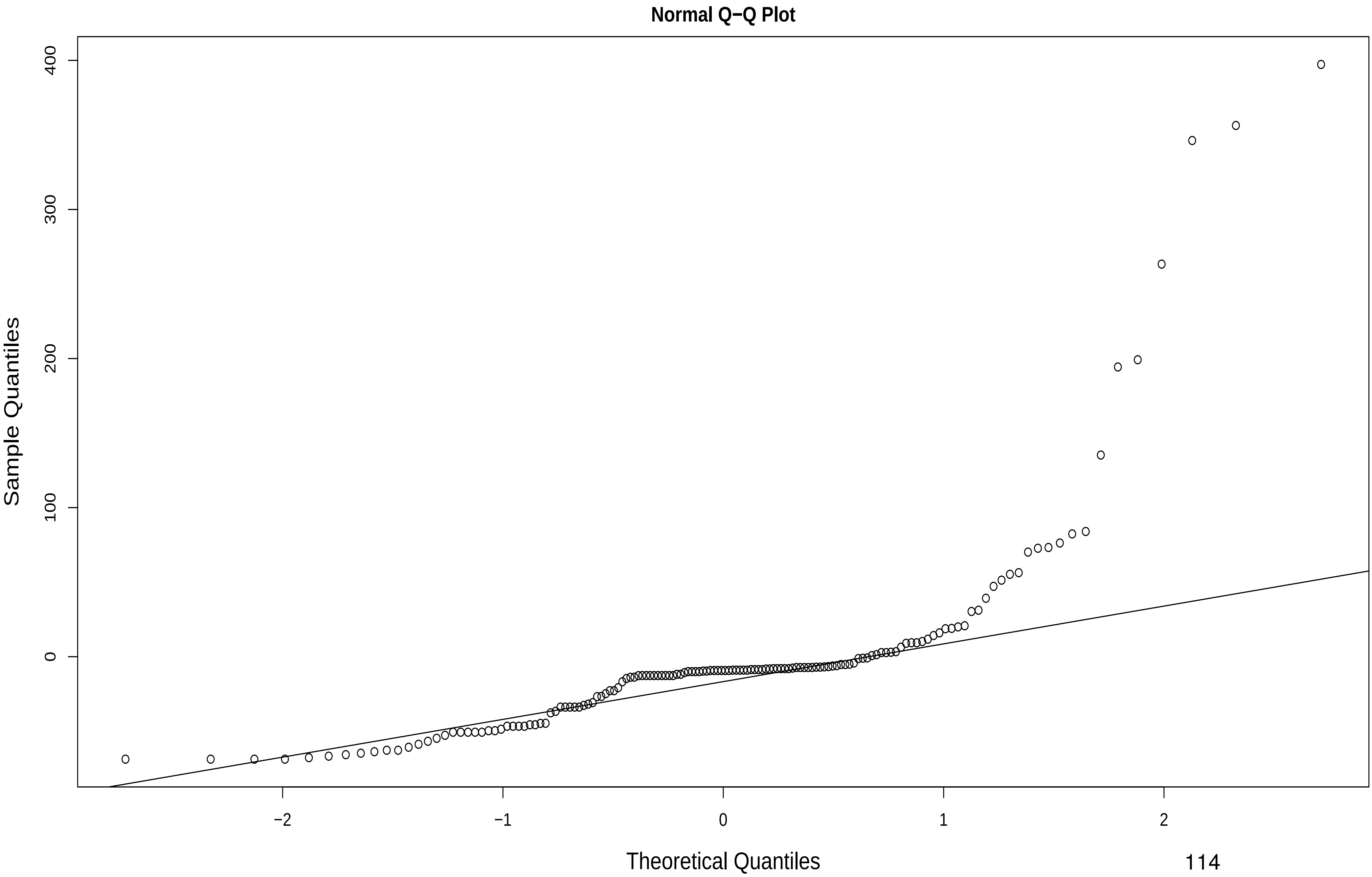


ANOVA for Hermit Crab Count Data

```
> out.lm=lm(crab$count~as.factor(crab$site))
> anova(out.lm)
Analysis of Variance Table

Response: crab$count
                      Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(crab$site)   5  76695   15339  2.9669 0.01401 *
Residuals            144 744493    5170
---
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



Checking for Homoscedasticity

The appropriate indicators for checking a constant variance over all t treatment groups would seem to be $s_1^2, \ldots, s_t^2$.

There are various rules of thumb involving $F_{\min} = \min(s_1^2, \ldots, s_t^2) / \max(s_1^2, \ldots, s_t^2)$.

For example, if $F_{\min} > 1/3$ the constant variance assumption should be OK while for $F_{\min} < 1/7$ we should deal with it.

Where the 1/3 or 1/7 come from and what to do in between is not clear.

With R it is simple enough to simulate the distribution for $F_{\min}$.

`Fmin.test`

The R function `Fmin.test` can be found on the class web site. It simulates the $F_{\rm min}$ distribution, assuming normal samples with equal variances. The sample sizes may vary. The documentation for `Fmin.test` is inside the function body.

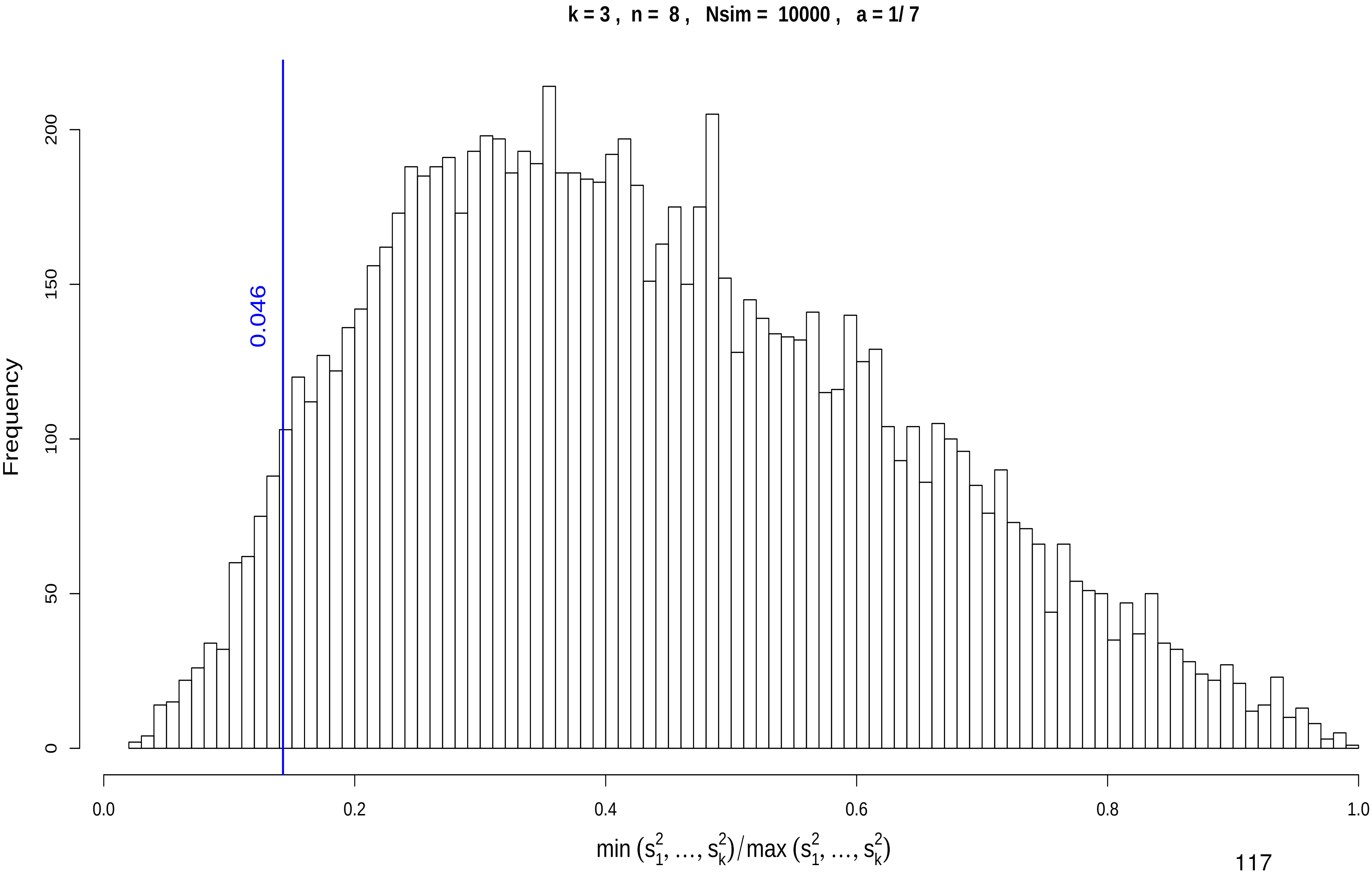
It can be used to explore any desired rule of thumb, by calculating the proportion of $F_{\rm min}$ values $\le$ to the rule of thumb value.

If $F_{\rm min,observed}$ is provided, it calculates the estimated p-value from this simulated distribution.

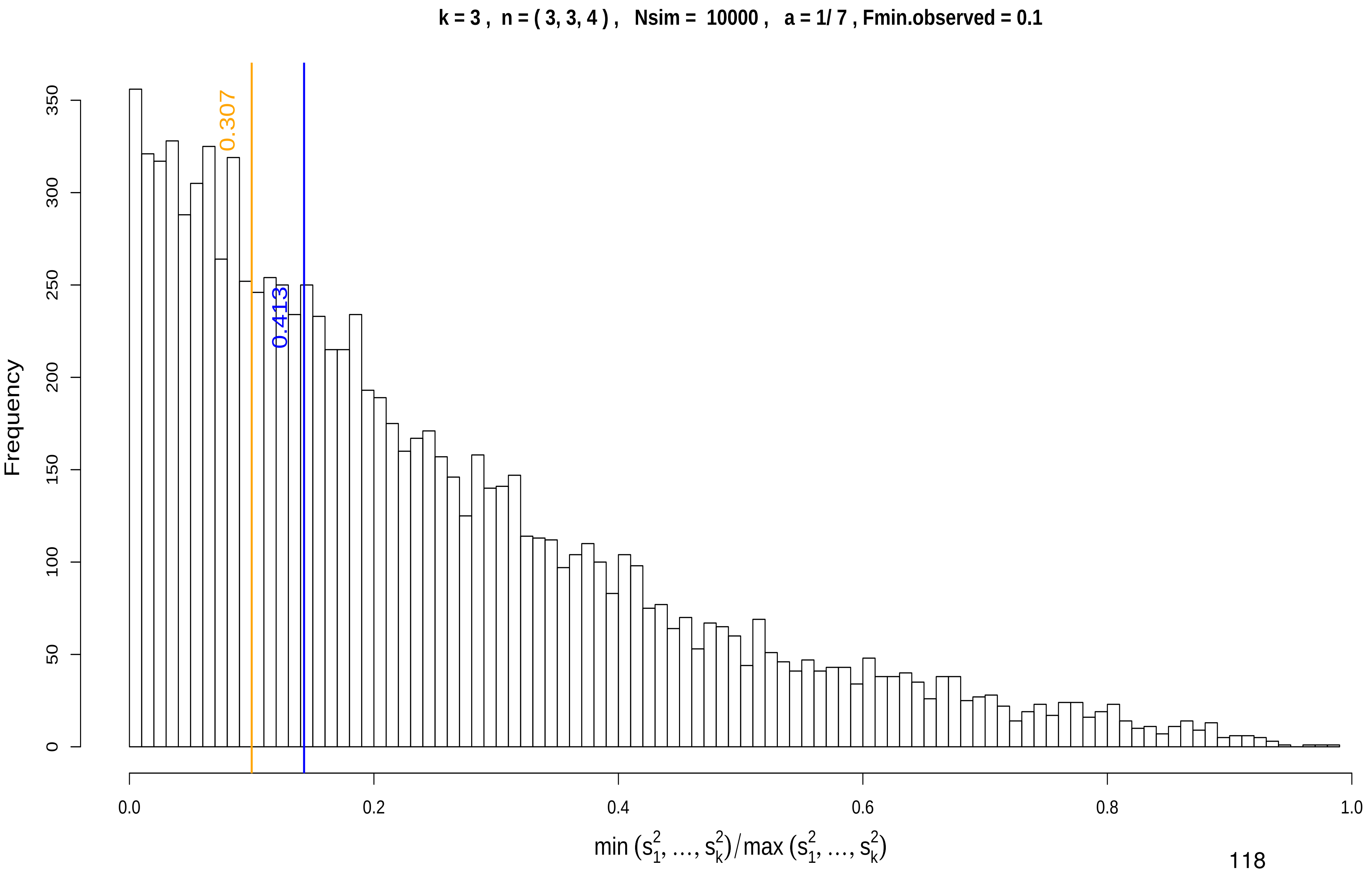
See the next two slides for examples.

Note however, that the validity of this test depends strongly on data normality.

Fmin.test(k=3,n=8,a.recip=7)



```
Fmin.test(k=3,n=c(3,3,4),a.recip=7,Fmin.observed=.1)
```



Diagnostic Plots for Checking Homoscedasticity

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

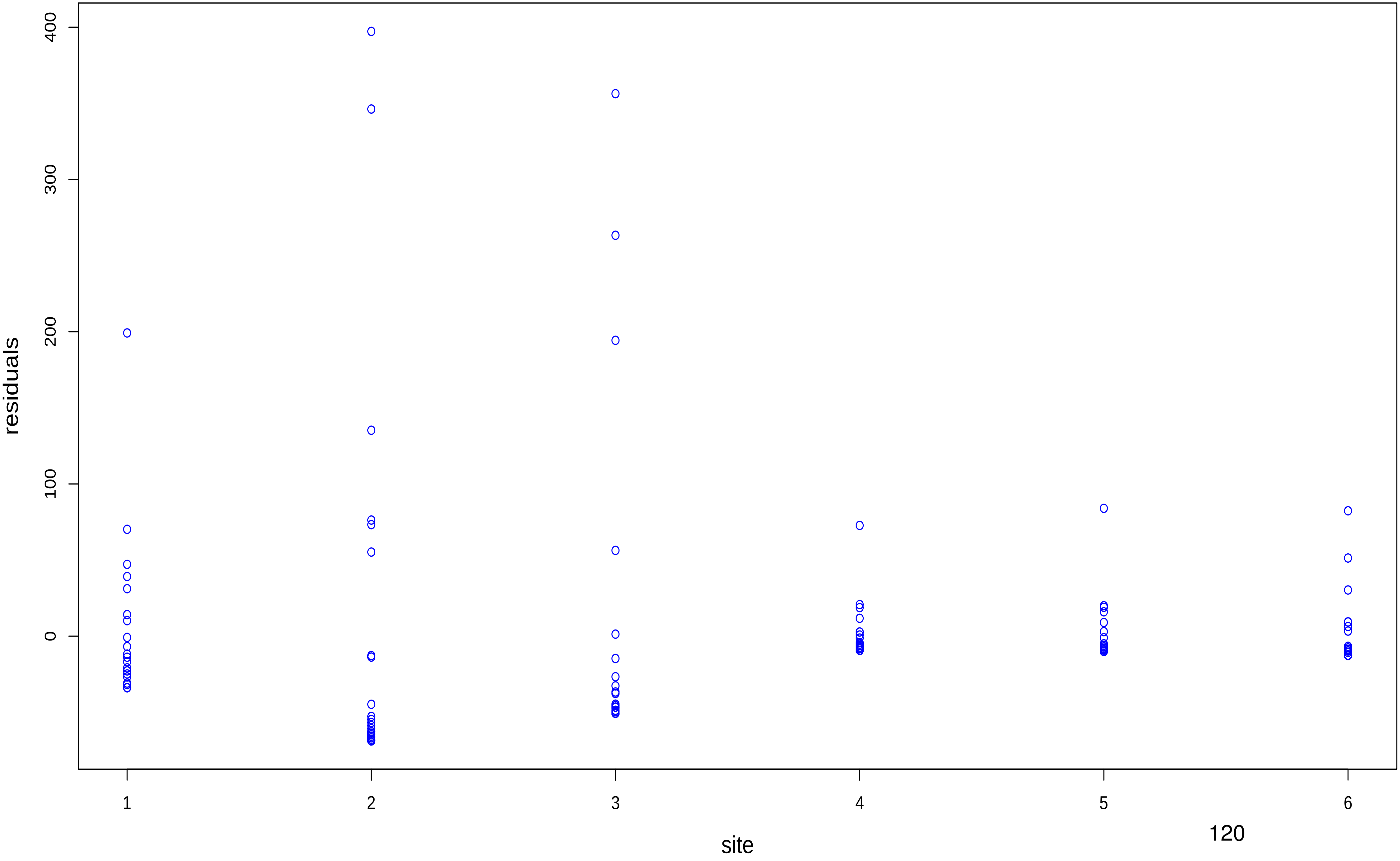
Compare the difference in information displayed in the next two plots.

The second display suggests that variability increases with fitted value.

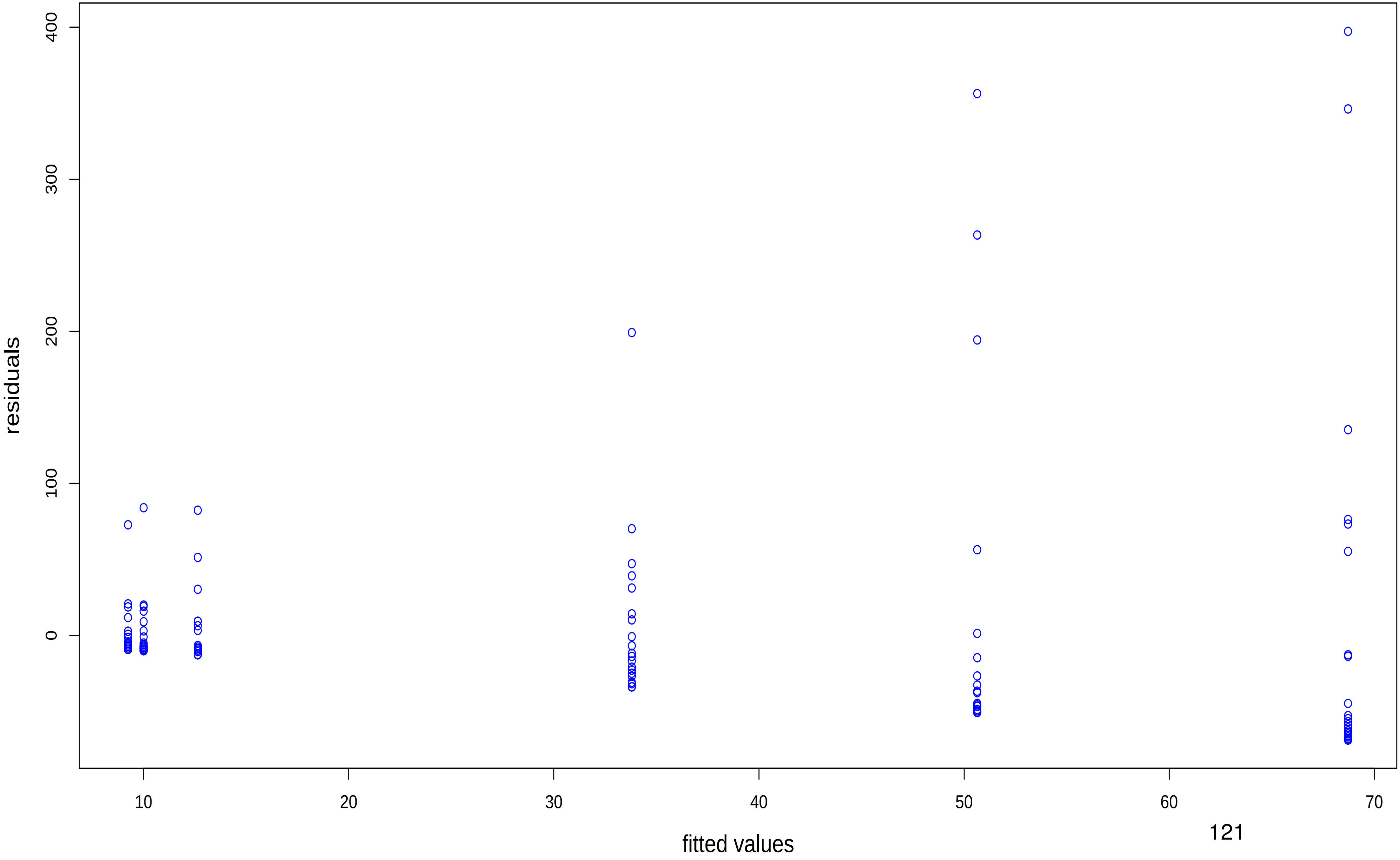
Often there is a relationship between variability and the mean.

There are ways to deal with this by using variance stabilizing transforms of the Y_{ij}.

```
plot(crab$site,out.lm$residuals,
          col="blue",xlab="site",ylab="residuals",cex.lab=1.3)
```



```
plot(out.lm$fitted.values,out.lm$residuals,col="blue",
          xlab="fitted values",ylab="residuals",cex.lab=1.3)
```



Levene’s Test for Homoscedasticity

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 $\bar{Y}_{i\bullet}$ 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.

The ANOVA F-test for means is not as sensitive to the normality assumption as the F-test or `Fmin.test` for comparing variances.

Levene’s Test for Crab Count Data

```
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  2.9278 0.01508 *
Residuals            144 699845    4860
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

A Multiplicative Error Model

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 ε has mean μ_ε and standard deviation σ_ε.

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

Variance Stabilization and Normality under $\log$-Transform

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

$\Longrightarrow \quad E(\log(Y)) = \log(\mu) + E(\log(\varepsilon))$ and $\mathrm{var}(\log(Y)) = \mathrm{var}(\log(\varepsilon))$

breaks the link, i.e., μ 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 ε 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:

$$V_{ij} = \log(Y_{ij}) = \tilde{\mu}_i + \tilde{\varepsilon}_{ij} \quad \text{with} \quad \tilde{\varepsilon}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

The Problem of Zero Counts

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.

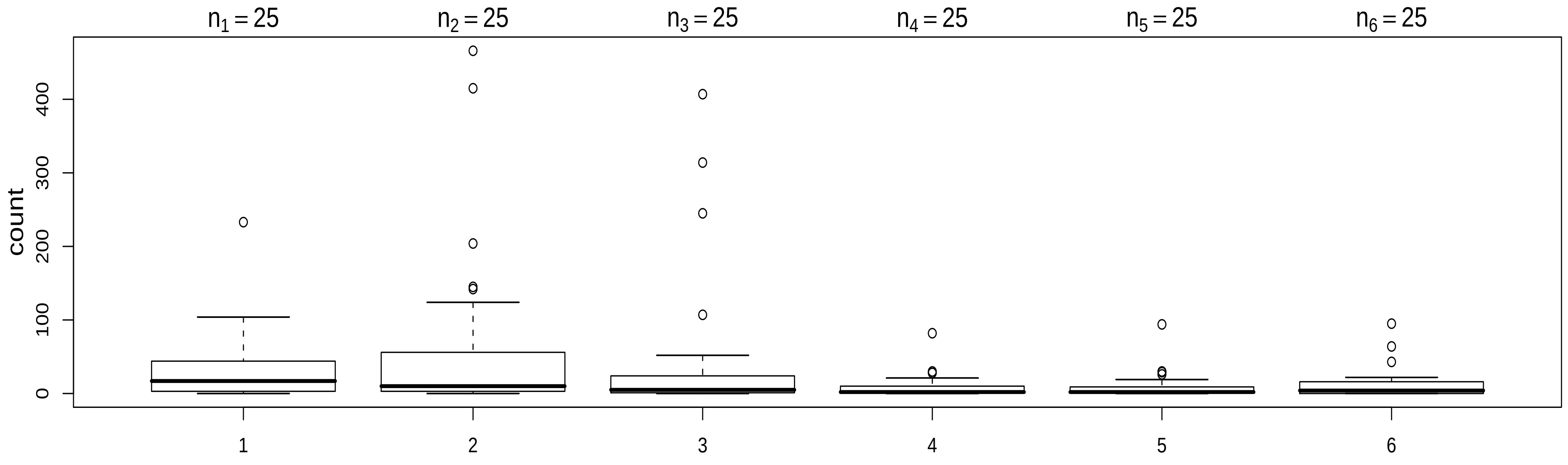
1. Adding a small fraction, say 1/6, to all counts. $(1/6 > 0$ is somewhat arbitrary)

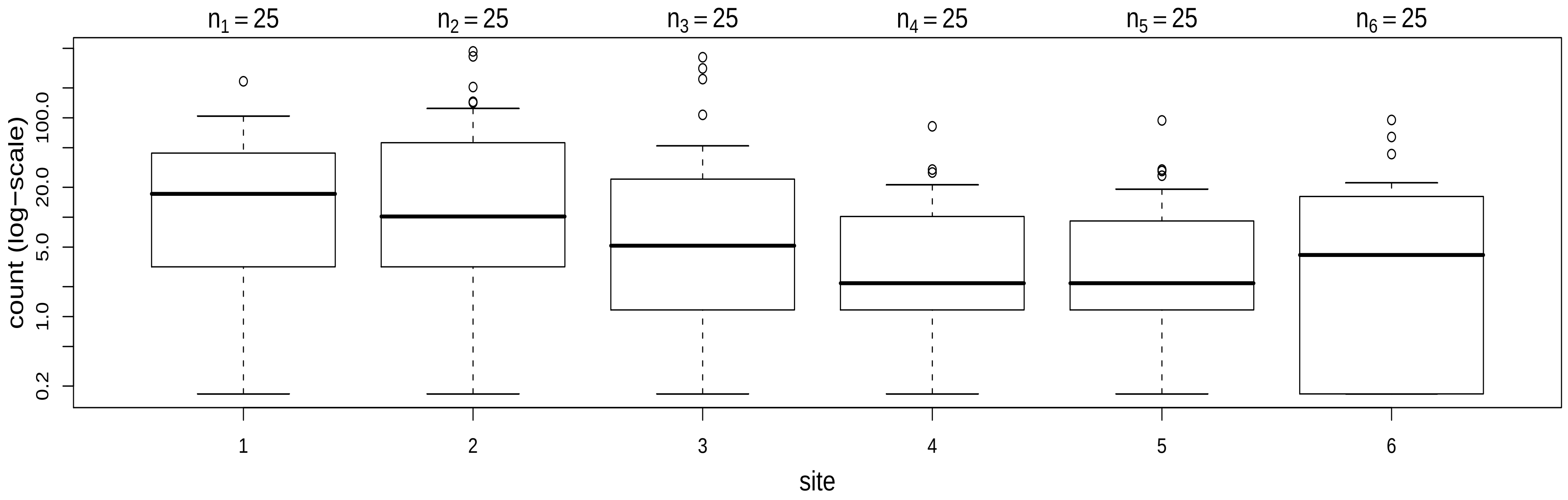
   This is a technical solution, keeping all the data.

2. Eliminate all zero counts.

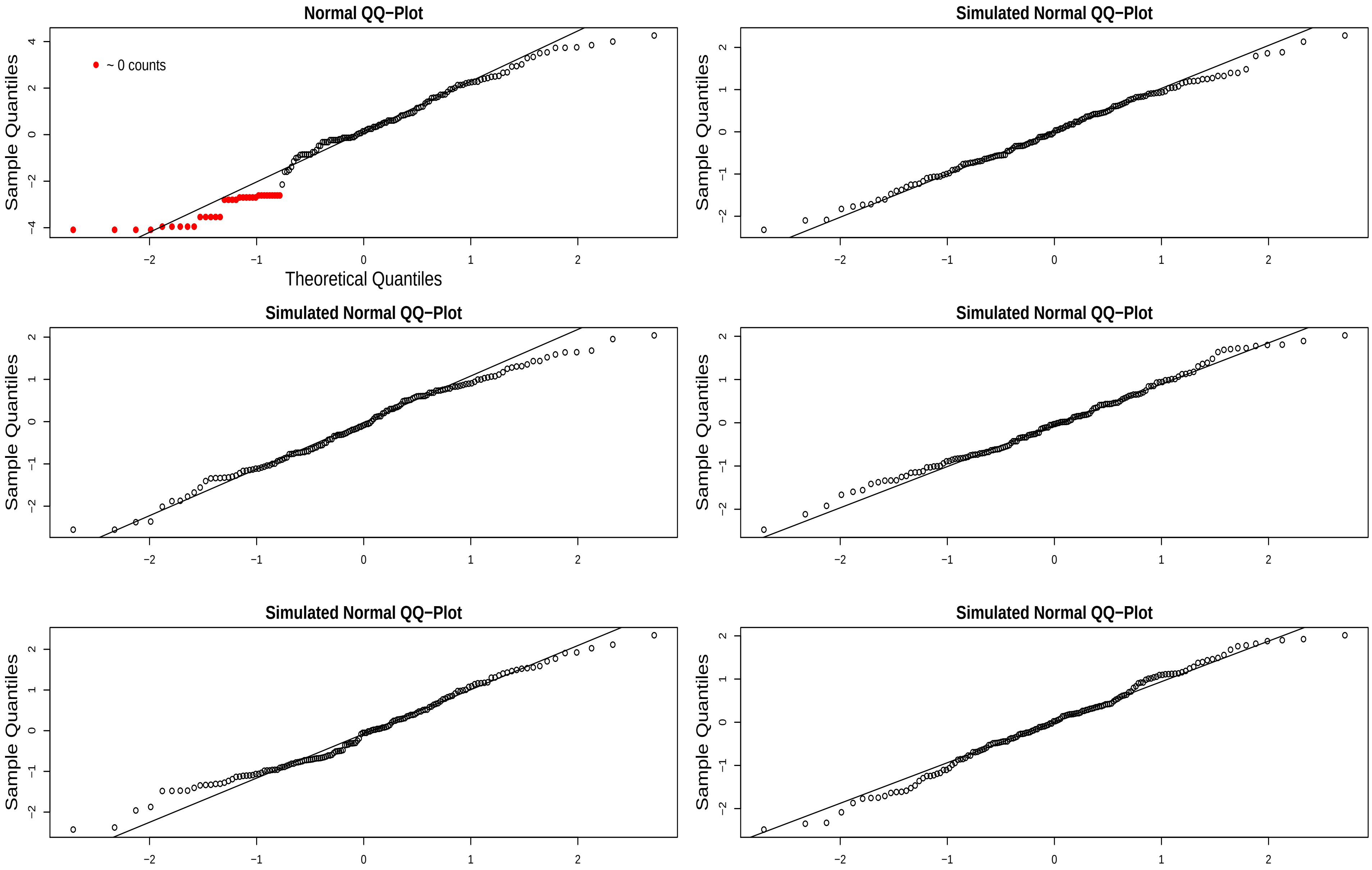
   This may be justified if a zero count just means that there were no crabs in that transect to begin with. It is not a matter of not seeing them because the population size is small. This reduces the count data to $150 - 33 = 117$ counts.

Box Plots for `count` and `log(count+1/6)`





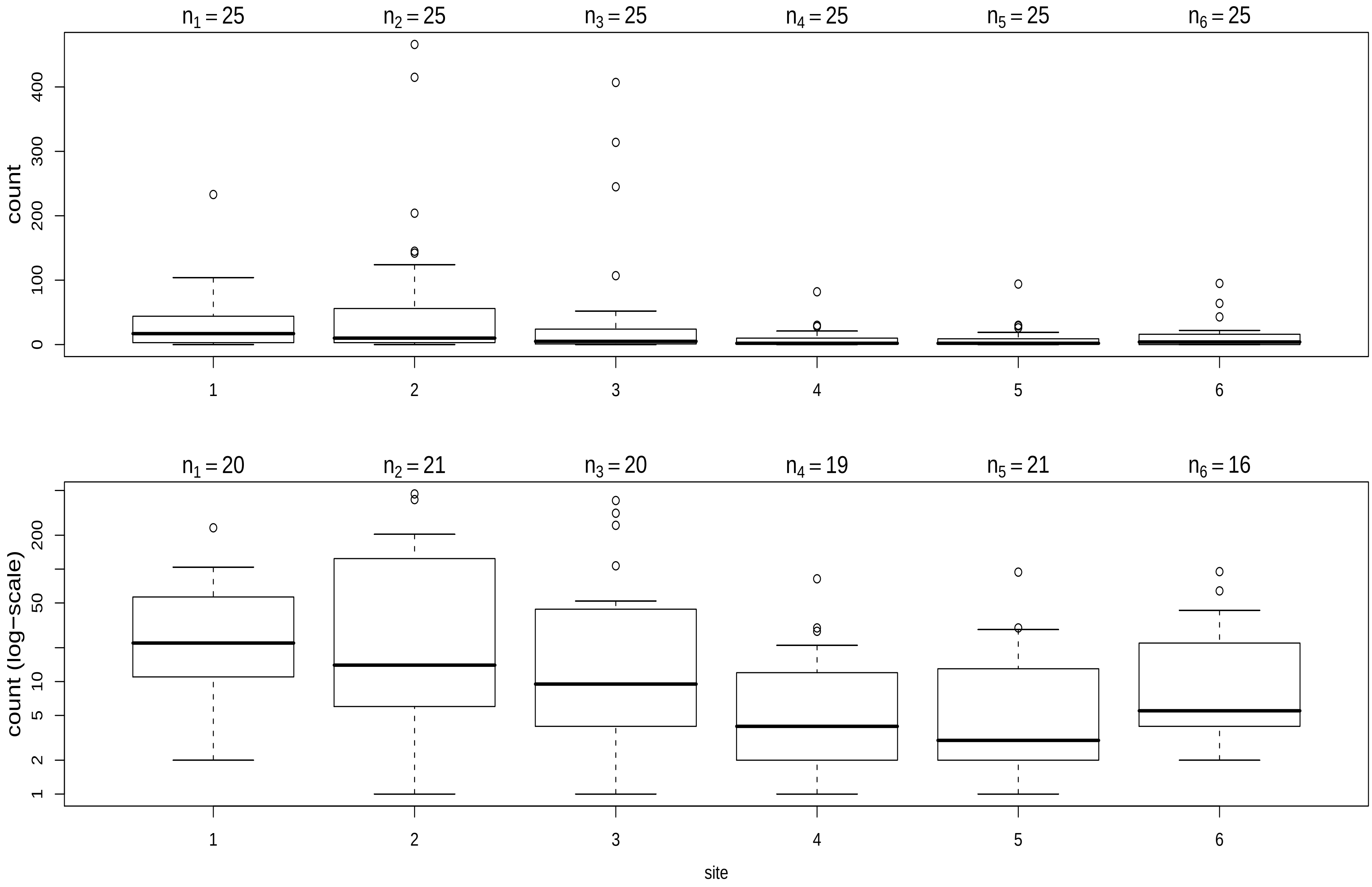
Normal QQ-Plots of 150 Residuals



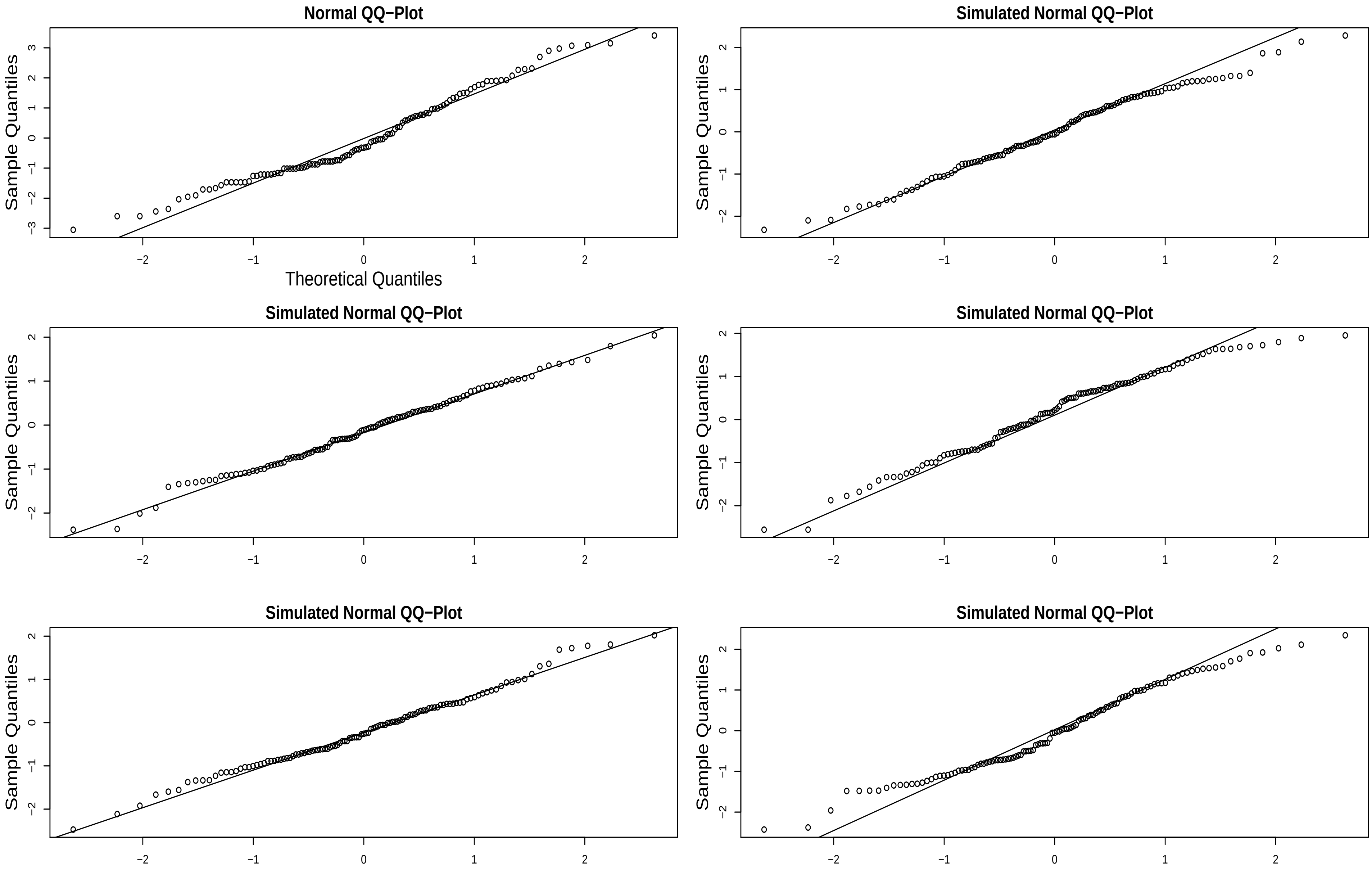
ANOVA for log(count+1/6)

```
Analysis of Variance Table

Response: log(count + 1/6)
                  Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(site)    5  54.73   10.95  2.3226 0.04604 *
Residuals        144 678.60    4.71
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Box Plots for `count` and `log(count[count>0])`



Normal QQ-Plots of 117 Residuals



ANOVA for `log(count[count>0])`

```
Analysis of Variance Table

Response: log(count[count > 0])
                            Df  Sum Sq Mean Sq F value   Pr(>F)
as.factor(site[count > 0])   5  47.905   9.581  4.3866 0.001107 **
Residuals                  111 242.440   2.184
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Levene Test for `log(count+1/6)` and `log(count[count>0])`

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

Response: d
                 Df  Sum Sq Mean Sq F value Pr(>F)
as.factor(site)   5   7.193   1.439  0.7513 0.5864
Residuals       144 275.748   1.915

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

Response: d
                 Df Sum Sq Mean Sq F value Pr(>F)
as.factor(site)   5  6.168   1.234  1.4711  0.205
Residuals       111 93.077   0.839
```

Comments: `log(count+1/6)` vs. `log(count[count>0])` Analysis

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 (see red dots).

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 σ across sites for either case.

Other Variance Stabilizing Transforms

For data with a multiplicative error model for Y_{ij} we showed $\sigma_i \propto \mu_i$ or $\sigma_\mu \propto \mu$ and we saw 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 μ?

A 1-term Taylor series expansion of f around $\mu = E(Y)$

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

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

$f'(\mu) = \dfrac{\tilde{c}}{\mu^\alpha}$ or $f(\mu) = \tilde{c}\,\dfrac{\mu^{1-\alpha}}{1-\alpha} + c^\star$ with $\alpha = 1 \Rightarrow f(\mu) = \log(\mu)$ as special case.

Finding the Variance Stabilizing Transform

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 α? Let the data speak for themselves.

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

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

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

$$\hat{\alpha} = \mathtt{lm}(\log(\mathtt{s_i}) \sim \log(\bar{\mathtt{Y}}_{\mathtt{i}\bullet}))\$\mathtt{coef}[2]$$

Then perform the ANOVA for $\tilde{Y}_{ij} = Y_{ij}^{1-\hat{\alpha}} = Y_{ij}^{\hat{\lambda}}$.

Variance Stabilizing Transforms

Relation $\sigma_Y \sim \mu_Y$	α	$\lambda = 1 - \alpha$	transform	$\tilde{Y}_{ij}$
$\sigma_Y \propto \text{const.}$	0	1	no transform!	Y_{ij}
$\sigma_Y \propto \mu_Y^{1/2}$	1/2	1/2	square root	$Y_{ij}^{1/2} = \sqrt{Y_{ij}}$
$\sigma_Y \propto \mu_Y$	1	0	log	$\log(Y_{ij})$
$\sigma_Y \propto \mu_Y^{3/2}$	3/2	-1/2	reciproc. of sqrt	$Y_{ij}^{-1/2} = 1/\sqrt{Y_{ij}}$
$\sigma_Y \propto \mu_Y^2$	2	-1	reciprocal	$1/Y_{ij}$

Box-Cox Transformations

All the above transformations can be captured in the following unified format known as the Box-Cox transformations

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

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.

Shifts b don't affect the SS's and scale factors a don't affect F-ratios of SS's.

Comments on Box-Cox Transformations

Don’t transform if $\min(s_1^2,\ldots,s_t^2)/\max(s_1^2,\ldots,s_t^2)$ is not sufficiently small

$\Longrightarrow$ `Fmin.test`.

Make sure the linear relationship between $\log(s_i)$ and $\log(\bar{Y}_{i\bullet})$ is strong.

Use simple exponents λ in the transformations, i.e., use $\lambda = 1/2$

rather than $\lambda = 1-\alpha = .473$, as possibly calculated from slope of the linear fit

of $\log(s_i) \approx \alpha \times \log(\bar{Y}_{i\bullet}) + b$.

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

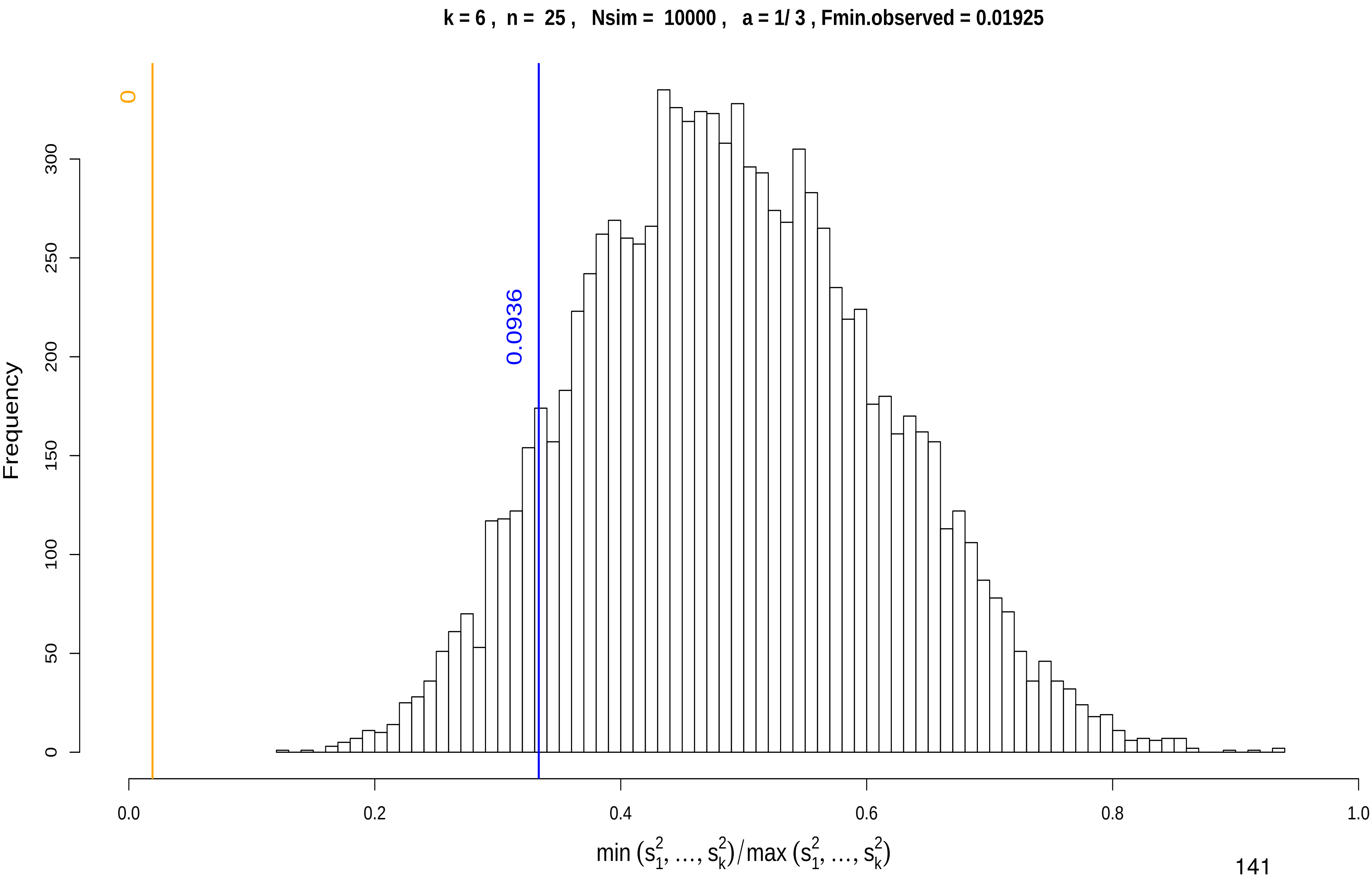
and show the transformed and untransformed data in graphical form.

$\log(s_i)$ vs $\log(\hat{\mu}_i)$ Analysis for Crab Data

site	s_i	$\hat{\mu}_i$	$\log(s_i)$	$\log(\hat{\mu}_i)$
4	17.39	9.24	2.86	2.22
5	19.84	10.00	2.99	2.30
6	23.01	12.64	3.14	2.54
1	50.39	33.80	3.92	3.52
3	107.44	50.64	4.68	3.92
2	125.35	68.72	4.83	4.23

$$F_{\text{min}} = \left(\frac{17.39}{125.35}\right)^2 = .01925$$

```
Fmin.test(k=6,n=25,a.recip=3,Fmin.observed=.01925)
```



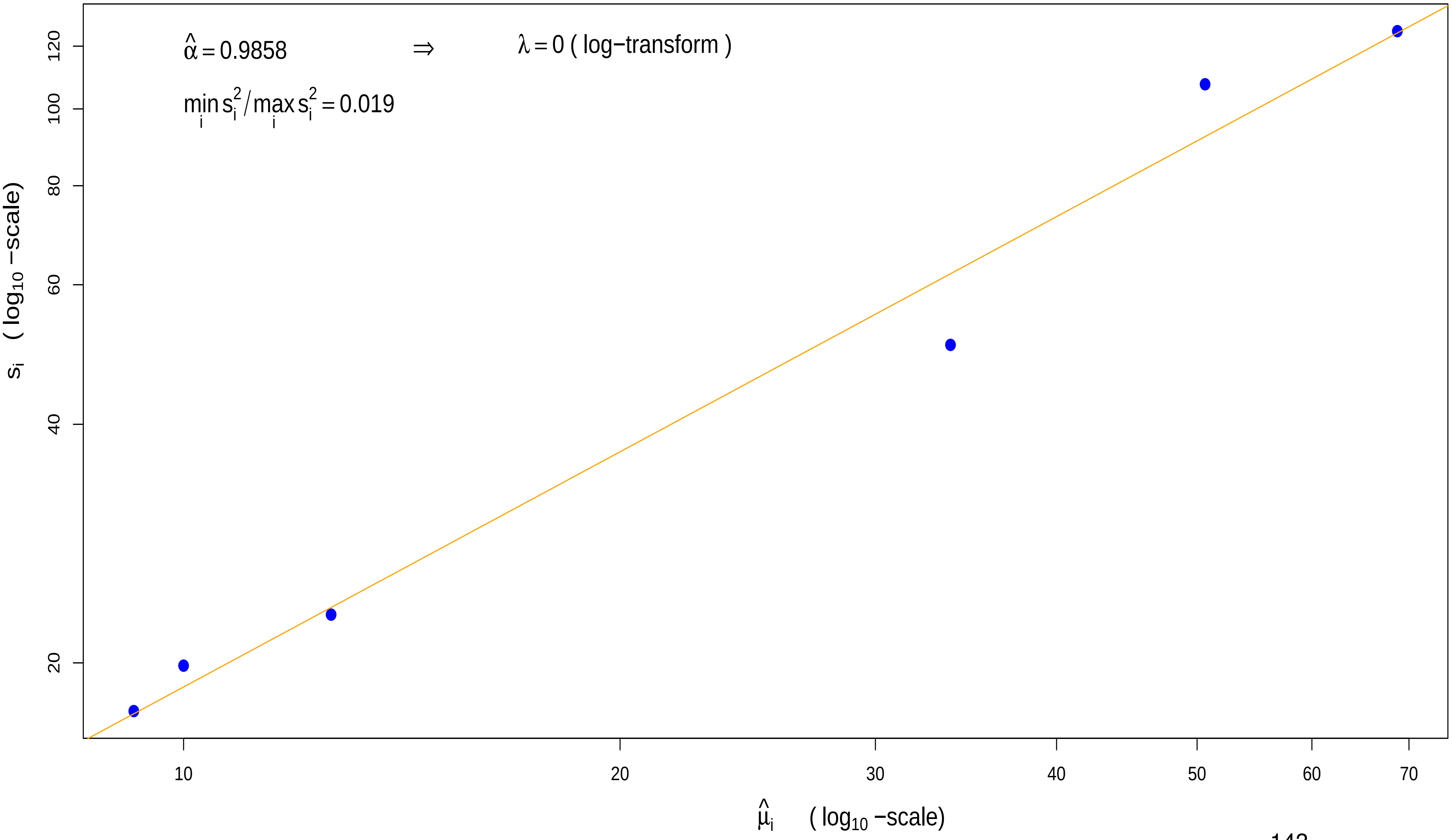
Some Comments

The p-value of 0 obtained by `Fmin.test` appears to be much stronger evidence against the hypothesis of homoscedasticity than the $.01508$ obtained by the Levene test.

However, recall the caution given for `Fmin.test`, that it is sensitive to the normality assumption.

The Levene test is more robust in that respect, thus the p-value of $.01508$ should be more relevant.

$\log(s_i)$ vs $\log(\hat{\mu}_i)$ Plot for Crab Data



Nonparametric k-Sample Tests

Let $Y_{11},\ldots,Y_{1n_1} \overset{\text{i.i.d.}}{\sim} F_1,\quad Y_{21},\ldots,Y_{2n_2} \overset{\text{i.i.d.}}{\sim} F_2\,,\quad \ldots,\quad Y_{k1},\ldots\,,Y_{kn_k} \overset{\text{i.i.d.}}{\sim} F_k$

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 should matter, i.e., one should only pay attention to their ranks $\Longrightarrow$ rank tests.

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.

Kruskal-Wallis k-Sample Test

Let $\bar{R}_{i\bullet} = \sum_{j=1}^{n_i} R_{ij}/n_i$ denote the average rank for the i^{th} sample

Note that the average $\bar{R}_{\bullet\bullet}$ 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., $\bar{R}_{\bullet\bullet} = (N+1)/2$.

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 $(N+1)/2$.

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

suggests itself as a reasonable test statistic.

ANOVA Analogy of the Kruskal-Wallis k-Sample Test

The notation SS_{Treat} and SS_{T} on the previous slide indicates the analogy to our previous use of these terms. All that is changed is that Y_{ij} is interchanged with R_{ij}.

The sum of squares decomposition $SS_{\text{T}} = SS_{\text{Treat}} + SS_{\text{E}}$ still holds.

$$\frac{H}{N-1} = \frac{SS_{\text{Treat}}}{SS_{\text{T}}} = \frac{SS_{\text{Treat}}}{SS_{\text{E}} + SS_{\text{Treat}}} = \frac{SS_{\text{Treat}}/SS_{\text{E}}}{1 + SS_{\text{Treat}}/SS_{\text{E}}} \quad \nearrow \quad \text{in} \quad SS_{\text{Treat}}/SS_{\text{E}}$$

$\Longrightarrow$ H is in 1-1 correspondence with the F-test applied to R_{ij} in place of the Y_{ij}.

Recall $$F = \frac{SS_{\text{Treat}}/(k-1)}{SS_{\text{E}}/(N-k)} \qquad (k \equiv t)$$

Null Distribution of H

$$\sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6} \quad \Longrightarrow$$

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{n_i}(R_{ij}-\bar{R}_{\bullet\bullet})^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}R_{ij}^2 - N\left(\frac{N+1}{2}\right)^2 \\
&= \frac{N(N+1)(2N+1)}{6} - N\left(\frac{N+1}{2}\right)^2 \\
&= \frac{N(N+1)(N-1)}{12} \qquad \Longrightarrow \qquad \frac{SS_{\mathrm{T}}}{N-1} = \frac{N(N+1)}{12}
\end{aligned}$$

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

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

has an approximate χ^2_{k-1} distribution as $N \longrightarrow \infty$.

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

Kruskal-Wallis Test for Flux3

```
> 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
```

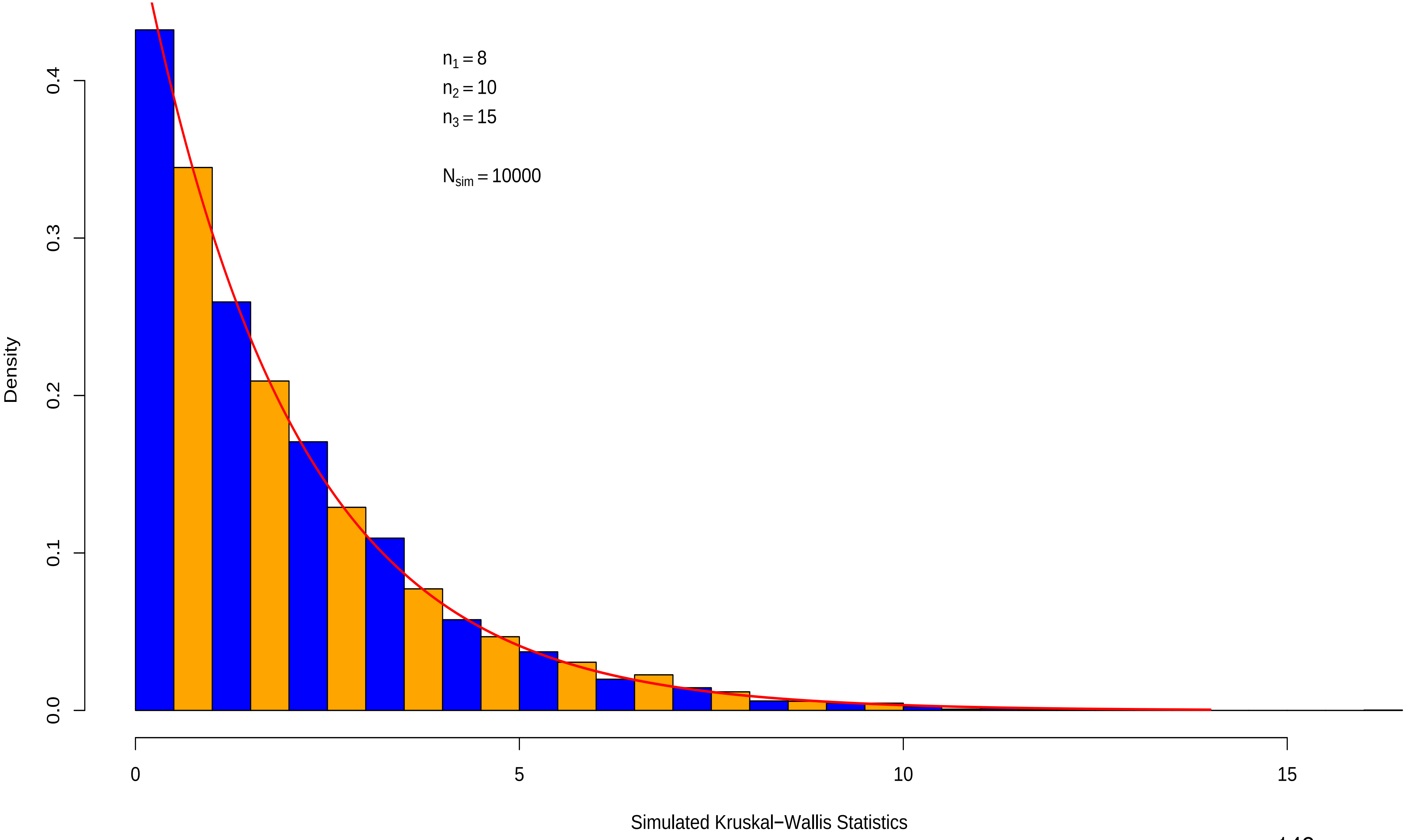
The p-value is not as small as in the normal ANOVA or randomization tests, i.e., $.05126$ from the F-distribution or $.04296$ from simulated randomization distribution.

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}.

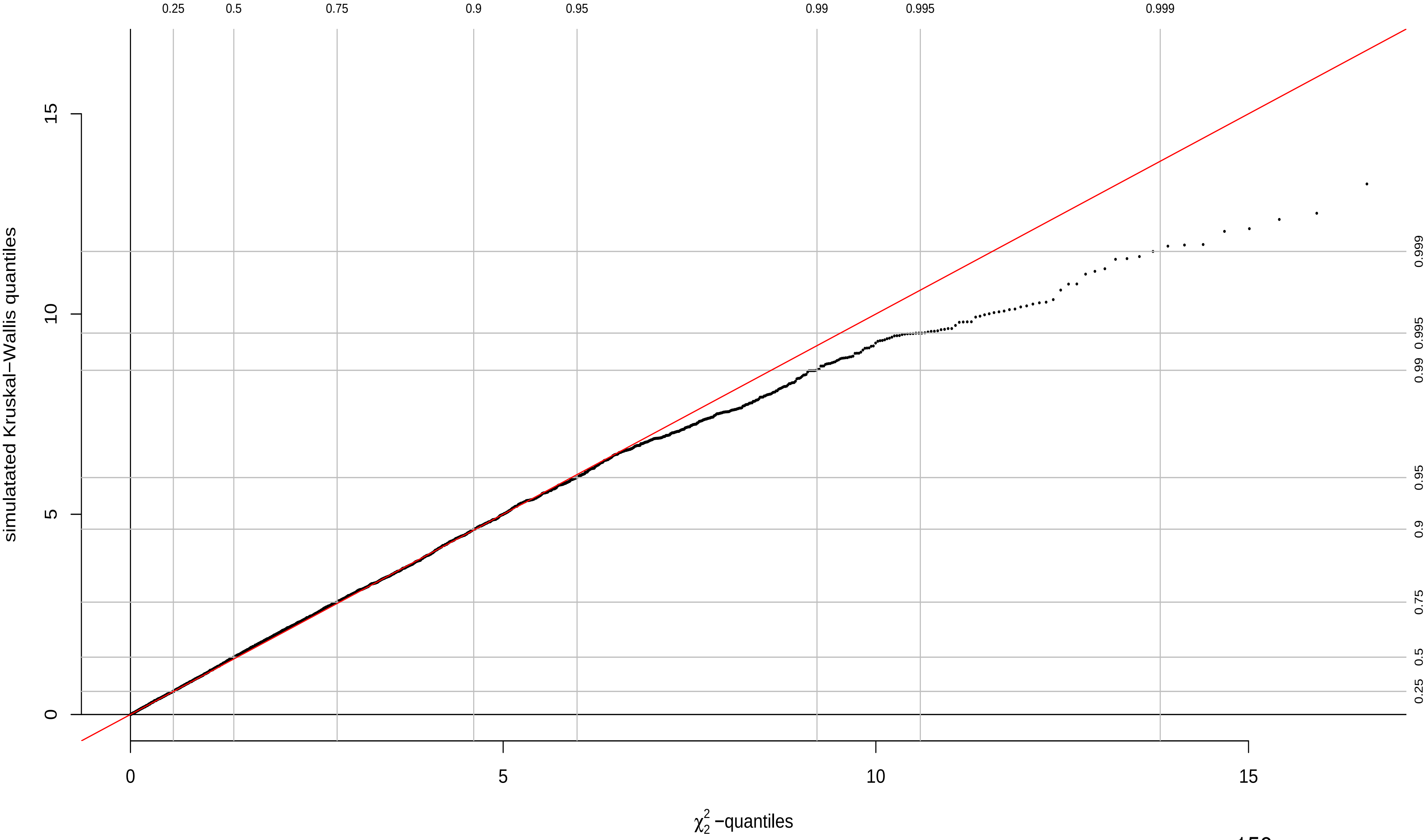
The K-W test is ineffective for changes in scale while locations are matched.

Look at the documentation of `kruskal.test` on how to use it.

How Good is the χ^2_{k-1} Approximation?



How Good is the χ^2_{k-1} Approximation?



How Good is the χ^2_{k-1} Approximation?

Histogram shows a good agreement with the approximating $\chi^2_{k-1} = \chi^2_2$ distribution.

The QQ-plot shows that the distributions agree fairly well up to and somewhat beyond the $.95$-quantile.

Above that the actual distribution of the Kruskal-Wallis statistic has a shorter tail than the approximating $\chi^2_{k-1} = \chi^2_2$ distribution.

This means that the approximating $\chi^2_{k-1} = \chi^2_2$ distribution will give p-values that are higher than they should be, in the range when the true p-value is less than .05.

kruskall.wallis.pvalue (on web)

```
kruskal.wallis.pvalue <- function (KW,nvec=c(8,10,15),nsim=1000){
# This function simulates the p-value of an observed Kruskal-Wallis
# statistic KW, computed from samples of sizes nvec.
# The p-value is based on nsim simulations.
#---------------------------------------------------
N<-sum(nvec)
k <- length(nvec)
nvec2<-cumsum(nvec)
nvec1<-c(0,nvec2[1:(k-1)])+1
out<- NULL
x <-list()
for(i in 1:nsim){
xx <- sample(1:N,replace=F)
for(j in 1:k){x[[j]]<-xx[nvec1[j]:nvec2[j]]}
out[i]<-kruskal.test(x)$statistic}
y<-mean(out>=KW)
names(y)<-"p-value"
y}
```

Kruskal-Wallis for Flux3 Revisited

```
kruskal.wallis.pvalue(4.263295,c(6,6,6),10000)
p-value
 0.1148
```

The simulated p-value agrees well with the $.1186$ from the χ^2_2 approximation.

This in turn agrees with our previous observations about the approximation.

However, note what we get for the more extreme $KW = 8$:

```
> kruskal.wallis.pvalue(8,c(6,6,6),10000)
p-value
 0.0108
> 1-pchisq(8,2)
[1] 0.01831564
```

The Anderson-Darling k-Sample Test

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 asses the difference between the $\hat{F}_i(x)$ and $\hat{F}(x)$ by the Anderson-Darling

discrepancy metric

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

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.

The AD_k Test Is a Rank Test

Assume that all N observation $Y_{i\ell}, \ell = 1, \ldots, n_i, \;\; i = 1, \ldots, k$ are distinct (no ties). From the second and computational form of AD_k one can see that it depends on the observations $Y_{i\ell}$ only through its ranks.

This becomes clear when looking at $\hat{F}_i(Z_r)$ which is the proportion of $Y_{i\ell}$ values that are $\le Z_r$, i.e., only the rank of the $Y_{i\ell}$ matters in such comparisons, since

$$Y_{i\ell} \le Z_r \iff \mathrm{rank}(Y_{i\ell}) \le \mathrm{rank}(Z_r) = r \iff R_{i\ell} \le r$$

Some thought makes clear that the argument stays the same in the case of ties.

For details on the approximate null distribution of AD_k see the class website Reference: K-Sample Anderson-Darling Tests (Scholz and Stephens, 1987) see under R Code for Lecture Examples.

For R code to carry out the AD_k test install package adk and see `?adk.test` after invoking `library(adk)` for each new R session.

Anderson-Darling Test for Flux3

```
> adk.test(Flux3$X,Flux3$Y,Flux3$Z)
Anderson-Darling k-sample test.

Number of samples:  3
Sample sizes: 6 6 6
Total number of values: 18
Number of unique values: 12

Mean of Anderson Darling Criterion: 2
Standard deviation of Anderson Darling Criterion: 0.94415

T = (Anderson Darling Criterion - mean)/sigma

Null Hypothesis: All samples come from a common population.

                    t.obs P-value extrapolation
not adj. for ties 1.22493 0.11073             0
adj. for ties     1.12515 0.12346             0
```

Comments on KW-Test and AD-Test

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 differences).

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 $\rightarrow 1$ 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 χ^2_{k-1} approximation reasonable.

The AD-test is often used to justify the pooling of data when H_0 is not rejected.

It pays special attention to behavior in the sample tails, when $[\hat{F}(x)(1 - \hat{F}(x))]^{-1}$ is large, thus giving larger weight to discrepancies $[\hat{F}_i(x) - \hat{F}(x)]^2$ there.

adk.pvalue

Although adk.test provides p-values, they are approximations based on a mix of large sample theory and simulations. In order to assess the p-value via simulations first hand we provide on the web the function adk.pvalue which is very similar to kruskal.wallis.pvalue. Honest answers for $n_i < 5$ & p-values $< .01$ or $> .25$.

```
> system.time(out<-adk.pvalue(1.22493,nvec=c(6,6,6),nsim=1000))
   user  system elapsed
  10.18    0.01   10.27
> out
p-value
   0.12
> system.time(out<-adk.pvalue(1.22493,nvec=c(6,6,6),nsim=10000))
   user  system elapsed
 101.67    0.21  107.93
 > out
p-value
 0.1155
```

Appendix A: Distribution of SS_{Treat}

The next three slides establish the noncentral $\chi^2_{t-1,\lambda}$ distribution for $SS_{\mathrm{Treat}}/\sigma^2$, with noncentrality parameter

$$\lambda = \sum_{i=2}^{t} \nu_i^2/\sigma^2 = \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu})^2/\sigma^2$$

Distribution of SS_{Treat}

$$\bar{Y}_{i\bullet} \sim \mathcal{N}(\mu_i, \sigma^2/n_i) \quad \Longrightarrow \quad \sqrt{n_i}\,\bar{Y}_{i\bullet} \sim \mathcal{N}(\sqrt{n_i}\,\mu_i, \sigma^2) \quad \Longrightarrow \quad \sqrt{n_i}\,\bar{Y}_{i\bullet} = \sqrt{n_i}\,\mu_i + \sigma Z_i$$

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

Via Gram-Schmidt get an orthonormal basis $\mathbf{g}_1, \ldots, \mathbf{g}_t$ with $\mathbf{g}_1' = (\sqrt{n_1/N}, \ldots, \sqrt{n_t/N})$

Then
$$\begin{aligned}(\sqrt{n_1}\,\bar{Y}_{1\bullet}, \ldots, \sqrt{n_t}\,\bar{Y}_{t\bullet}) &= \sqrt{n_1}\,\bar{Y}_{1\bullet}\,\mathbf{e}_1' + \ldots + \sqrt{n_t}\,\bar{Y}_{t\bullet}\,\mathbf{e}_t' \\ &= V_1\mathbf{g}_1' + \ldots + V_t\mathbf{g}_t'\end{aligned}$$

The latter is the representation of $(\sqrt{n_1}\,\bar{Y}_{1\bullet}, \ldots, \sqrt{n_t}\,\bar{Y}_{t\bullet})$ in terms of the orthonormal

basis vectors $\mathbf{g}_i$ with random coefficients

$$V_i = (V_1\mathbf{g}_1' + \ldots + V_t\mathbf{g}_t')\mathbf{g}_i = (\sqrt{n_1}\,\bar{Y}_{1\bullet}, \ldots, \sqrt{n_t}\,\bar{Y}_{t\bullet})\mathbf{g}_i\ .$$

In particular

$$V_1 = (\sqrt{n_1}\,\bar{Y}_{1\bullet}, \ldots, \sqrt{n_t}\,\bar{Y}_{t\bullet})\mathbf{g}_1 = \sum_{i=1}^{t} \sqrt{n_i}\bar{Y}_{i\bullet} \times \sqrt{n_i/N} = \sum_{i=1}^{t} n_i\bar{Y}_{i\bullet}/\sqrt{N} = \sqrt{N}\bar{Y}_{\bullet\bullet}$$

Distribution of SS_{Treat}

and $$\sum_{i=1}^{t}(\sqrt{n_i}\bar{Y}_{i\bullet})^2 = \sum_{i=1}^{t} n_i\bar{Y}_{i\bullet}^2 = (V_1\mathbf{g}_1' + \ldots + V_t\mathbf{g}_t')(V_1\mathbf{g}_1 + \ldots + V_t\mathbf{g}_t) = \sum_{i=1}^{t} V_i^2$$

Thus $$\sum_{i=2}^{t} V_i^2 = \sum_{i=1}^{t} n_i\bar{Y}_{i\bullet}^2 - V_1^2 = \sum_{i=1}^{t} n_i\bar{Y}_{i\bullet}^2 - (\sqrt{N}\bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{t} n_i\bar{Y}_{i\bullet}^2 - N\bar{Y}_{\bullet\bullet}^2 = SS_{\text{Treat}}$$

$$\begin{aligned} V_i &= (\sqrt{n_1}\,\bar{Y}_{1\bullet},\ldots,\sqrt{n_t}\,\bar{Y}_{t\bullet})\mathbf{g}_i = (\sqrt{n_1}\mu_1,\ldots,\sqrt{n_t}\mu_t)\mathbf{g}_i + \sigma(Z_1,\ldots,Z_t)\mathbf{g}_i \\ &= \nu_i + \sigma U_i \quad \text{where} \quad \nu_i = (\sqrt{n_1}\mu_1,\ldots,\sqrt{n_t}\mu_t)\mathbf{g}_i \quad \text{and} \quad U_i = (Z_1,\ldots,Z_t)\mathbf{g}_i \end{aligned}$$

$\sum_{i=1}^{t} U_i\mathbf{g}_i'$ is the representation of $(Z_1,\ldots,Z_t)$ in terms of the $\mathbf{g}_i$ basis.
$\sum_{i=1}^{t} \nu_i\mathbf{g}_i'$ is the representation of $(\sqrt{n_1}\mu_1,\ldots,\sqrt{n_t}\mu_t)$ in terms of the $\mathbf{g}_i$ basis.

$$\sum_{i=1}^{t}\nu_i\mathbf{g}_i' \times \sum_{j=1}^{t}\nu_j\mathbf{g}_j = \sum_{i=1}^{t}\nu_i^2 = \sum_{i=1}^{t}(\sqrt{n_i}\mu_i)^2 = \sum_{i=1}^{t} n_i\mu_i^2$$

As argued previously, $Z_1,\ldots,Z_t$ i.i.d. $\mathcal{N}(0,1) \implies U_1,\ldots,U_t$ i.i.d. $\mathcal{N}(0,1)$.

Distribution of SS_{Treat}

$$\nu_1 = \sum_{i=1}^{t} \sqrt{n_i}\mu_i \times \sqrt{n_i/N} = \sum_{i=1}^{t} n_i\mu_i/\sqrt{N} = \sqrt{N}\bar{\mu}$$

$$\sum_{i=2}^{t} \nu_i^2 = \sum_{i=1}^{t} n_i\mu_i^2 - \nu_1^2 = \sum_{i=1}^{t} n_i\mu_i^2 - N\bar{\mu}^2 = \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu})^2$$

$$SS_{\text{Treat}}/\sigma^2 = \sum_{i=2}^{t} V_i^2/\sigma^2 = \sum_{i=2}^{t} (U_i + \nu_i/\sigma)^2 \sim \chi^2_{t-1,\lambda}$$

with $$\lambda = \sum_{i=2}^{t} \nu_i^2/\sigma^2 = \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu})^2/\sigma^2$$

Appendix B: F-Test Power Monotonicity

The next two slides establish the “intuitively obvious” fact that the power function of the F-test is monotonically increasing in the noncentrality parameter λ.

A Monotonicity Property of Coverage Probability

Theorem: If a r.v. $X \sim f(x) = F'(x)$ with $f(x) = f(-x)$ and if $f(x)$ is (strictly) monotone decreasing in $x \geq 0$, then $H(a) = P(|X-a| \leq x)$ (strictly) $\searrow$ in $|a|$.

Proof: $H(a) = P(|X-a| \leq x) = P(|-X-a| \leq x) = P(|X+a| \leq x) = H(-a)$, and thus it suffices to show $H(a) \searrow$ for $a \geq 0$. Also, only the case $x > 0$ matters.

$$H(a) = P(a-x \leq X \leq a+x) = F(a+x) - F(a-x)$$

with $$\frac{\partial H(a)}{\partial a} = f(a+x) - f(a-x) = f(a+x) - f(x-a) \leq 0 \ (< 0)\,,$$

since either $0 \leq a-x < a+x \implies f(a+x) - f(a-x) \leq 0 (< 0)$ or

$0 \leq x-a < x+a \implies f(a+x) - f(x-a) \leq 0 (< 0)$.

Corollary: $P(|X-a| \geq x) = 1 - H(a)$ (strictly) $\nearrow$ in $|a|$.

Monotonicity of the Power Function

The noncentral F tail probability is strictly $\nearrow$ in λ, i.e., $\beta(\lambda)$ strictly $\nearrow$ in λ.

With $Z_i, \tilde{Z}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ the monotonicity in λ follows from

$$\begin{aligned}
\beta(\lambda) = P(F_{t-1,N-t,\lambda} \geq F_{\text{crit}}) &= P\left(\frac{\left(Z_1 + \sqrt{\lambda}\right)^2 + \sum_{i=2}^{t-1} Z_i^2}{t-1} \geq F_{\text{crit}} \frac{\sum_{j=1}^{N-t} \tilde{Z}_j^2}{N-t}\right) \\
&= P\left(\left(Z_1 + \sqrt{\lambda}\right)^2 \geq F_{\text{crit}} \sum_{j=1}^{N-t} \tilde{Z}_j^2 \frac{t-1}{N-t} - \sum_{i=2}^{t-1} Z_i^2\right) \\
&= \int_{-\infty}^{\infty} P\left(\left(Z_1 + \sqrt{\lambda}\right)^2 \geq y\right) g(y) dy \quad \text{strictly} \nearrow \text{in } \lambda
\end{aligned}$$

applying the previous theorem with $f(x) = \varphi(x)$, the standard normal density.

Here $g(y)$ is the density of $Y = F_{\text{crit}} \sum_{j=1}^{N-t} \tilde{Z}_j^2 \, (t-1)/(N-t) - \sum_{i=2}^{t-1} Z_i^2$.