Stat 502
Design and Analysis of Experiments
One-Factor ANOVA

Fritz Scholz

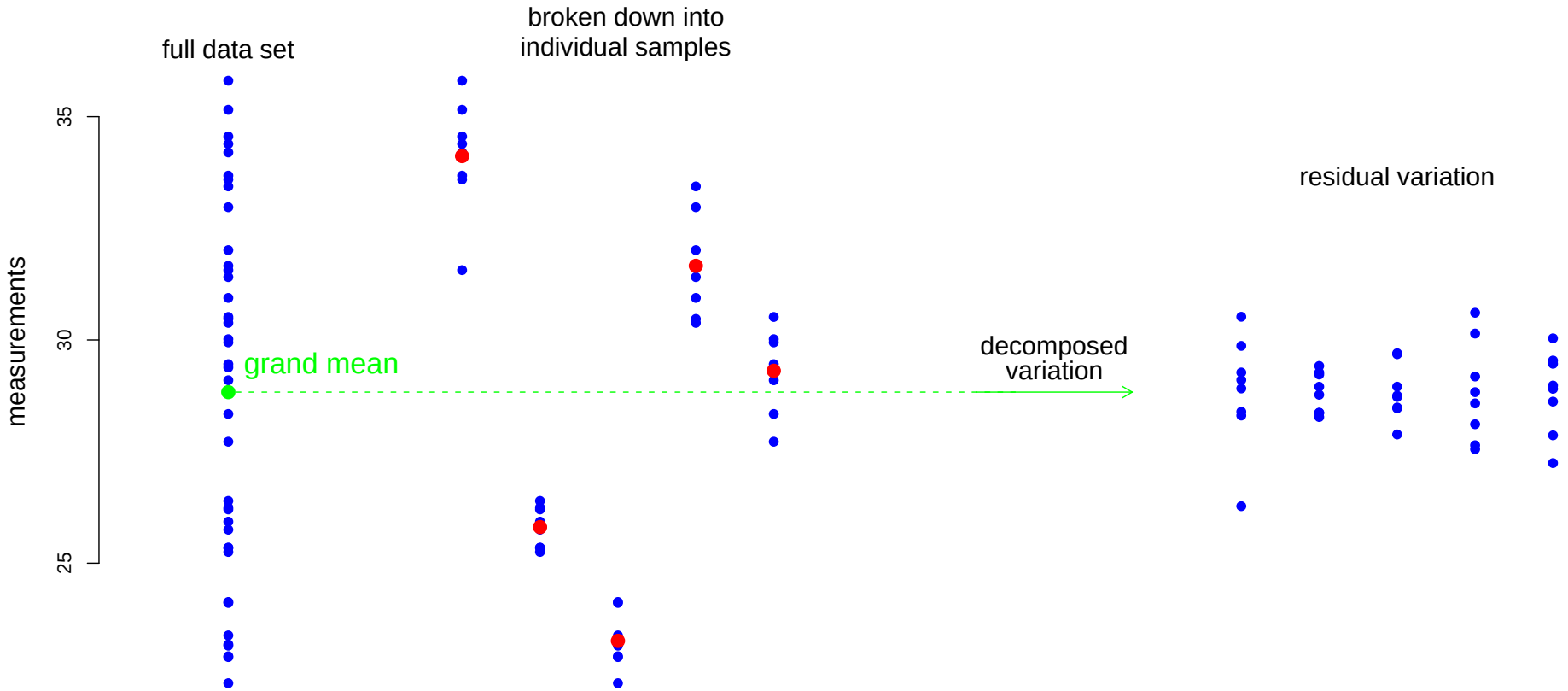
Department of Statistics, University of Washington

January 7, 2014

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
- Variance in ANOVA alludes to the analysis technique.
- The overall data variation is decomposed into several variation 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

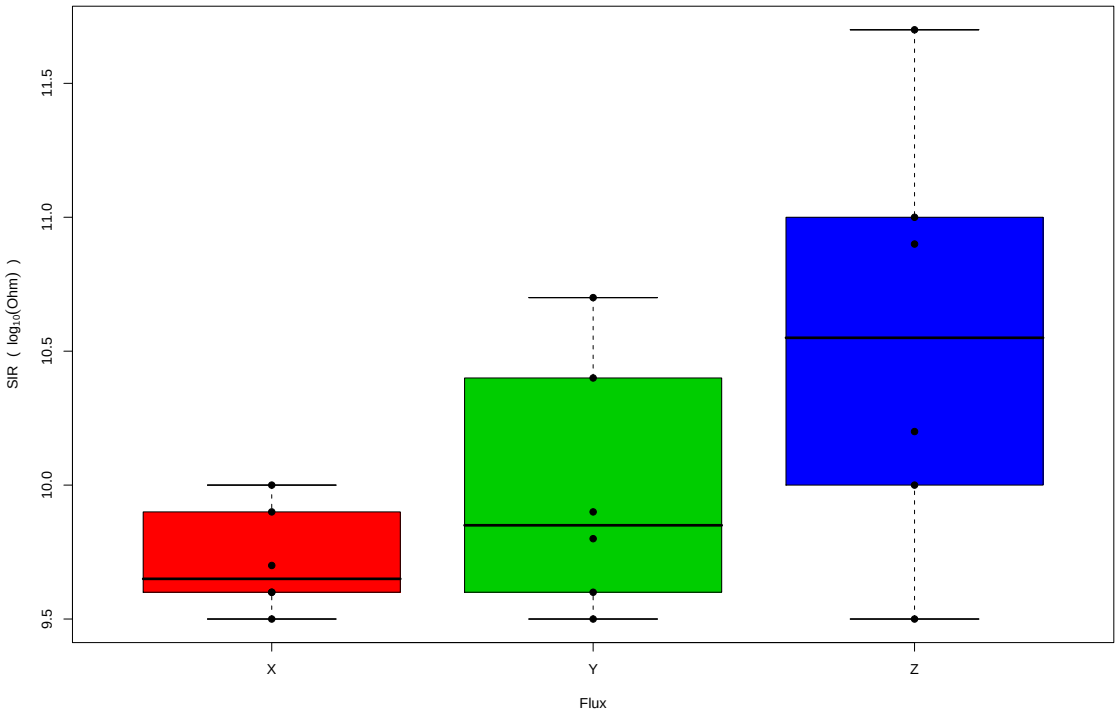
- Again we deal with circuit boards.
- Now we investigate 3 types of fluxes: X, Y, Z.
- With 18 circuit boards, randomly assign each flux to 6 boards.
- In principle, this gives us the randomization reference distribution and thus a logical basis for a test of the hypothesis H_0 : no flux differences.
- Randomize the order of soldering/cleaning, coating, and humidity chamber slots. These randomizations avoid unintended biases from hidden factors (dimensions).
- $\binom{18}{6}\binom{12}{6}\binom{6}{6} = 18,564 \cdot 924 \cdot 1 = 17,153,136$ flux allocations.
- Note growth in number of splits, 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.
- Denote these means by $\mu_1 = \mu_X, \quad \mu_2 = \mu_Y, \quad \text{and} \quad \mu_3 = \mu_Z.$
- Mathematically, $X \equiv Y \quad \text{and} \quad Y \equiv Z \implies X \equiv Z.$
- It would seem that testing $H_{0,XY} : X \equiv Y \quad \text{and} \quad H_{0,YZ} : Y \equiv Z$ might suffice.
- Statistically, $X \approx Y \quad \text{and} \quad Y \approx Z$ allows for the possibility that X and Z are sufficiently different.
- To guard against this we could perform all 3 possible 2-sample tests for the respective hypothesis testing problems:

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

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 these 3 tests are independent of each other we would have

P_0(Overall Type I Error)

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

- 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 in flight shutdown = 1/10000, the chance of at least one shutdown on a flight with k engines is

$$P(\text{at least one shutdown}) = 1 - (1 - p_F)^k \approx k \times p_F \ .$$

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

- $(1 - p_F)^k$ assumes that engine shutdowns are independent.
- This independence is the goal of ETOPS (Extended-range Twin-engine Operational Performance Standards) http://en.wikipedia.org/wiki/ETOPS
- E.g., different engines are serviced by different mechanics, or at separate time maintenance events.

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?
- Calculation based on independence is not correct.
- Any 2 comparisons involve a common sample $\Rightarrow$ dependence.
- Boole's inequality bounds the overall type I error probability:

$$\begin{aligned}
\pi_0 = P_0(\text{Overall Type I Error }) & \\
&= P_0(\text{reject } H_{0,XY} \cup \text{reject } H_{0,XZ} \cup \text{reject } H_{0,YZ}) \\
&\leq P_0(\text{reject } H_{0,XY}) + P_0(\text{reject } H_{0,XZ}) + P_0(\text{reject } H_{0,YZ}) \\
&= 3\alpha = .15 \quad \text{when} \quad \alpha = .05 \\
&= \text{expected number of false rejections}
\end{aligned}$$

- How much smaller than this upper bound is the true π_0?

Overall Type I Error Probability

- Evaluate it for 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.
- Take the difference of averages for the first two groups.
- Do this by simulation: `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 `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 `typeIerror.rateRand` is posted on web.

A Global Testing View

- Rather than using all 3 pairwise tests statistics $|\bar{X} - \bar{Y}|$, $|\bar{X} - \bar{Z}|$, and $|\bar{Y} - \bar{Z}|$ separately, 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 Flux3 data: $t = 3$, $n_1 = n_2 = n_3 = 6$, a balanced design.
- When the n_i are not all the same $\Longrightarrow$ 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:

$$Y_{ij} = \mu_i + \epsilon_{ij} \text{ with } E(\epsilon_{ij}) = 0 \text{ and } \text{var}(\epsilon_{ij}) = \sigma^2$$

- View ϵ_{ij} (i.i.d.) as response variation/error/noise.
- Treatment Effects Model:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \text{ with } E(\epsilon_{ij}) = 0 \text{ and } \text{var}(\epsilon_{ij}) = \sigma^2$$

- $\mu = \bar{\mu} = \sum_{ij} \mu_i / N = \sum_i n_i \mu_i / N =$ grand mean or n_i/N-weighted average of the μ_i
- $\tau_i = \mu_i - \mu = \mu_i - \bar{\mu}$ is the i^{th} treatment effect
- ϵ_{ij} (i.i.d.) = within treatment variation, $E(\epsilon_{ij}) = 0$, $\text{var}(\epsilon_{ij}) = \sigma^2$.
- Note: 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:

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

  i.e., there is no variation or change due to treatments.
- The reduced model corresponds to our stated hypothesis

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

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

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

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

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

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

- least squares estimates (LSE) $\hat{\mu}_i = \bar{Y}_{i\bullet}$ minimize $SS(\mu_1, \ldots, \mu_t) \Longrightarrow SS(\hat{\mu}_1, \ldots, \hat{\mu}_t) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2$.

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 $a_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n,$ write

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

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

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

$$\bar{Y}_{\cdot\cdot} = \sum_{ij} Y_{ij} / \sum_{i} n_i = \sum_{i} (n_i/N) \bar{Y}_{i\cdot} \implies \sum_{i} \sum_{j} (Y_{ij} - \bar{Y}_{\cdot\cdot}) = 0$$

$$\begin{aligned}
\implies SS(\mu) &= \sum_{ij} (Y_{ij} - \mu)^2 = \sum_{ij} (Y_{ij} - \bar{Y}_{\cdot\cdot} + \bar{Y}_{\cdot\cdot} - \mu)^2 \\
&= \sum_{ij} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2 + \sum_{ij} (\bar{Y}_{\cdot\cdot} - \mu)^2 \\
&\quad + 2 \sum_{ij} (Y_{ij} - \bar{Y}_{\cdot\cdot})(\bar{Y}_{\cdot\cdot} - \mu) \\
&= \sum_{ij} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2 + \sum_{ij} (\bar{Y}_{\cdot\cdot} - \mu)^2 \geq \sum_{ij} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2
\end{aligned}$$

- The least squares estimate (LSE) $\hat{\mu} = \bar{Y}_{\cdot\cdot}$ minimizes $SS(\mu)$ $\implies SS(\hat{\mu}) = \sum_{ij} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2$.

Means and Variances of Least Squares Estimates

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

Sum of Squares (SS) Decomposition

Using $\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot}) = 0 \implies$ sum of squares decomposition

$$
\begin{aligned}
SS_T = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2 &= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot} + \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2 + \sum_{i=1}^{t}\sum_{j=1}^{n_i}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \\
&\quad + 2\sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}) \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2 + \sum_{i=1}^{t}\sum_{j=1}^{n_i}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \\
&= SS_E + SS_{Treat}
\end{aligned}
$$

ANOVA decomposition of total SS variation SS_T:

$SS_T = SS_E + SS_{Treat} = SS_W + SS_B$.

$SS_T =$ error variation+treatment variation

$SS_T =$ variation within samples + variation between samples.

How to Compare the Model Fits?

- How should we compare the two model fits

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

- 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

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

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

- For fair comparison make allowances for this extra freedom.
- 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 $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2.

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

- Using $E(Y^2) = \mathrm{var}(Y) + [E(Y)]^2$

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

$E(MS_E) = \sigma^2$

$$s_i^2 = \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2/(n_i - 1) \implies \sum_{i=1}^{t} (n_i - 1)s_i^2 = SS_E$$

and the result from the previous slide shows

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

or the Mean Square for Error

$MS_E = \dfrac{SS_E}{N - t} = \dfrac{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2}{N - t}$ is an unbiased estimate for σ^2

True whether $H_0 : \mu_1 = \ldots = \mu_t$ holds or not (without normality).

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

$$
\begin{aligned}
SS_{Treat} &= \sum_{i=1}^{t} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{t} n_i(\bar{Y}_{i\cdot}^2 - 2\bar{Y}_{i\cdot}\bar{Y}_{\cdot\cdot} + \bar{Y}_{\cdot\cdot}^2) \\
&= \sum_{i=1}^{t} n_i \bar{Y}_{i\cdot}^2 - N\bar{Y}_{\cdot\cdot}^2 \\
\Longrightarrow E(SS_{Treat}) &= \sum_{i=1}^{t} n_i E(\bar{Y}_{i\cdot}^2) - N\, E(\bar{Y}_{\cdot\cdot}^2) \qquad \text{(with/without normality)} \\
&= \sum_{i=1}^{t} n_i(\mathrm{var}(\bar{Y}_{i\cdot}) + [E(\bar{Y}_{i\cdot})]^2) - N(\mathrm{var}(\bar{Y}_{\cdot\cdot}) + [E(\bar{Y}_{\cdot\cdot})]^2) \\
&= \sum_{i=1}^{t} n_i(\sigma^2/n_i + \mu_i^2) - N(\sigma^2/N + \bar{\mu}^2) \\
&= (t-1)\sigma^2 + \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu})^2
\end{aligned}
$$

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

A Test Statistic for H_0

- Under H_0 both MS_{Treat} and MS_E are unbiased estimates of σ^2
- H_0 is false
  $\Longrightarrow \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu})^2/(t-1) > 0 \implies E(MS_{Treat}) > E(MS_E)$
- MS_{Treat} will generally be somewhat larger than MS_E.
- More so when the μ_i are more dispersed.
- The n_i act as magnifiers!
- This suggests $F = MS_{Treat}/MS_E$ as a plausible test statistic.
- Looking at ratio makes more sense than looking at difference.
- Any such difference should be viewed relative to MS_E.
- Use this test statistic also in our randomization test.

Equivalent Form for the F-Statistic under Randomization

- In $SS_T = SS_{Treat} + SS_E$ the sum SS_T stays constant over all partitions of the full data set into t groups of sizes $n_1, \ldots, n_t$.

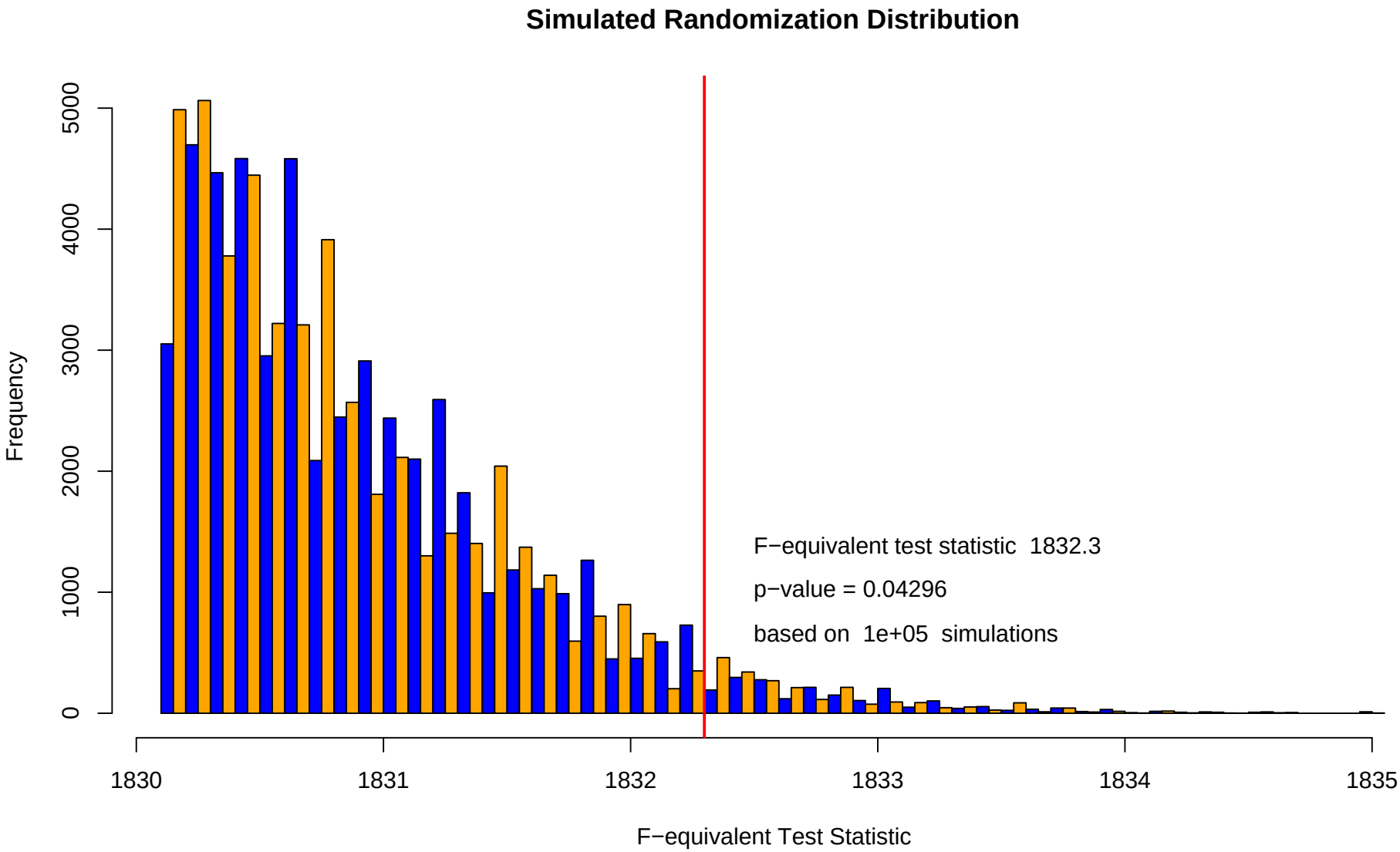
$$SS_{Treat} = \sum_{i=1}^{t} n_i \bar{Y}_{i\bullet}^2 - N\bar{Y}_{\bullet\bullet}^2 = F_{equiv} - N\bar{Y}_{\bullet\bullet}^2$$

- $F_{equiv} = \sum_{i=1}^{t} n_i \bar{Y}_{i\bullet}^2$ varies with partition splits.
- $\bar{Y}_{\bullet\bullet}$ also stays constant over all such partition splits.

$$\begin{aligned} F = \frac{N-t}{t-1}\frac{SS_{Treat}}{SS_E} &= \frac{N-t}{t-1}\frac{SS_{Treat}}{SS_T - SS_{Treat}} \\ &= \frac{N-t}{t-1}\frac{F_{equiv} - N\bar{Y}_{\bullet\bullet}^2}{SS_T - (F_{equiv} - N\bar{Y}_{\bullet\bullet}^2)} \quad \nearrow \text{ in } F_{equiv} \end{aligned}$$

- Thus the randomization distribution of F is in 1-1 correspondence with the randomization distribution of F_{equiv}.
- We can then take it 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){
  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 <- numeric(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!

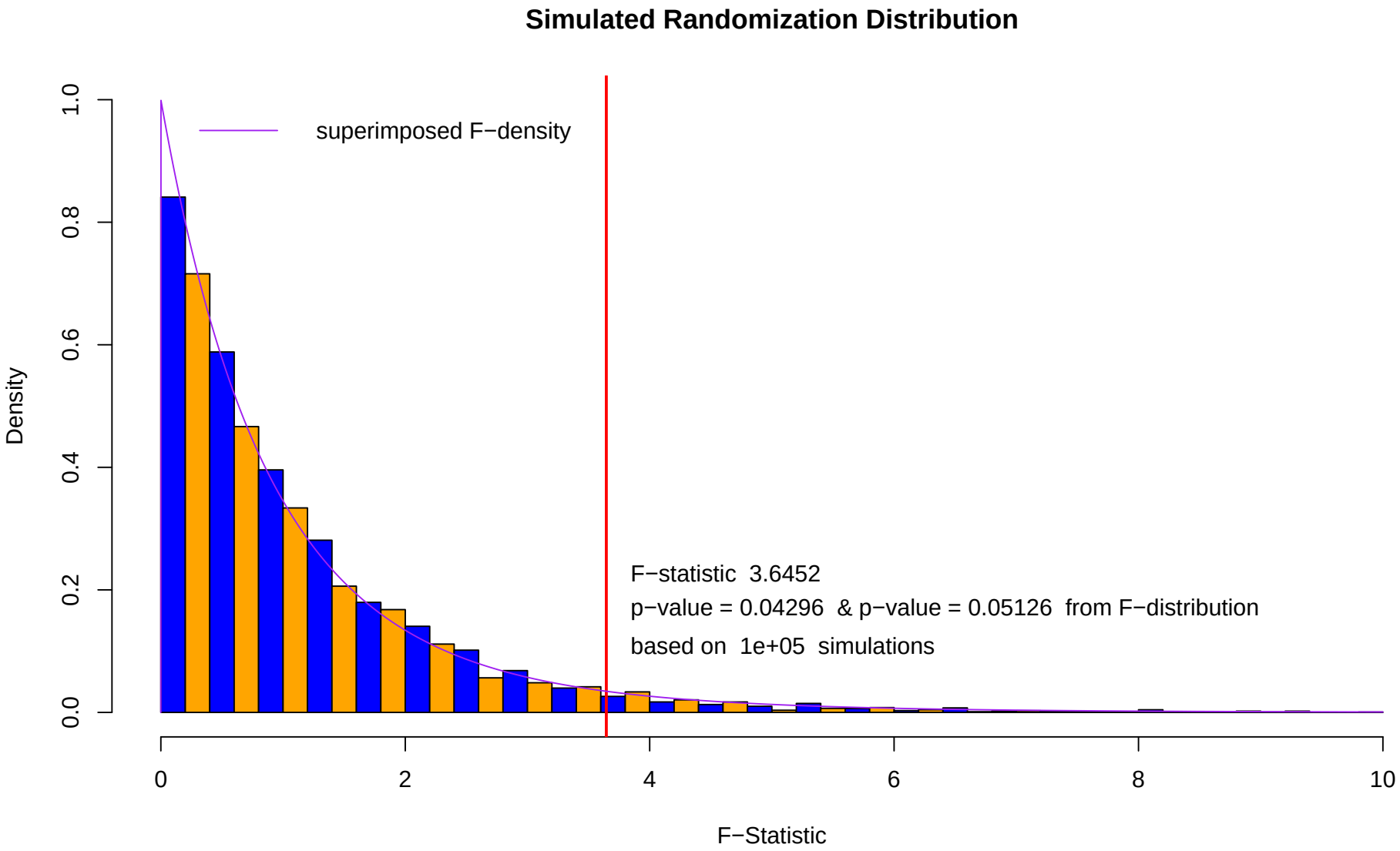
F-Distribution as Approximation to the Randomization Distribution

- 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

$$F = \frac{N-t}{t-1} \frac{F_{equiv} - N\bar{Y}_{\cdot\cdot}^2}{SS_T - (F_{equiv} - N\bar{Y}_{\cdot\cdot}^2)}$$

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

Randomization Distribution for Flux3



Assuming Normality

- We now assume that the Y_{ij} are independent, normal r.v.'s with the previously indicated model parameters.
- For $H_0 : \mu_1 = \ldots = \mu_t$ true or not $\Rightarrow$ $(n_i - 1)s_i^2 \sim \sigma^2 \chi^2_{n_i - 1}$.
- Further, $s_1^2, \ldots, s_t^2$ are independent and thus

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

- 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 $SS_{Treat} = \sum_{i=1}^{t} n_i \bar{Y}_{i\bullet}^2 - N\bar{Y}_{\bullet\bullet}^2 \sim \sigma^2 \chi^2$?
- What degrees of freedom f?
- Under H_0 we expect $f = t - 1$ since $E(MS_{Treat}) = E(SS_{Treat}/(t-1)) = \sigma^2$.

The Distribution of F

- The previous slide and Appendix A, slide 148, establish the following: SS_E and SS_{Treat} are independent and

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

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

- For $H_0 : \mu_1 = \ldots = \mu_t$ this becomes the $F_{t-1,N-t}$ distribution.
- We reject H_0 whenever
  $F \geq F_{t-1,N-t}(1-\alpha) = \texttt{Fcrit} = \texttt{qf}(1-\alpha, \texttt{t}-1, \texttt{N}-\texttt{t})$
  $= (1-\alpha)$-quantile of the $F_{t-1,N-t}$ distribution.
- Power function: $\beta(\lambda) = P_\lambda(F \geq F_{t-1,N-t}(1-\alpha)) =$
  $1 - \texttt{pf}(\texttt{Fcrit}, \texttt{t}-1, \texttt{N}-\texttt{t}, \lambda)$

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
[11]  9.8  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"
[13] "Z" "Z" "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, slide 151).
- Let us consider the drivers in λ.
- $\lambda \nearrow$ as $\sigma \searrow$, provided the μ_i are not all the same.
- The more difference between the μ_i the higher λ.
- Increasing the sample sizes will magnify $n_i(\mu_i - \bar{\mu})^2$.
- For fixed σ and μ_i not all equal, the power increases strictly

$$\frac{\partial \lambda \sigma^2}{\partial n_i} = (\mu_i - \bar{\mu})^2 - \sum_j 2n_j \frac{(\mu_j - \bar{\mu})(\mu_i - \bar{\mu})}{N} = (\mu_i - \bar{\mu})^2 > 0$$

$$\text{since} \quad \frac{\partial \bar{\mu}}{\partial n_i} = \frac{\partial \left(\sum_j n_j \mu_j / \sum_j n_j \right)}{\partial n_i} = \frac{(\mu_i - \bar{\mu})}{N}$$

- The sample sizes we can plan for.
- Later: Blocking units into homogeneous groups $\Rightarrow$ smaller σ.

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.
- Then we would not emphasize one treatment over any other.
- Optimality criterion that could be used as justification?
- Plan for a balanced design upfront. How large should n be?
- 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

$$\lambda_1 = \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu})^2/\sigma^2 = \frac{N\Delta^2}{\sigma^2}\,\frac{n_1}{N}\left(1 - \frac{n_1}{N}\right) = \frac{N\Delta^2}{\sigma^2}R_1$$

similarly $\lambda_i = \dfrac{N\Delta^2}{\sigma^2}\,\dfrac{n_i}{N}\left(1 - \dfrac{n_i}{N}\right) = \dfrac{N\Delta^2}{\sigma^2}R_i$ for the other cases.

The Max-Min Solution

- For fixed $\sigma > 0$ and $\Delta \neq 0$ the following max min power

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

  is achieved when $n_1 = \ldots = n_t$. Here $R_i = (n_i/N)(1 - n_i/N)$.
- Reason: $R_i = (n_i/N)(1 - n_i/N)$ increases for $n_i/N \leq 1/2$.
- Since $n_1 + \ldots + n_t = N = nt$ is fixed, can increase the smallest R_i only at the expense of lowering some higher R_j.
- This increase only happens when something is left to lower.

$$\begin{aligned} \Rightarrow \quad \max_{n_1,\ldots,n_t} \min_{1\leq i\leq t} [\lambda_i] &= \frac{N\Delta^2}{\sigma^2}\frac{n}{N}\left(1 - \frac{n}{N}\right) \text{ for } n = n_1 = \ldots = n_t \\ &= n\frac{\Delta^2}{\sigma^2}\left(1 - \frac{n}{nt}\right) = n\frac{\Delta^2}{\sigma^2}\frac{t-1}{t} = n \cdot \lambda_0 \ . \end{aligned}$$

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

An Alternate Rationale (Dean and Voss, p. 52)

- Let $C_\Delta = \{\boldsymbol{\mu} = (\mu_1, \ldots, \mu_t) : \max(\boldsymbol{\mu}) - \min(\boldsymbol{\mu}) \geq \Delta\}$ for $\Delta > 0$.
- For fixed $\sigma > 0, \Delta > 0$ find sample sizes $n_1, \ldots, n_t$ (with $\sum n_i = nt = N$ fixed), to maximize the power, i.e.,

$$\min_{\boldsymbol{\mu} \in C_\Delta} \lambda(\boldsymbol{\mu}) = N \min_{\boldsymbol{\mu} \in C_\Delta} \frac{\sum_{i=1}^t p_i(\mu_i - \bar{\mu})^2}{\sigma^2} \quad \text{with} \quad p_i = \frac{n_i}{N}$$

- It can again be shown that equal sample size allocation, i.e., $n_1 = \ldots = n_t = n$, is the optimal (max-min) strategy.
- Suppose $\mu_1 \leq \mu_2, \ldots, \mu_{t-1} \leq \mu_t = \mu_1 + \Delta$, then $\lambda(\boldsymbol{\mu})$ is minimized over the restricted $\mu_2, \ldots, \mu_{t-1}$ when these are $= \bar{\mu}$

$$\bar{\mu} = (p_2 + \ldots + p_{t-1})\bar{\mu} + p_1\mu_1 + p_t(\mu_1 + \Delta)$$

$$\Longrightarrow \quad \bar{\mu} = \frac{p_1}{p_1 + p_t}\mu_1 + \frac{p_t}{p_1 + p_t}(\mu_1 + \Delta) = \tilde{p}_1\mu_1 + \tilde{p}_t(\mu_1 + \Delta)$$

$$\Longrightarrow \quad \mu_t - \bar{\mu} = \Delta\tilde{p}_1 \quad \text{and} \quad \mu_1 - \bar{\mu} = -\tilde{p}_t\Delta$$

$$\min_{\boldsymbol{\mu} \in C_\Delta} \lambda(\boldsymbol{\mu}) = \frac{N(p_1 + p_t)\Delta^2}{\sigma^2}(\tilde{p}_1\tilde{p}_t^2 + \tilde{p}_1^2\tilde{p}_t) = \frac{N\Delta^2}{\sigma^2}\frac{p_1 p_t}{p_1 + p_t}$$

- For fixed $p_1 + p_t$ maximized by $p_1 = p_t$, i.e., $n_1 = n_t$ q.e.d.

Some Discussion of max min Results

- The previous two max min situations are different.
- In the first we stipulated one mean different from the others, the latter assumed without treatment effect, i.e., the same.
- In the second we assumed two means to differ by $\Delta > 0$, while the others were somewhere in between.
- i.e., we focus on maximum treatment difference.
- In either case nothing was known about the indices of the differing treatment effects.
- My hunch is that more general results like this can be formulated and solved with the same resolution.
- Research problem for formulation and resolution!?

`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 have a reasonable upper bound σ_u for σ, 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 given on 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 an optional value for the
# target power that will be highlighted on the plot.
....
}
```

Example Usage of `sample.sizeANOVA`

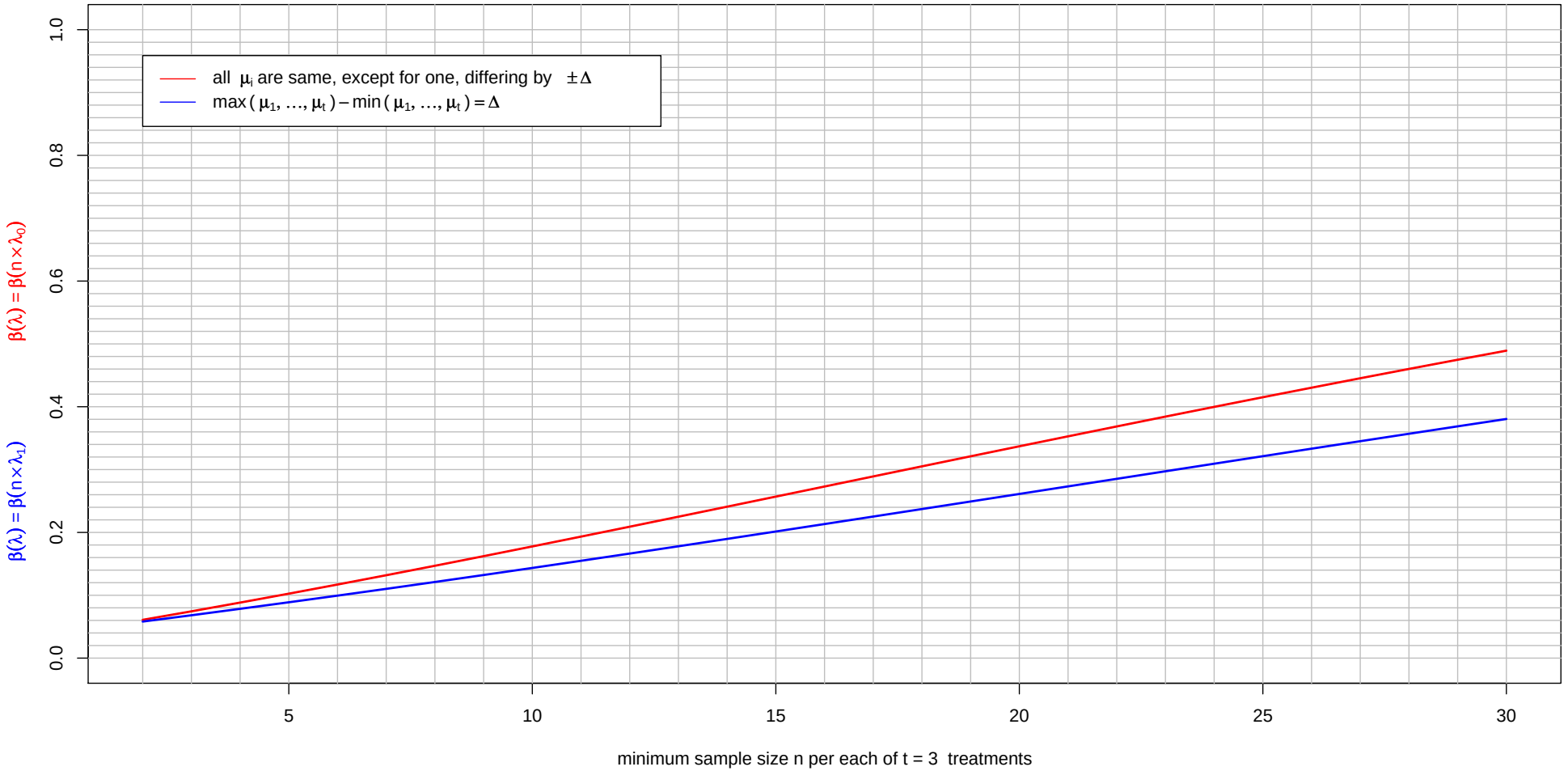
- The following three function calls invoke the default `t.treat=3` to produce the plots on the next three slides.

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

- $n = 77$ the minimal sample size under the first rationale.
- $n = 103$ the minimal sample size under the alternate rationale.

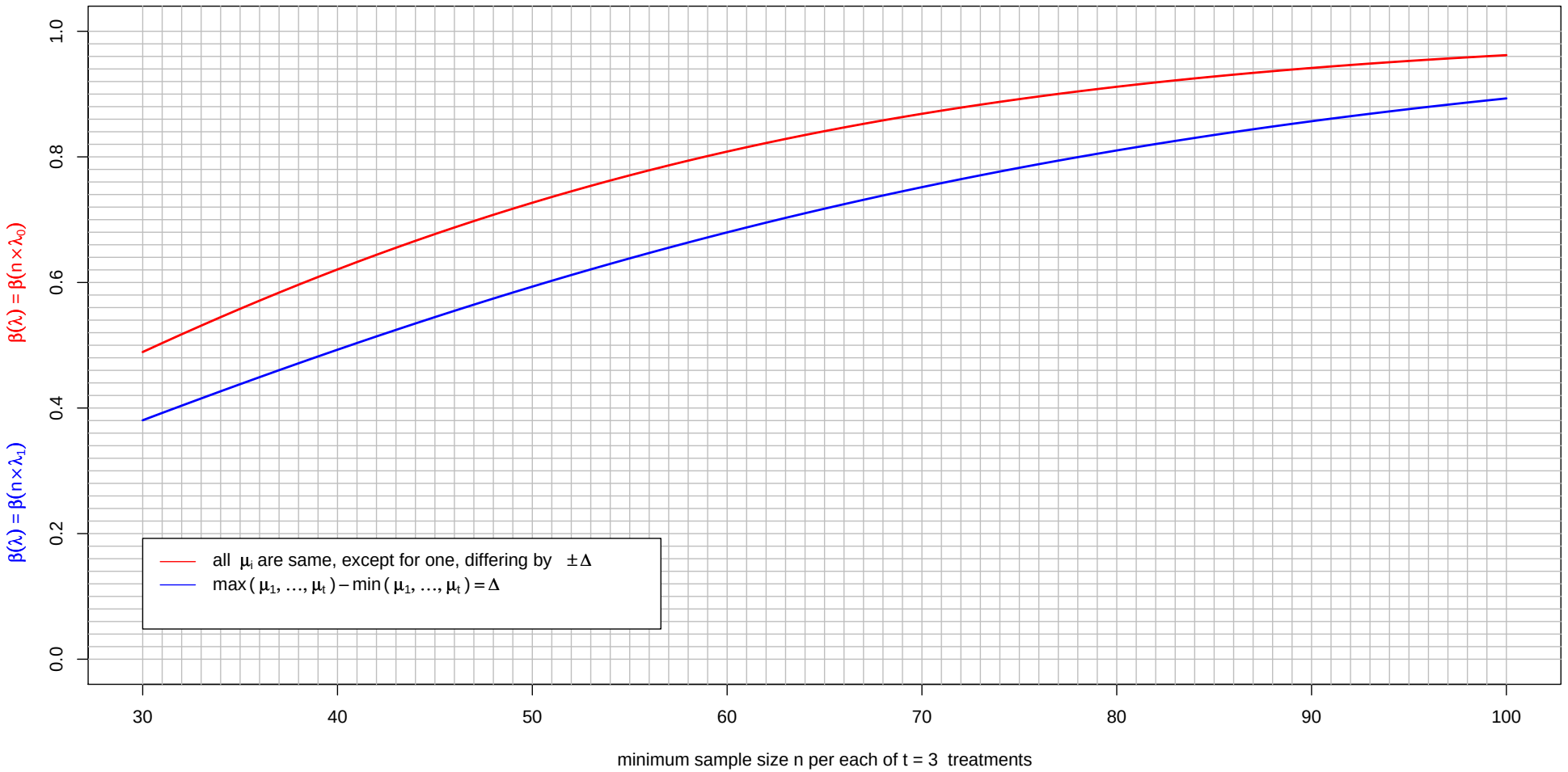
Sample Size Determination

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



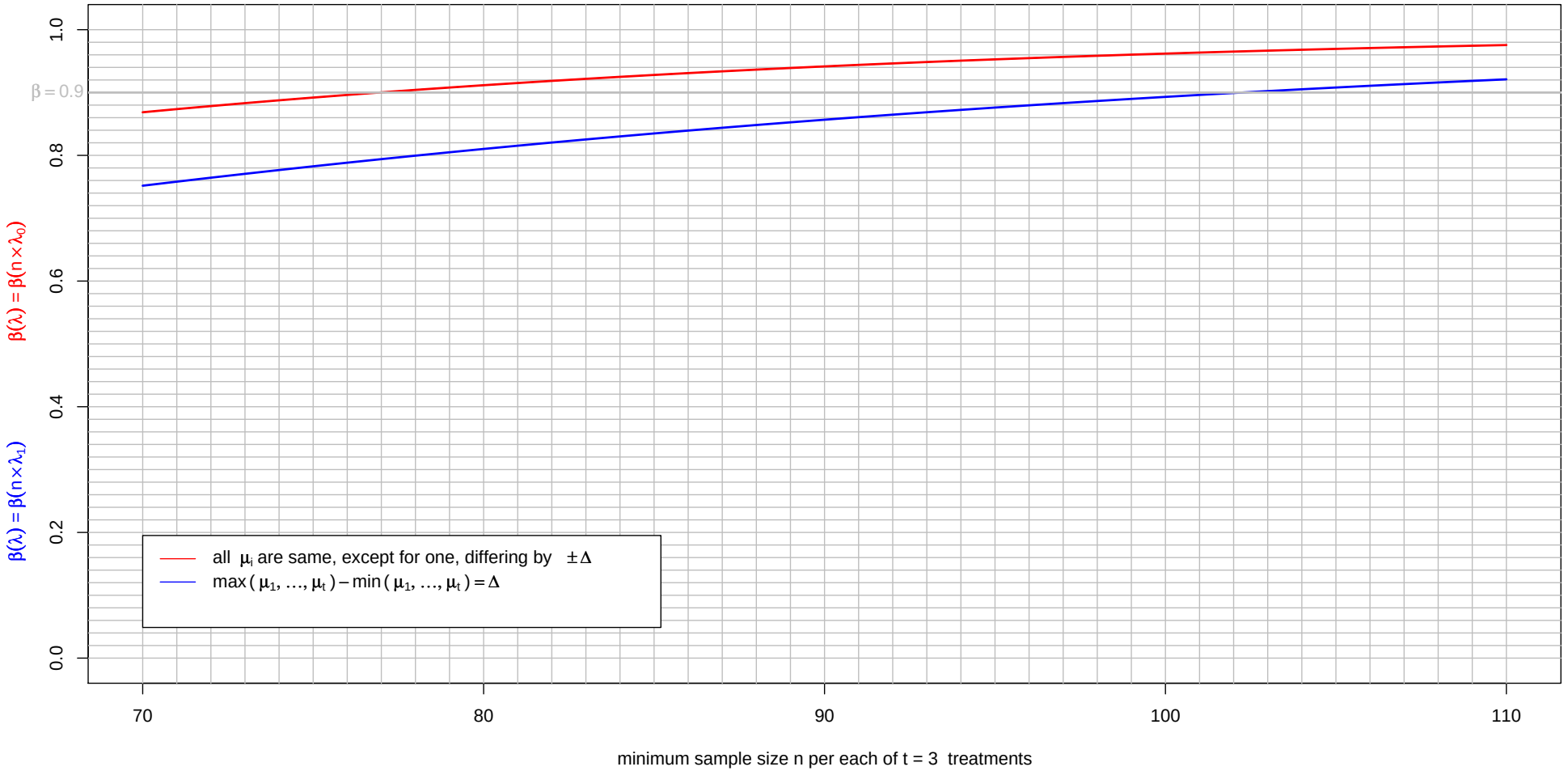
Sample Size Determination (increased n)

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



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

```
tcrit <- qf(1-alpha,t-1,n*t-t)}
1-pf(tcrit,t-1,n*t-t,ncp)}
```

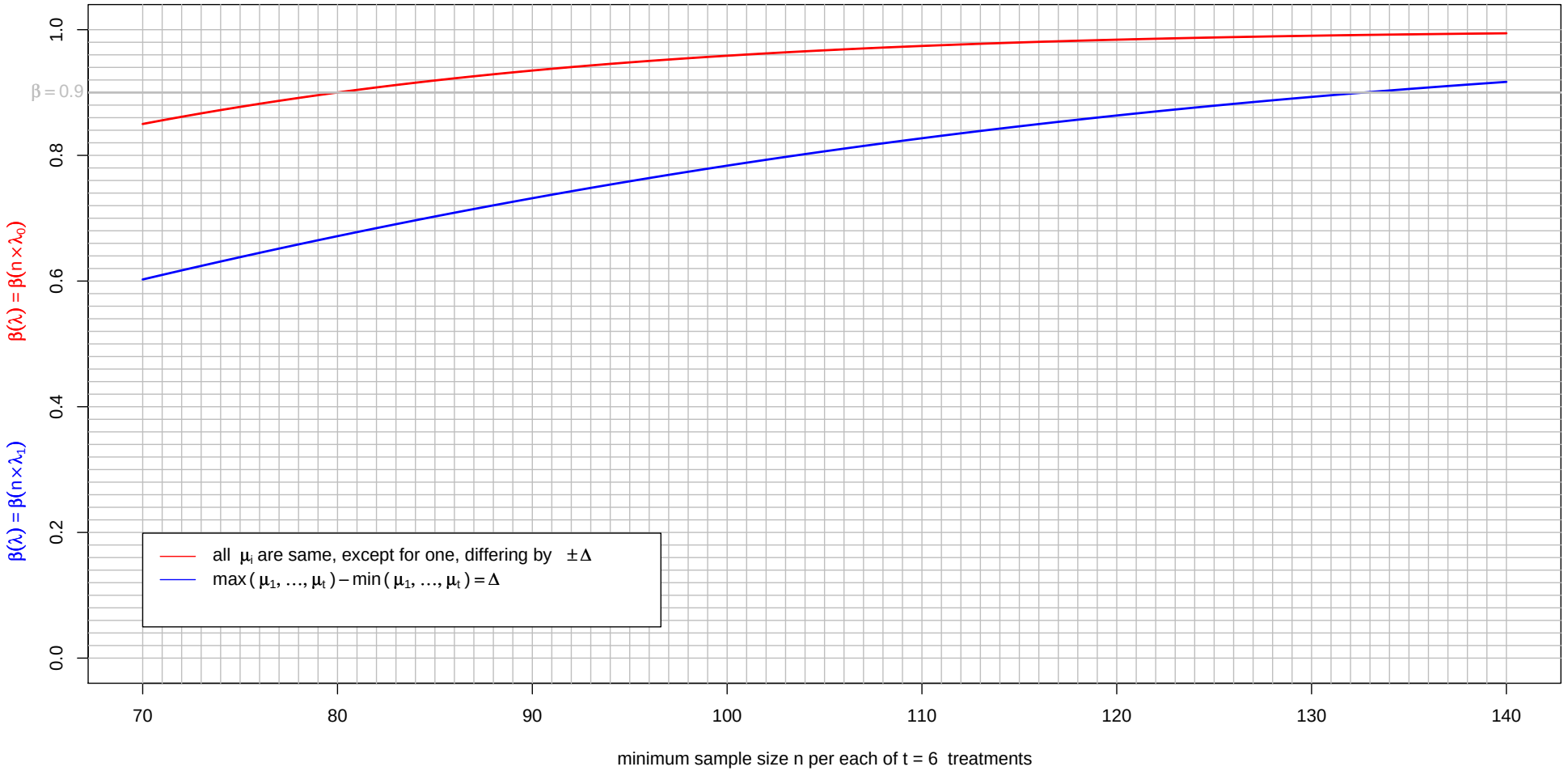
- Thus the choice of n is very much affected, as can be seen in the following slide produced with $t = 6$

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

- $n = 81$ minimum sample size under the first rationale.
- $n = 133$ under the alternate rationale.

Sample Size Determination (magnified)

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



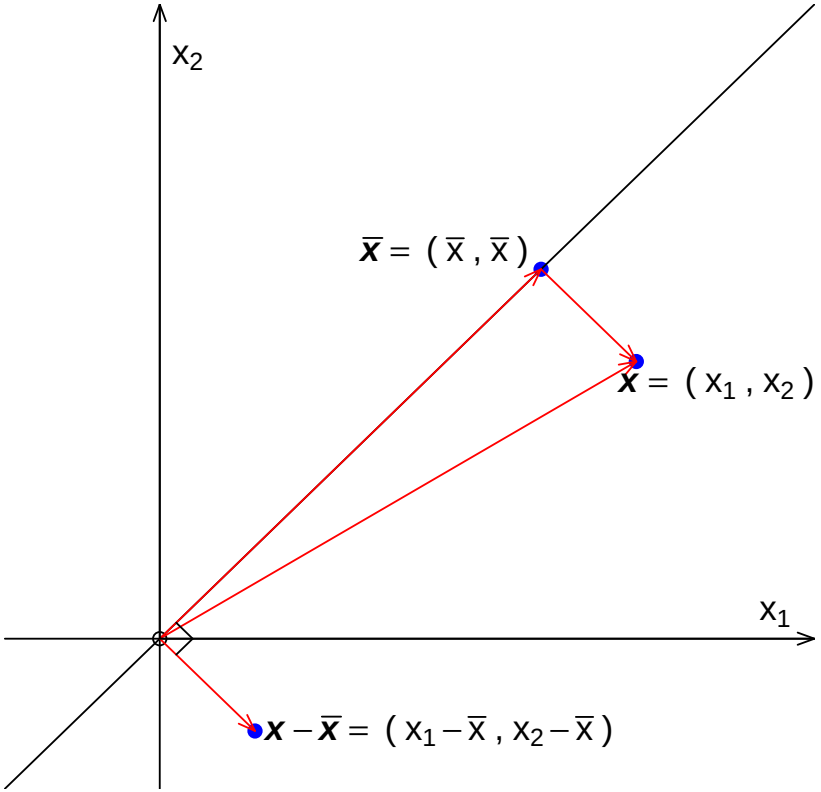
Degrees of Freedom and Geometry – Single Sample

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

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

- $(\bar{X}, \ldots, \bar{X})$ varies in one dimension, along $\mathbf{1}' = (1, \ldots, 1)$
- 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



$$\text{Pythagoras: } |\mathbf{x}|^2 = |\bar{\mathbf{x}}|^2 + |\mathbf{x} - \bar{\mathbf{x}}|^2 = \sum_i \bar{x}^2 + \sum_i (x_i - \bar{x})^2$$

$$= n\bar{x}^2 + \sum_i (x_i - \bar{x})^2 \text{ our previous SS decomposition}$$

Degrees of Freedom and Geometry in t Samples

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

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

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

$$\begin{aligned} \sum_i \sum_j Y_{ij}^2 &= \sum_i \sum_j \bar{Y}_{\cdot\cdot}^2 + \sum_i \sum_j (Y_{ij} - \bar{Y}_{\cdot\cdot})^2 \\ &= \sum_i \sum_j \bar{Y}_{\cdot\cdot}^2 + \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2 + \sum_i \sum_j (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \end{aligned}$$

Orthogonalities

$$\begin{aligned}\sum_i \sum_j \bar{Y}_{\cdot\cdot}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}) &= \bar{Y}_{\cdot\cdot} \sum_i n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}) \\ &= \bar{Y}_{\cdot\cdot}(\sum_i \sum_j Y_{ij} - N\bar{Y}_{\cdot\cdot}) = 0\end{aligned}$$

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

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

Dimensions of Subspaces or Degrees of Freedom

Let $\mathbf{1}_n' = (1, 1, \ldots, 1)$ denote an n-vector of 1's. The vectors

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

$$= (\bar{Y}_{1\bullet} - \bar{Y}_{\bullet\bullet})\mathbf{E}_1 + \ldots + (\bar{Y}_{t\bullet} - \bar{Y}_{\bullet\bullet})\mathbf{E}_t = \mathbf{D}$$

span a $(t-1)$-dimensional subspace $\boldsymbol{M} \subset R^N$, because $\mathbf{E}_1, \ldots, \mathbf{E}_t$ span a t-dimensional subspace of R^N and $\mathbf{D}$ is always orthogonal to $\mathbf{1}_N \in \boldsymbol{M}$, since $\mathbf{1}_N'\mathbf{D} = \sum_{i=1}^t n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) = 0$.

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

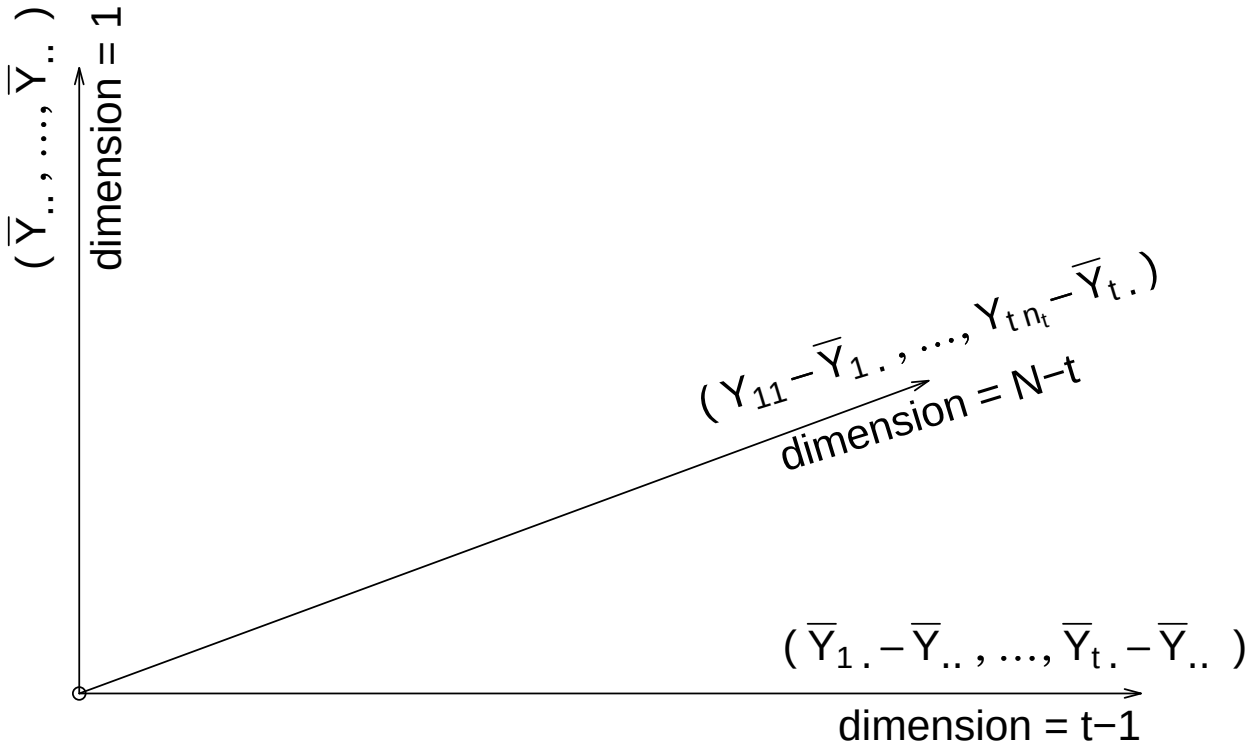
More on Dimensions and Degrees of Freedom

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

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

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 $\mathbf{v} \in M_{N-t}$ is of form $\mathbf{v} = a_{11}\mathbf{e}_{11} + \ldots + a_{tn_t}\mathbf{e}_{tn_t}$ 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



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

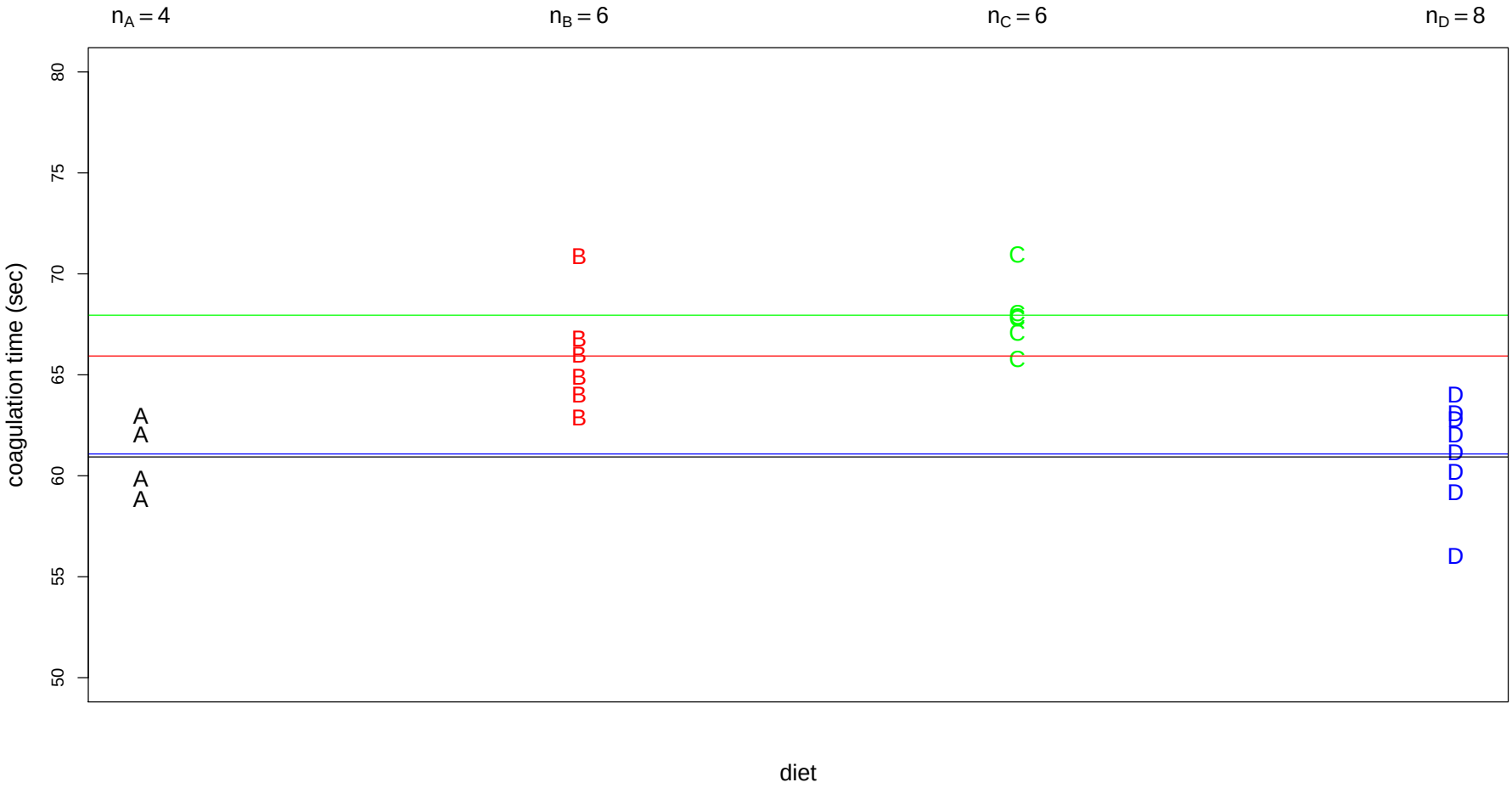
$$\sum_i \sum_j Y_{ij}^2 = \sum_i \sum_j \bar{Y}_{\bullet\bullet}^2 + \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\bullet})^2 + \sum_i \sum_j (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2$$

Coagulation Example

- Import data: `coag <- read.csv("coag.csv")`
- To understand the blood coagulation behavior in relation to various diets, 24 lab animals were given 4 different diets.
- Their respective coagulation times were measured in seconds.
- The lab animals were assigned randomly to the various diets.
- The results were as follows:

```
> coag$ctime
 [1] 59 60 62 63 63 64 65 66 67 71 66 67 68 68 68
[16] 71 56 59 60 61 62 63 63 64
> coag$diet
 [1] A A A A B B B B B B C C C C C C D D D D D D D
[24] D
Levels: A B C D
```

Plot for Coagulation Example



ANOVA for Coagulation Example

- The plot used `jitter(coag$ctime)` to plot `ctime` in the vertical direction and to plot its horizontal mean lines.
- This perturbs observations a small random amount to make tied observations more visible.
- The means for diet A and D would have coincided otherwise.

```
> anova(lm(ctime~diet,data=coag))
# assumes data frame coag with variables ctime and diet
# is in the work space
Analysis of Variance Table

Response: ctime
          Df Sum Sq Mean Sq F value    Pr(>F)
diet       3    228    76.0  13.571 4.658e-05 ***
Residuals 20    112     5.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

lm for Coagulation Example

```
> out <- lm(ctime~diet,data=coag)
   # 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)         dietB         dietC         dietD
 6.100000e+01  5.000000e+00  7.000000e+00 -2.515253e-15
```

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).

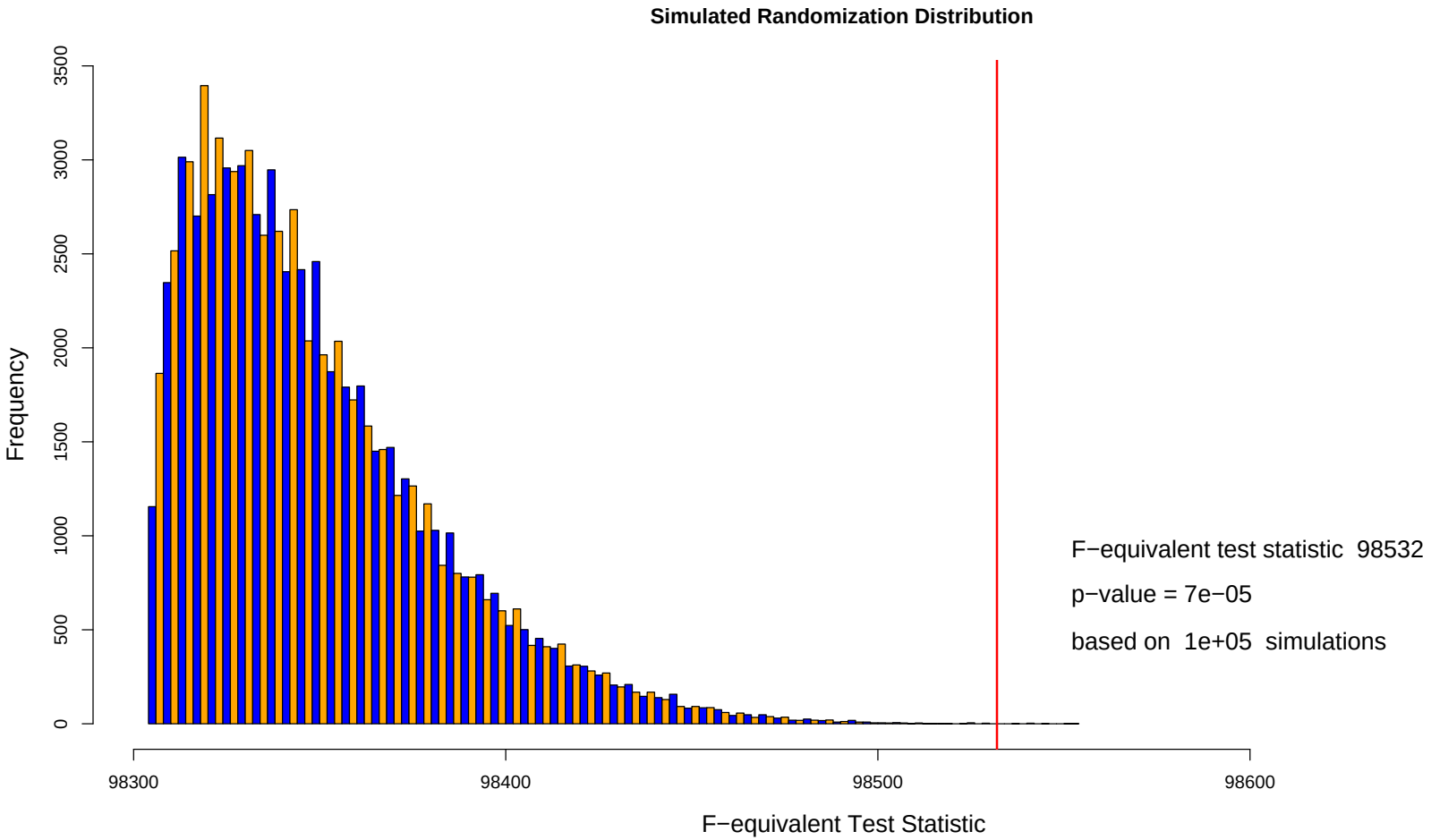
Rounded Residuals from `lm` for Coagulation Example

```
>  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
```

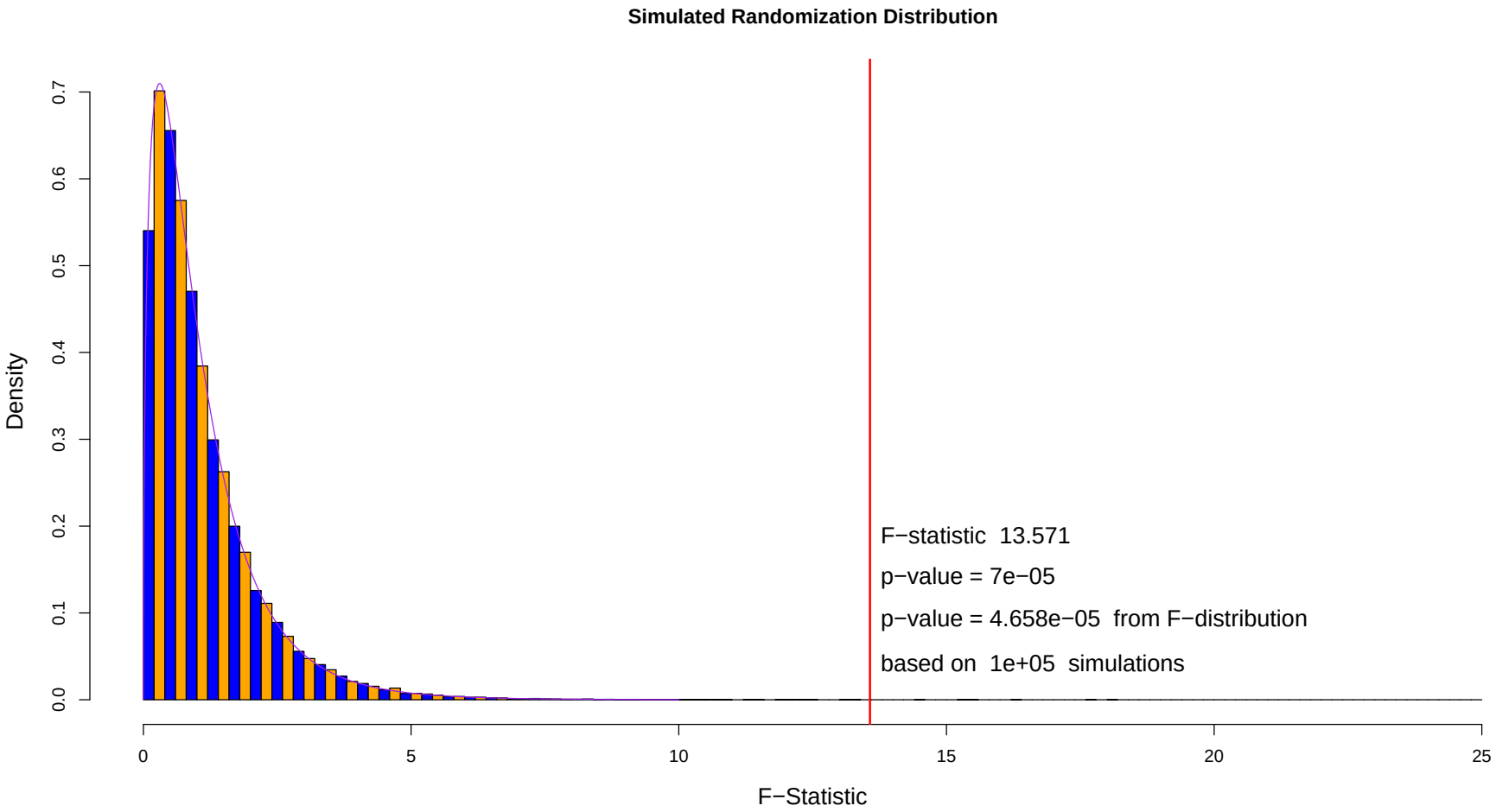
Fitted Values from `lm` for Coagulation Example

```
> 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 $\bar{Y}_{i\cdot}$

- When $H_0 : \mu_1 = \ldots = \mu_t$ is not rejected at level α, there is little purpose looking closer at differences between the $\bar{Y}_{i\cdot}$ 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 would be 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 $\hat{\mu}_i = \bar{Y}_{i\bullet}$ and their confidence intervals.
- 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

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

  as our $100(1-\alpha)\%$ confidence interval for μ_i.
- This follows as before 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 earlier discussion of this issue for 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., $\hat{\sigma}_{\hat{\theta}} = \sqrt{g(\hat{\theta}, \hat{\psi})} = SE(\hat{\theta})$, is the standard error of $\hat{\theta}$.
- 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.
- Example 1: $\theta = \mu$ and $\psi = \sigma^2$ and we only use $\hat{\psi}$ in $SE(\hat{\theta})$.
- Example 2: $\theta = \sigma^2$ and there is no ψ. We only use $\hat{\theta}$ in $SE(\hat{\theta})$.

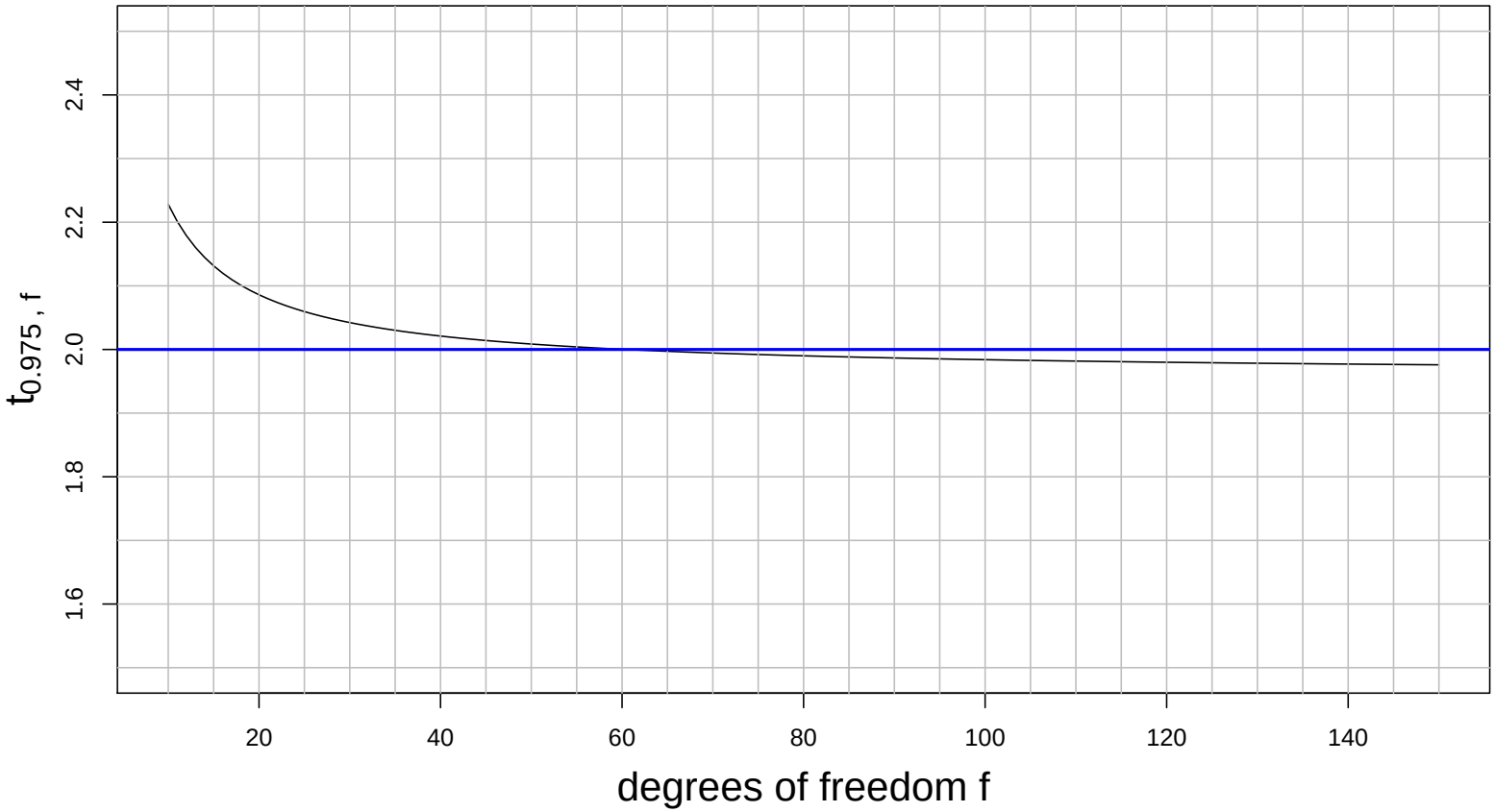
95%-Rule of Thumb Using *SE*s

- If $\hat{\theta} \approx \mathcal{N}(\theta, \sigma^2_{\hat{\theta}}) \approx \mathcal{N}(\theta, SE^2(\hat{\theta}))$, as is often the case, then $\hat{\theta} \pm 2 \times SE(\hat{\theta})$ is an approximate 95% confidence interval for θ
- This results from $z_{.975} = \texttt{qnorm}(.975) = 1.959964 \approx 2$.
- This works especially well for Student-t based intervals

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

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

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



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 $\implies$ approximate normality for s^2 since

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

$$\implies s^2 \approx \mathcal{N}\left(\sigma^2, 2\sigma^4/(n-1)\right)$$

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

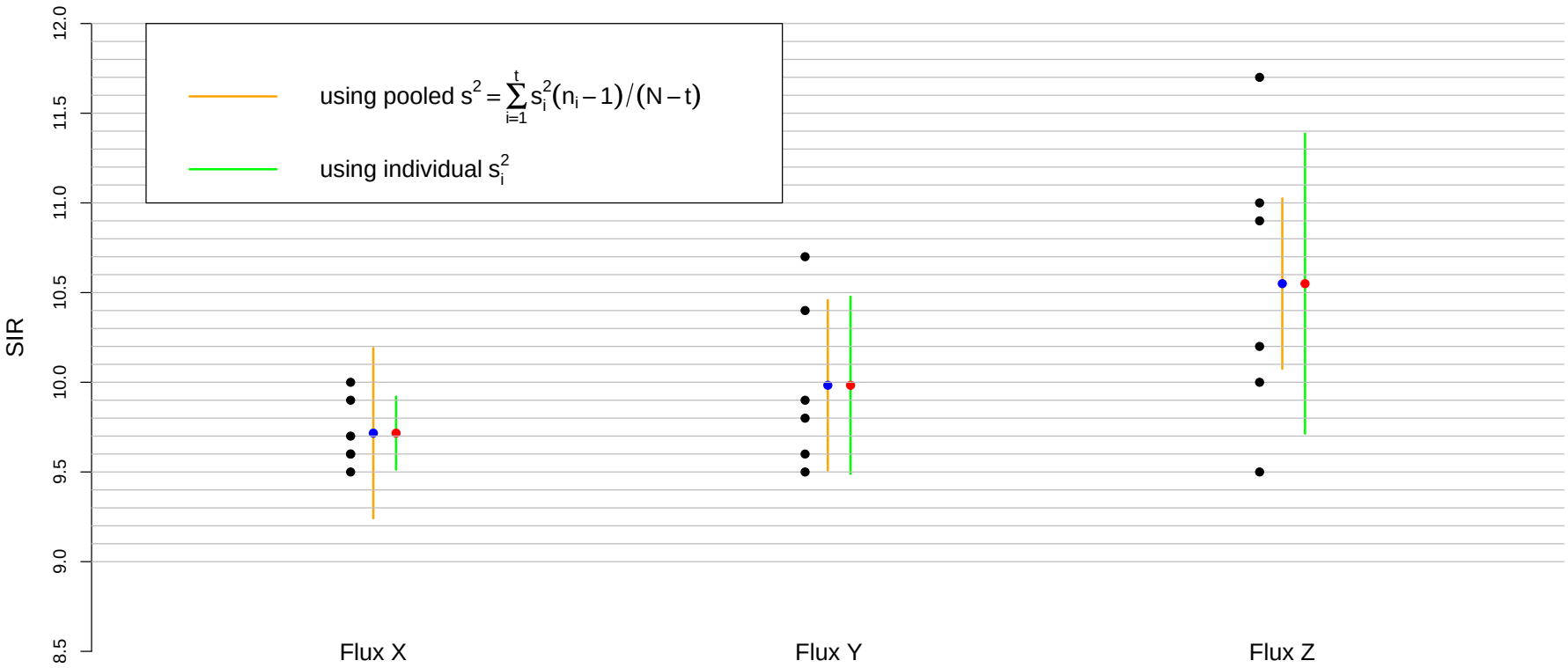
since $SE(s^2) = s^2\sqrt{2/(n-1)}$ is the estimate of $\sigma^2\sqrt{2/(n-1)}$, the sampling standard deviation of s^2.

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, 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

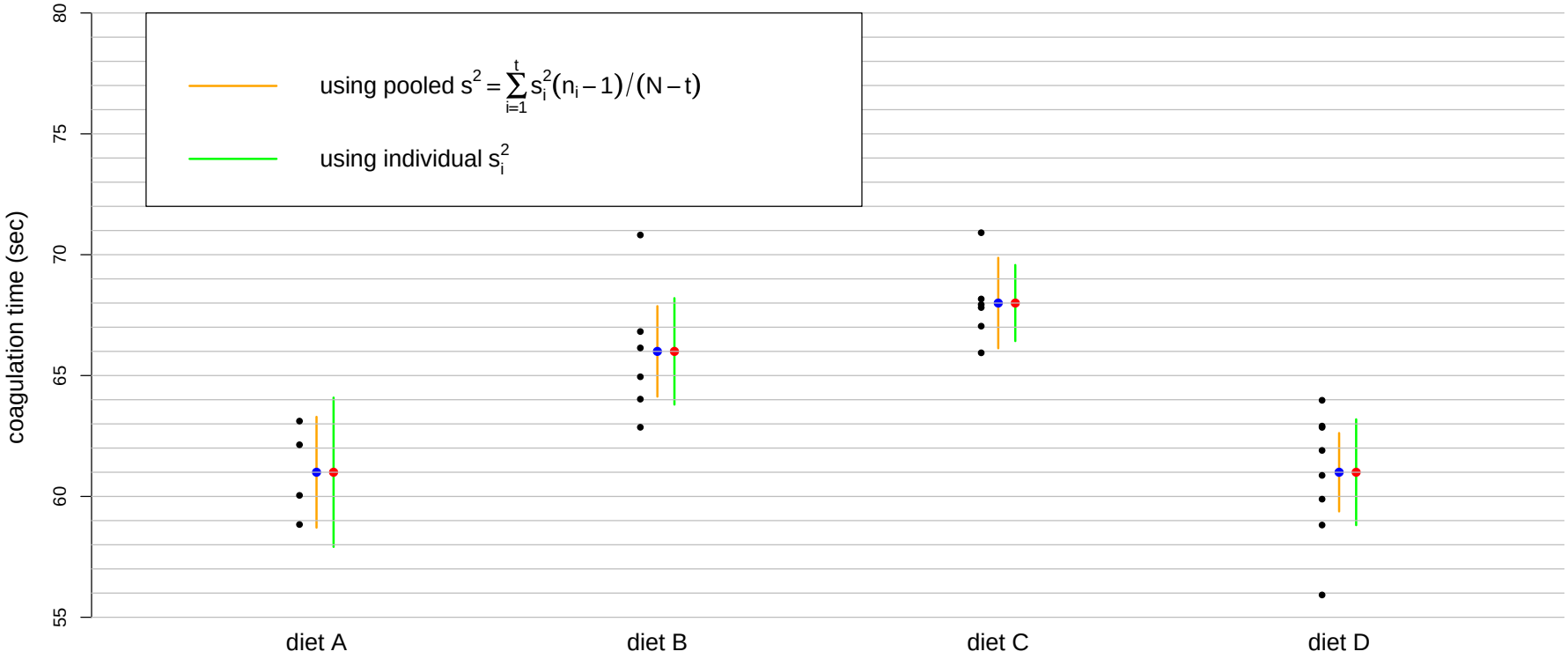


Tables of Confidence Intervals for the Coagulation Data

- 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:

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

  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

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

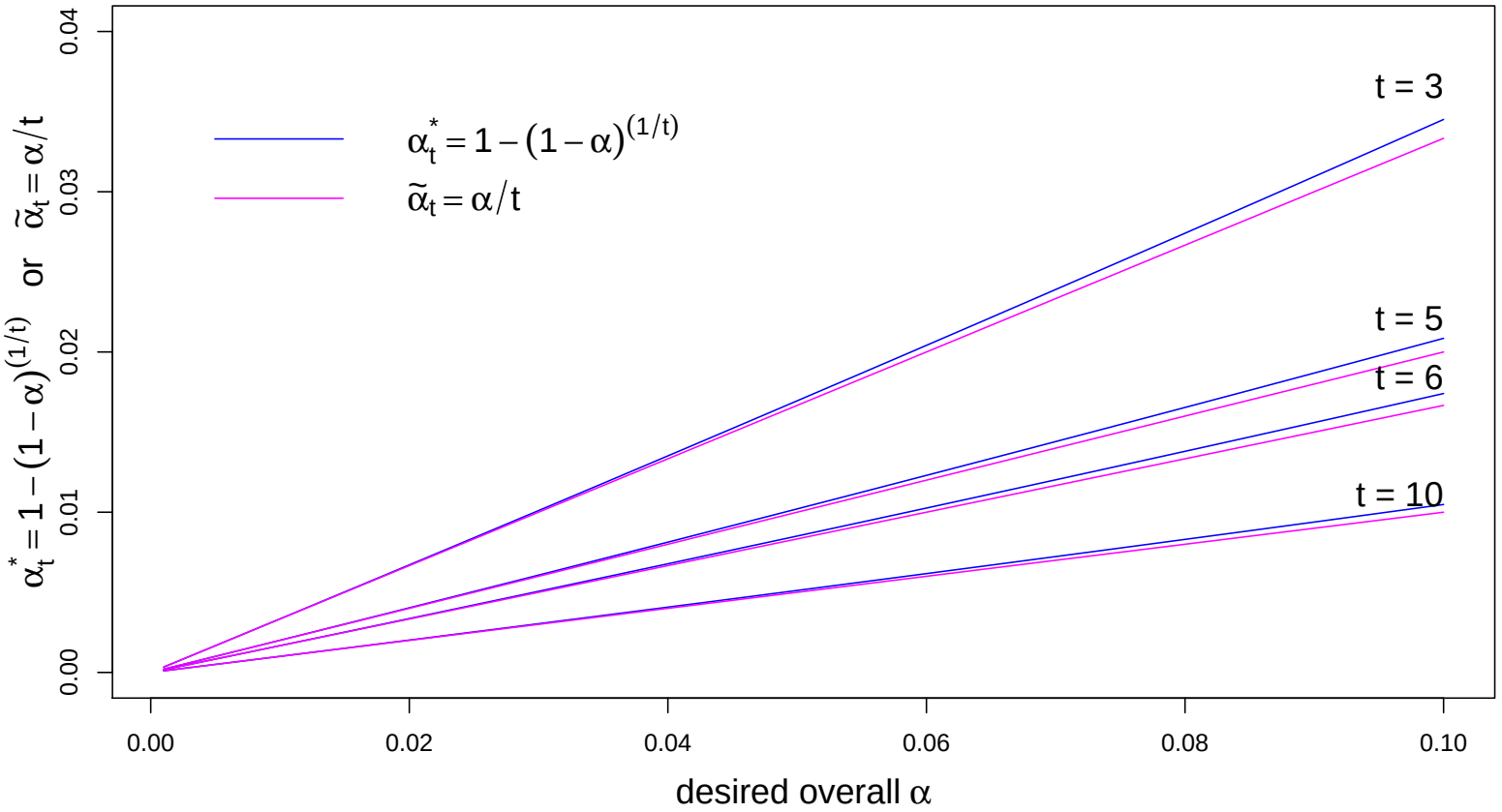
$$= \prod_{i=1}^{t} P\left(\mu_i \in \hat{\mu}_i \pm t_{n_i-1,1-\alpha/2}\frac{s_i}{\sqrt{n_i}}\right) = (1-\alpha)^t < 1-\alpha.$$

- 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 in using a pooled estimate s: No independence!

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



Dealing with Dependence from Using Pooled s

- Using a pooled estimate s for the standard deviation σ, the previous confidence intervals are no longer independent.
- However, it can be shown that

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

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

- This comes from the positive dependence between confidence intervals through s.
- If one interval is more (less) likely to cover μ_i due to s, so are the other intervals more (less) likely to cover their 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

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

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

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

- The statements are equivalent by taking complements.
- If $E_i = \left\{\mu_i \in \hat{\mu}_i \pm t_{N-t,1-\tilde{\alpha}/2}\frac{s}{\sqrt{n_i}}\right\}$ with $P(E_i) = 1 - \tilde{\alpha}$, then the simultaneous coverage probability is bounded from below

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

- 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 $\boldsymbol{\mu}$

- The μ_i variation is best understood via familiar decomposition:

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

- The two vectors on the right are orthogonal to each other.
- The first vector represents the projection of $\boldsymbol{\mu}$ onto $\mathbf{1}_N$.
- The second represents the projection of $\boldsymbol{\mu}$ onto V_{t-1}, a $(t-1)$-dimensional subspace of the $(N-1)$-dimensional orthogonal complement V_{N-1} to $\mathbf{1}_N$. See slide 51.
- The second vector captures all aspects of variation in $\boldsymbol{\mu}$.

Motivating Contrasts

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

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

$$= \sum_{i=1}^{t} a_i \mu_i - \sum_{i=1}^{t} \frac{n_i}{N} \mu_i \sum_{j=1}^{t} a_j = \sum_{i=1}^{t} c_i \mu_i$$

$$\text{with} \quad c_i = a_i - \frac{n_i}{N} \sum_{j=1}^{t} a_j$$

$$\text{where} \quad \sum_{i=1}^{t} c_i = \sum_{i=1}^{t} a_i - \sum_{i=1}^{t} \frac{n_i}{N} \sum_{j=1}^{t} a_j = \sum_{i=1}^{t} a_i - \sum_{j=1}^{t} a_j = 0 \, .$$

- 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

- Say 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 $C_{1.234} = \mu_1 - \frac{\mu_2+\mu_3+\mu_4}{3}$ with $\mathbf{c}' = (1, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3})$.
- Sometimes the first 2 treatments share something in common and so do the last 2. One might then try: $C_{12.34} = \frac{\mu_1+\mu_2}{2} - \frac{\mu_3+\mu_4}{2}$ with $\mathbf{c}' = (\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{1}{2})$ difference of average treatment effect between between the 2 camps.

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

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

$$\Longrightarrow \quad \hat{C} \pm t_{N-t,1-\alpha/2} \cdot s \cdot \sqrt{\sum_{i=1}^{t} c_i^2 / n_i}$$

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

$$\hat{C} \pm t_{N-t,1-\alpha/2} \cdot SE(\hat{C}) \quad \text{where} \quad SE(\hat{C}) = s \cdot \sqrt{\sum_{i=1}^{t} c_i^2 / n_i} \; .$$

- $SE(\hat{C})$ is the standard error of $\hat{C}$.

Simultaneous Confidence Intervals for Contrasts

- 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 interval.
- We will introduce/compare several such procedures, although there are still others.
- Multiple comparisons is a very active research area.

Simultaneous Statistical Inference by Miller (1966)
Multiple Comparison Procedures by Hochberg and Tamhane (1987)
Multiple Comparisons: Theory and Methods by Hsu (1996)
Multiple Comparisons and Multiple Tests by Westfall (2000)
Multiple Comparisons Using R by Bretz, Hothorn, Westfall (2011)

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 $C_{ij} = \mu_i - \mu_j$:

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

  and declare all $\mu_i - \mu_j \neq 0$ for which $\hat{I}_{ij}$ does not contain zero.

- Here $LSD = t_{N-t,1-\alpha/2} \cdot s \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$ is the Least Significant Difference.

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

$$Q_{t,f} = \max_{1 \le i < j \le t} \left\{ \frac{|Z_i - Z_j|}{V} \right\}$$

$Z_1, \ldots, Z_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $f \cdot V^2 \sim \chi^2_f$ (independent of Z_i)

- Its cdf and quantile function are given in R as `ptukey(q,nmeans,df)` and `qtukey(p,nmeans,df)`, `nmeans` $= t$ is the number of means, `df` $= f = N - t =$ degrees of freedom in $s^2 = MS_E$, where $V^2 = s^2/\sigma^2$ above.
- 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.

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

Tukey-Kramer Method: Unequal Sample Sizes

- The simultaneous intervals for all pairwise mean differences was due to Tukey.
- 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 $1/\sqrt{n}$ by $n^{\star}_{ij}$, where $n^{\star}_{ij}$ is the harmonic mean of n_i and n_j, i.e., $1/n^{\star}_{ij} = (1/n_i + 1/n_j)/2 \quad \text{or} \quad n^{\star}_{ij} = 2n_i n_j/(n_i + n_j)$.
- 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 (continued)

```
  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 $\mu_3 - \mu_4$.
- Under H_0 the risk of declaring any significant differences $\leq \alpha$.

$$P_{H_0}\left(0 \notin \hat{\mu}_i - \hat{\mu}_j \pm q_{t,f,1-\alpha}\ s/\sqrt{n_{ij}^{\star}} \text{ for some } i < j\right) \leq \alpha$$

Bonferroni Confidence Intervals for Pairwise Contrasts

- Applying Bonferroni's method for simultaneous confidence statement, use $\tilde{\alpha} = \alpha/\binom{t}{2}$ for individual confidence statements

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

- The individual coverage probability is $1 - \tilde{\alpha}$.
- The joint coverage probability for all pairwise contrasts is

$$P(\mu_i - \mu_j \in \hat{\mu}_i - \hat{\mu}_j \pm t_{N-t,1-\tilde{\alpha}/2} \cdot s \ \forall i < j)$$

$$\geq 1 - \binom{t}{2}(1 - (1 - \tilde{\alpha})) = 1 - \binom{t}{2}\tilde{\alpha} = 1 - \alpha$$

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 $\mathbf{c}'\boldsymbol{\mu} = \sum_{i=1}^{t} c_i \mu_i$:

$$P\left(\mathbf{c}'\boldsymbol{\mu} \in \mathbf{c}'\hat{\boldsymbol{\mu}} \pm \sqrt{(t-1)\cdot F_{t-1,N-t,1-\alpha}} \cdot s \cdot \left(\sum_{i=1}^{t} c_i^2/n_i\right)^{1/2} \quad \forall \ \mathbf{c}\right)$$

$$= 1-\alpha$$

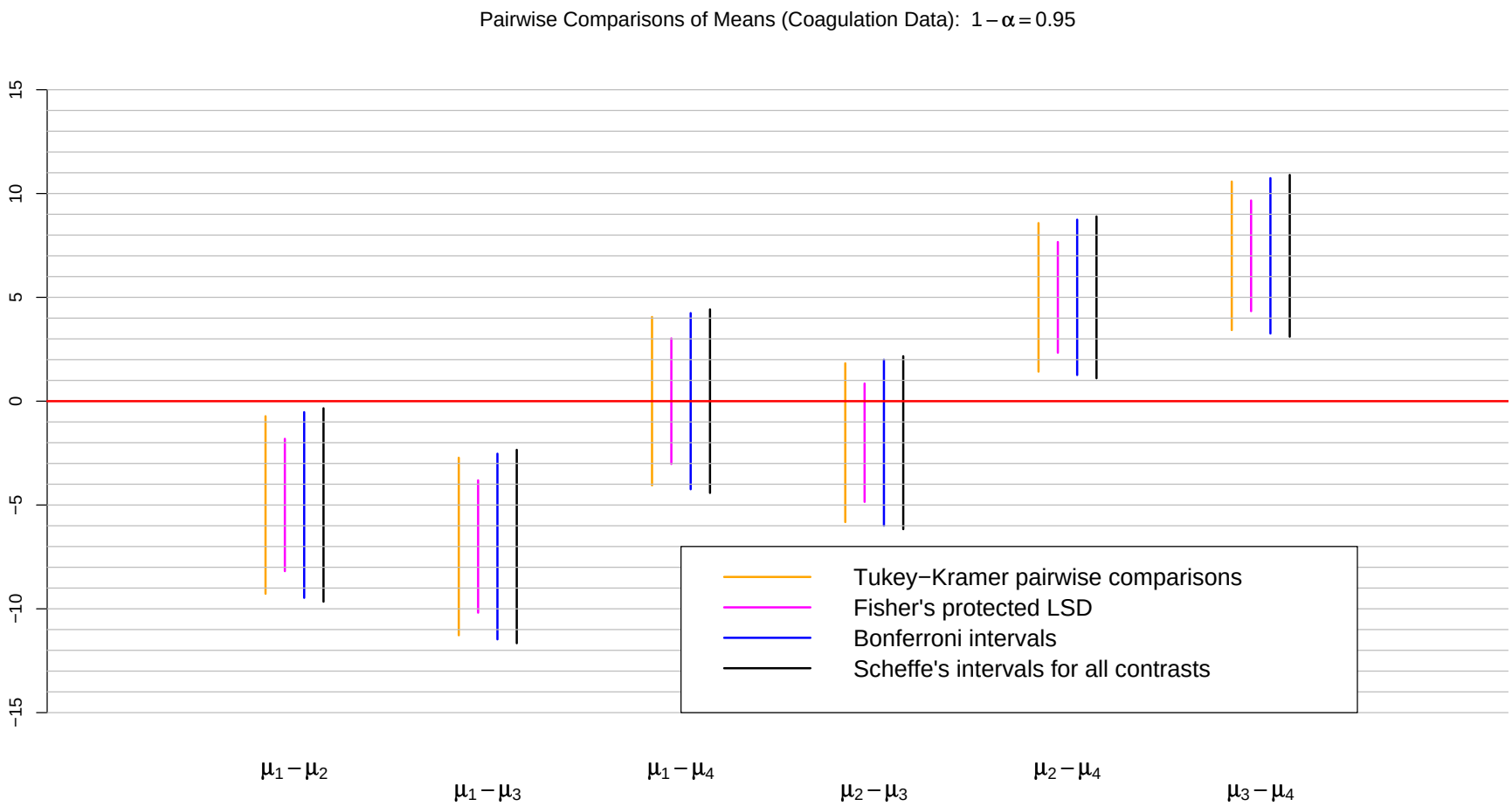
- Coverage statement for an infinite number of contrasts.
- It can be applied conservatively to all pairwise contrasts.
- The resulting intervals tend to be quite conservative.
- But it compares well with Bonferroni type intervals if applied to many contrasts.

Pairwise Comparison Intervals for Coagulation Data

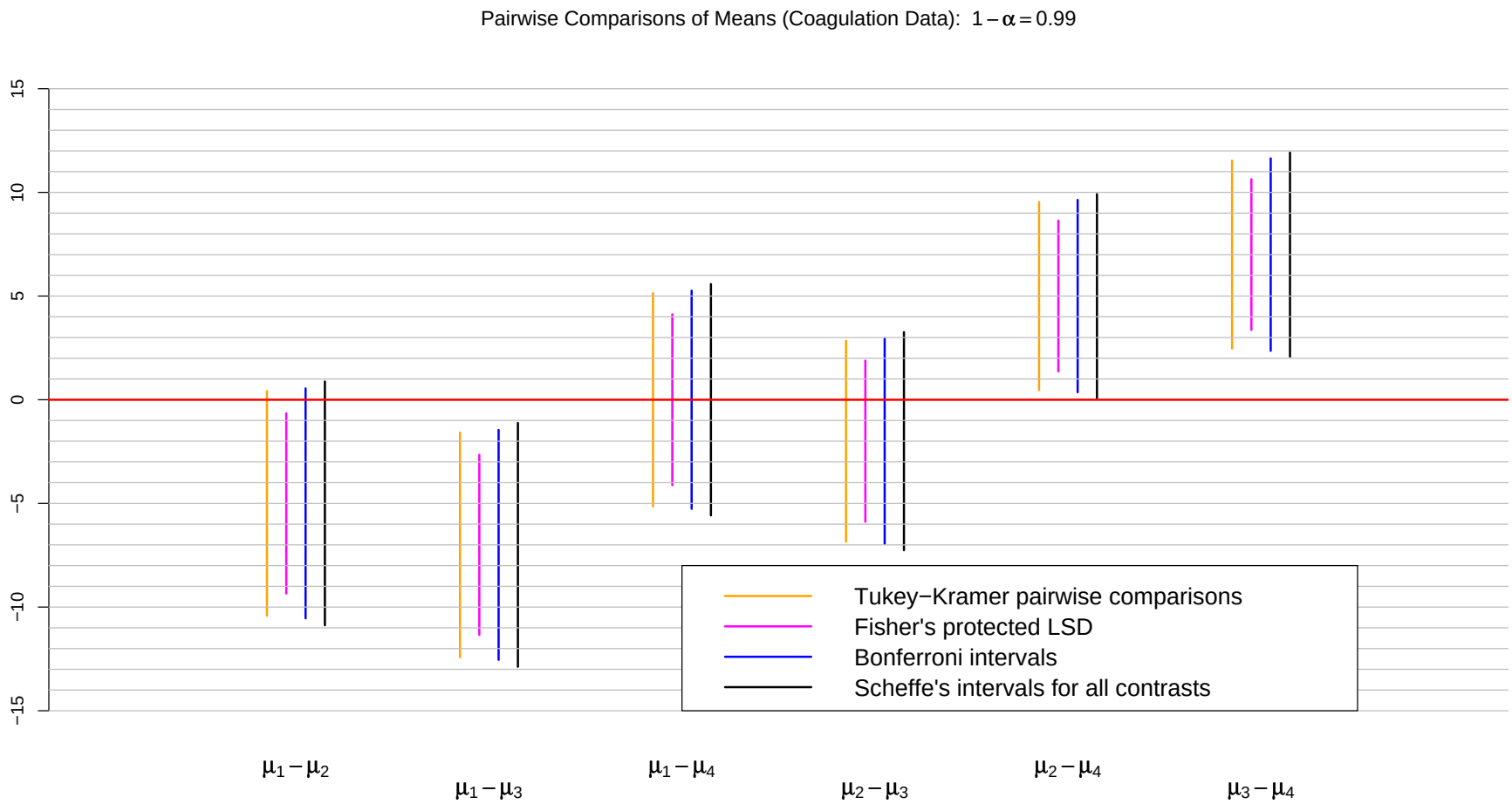
	(simultaneous) 95%-Intervals							
mean difference	Tukey-Kramer		Fisher's protected LSD		Bonferroni inequality		Scheffé's all contrasts method	
$\mu_1 - \mu_2$	-9.28	-0.72	-8.19	-1.81	-9.47	-0.53	-9.66	-0.34
$\mu_1 - \mu_3$	-11.28	-2.72	-10.19	-3.81	-11.47	-2.53	-11.66	-2.34
$\mu_1 - \mu_4$	-4.06	4.06	-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

Using any of the four methods declare significant differences in $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, $\mu_2 - \mu_4$, and $\mu_3 - \mu_4$.

Simultaneous Paired Comparisons (95%)



Simultaneous Paired Comparisons (99%)



Model Diagnostics

- **Model:** $Y_{ij} = \mu_i + \epsilon_{ij}, \quad j = 1, \ldots, n_i \ , \quad i = 1, \ldots, t.$
- We made the following assumptions:
  - A1: $\{\epsilon_{ij}\}$ are independent;
  - A2: $\mathrm{var}(\epsilon_{ij}) = \mathrm{var}(Y_{ij}) = \sigma^2$ for all i, j (homogeneity of variances or homoscedasticity);
  - A3: $\{\epsilon_{ij}\}$ are normally distributed.
- These assumption allow us to
  - i perform the F-test for homogeneity of means,
  - ii do power calculations,
  - iii plan sample sizes to achieve a desired power,
  - iv obtain simultaneous confidence intervals for means/contrasts.
- We will examine A2 and A3.
- Won't deal with A1. Use judgment, examine serial correlation?

Checking Normality

- Here we would like to check normality of $\epsilon_{ij} = Y_{ij} - \mu_i,\ j = 1, \ldots, n_i,\ i = 1, \ldots, t.$
- Not knowing μ_i we estimate the error term ϵ_{ij} via $\hat{\epsilon}_{ij} = Y_{ij} - \hat{\mu}_i = Y_{ij} - \bar{Y}_{i\bullet}.$
- 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 now.

Checking Normality by Simulation

- Can adapt the KS, CvM, and AD EDF test of fit criteria and simulate their null distribution.
- Limiting results by Pierce (1978) support this.
- Use them to judge significant non-normality in 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 $U_{(1)} \leq \ldots \leq U_{(N)}$ 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).

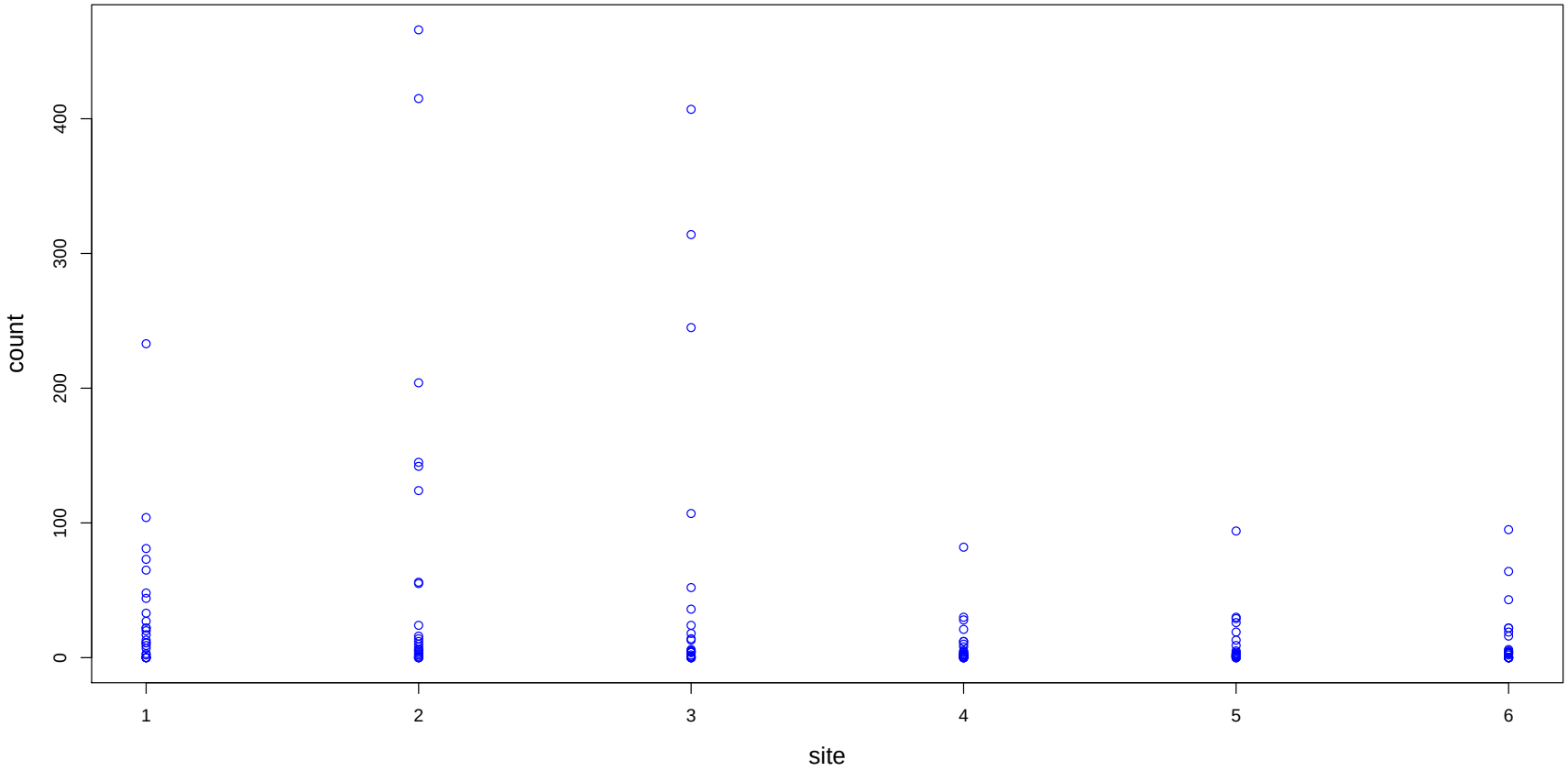
Hermit Crab Count Data

- Hermit Crab counts were obtained at 6 different coastline sites.
- Each site obtained counts at 25 randomly selected transects.
- Download the data file `crab.csv` from the web into your work directory. `crab <- read.csv("crab.csv")`.
- Count data: $\Rightarrow$ don't 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

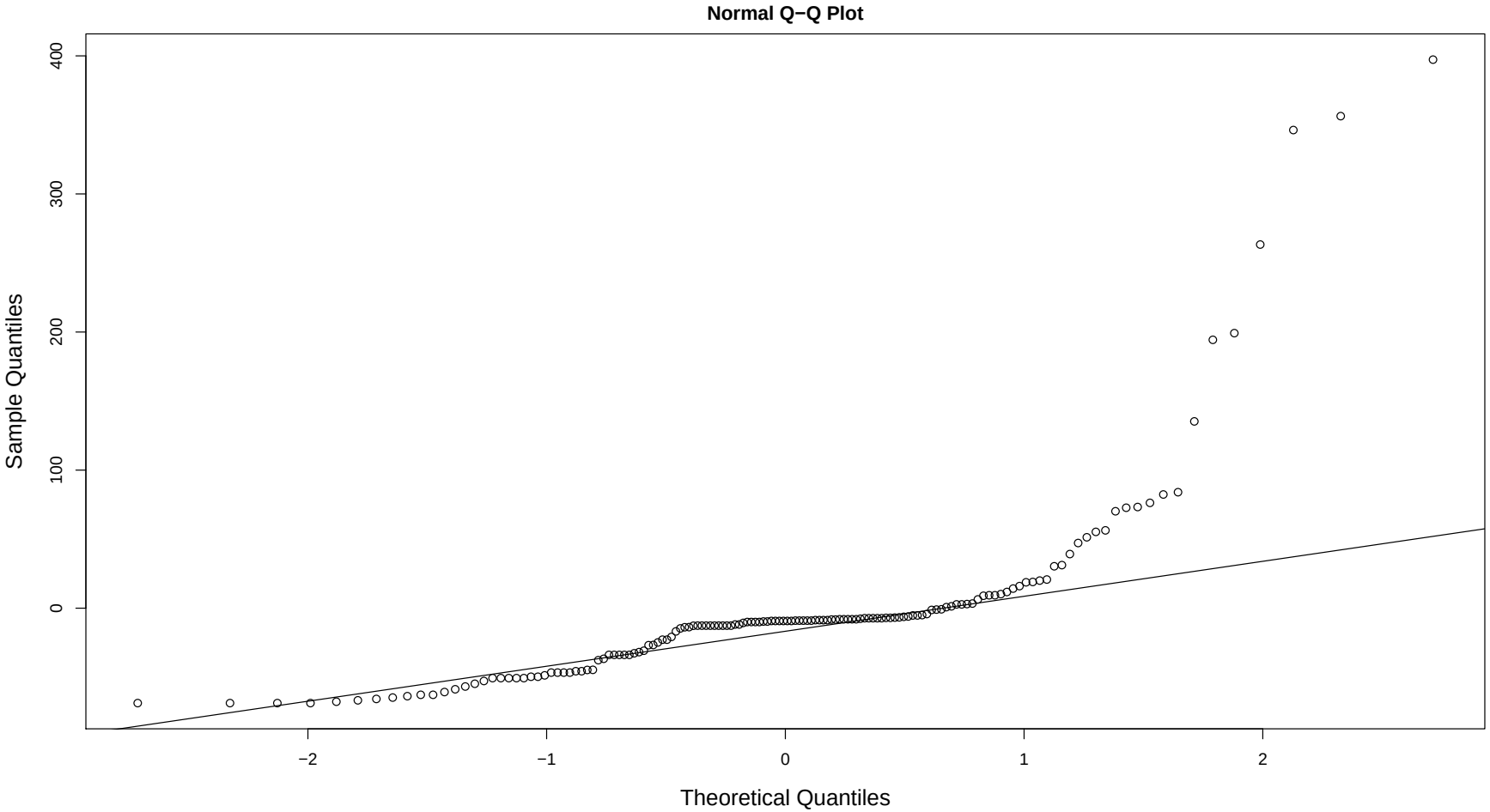


ANOVA for Hermit Crab Count Data

```
> 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



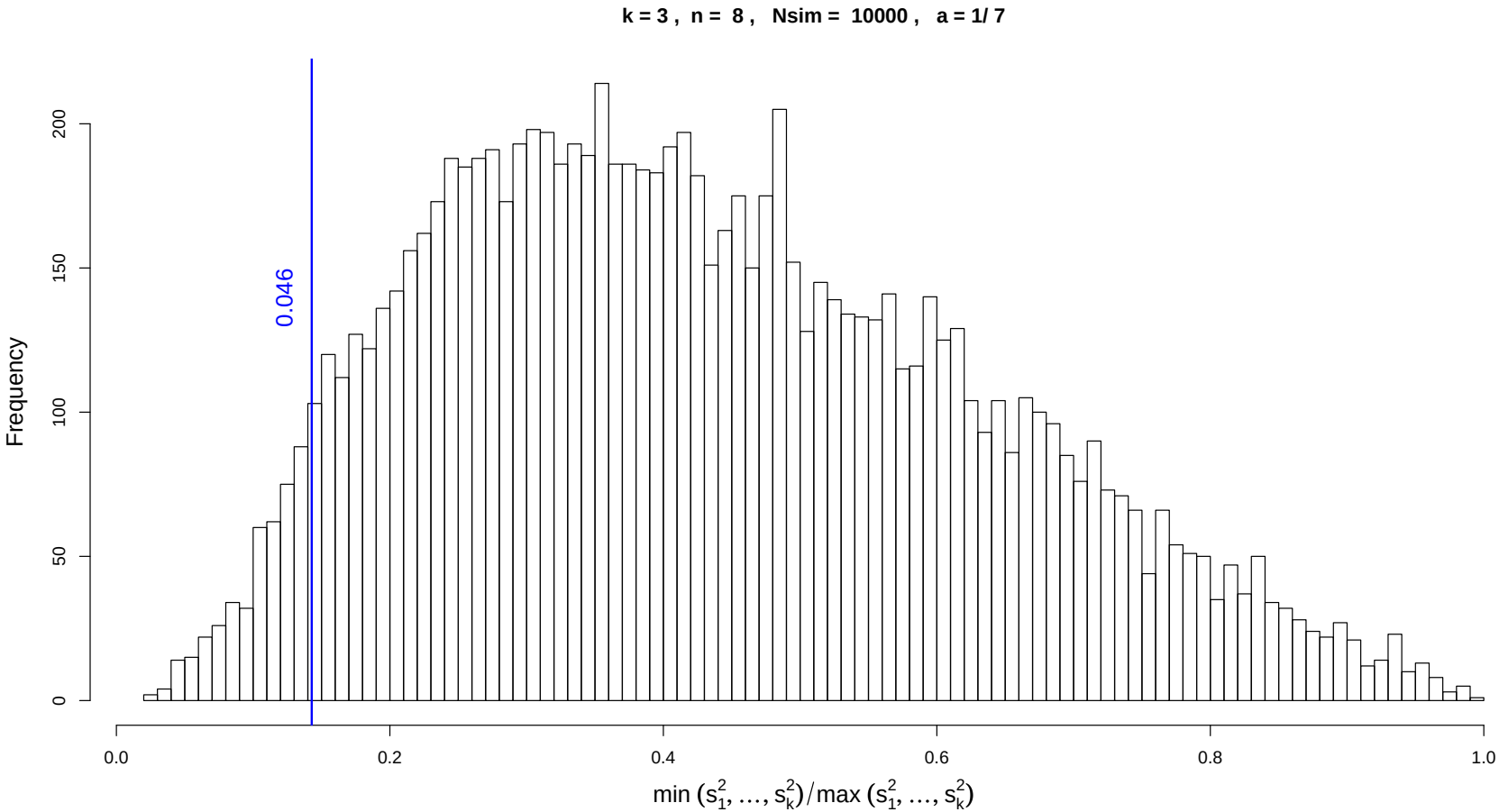
Checking for Homoscedasticity

- 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 to simulate the distribution for $F_{\min}$.

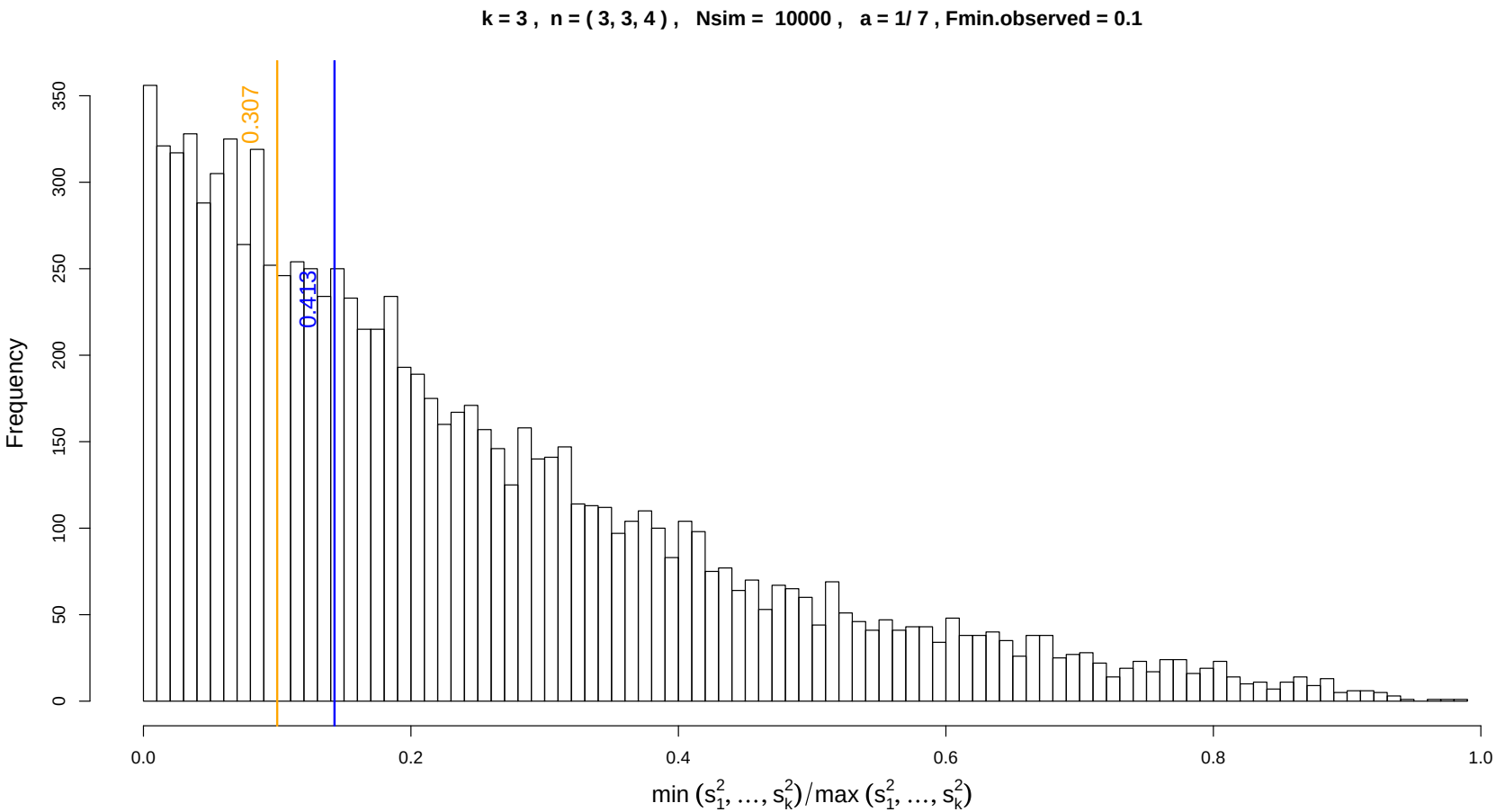
Fmin.test

- The R function `Fmin.test` $\rightarrow$ on the class web site.
- It simulates the F_{min} distribution, assuming normal samples with equal variances.
- The sample sizes may vary.
- The documentation for `Fmin.test` inside function body.
- Use it to explore any desired rule of thumb, calculating the proportion of F_{min} values $\leq$ to the rule of thumb value.
- If $F_{\text{min,observed}}$ is provided, it calculates the estimated p-value from this simulated distribution.
- See the next two slides for examples.
- The validity of this test depends strongly on data normality.

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



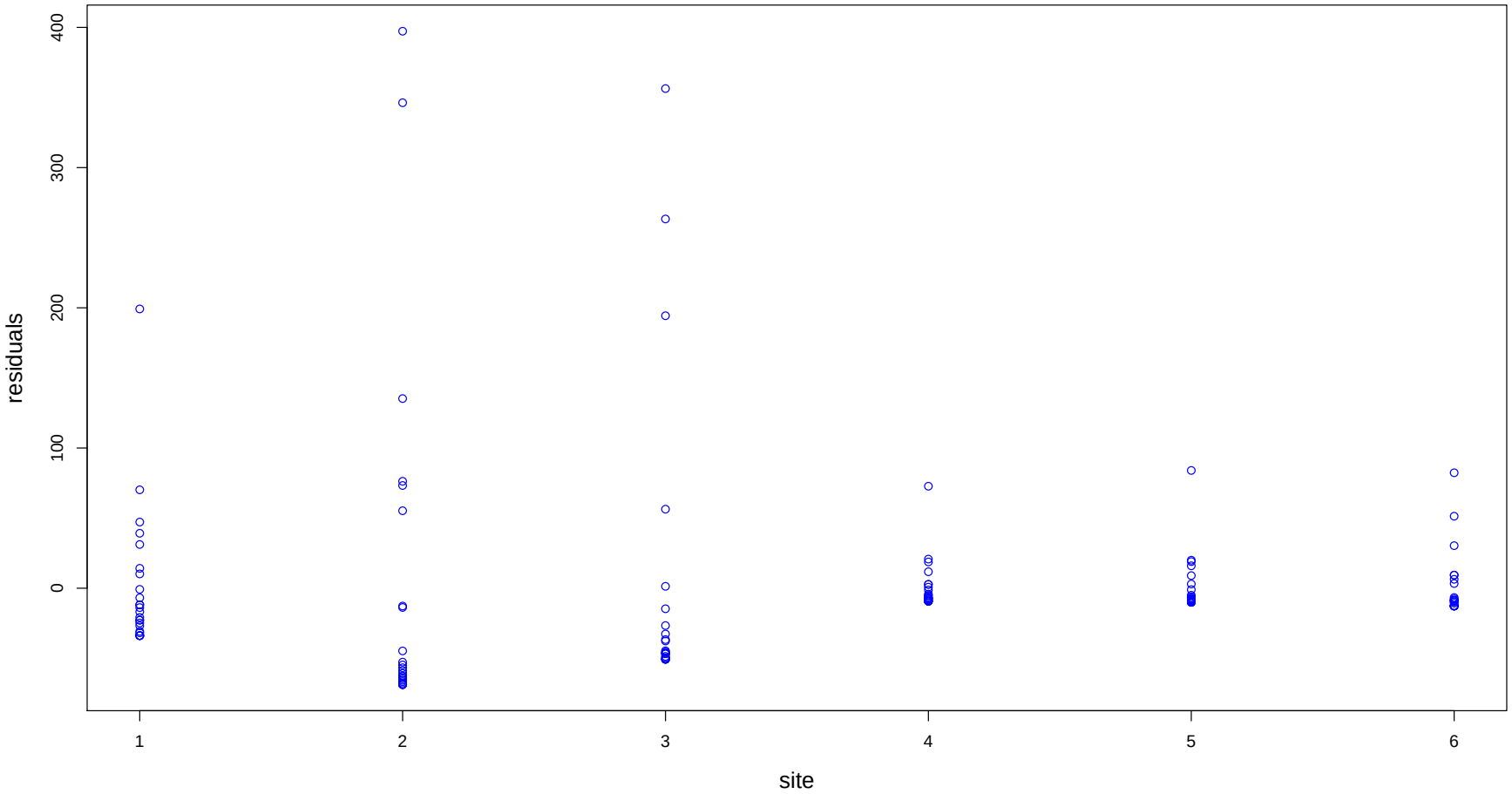
```
Fmin.test(k=3,n=c(3,3,4),a.recip=7,
                              Fmin.observed=.1)
```



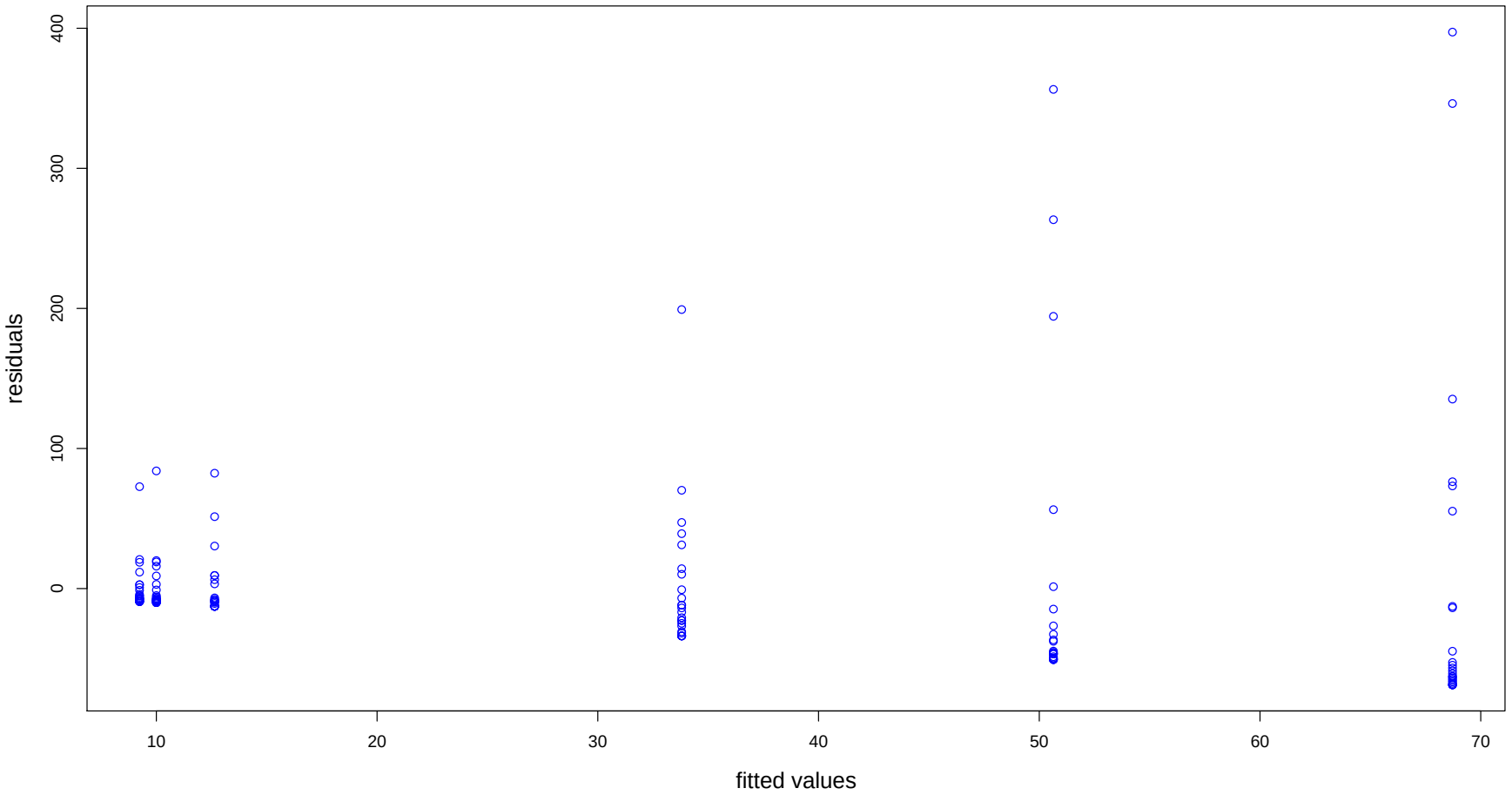
Diagnostic Plots for Checking Homoscedasticity

- One first diagnostic is to plot the residuals $Y_{ij} - \bar{Y}_{i\bullet}$ versus the corresponding fitted values $\bar{Y}_{i\bullet}$ for $j = 1, \ldots, n_i, \quad i = 1, \ldots, t$.
- Compare the difference in information in the next two plots.
- The second display: $\Rightarrow$ 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 $X_{ij} = |Y_{ij} - \tilde{Y}_i|$, where $\tilde{Y}_i$ denotes the median of the i^{th} treatment sample.
- Originally used $\bar{Y}_{i\cdot}$ 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}.
- 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

- Variability in the crab count data seemed proportional to the count averages.
- 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 size X_0.
- This group produced a random number $X_0 \cdot X_1$ of new crabs, where X_1 represents the reproduction rate per crab.
- This rate is variable or random.
- The next generation would have $X_0 \cdot X_1 \cdot X_2$ crabs, and so on.
- This motivates the following variation model: $Y = \mu \times \epsilon = \mu \cdot (X_1 \cdot X_2 \cdot \ldots)$, where the random term ϵ has mean μ_ϵ and standard deviation σ_ϵ.
- $\Rightarrow \mathrm{var}(Y) = \mu^2 \cdot \mathrm{var}(\epsilon)$ and $\mu_Y = E(Y) = \mu \cdot E(\epsilon)$.
- σ_Y is proportional to μ_Y since both are proportional to μ.

Variance Stabilization and Normality under log-Transform

- Multiplicative error model $\Longrightarrow \sigma \propto \mu$.
- Using $\log(Y) = \log(\mu) + \log(\epsilon)$ breaks the link

$$E(\log(Y)) = \log(\mu) + E(\log(\epsilon)) \text{ and } \operatorname{var}(\log(Y)) = \operatorname{var}(\log(\epsilon))$$

- μ affects the mean $E(\log(Y))$ but no longer $\operatorname{var}(\log(Y))$.
- 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 $\log(Y)$:

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

  we can appeal to the CLT for the sum of the $\log(X_i)$ terms.
- This justifies a normal additive error model for V, i.e., $V = \tilde{\mu} + \tilde{\epsilon}$ with $\tilde{\epsilon} \sim \mathcal{N}(0, \sigma^2)$.
- Apply this to the count data $\Rightarrow$ the following familiar model:

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

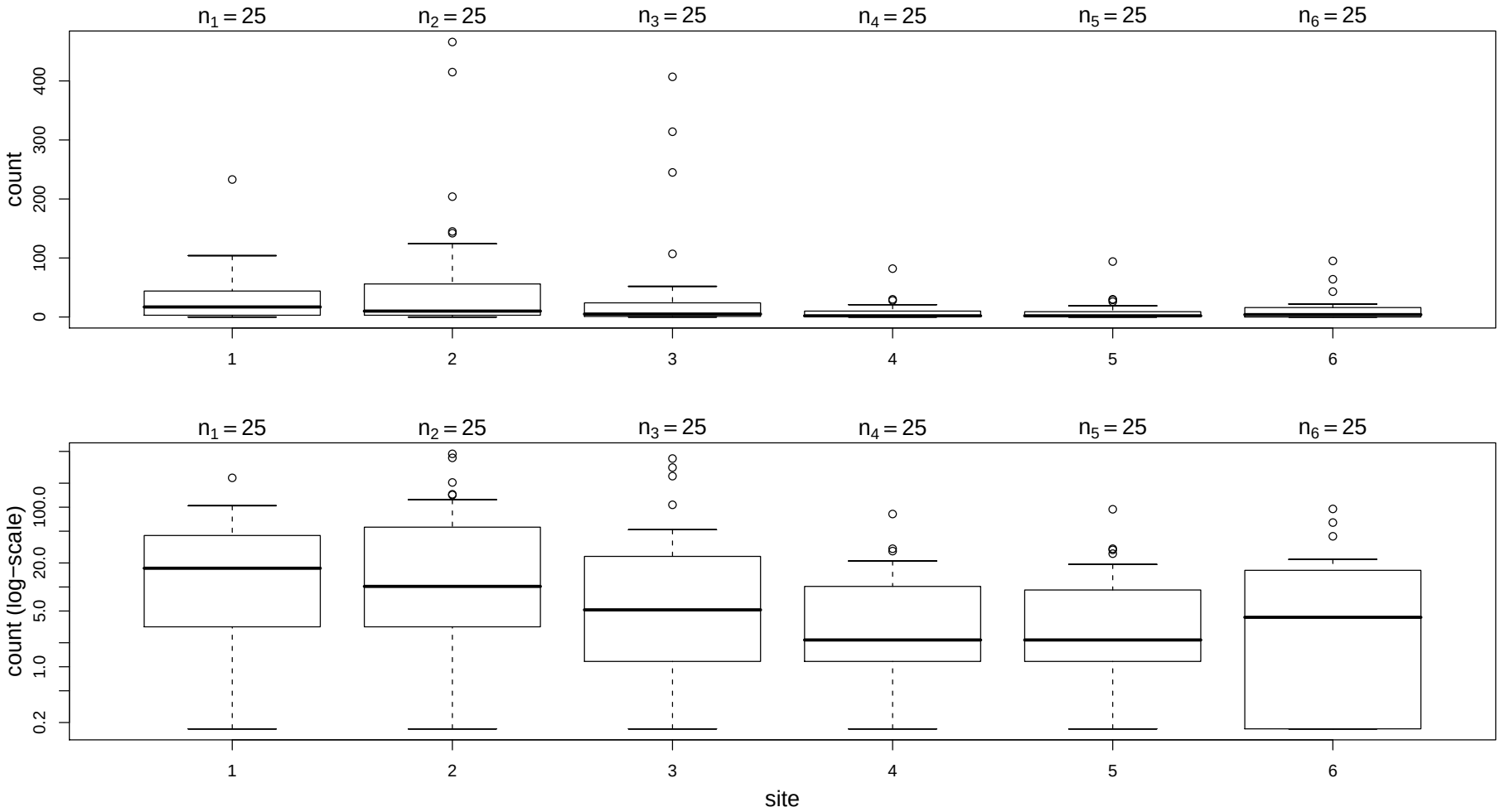
The Problem of Zero Counts

- Some observed counts are zero $\Rightarrow$ problem of $\log(0)$.
- We look at two ways of dealing with it.
  1. Adding a small fraction, say $1/6$, to all counts.

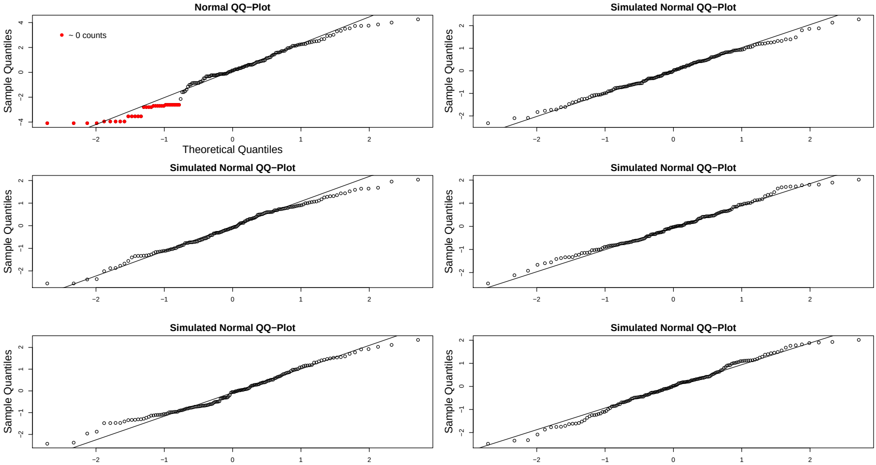
     $1/6 > 0$ is somewhat arbitrary.
     This is a technical solution, keeping all the data.
  2. Eliminate all zero counts.

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

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



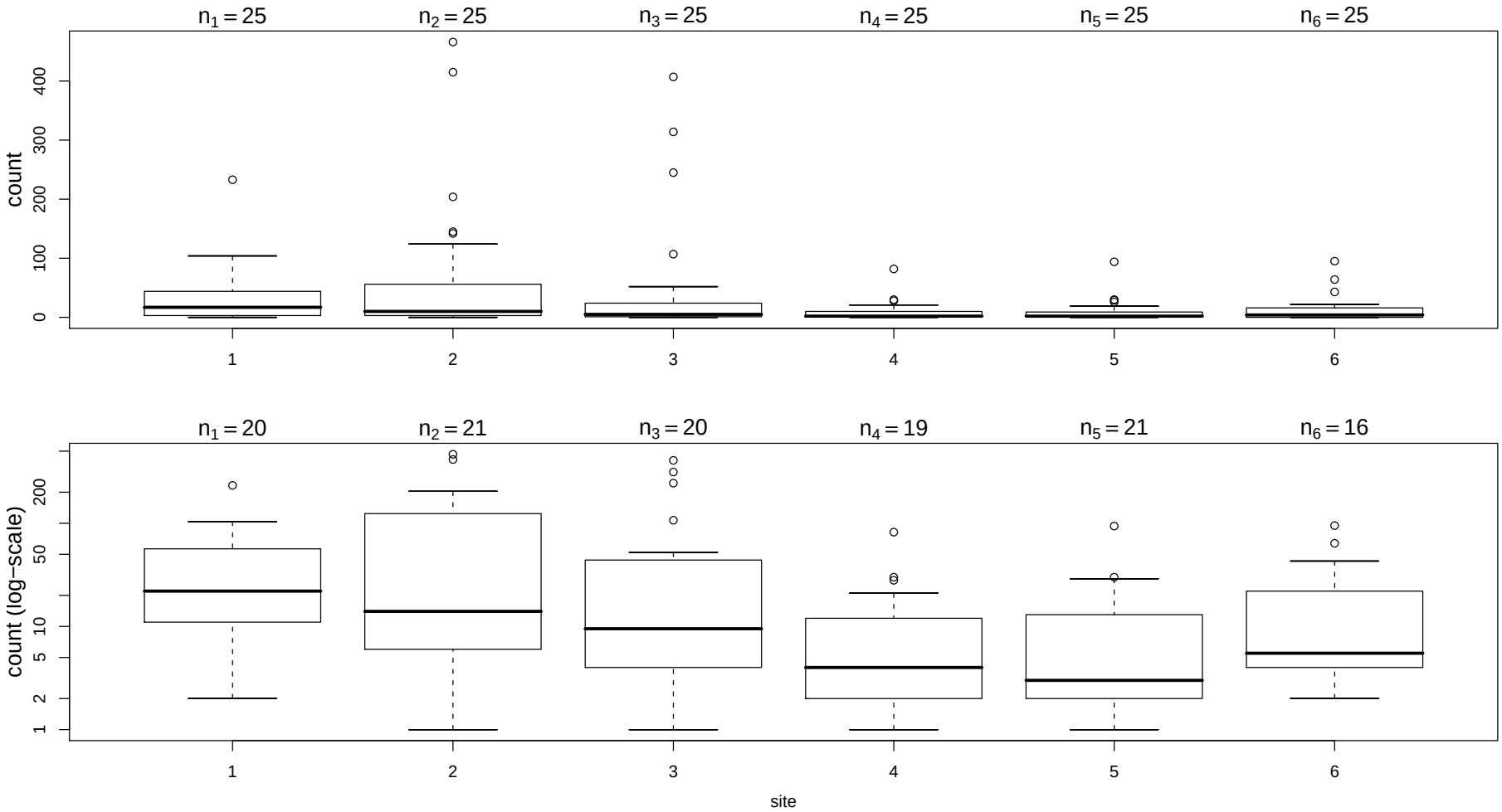
Normal QQ-Plots of 150 Residuals



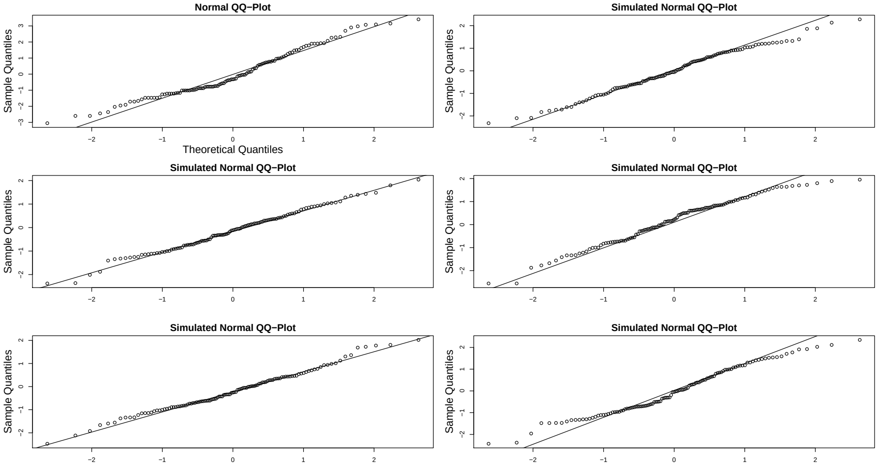
ANOVA for log(count+1/6)

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

- The `log(count[count>0])` analysis shows stronger evidence of site differences, 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 `log(count[count>0])` seem better regularized than those of `log(count+1/6)` (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

- Multiplicative error model for $Y_{ij} \Rightarrow \sigma_\mu \propto \mu$.
- log-transform had a variance stabilizing effect.
- Suppose $\sigma_\mu = k \cdot \mu^\alpha$, somewhat more general than $\sigma_\mu \propto \mu$.
- Find $V = f(Y) \Rightarrow$ the variance no longer depends on μ?
- A 1-term Taylor series expansion of f around $\mu = E(Y)$

$$f(Y) \approx f(\mu) + (Y - \mu) f'(\mu)$$

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

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

$$f'(\mu) = \frac{\tilde{c}}{\mu^\alpha} \;\text{ or }\; f(\mu) = \tilde{c} \; \frac{\mu^{1-\alpha} - 1}{1 - \alpha} + c^\star = \tilde{c} \; \frac{\exp((1-\alpha)\log(\mu)) - 1}{1 - \alpha} + c^\star$$

with $\alpha = 1 \;\Rightarrow\; f(\mu) = \log(\mu)$ as special case (L'Hospital's rule)

Finding the Variance Stabilizing Transform

- If $\sigma_\mu = k\mu^\alpha$ analyze the transformed data $\tilde{Y} = f(Y) = Y^{1-\alpha}$ when $\alpha \neq 1$ and $\tilde{Y} = \log(Y)$ when $\alpha = 1$.
- But what is the correct α? Let the data speak for themselves.

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

- Look for 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.

$$\hat{\alpha} = \texttt{lm}(\log(\texttt{s}_\texttt{i}) \sim \log(\overline{\texttt{Y}}_\texttt{i}\texttt{.}))\$\texttt{coef}[2]$$

- 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

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

- 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

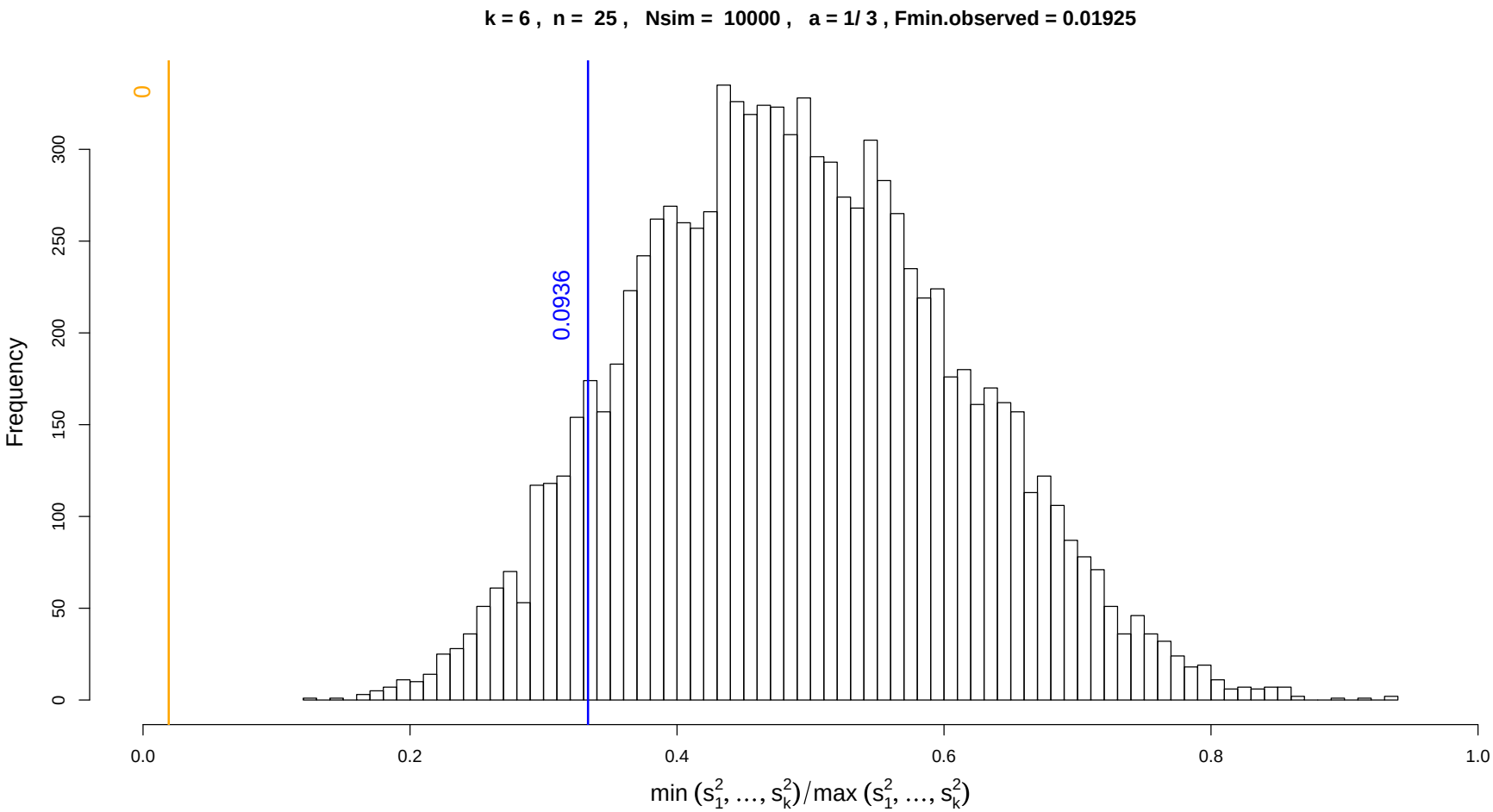
- Don't transform if $\min(s_1^2, \ldots, s_t^2)/\max(s_1^2, \ldots, s_t^2)$ is not sufficiently small $\Longrightarrow$ `Fmin.test`.
- Linear relationship $\log(s_i) \sim \log(\bar{Y}_{i\cdot})$ should be strong.
- Use simple exponents λ in the transformations, i.e., use $\lambda = 1/2$ rather than $\lambda = 1 - \alpha = .473$, as possibly calculated from slope of the linear fit of $\log(s_i) \approx \alpha \cdot \log(\bar{Y}_{i\cdot}) + b$.
- 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

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

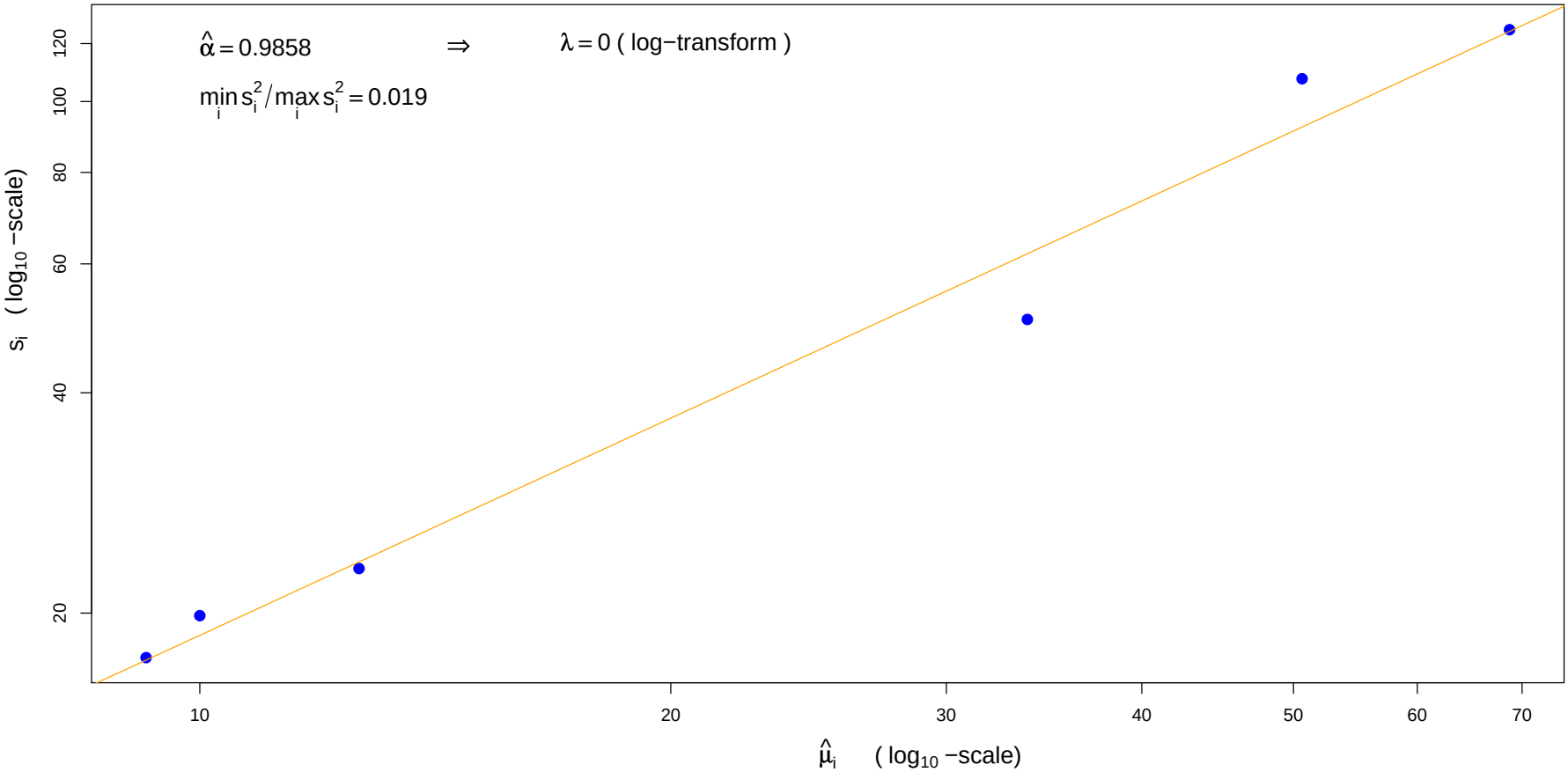
Fmin.test(k=6,n=25,a.recip=3,Fmin.observed=.01925)



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

- $Y_{i1}, \ldots, Y_{in_i} \overset{\text{i.i.d.}}{\sim} F_i$, $i = 1, \ldots, k$ be independent samples.
- Test $H_0 : F_1 = \ldots = F_k$, 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.
- We should only pay attention to their ranks $\Longrightarrow$ rank tests.
- R_{ij} be the rank of observation Y_{ij} among all N observations $Y_{11}, \ldots, Y_{kn_k}$, i.e., the smallest Y_{ij} gets rank 1, the second smallest gets rank 2, ..., and the largest gets rank N.
- For ties assign the same average rank (midrank) to all observations tied at the same value.

Kruskal-Wallis k-Sample Test

- $\bar{R}_{i\bullet} = \sum_{j=1}^{n_i} R_{ij}/n_i =$ average rank for the i^{th} sample.
- The average $\bar{R}_{\bullet\bullet}$ of all N ranks $= (N+1)/2$, midpoint between 1 and N.
- 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 $\bar{R}_{\bullet\bullet}$.

$$
\begin{aligned}
KW &= \frac{SS_{Treat}}{SS_T/(N-1)} = \frac{\sum_{i=1}^{k} n_i(\bar{R}_{i\bullet}^2 - \bar{R}_{\bullet\bullet})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(R_{ij} - \bar{R}_{\bullet\bullet})^2/(N-1)} \\
&= \frac{\sum_{i=1}^{k} n_i \bar{R}_{i\bullet}^2 - N\bar{R}_{\bullet\bullet}^2}{[\sum_{i=1}^{k}\sum_{j=1}^{n_i} R_{ij}^2 - N\bar{R}_{\bullet\bullet}^2]/(N-1)}
\end{aligned}
$$

- This suggests itself as a reasonable test statistic.

ANOVA Analogy of the Kruskal-Wallis k-Sample Test

- SS_{Treat} and SS_T suggest an ANOVA analogy.
- R_{ij} takes the place of Y_{ij}.
- The SS decomposition $SS_T = SS_{Treat} + SS_E$ still holds.

$$\frac{KW}{N-1} = \frac{SS_{Treat}}{SS_T} = \frac{SS_{Treat}}{SS_E + SS_{Treat}}$$

$$= \frac{SS_{Treat}/SS_E}{1 + SS_{Treat}/SS_E} \quad \nearrow \text{ in } \quad SS_{Treat}/SS_E$$

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

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

Null Distribution of KW

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

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

- Kruskal-Wallis showed: Under H_0 (all rankings equally likely)

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

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

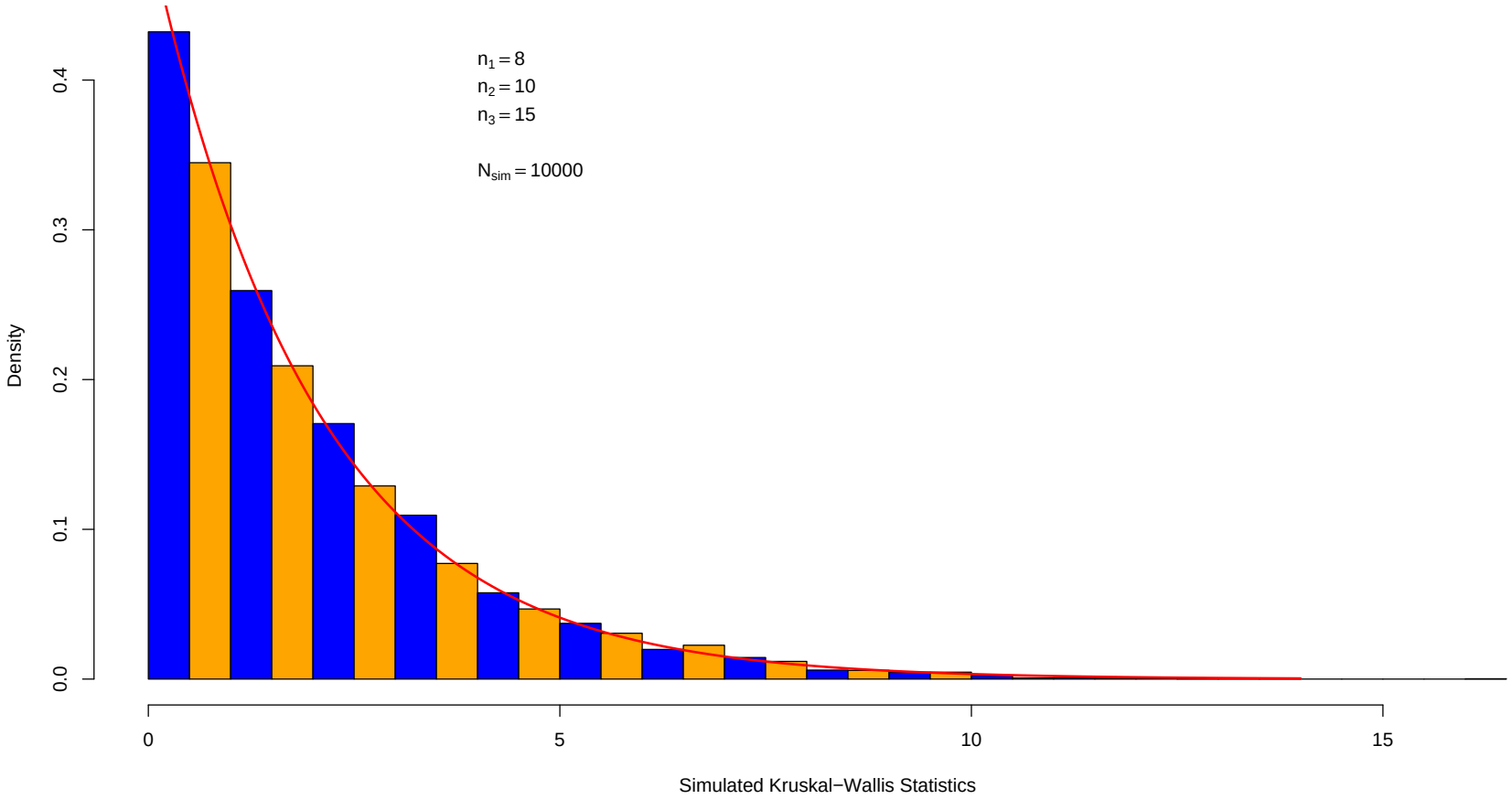
- Reject H_0 whenever
  $KW \geq \chi^2_{k-1,1-\alpha} = \texttt{qchisq}(1-\alpha, \texttt{k}-1)$.

Kruskal-Wallis Test for Flux3

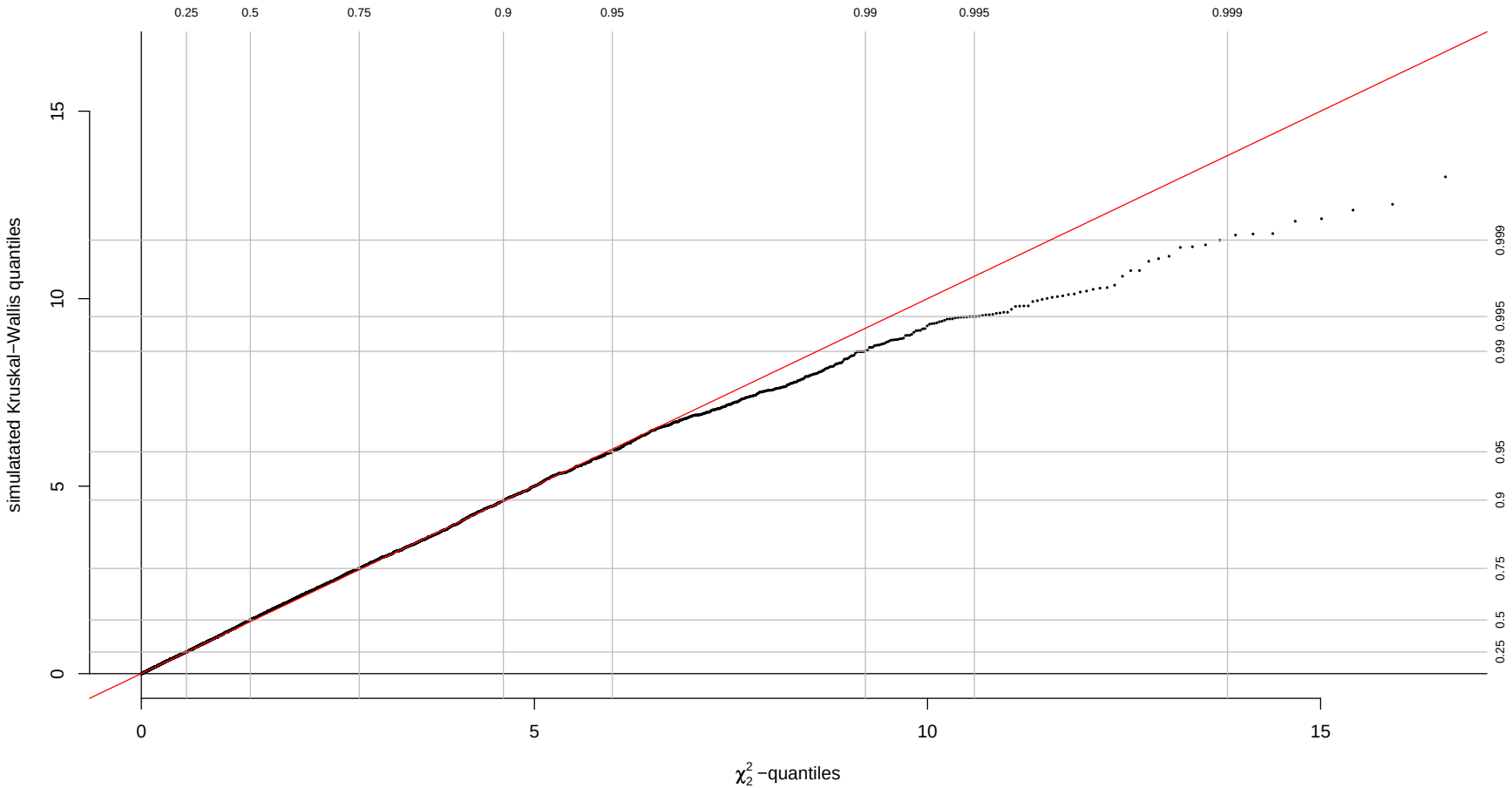
```
> kruskal.test(list(Flux3$X,Flux3$Y,Flux3$Z))

        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.
- We no longer assume normality.
- We used R_{ij} in place of the more informative Y_{ij}.
- The KW test is ineffective for changes in scale while locations are matched.
- Look at the documentation of `kruskal.test` on usage.

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 $\chi^2_{k-1} = \chi^2_2$ (exponential) 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 KW 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.
- Thus occurs in the range where the true p-value $< .05$.

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 <- numeric(nsim)
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 $\approx .1186$ from the χ^2_2 approximation.
- Agrees with 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
```

Kruskal-Wallis in Case of Ties

- Suggested using midranks, R_i^*, when observations are tied.
- The expression of KW needs to be adjusted to

$$KW^* = \frac{[12/N(N+1)]\sum R_i^{*2}/n_i - 3(N+1)}{1 - \sum(d_i^3 - d_i)/(N^3 - N)}$$

- d_i = multiplicity of the i^{th} smallest distinct observation.
- For large samples the χ^2_{k-1} approximation still applies.
- For details, see Lehmann (2006)
  Nonparametrics, Statistical Methods Based on Ranks, 2006.

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 $\hat{F}_i(x) \approx \hat{F}(x)$.
- Assess the differences $\hat{F}_i(x) - \hat{F}(x)$ across all x by the Anderson-Darling discrepancy metric

$$\begin{aligned} AD_k &= \sum_{i=1}^{k} n_i \int_B \frac{[\hat{F}_i(x) - \hat{F}(x)]^2}{\hat{F}(x)(1-\hat{F}(x))}\, d\hat{F}(x) \\ &= \sum_{i=1}^{k} \frac{n_i}{N} \sum_{r=1}^{N-1} \frac{[\hat{F}_i(Z_r) - \hat{F}(Z_r)]^2}{\hat{F}(Z_r)(1-\hat{F}(Z_r))} \quad \text{computational formula} \end{aligned}$$

- B denotes the set of all x for which $\hat{F}(x) < 1$.
- $Z_1 < \ldots < Z_N$ 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 $Y_{i\ell}, \ell = 1, \ldots, n_i, \quad i = 1, \ldots, k$ are distinct (no ties).
- From the 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 $\leq Z_r$, i.e., only the rank of the $Y_{i\ell}$ matters in such comparisons, since

$$Y_{i\ell} \leq Z_r \iff \text{rank}(Y_{i\ell}) \leq \text{rank}(Z_r) = r \iff R_{i\ell} \leq r$$

- The argument stays the same in the case of ties.
- For details on the approximate null distribution of AD_k see Scholz and Stephens (1987).
- To use `ad.test` install package kSamples, see `?ad.test`
- Invoke `library(kSamples)` for each new R session.

Anderson-Darling Test for Flux3

```
> ad.test(Flux3$X,Flux3$Y,Flux3$Z)


 Anderson-Darling k-sample test.

Number of samples:  3
Sample sizes:  6, 6, 6
Number of ties: 6

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

T.AD = ( Anderson-Darling  Criterion - mean)/sigma

Null Hypothesis: All samples come from a common population.

                AD   T.AD  asympt. P-value
version 1: 3.1565 1.2249          0.10822
version 2: 3.0600 1.1251          0.12051
```

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.
- Not the case for KW-test (immune to 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 $\to 1$ for any alternative to H_0. Not always true for the KW-test.
- The AD-test is often used to justify the pooling of data, when H_0 is not rejected. Original motivation.
- The AD-test emphasizes discrepancies in the sample tails.

Appendix A: Distribution of SS_{Treat}

The next two slides establish the noncentral $\chi^2_{t-1,\lambda}$ distribution for SS_{Treat}/σ^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}

$$\bar{Y}_{i\cdot} \sim \mathcal{N}(\mu_i, \sigma^2/n_i) \;\Rightarrow\; \sqrt{n_i}\;\bar{Y}_{i\cdot} = \sqrt{n_i}\;\mu_i + \sigma Z_i = \tilde{\mu}_i + \sigma Z_i$$

with $\boldsymbol{Z} = (Z_1, \ldots, Z_t)'$ $i.i.d.$ $\sim \mathcal{N}(0,1)$ and $\tilde{\mu}_i = \sqrt{n_i}\;\mu_i$.

Via Gram-Schmidt get an orthonormal basis $\mathbf{g}_1, \ldots, \mathbf{g}_t$ with
$\mathbf{g}_1' = (\sqrt{n_1/N}, \ldots, \sqrt{n_t/N}) \;\Rightarrow\; \mathbf{g}_1'\mathbf{g}_1 = 1.$

Let $G = (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_t)$, $\boldsymbol{a}' = (\sqrt{n_1}\;\bar{Y}_{1\cdot}, \ldots, \sqrt{n_t}\;\bar{Y}_{t\cdot}) = \tilde{\boldsymbol{\mu}}' + \sigma \boldsymbol{Z}'$,
$\boldsymbol{V}' = (V_1, \ldots, V_t) = \boldsymbol{a}'G$ or $\boldsymbol{V} = G'\boldsymbol{a} = G'\tilde{\boldsymbol{\mu}} + \sigma G'\boldsymbol{Z} = \boldsymbol{\nu} + \sigma \boldsymbol{U}$.

As shown previously, we have $\boldsymbol{U}' = (U_1, \ldots, U_t)$ $i.i.d.$ $\sim \mathcal{N}(0,1)$.

$$\Rightarrow\; V_1 = \boldsymbol{g}_1'\boldsymbol{a} = \sqrt{N}\bar{Y}_{\cdot\cdot},\;\; \nu_1 = \boldsymbol{g}_1'\tilde{\boldsymbol{\mu}} = \sqrt{N}\bar{\mu},\;\; |\boldsymbol{a}|^2 = \sum_{i=1}^{t} a_i^2 = \sum_{i=1}^{t} V_i^2 = |\boldsymbol{V}|^2$$

$$SS_{Treat} = \sum_{i=1}^{t} n_i Y_{i\cdot}^2 - N\bar{Y}_{\cdot\cdot}^2 = \sum_{i=1}^{t} a_i^2 - V_1^2 = \sum_{i=2}^{2} V_i^2$$

Distribution of SS_{Treat}

$$\begin{aligned}
\nu_1 &= \sum_{i=1}^{t} \sqrt{n_i}\mu_i\sqrt{n_i/N} = \sum_{i=1}^{t} n_i\mu_i/\sqrt{N} = \sqrt{N}\bar{\mu} \\
\sum_{i=2}^{t} \nu_i^2 &= \sum_{i=1}^{t} n_i\mu_i^2 - \nu_1^2 = \sum_{i=1}^{t} n_i\mu_i^2 - N\bar{\mu}^2 = \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu})^2 \\
SS_{Treat}/\sigma^2 &= \sum_{i=2}^{t} V_i^2/\sigma^2 = \sum_{i=2}^{t}(U_i + \nu_i/\sigma)^2 \sim \chi^2_{t-1,\lambda} \\
\text{with} \quad \lambda &= \sum_{i=2}^{t} \nu_i^2/\sigma^2 = \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu})^2/\sigma^2
\end{aligned}$$

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

Proof:

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

It suffices to show $H(a) \searrow$ for $a \geq 0$. Only the case $x > 0$ matters.

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

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

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 $\beta(\lambda)$ is 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

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

by previous theorem with $f(x) = \varphi(x)$, density of $\mathcal{N}(0,1)$.

Here $g(y)$ is the density of

$Y = F_{\text{crit}} \sum_{j=1}^{N-t} \tilde{Z}_j^2 \, (t-1)/(N-t) - \sum_{i=2}^{t-1} Z_i^2$.