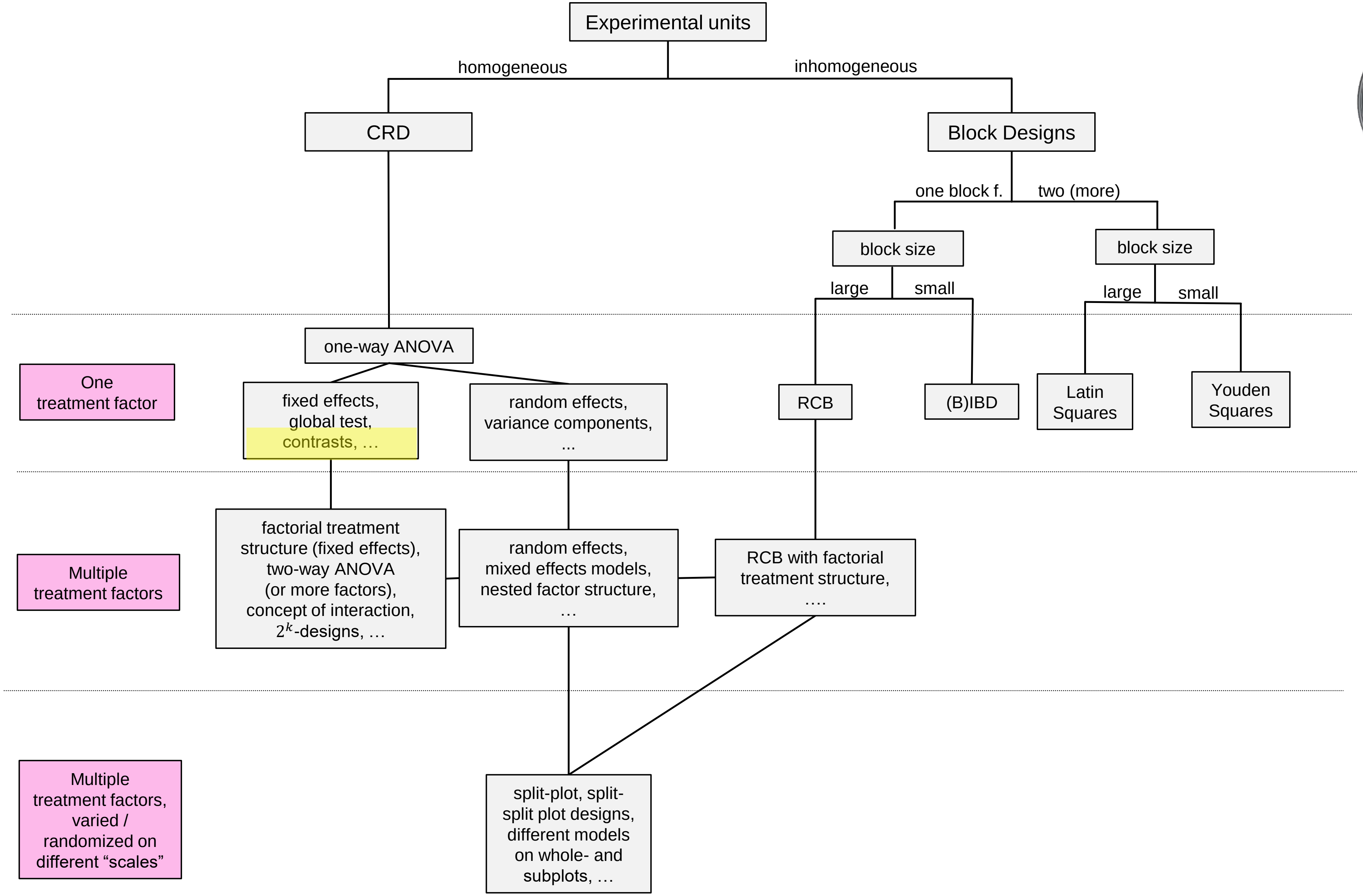
Specific Differences (Contrasts) and Multiple Testing



Experimental units

homogeneous

inhomogeneous

CRD

Block Designs

one block f.

two (more)

block size

block size

large

small

large

small

One treatment factor

one-way ANOVA

fixed effects, global test, contrasts, …

random effects, variance components, ...

RCB

(B)IBD

Latin Squares

Youden Squares

Multiple treatment factors

factorial treatment structure (fixed effects), two-way ANOVA (or more factors), concept of interaction, 2^k-designs, …

random effects, mixed effects models, nested factor structure, …

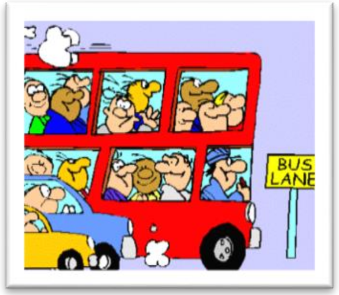
RCB with factorial treatment structure, ….

Multiple treatment factors, varied / randomized on different "scales"

split-plot, split-split plot designs, different models on whole- and subplots, …

Similar to Lawson (2015)

Problem with Global F-test



- Problem: Global F-test (aka **omnibus F-test**) is very **unspecific.**
- Typically: Want a **more precise answer** (or have a **more specific question**) on **how** the group means differ.
- Examples
  - Compare all new treatments with **control** treatment (reference treatment).
  - Do pairwise comparisons between **all** treatments.
  - ….
- A specific question can typically be formulated as an appropriate **contrast**.

Contrasts: Simple Example

- Want to compare group 2 with group 1 (don't care about the remaining groups for the moment).

- $H_0\colon \mu_1 = \mu_2$ vs. $H_A\colon \mu_1 \neq \mu_2$.

- Equivalently: $H_0\colon \mu_1 - \mu_2 = 0$ vs. $H_A\colon \mu_1 - \mu_2 \neq 0$.

- The corresponding contrast would be $c = (1, -1, 0, 0, \dots, 0)$.

- A **contrast** $c \in \mathbb{R}^g$ is a **vector** that **encodes** the **null hypothesis** in the sense that

$$H_0\colon \sum_{i=1}^{g} c_i \cdot \mu_i = 0$$

- A **contrast** is nothing else than an **encoding of your research question**.

Contrasts: Formal Definition

- Formally, a **contrast** is a g-dimensional **vector**

$$c = (c_1, c_2, \dots, c_g) \in \mathbb{R}^g$$

  with the **constraint** that $\sum_{i=1}^{g} c_i = 0$.

- The constraint reads: “contrast coefficients add to zero”.
- The side constraint ensures that the contrast is about **differences** between group means and **not** about the **overall** level of our response.
- Mathematically speaking, c is **orthogonal** to $(1, 1, \dots, 1)$ or $(1/g, 1/g, \dots, 1/g)$ which is the **overall mean**.
- This means: contrasts do not care about the overall mean, just about differences between groups.

More Examples using Meat Storage Data

- Treatments were:
  1) Commercial plastic wrap (ambient air)
  2) Vacuum package
  3) 1% CO, 40% O_2, 59% N
  4) 100% CO_2

Current techniques (control groups)

New techniques

- Possible questions and their corresponding contrasts

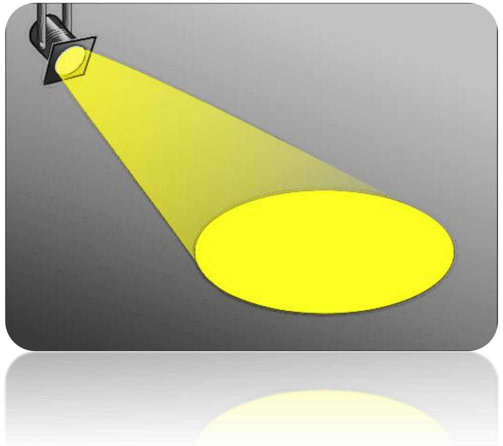
Comparison	Corresponding contrast $c \in \mathbb{R}^4$
New vs. Old	$\left(-\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$
New vs. Vacuum	$\left(0, -1, \frac{1}{2}, \frac{1}{2}\right)$
CO_2 vs. Mixed	$(0, 0, -1, 1)$
Mixed vs. Commercial	$(-1, 0, 1, 0)$

Global F-Test vs. Contrasts

As explained in Oehlert (2010):

- "ANOVA is like background lighting that **dimly illuminates the data** but not giving enough light to see details."
- "A contrast is like using a **spotlight**; it enables us to **focus in on a specific**, **narrow feature** of the data […] but it does **not** give the overall picture."
- Intuitively: "By using several contrasts we can move our focus around and see more features of the data."

vs.



Estimation and Inference for Contrasts

- We estimate the value

$$\sum_{i=1}^{g} c_i \cdot \mu_i$$

  with

$$\sum_{i=1}^{g} c_i \cdot \bar{y}_{i\cdot}$$

  i.e. we simply replace μ_i by its estimate $\hat{\mu}_i = \bar{y}_{i\cdot}$

- The corresponding standard error can be easily derived.
- This information allows us to construct **tests** and **confidence intervals**.
- See blackboard for details.

Sum of Squares of a Contrast

- We can also compute an **associated sum of squares**

$$SS_c = \frac{\left(\sum_{i=1}^{g} c_i \, \bar{y}_{i\cdot}\right)^2}{\sum_{i=1}^{g} \frac{c_i^2}{n_i}}$$

  having **one** degree of freedom, hence $MS_c = SS_c$.

- This looks unintuitive at first sight but it is nothing else than the **square** of the t-statistic of our null hypothesis $H_0 \colon \sum_{i=1}^{g} c_i \cdot \mu_i = 0$ (without the MS_E factor).

- Hence, $\frac{MS_c}{MS_E} \sim F_{1,\, N-g}$ under H_0.

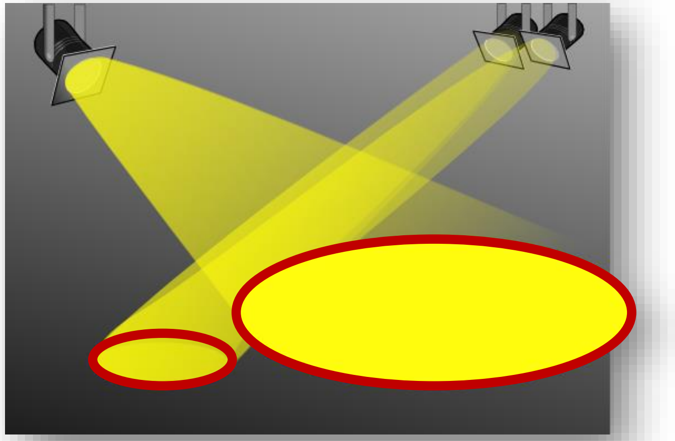
- Again: Nothing else than a **squared version** of the ***t*-test**.

Contrasts in R

- Multiple options:
  - Directly in R (not very user-friendly)
  - Package **multcomp** (will also be very useful later)
  - Many more…
- See the corresponding R-script for details.

Orthogonal Contrasts

- Two contrasts c and c^* are called **orthogonal**, if $\sum_{i=1}^{g} c_i \cdot c_i^* / n_i = 0$.
- Orthogonal contrasts contain **independent** information.
- If there are g groups, one can find $g - 1$ different orthogonal contrasts (1 dimension already used by global mean $(1, \ldots, 1)$).
- However, infinitely many possibilities…



Decomposition of Sum of Squares

- A set of **orthogonal** contrasts **partitions** the treatment sum of squares.
- It means: the sum of the contrast sum of squares is SS_{Trt}, i.e. for orthogonal contrasts $c^{(1)}, c^{(2)}, \ldots, c^{(g-1)}$ it holds that

$$SS_{c^{(1)}} + SS_{c^{(2)}} + \cdots + SS_{c^{(g-1)}} = SS_{Trt}$$

- Intuition: “We get all the information about the treatment by pointing the spotlight at all directions.”

It’s your **research hypotheses** that define the contrasts,
not the orthogonality criterion.

Multiple Testing

Multiple Comparisons

- The more tests we perform, the more likely we are doing at least one **type I error** (i.e., falsely rejecting H_0).
- More formally: Perform m tests $H_{0,j}, j = 1, \dots, m$.
- If all $H_{0,j}$ are true and if all tests are **independent**:

  Probability to make **at least one** false rejection is given by

  $$1 - (1 - \alpha)^m$$

  where α is the (**individual**) significance level.
- For $\alpha = 0.05$ and $m = 50$ this is 0.92 (!)

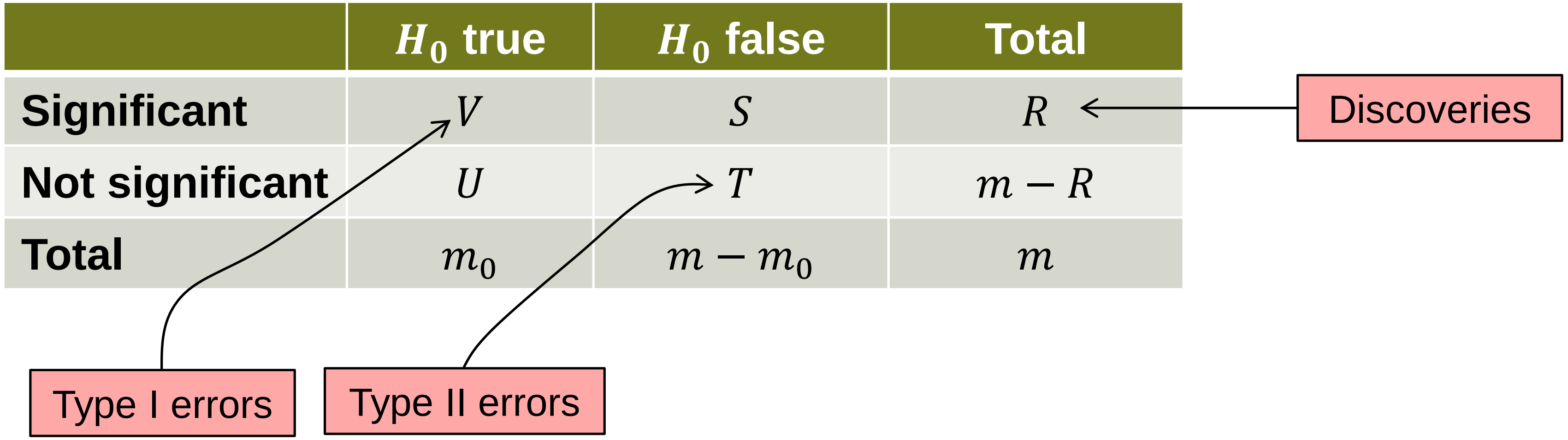
Multiple Comparisons

- The **more tests** we perform, the more likely we are getting some **significant result**.
- If we test many null hypotheses, we expect to reject some of them, even if they are all true.
- If we start **data-fishing** (i.e., screening data for “special” patterns) we (implicitly) do **a lot** of tests.

Different Error Rates

- Consider testing m hypotheses, whereof m_0 are true.

- These are the potential outcomes (numbers):



	H_0 true	H_0 false	Total
Significant	V	S	R
Not significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

Discoveries

Type I errors

Type II errors

- **Comparisonwise error rate** is type I error rate of an **individual** test.

- **Family-wise (FWER)** (or **experimentwise**) **error rate** is the probability of rejecting **at least one** of the true H_0's:

$$\text{FWER} = P(V \geq 1)$$

Different Error Rates

- A procedure is said to **control** the FWER at level α in the **strong** sense, if

  $$\mathrm{FWER} \leq \alpha$$

  for **any configuration** of true and non-true null hypotheses.

- The **false discovery rate (FDR)** is the expected fraction of false discoveries

  $$\mathrm{FDR} = E\left[\frac{V}{R}\right]$$

  false discovery fraction

Confidence Intervals

- Quite often, each H_0 corresponds to a parameter.
- We can construct **confidence intervals** for each of them.
- We call these confidence intervals **simultaneous** at level $(1 - \alpha)$ if the probability that **all** intervals cover the corresponding true parameter is $1 - \alpha$.
- Intuition: Can look at all confidence intervals **at the same time** and get the correct “big picture” with probability $1 - \alpha$.
- Remember: For 20 **individual** 95% confidence intervals it holds that on average one doesn’t cover the true value.

Overview of Multiple Testing Procedures

Control of Family-Wise Error Rate (FWER)

- Bonferroni (conservative)
- Bonferroni-Holm (better version of Bonferroni)
- Scheffé (for search over all possible contrasts, conservative)
- Tukey-HSD (for pairwise comparisons)
- Multiple Comparison with a Control

False Discovery Rate (FDR): see book

- Benjamini-Hochberg
- Benjamini-Yekutieli
- Others

Bonferroni

- Use **more restrictive** significance level $\alpha^* = \frac{\alpha}{m}$.
- That's it!
- This controls the family-wise error rate. No assumption regarding independence required (see blackboard) .
- Equivalently: Multiply all p-values by m and keep using the original α.
- Can get quite conservative if m is large.
- The corresponding confidence intervals (based on the adjusted significance level) are **simultaneous**.

Bonferroni-Holm

- **Less conservative** and hence (uniformly) **more powerful** than Bonferroni.
- Sort p-values from **small** to **large**: $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ where

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$$

- For $j = 1, 2, \ldots$: Reject null hypothesis if $p_{(j)} \leq \frac{\alpha}{(m-j+1)}$.
- **Stop** when you reach the **first** non-significant p-value.
- Only the **smallest** p-value has the traditional Bonferroni correction, hence the method is more powerful than Bonferroni.
- R: `p.adjust` etc.
- This is a so called **step-down procedure** (“stepping-down the sequence of hypotheses”).

Scheffé

- A method which controls for the **search over any** possible contrast …
- This means:
  You are even allowed to perform data-fishing and test the most extreme contrast you'll find (really!).
- These p-values are honest (really!).
- Sounds too good to be true!
- Theory:
  - $SS_c \leq (g-1)MS_{Trt}$ for **any** contrast c (because $SS_{Trt} = SS_c + \cdots$)
  - Hence, $\frac{SS_c}{MS_E} \leq (g-1)\frac{MS_{Trt}}{MS_E}$ for **any** contrast c.
  - Therefore, $\max\limits_{c} \frac{SS_c\,/(g-1)}{MS_E} \leq \frac{MS_{Trt}}{MS_E} \sim F_{g-1,\,N-g}$ under $H_0\colon \mu_1 = \cdots = \mu_g$.

Scheffé

- The price for the nice properties are **low power** (meaning: test will **not** reject often when H_0 is **not** true).

- If F-test is **not** significant: Don’t even have to start searching!

- R:
  - Calculate F-ratio (MS_c / MS_E) as if “ordinary” contrast.
  - Use $(g-1) \cdot F_{g-1,\, N-g, 1-\alpha}$ as critical value (instead of $F_{1,\, N-g, 1-\alpha}$).

Pairwise Comparisons

- A pairwise comparison is nothing else than comparing two specific treatments (e.g., “Vacuum” vs. “CO_2”).

- This is a **multiple testing** problem because there are

  $$g \cdot \frac{g-1}{2}$$

  possible comparisons (basically a lot of two-sample t-tests).

- Hence, we need a method which adjusts for this multiple testing problem in order to control the family-wise error rate.

- Simplest solution: Apply **Bonferroni** correction.

- Better (more powerful): Tukey Honest Significant Difference.

Tukey Honest Significant Difference (HSD)

- Start with statistics of t-test (here for the balanced case where HSD gives exact p-values)

$$\frac{\left|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}\right|}{\sqrt{MS_E}\sqrt{\left(\frac{1}{n} + \frac{1}{n}\right)}}$$

- Use the distribution of

$$\max_i \frac{\bar{y}_{i\cdot}}{\sqrt{MS_E 1/n}} - \min_j \frac{\bar{y}_{j\cdot}}{\sqrt{MS_E 1/n}}$$

  (the so called **studentized range**) for critical values.

- Means: “How does the **maximal difference** between groups behave?”

- If all the means are equal (H_0), this follows the so called **studentized range distribution** (R: `ptukey`).

Tukey Honest Significant Difference (HSD)

- Tukey honest significant difference uses this studentized range distribution to construct **simultaneous confidence intervals** for differences **between all pairs**.

- … and calculates p-values such that the family-wise error rate is controlled.

- R: `TukeyHSD` or package `multcomp` (see R-file for demo).

- Tukey HSD is better (more powerful) than Bonferroni if **all** pairwise comparisons are of interest.

- If only a subset: Re-consider Bonferroni.

Interpreting and Displaying the Results

- A non-significant difference does **not** imply equality.
- Reason:

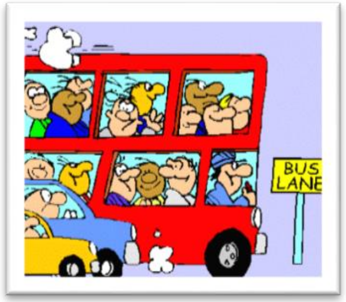
  “Absence of evidence is not evidence of absence”.

- Results can be displayed using
  - Same letters/numbers for treatments with non-significant difference.
  - Matrix (upper or lower triangle) with p-values
  - …

Multiple Comparison with a Control (MCC)

- Often: Compare all treatments with a (specific) **control treatment**.
- Hence, do $g - 1$ (pairwise) comparisons with the control group.
- **Dunnett procedure** constructs simultaneous confidence intervals for the differences $\mu_i - \mu_g, i = 1, \dots, g - 1$ (assuming group g is control group).
- R: Use package `multcomp`.

What About F-Test?



- Can I only do pairwise comparisons etc. if the omnibus F-test is significant?
- **No**, although many textbooks recommend this (!)
- The presented procedures have a **built-in** multiple-testing correction.
- Conditioning on a significant F-test makes them **over-conservative.**
- Moreover, the **conditional error** or **coverage rates** can be (very) bad.

Statistical Significance vs. Practical Relevance

- An effect that is statistically significant is **not** necessarily of practical relevance.
- Instead of simply reporting p-values, one should always consider the corresponding confidence intervals.
- **Background knowledge** should be used to judge when an effect is potentially **relevant**.

Recommendations

- Planned contrasts: Bonferroni
- All pairwise comparisons: Tukey HSD
- Comparison with a control: Dunnett
- Unplanned contrasts: Scheffé