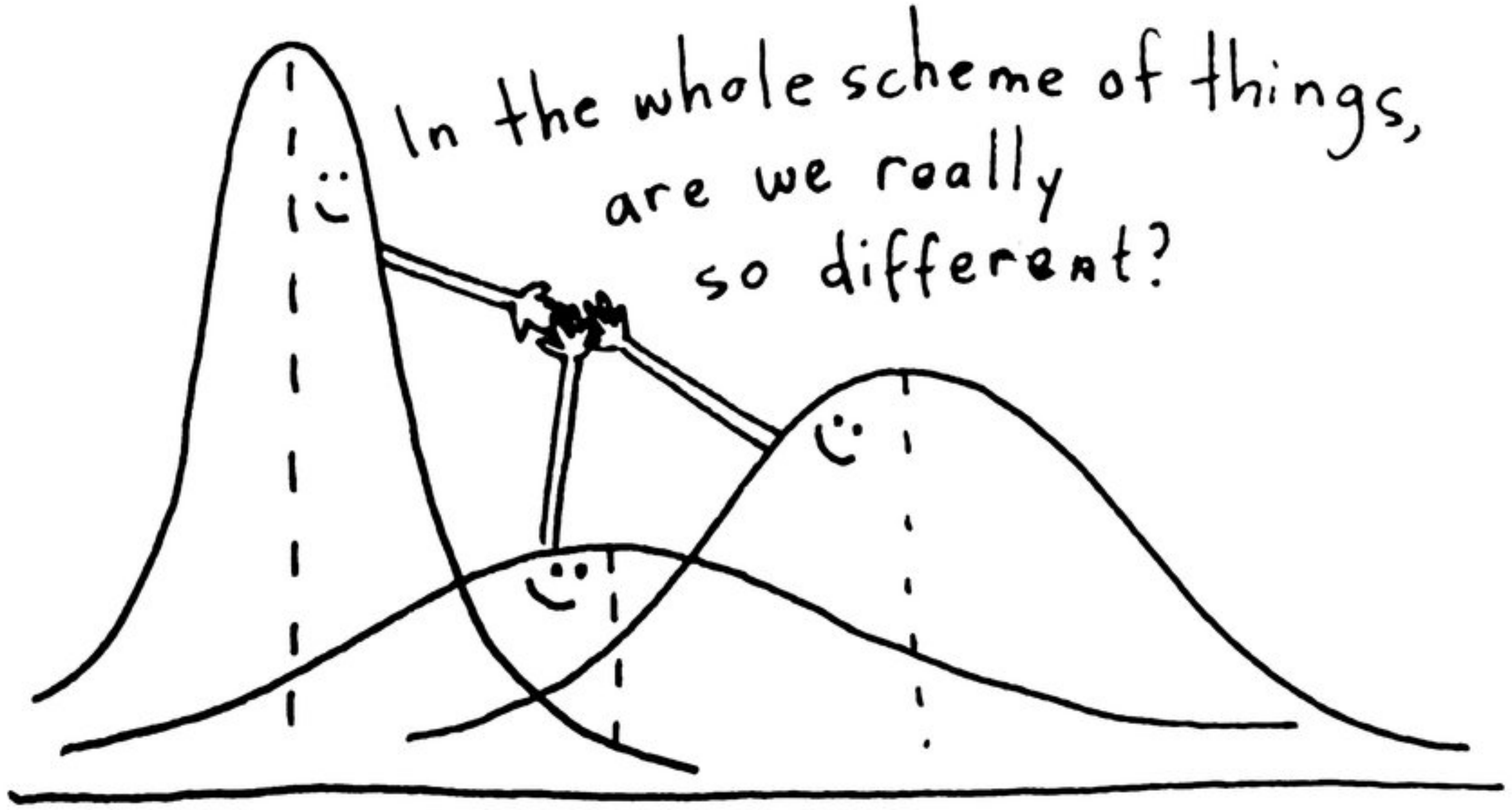
Foundational Statistics

Introduction to Analysis of Variance



From: Questionpro

```
Analysis of Variance Table

Response: Sepal.Length
           Df Sum Sq Mean Sq F value    Pr(>F)
Species     2 63.212  31.606  119.26 < 2.2e-16 ***
Residuals 147 38.956   0.265
```

General Linear Models for a continuous response and a categorical predictor

Regression linear model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Categorical predictor with *i* levels: $y_{ij} = \mu + \beta_1 (level_1)_{ij} + \beta_2 (level_2)_{ij} + \ldots + \varepsilon_{ij}$

Response variable

Overall mean of *y*

Effect of level *i*

Binary encoder

Error (var. in *y* unexplained by model)

Simplified linear model notation:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

General Linear Models for a continuous response and categorical predictors

- ANOVA: **An**alysis **o**f **Va**riance
- Fundamental statistical procedure in biology, developed in the early 20th century
- The core idea is to ask how much variation exists **within** vs. **among** groups
- The **categorical predictors** are also called **factors**, and can have two or more **factor levels**
- Each factor in an ANOVA model can have a hypothesis test, and levels within a factor can be contrasted
- Diversity of ANOVA model complexity: (e.g. nested, factorial, etc.)

ANOVA - an experimental example

Percent time male mice experiencing discomfort spent “stretching”.

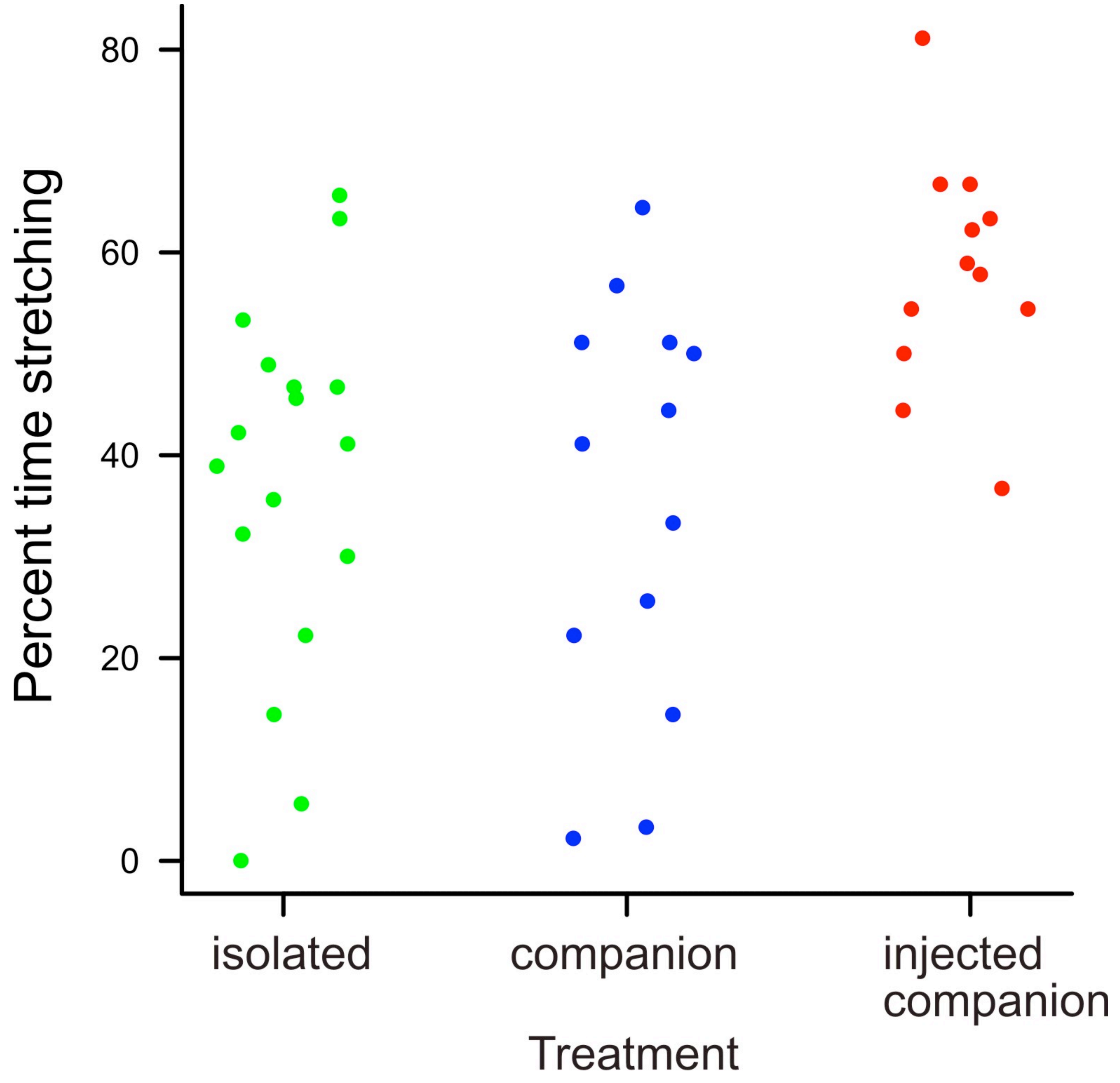
Data are from an experiment in which mice experiencing mild discomfort (result of injection of 0.9% acetic acid into the abdomen) were kept in:

(1) isolation,
(2) with a companion mouse not injected, or
(3) with a companion mouse also injected and exhibiting “stretching” behaviors associated with discomfort.

The results suggest that mice stretch the most when a companion mouse is also experiencing mild discomfort. Mice experiencing pain appear to “empathize” with co-housed mice also in pain.

Langford, D. J.,et al. 2006. Science 312: 1967-1970

ANOVA - an experimental example



In words:

stretching = intercept + treatment

The model statement includes a response variable, a constant (intercept), and an explanatory variable, which is categorical

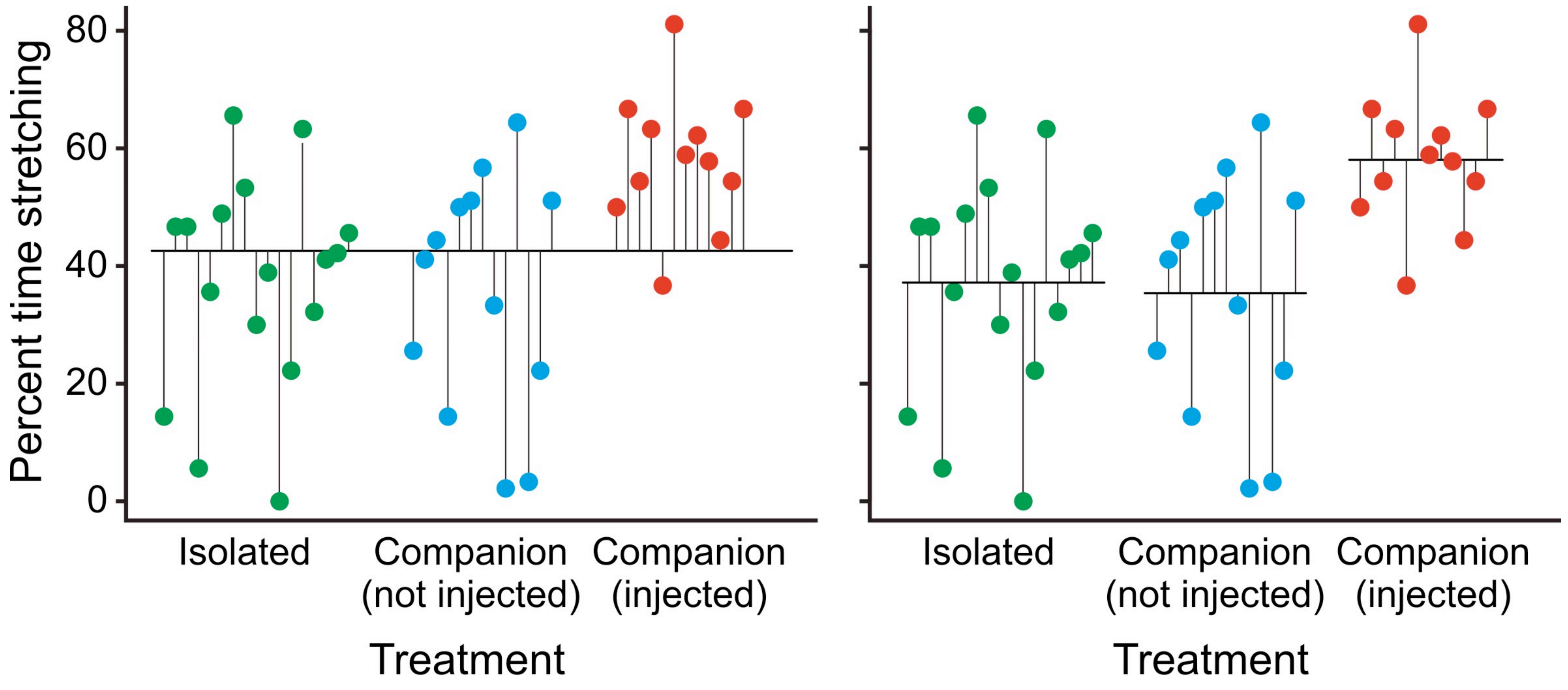
Langford, D. J.,et al. 2006. Science 312: 1967-1970

ANOVA is a linear model, like regression

As before, `anova` compares the fit of “reduced” and “full” models:

`stretching ~ 1`   `stretching ~ treatment`

Reduced Model: One overall mean



Full Model: Different Means

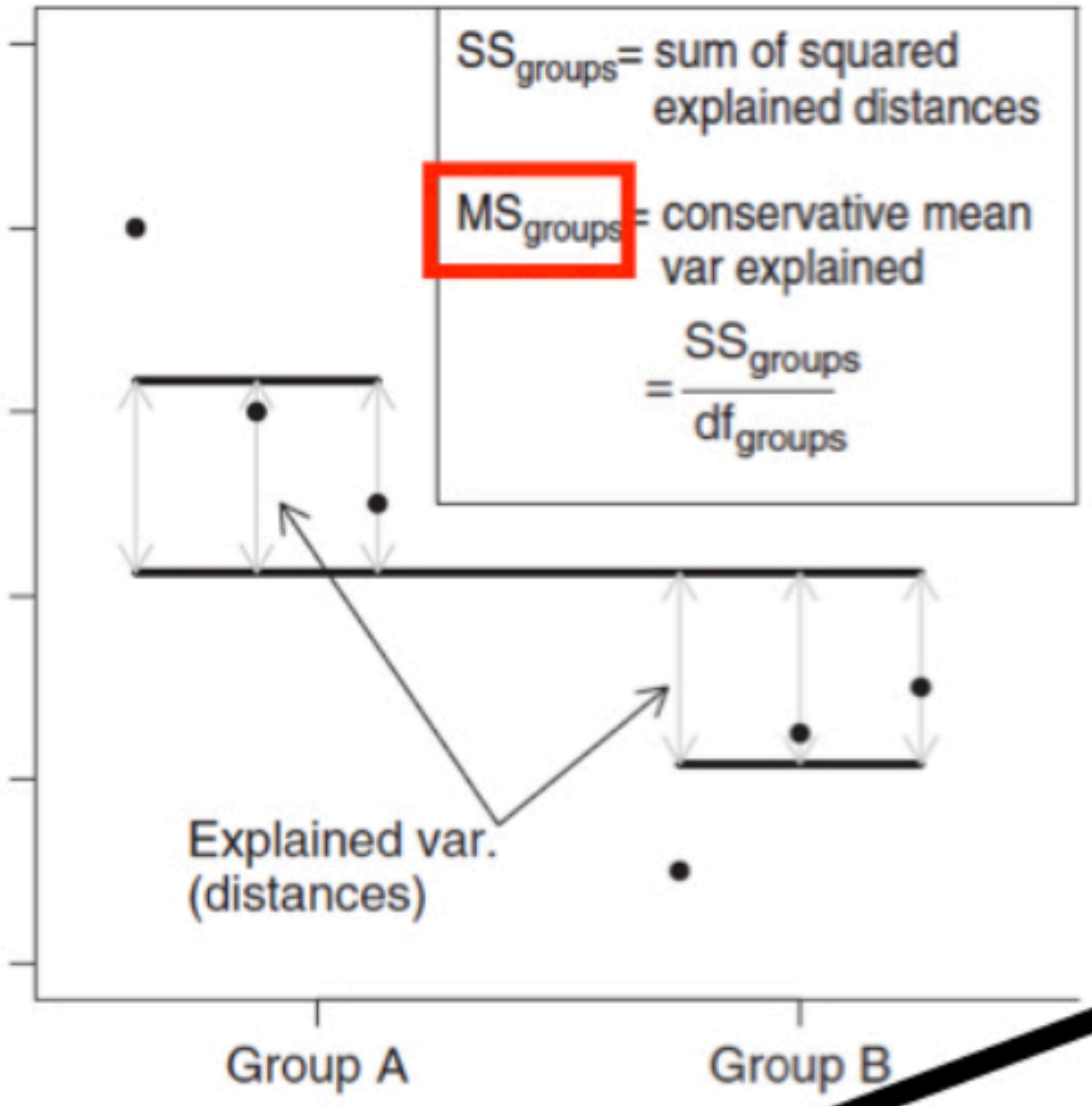
Single factor ANOVA - getting the F-ratio

Table 8.2 ANOVA table for single factor linear model showing partitioning of variation

Source of	SS	df	MS	
Between groups	$\sum_{i=1}^{p} n_i(\bar{y}_i - \bar{y})^2$	$p-1$	$\dfrac{\sum_{i=1}^{p} n_i(\bar{y}_i - \bar{y})^2}{p-1}$	**Var. explained by groupings**
Residual	$\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2$	$\sum_{i=1}^{p} n_i - p$	$\dfrac{\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2}{\sum_{i=1}^{n} n_i - p}$	**Var. unexplained by groupings**
Total	$\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij} - \bar{y})^2$	$\sum_{i=1}^{p} n_i - 1$		

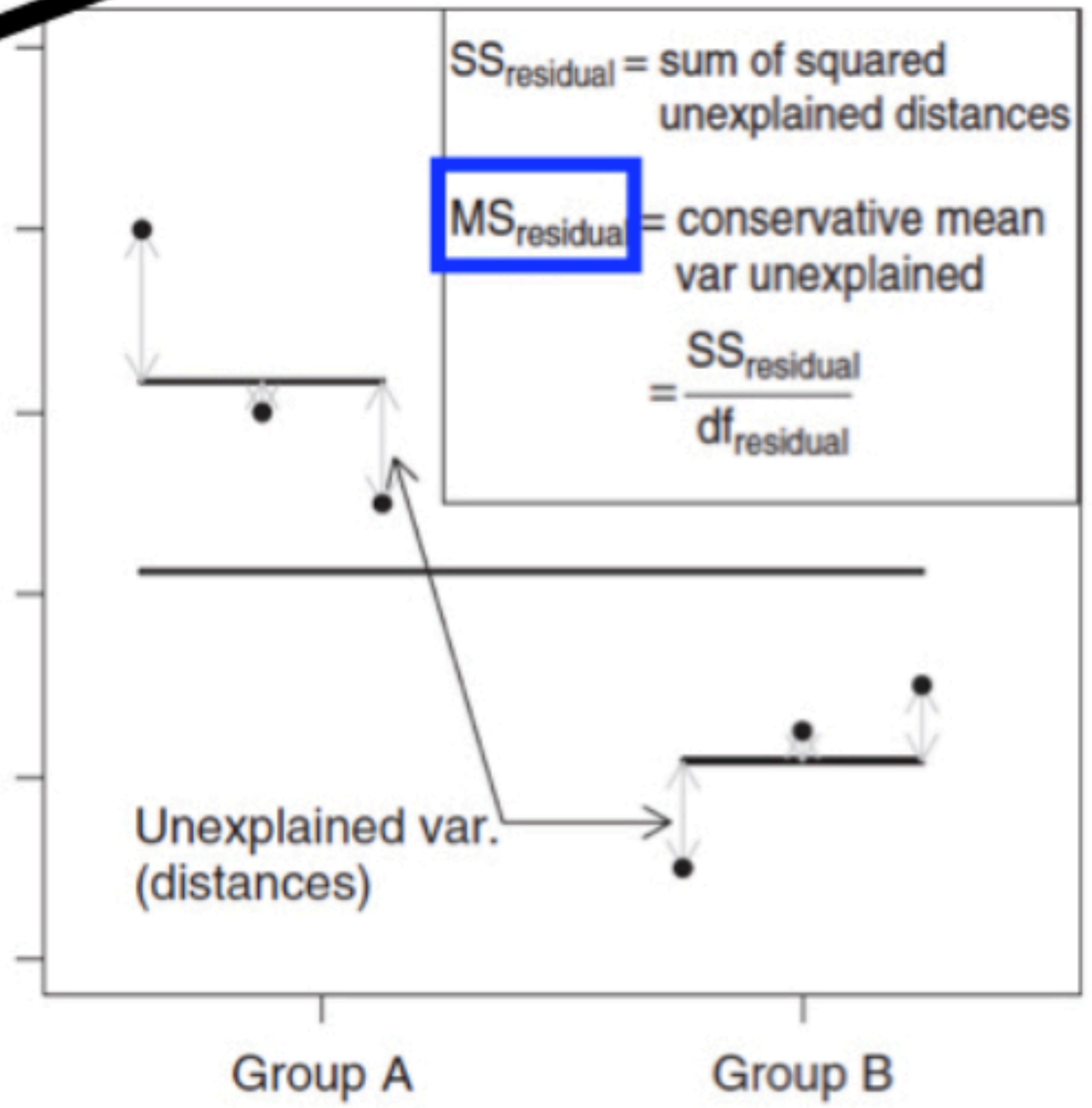
$$\text{F-ratio} = \frac{MS_{groups}}{MS_{residuals}}$$

Single factor ANOVA - getting the *F*-ratio



$$\boxed{\frac{\sum_{i=1}^{p} n_i(\bar{y}_i - \bar{y})^2}{p-1}}$$

n* = num. obs. in group *i
$\bar{y}$ = overall mean
$\bar{y}_i$ = mean of group *i*
***p* = number of groups**



$$\boxed{\frac{\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2}{\sum_{i=1}^{p} n_i - p}}$$

y_{ij} = obs. *j* in group *i*

Single factor ANOVA **Hypotheses**

$H_0 : \alpha_i = 0$ **No effect (all group means are equal)**

$H_A : \alpha_i \neq 0$ **A non-zero effect (at least 2 group means are different)**

These are for “fixed” effects (factors)

Single factor ANOVA **Hypotheses** (random effects)

$$H_0 : \sigma_\alpha^2 = 0$$

No additional variance introduced by the factor levels

$$H_A : \sigma_\alpha^2 > 0$$

Additional variance contributions from the factor levels

These are for "random" effects (factors)

Single factor ANOVA **Assumptions**

1. Response variable normally dist. in all groups

(Check using histograms, boxplots, etc.)

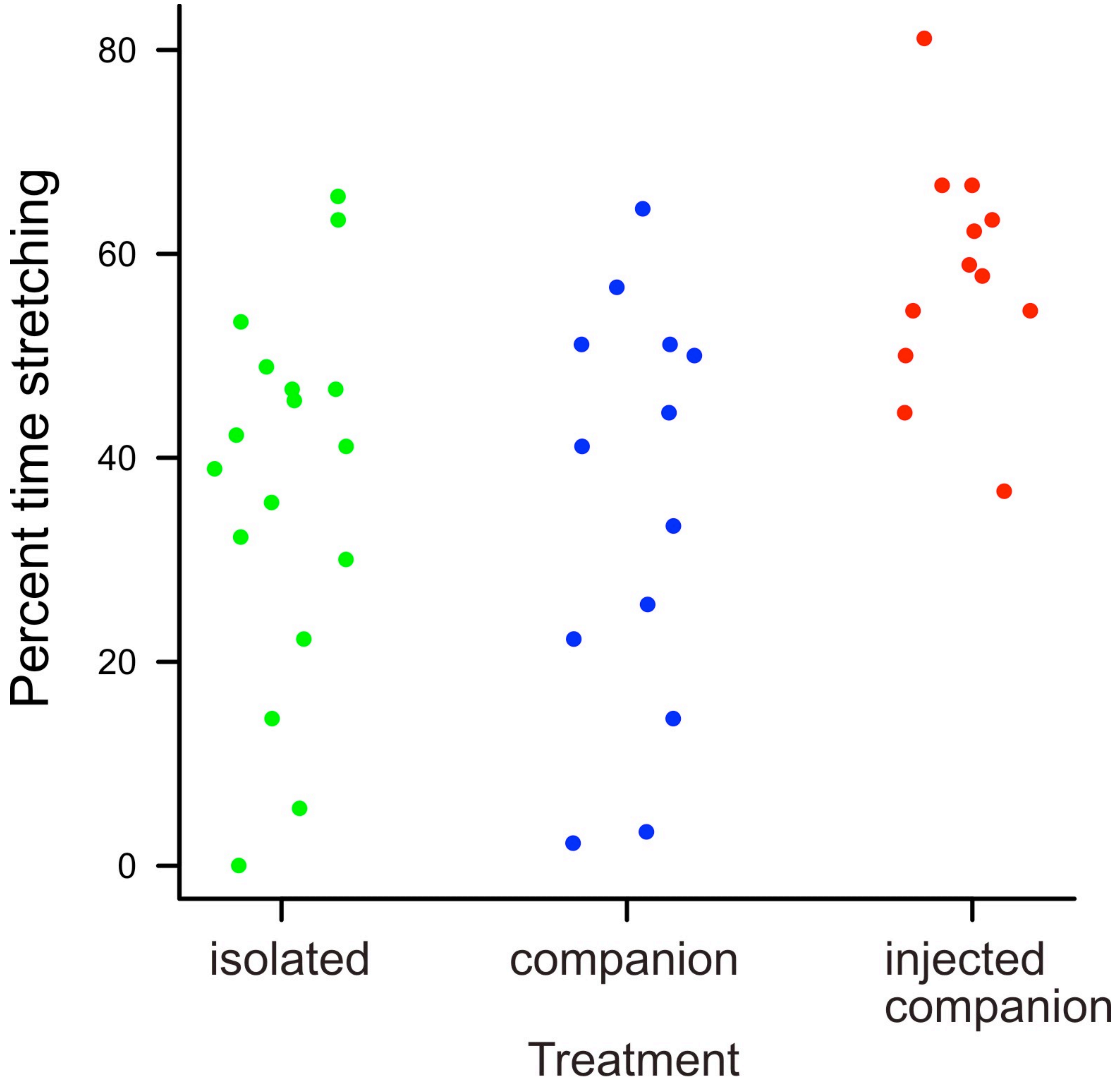
2. Variances equal among groups (no strong mean-var. or sample size-relationships)

(Check using histograms, boxplots, mean vs. var. plots, etc.)

3. Observations within groups are independent, random samples

(Your experimental design needs to ensure this)

Post-hoc comparisons among factor levels



Post-hoc comparisons test all group differences and correct for multiple hypothesis tests.

Tukey tests: compare all pairs of means

Scheffé contrasts: compare all combinations of means

Which of these 3 groups are different from one another?