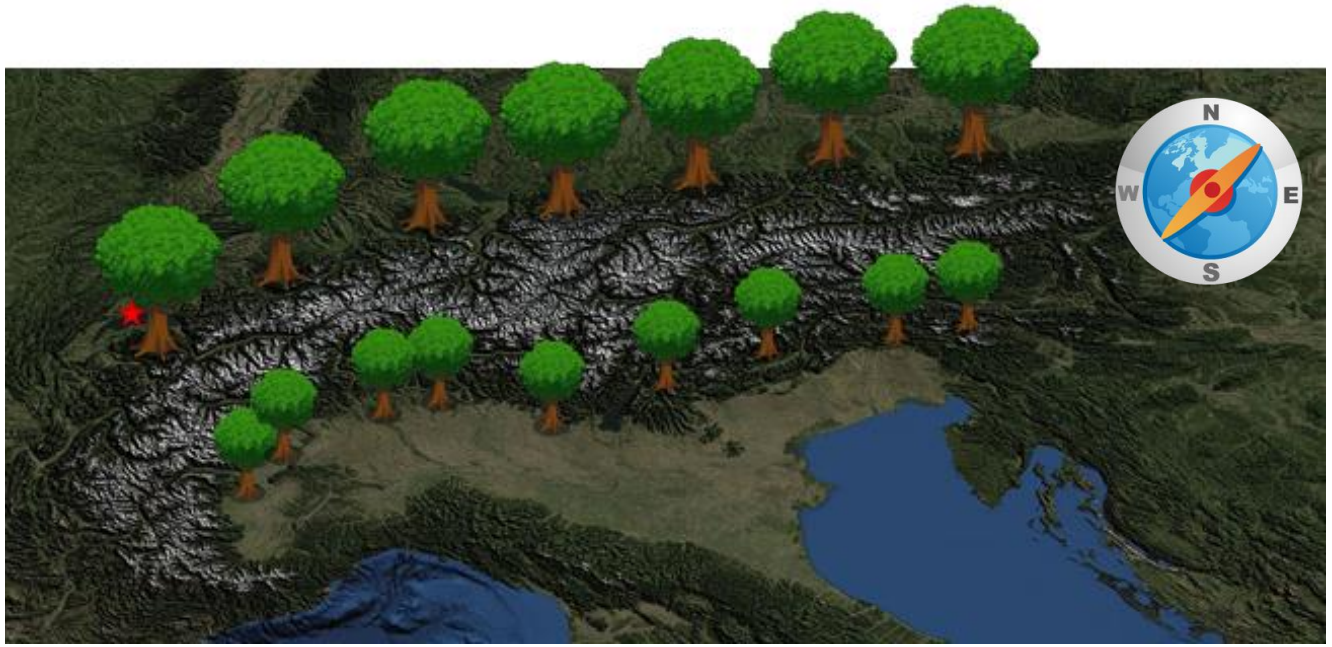
UNDERSTANDING STATISTICS & EXPERIMENTAL DESIGN

Content

1. Basic Probability Theory
2. Signal Detection Theory (SDT)
3. SDT and Statistics I and II
4. Statistics in a nutshell
5. Multiple Testing
6. ANOVA
7. Experimental Design & Statistics
8. Correlations & PCA
9. Meta-Statistics: Basics
10. Meta-Statistics: Too good to be true
11. Meta-Statistics: How big a problem is publication bias?
12. Meta-Statistics: What do we do now?

Example

$$\bar{x}(North) = \frac{1}{n}\sum_{i=1}^{n} x_i$$



$$\bar{x}(South) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Bergmann's rule states that "populations and species of **larger size** are found in **colder environments**, and species of **smaller size** are found in **warmer regions**" (Wikipedia.org).

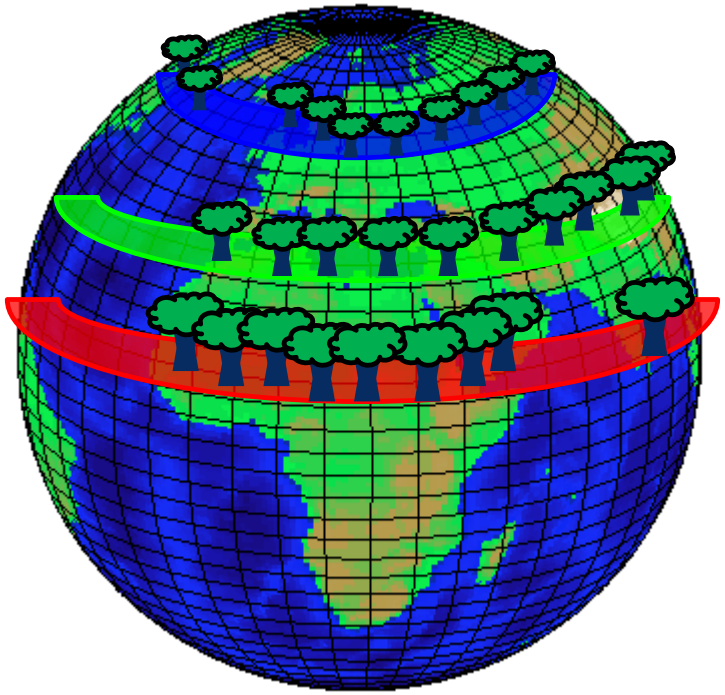
Compare mean height of oaks in the North and in the South

μ(North) > μ(South)

Tree Heights

- Suppose we want to compare tree heights over the three regions shown on the right in the red, green, and blue bands.

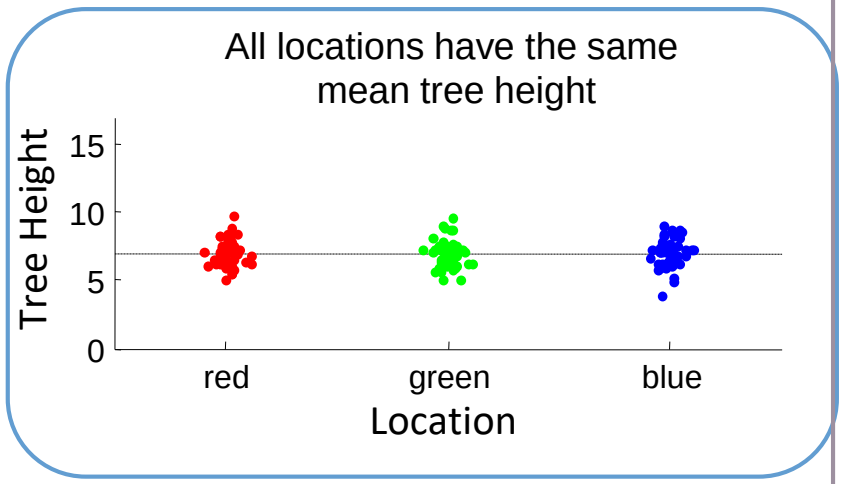
- Are the tree heights all the same, or does at least one region contain trees whose heights differ from those in the other regions?



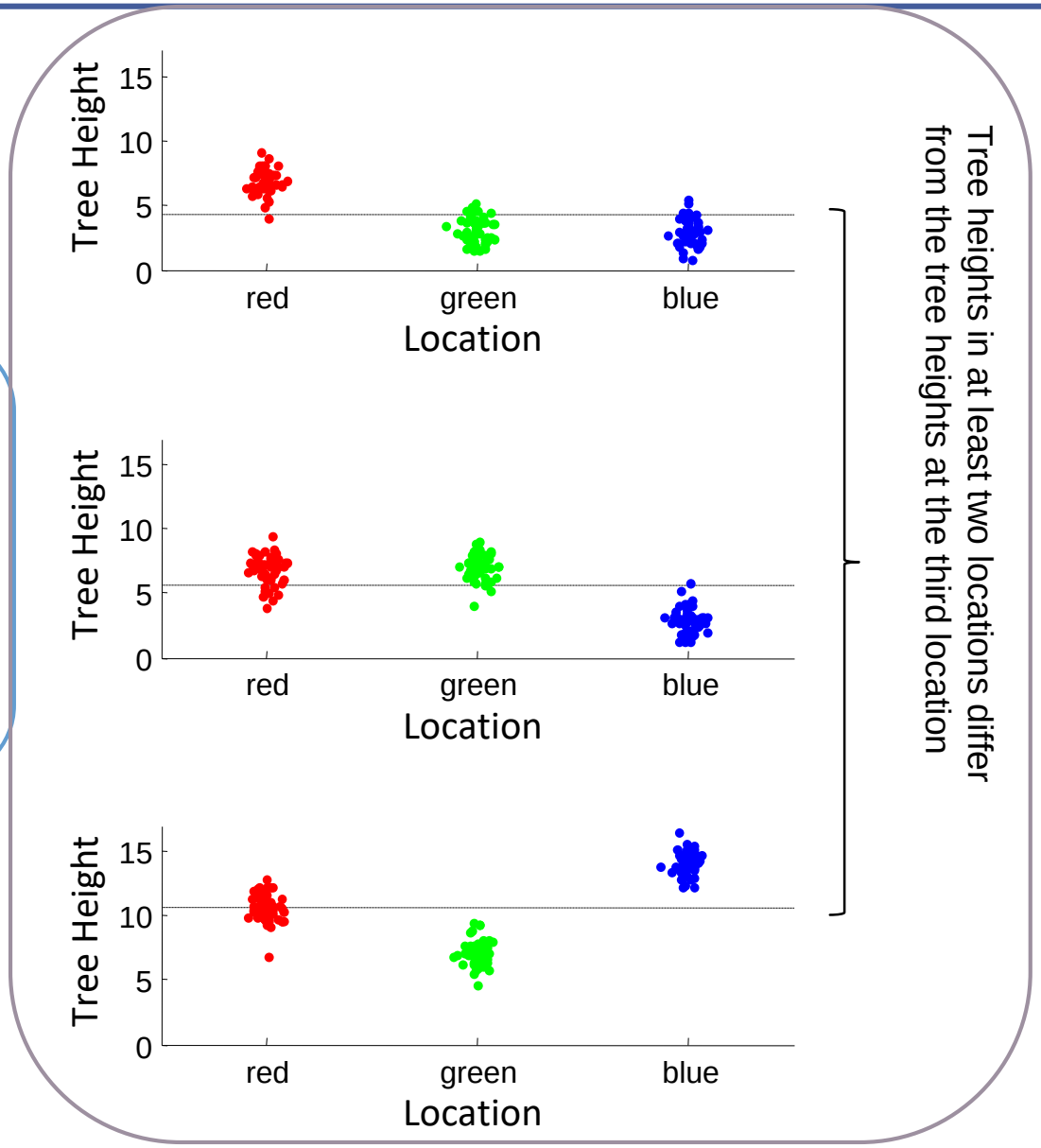
ANOVA: Analysis Of Variance

One Way ANOVA

Possible Outcomes



Null Hypothesis



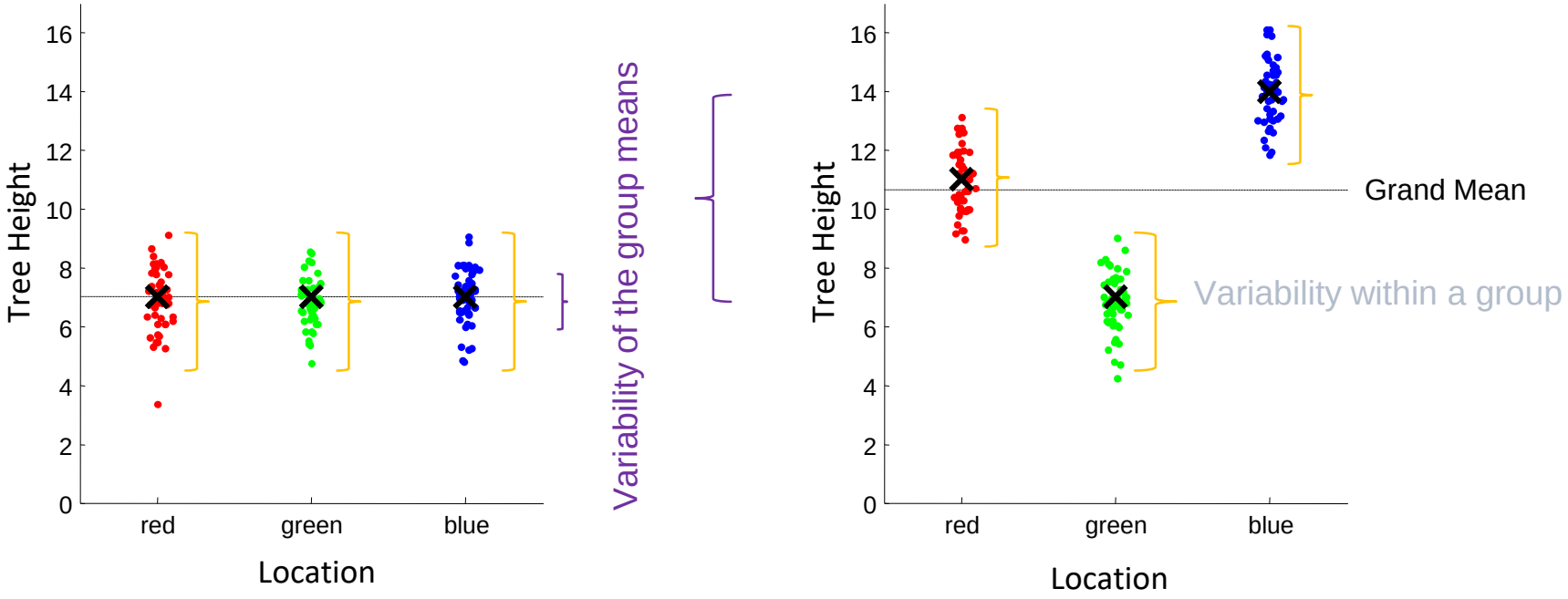
Alternative Hypotheses

Problem

How can we distinguish between the null (H0) and alternative hypothesis (H1) outcomes in a principled way?

Strategy

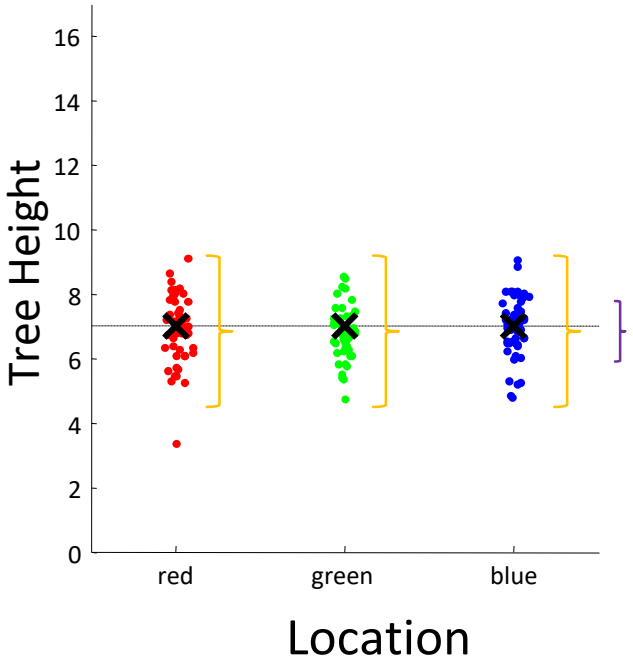
Measure the average variability of tree heights within each region and compare it to the variability of the mean tree heights around the grand mean.



Strategy

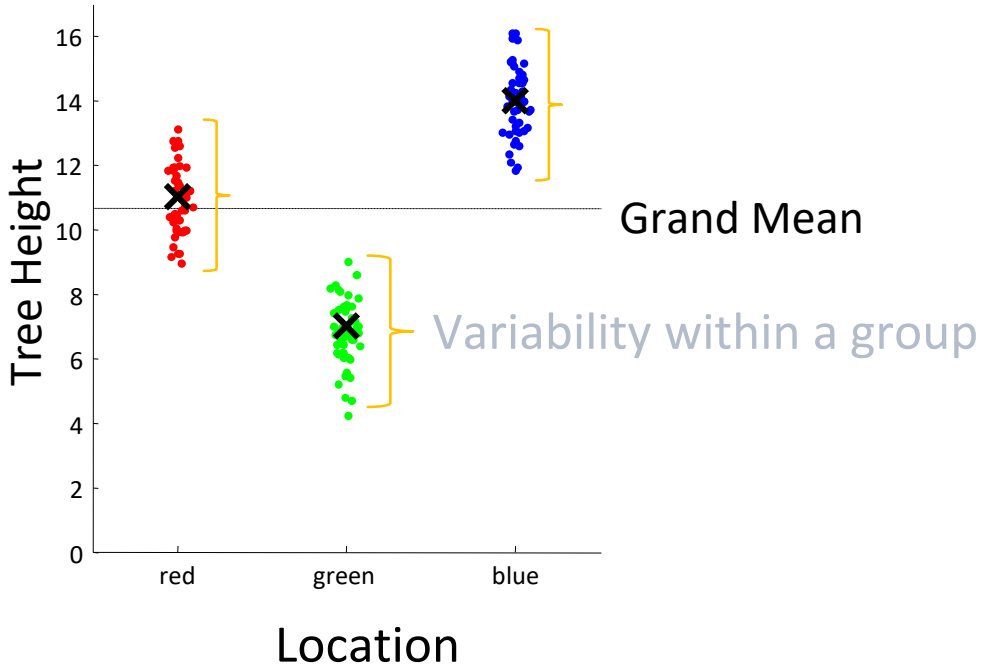
$$F = \frac{\text{Variability between group means}}{\text{Variability within groups}}$$

$$\frac{\text{Small number}}{\text{Large number}}$$



Variability of the group means

$$\frac{\text{Large number}}{\text{Small number}}$$

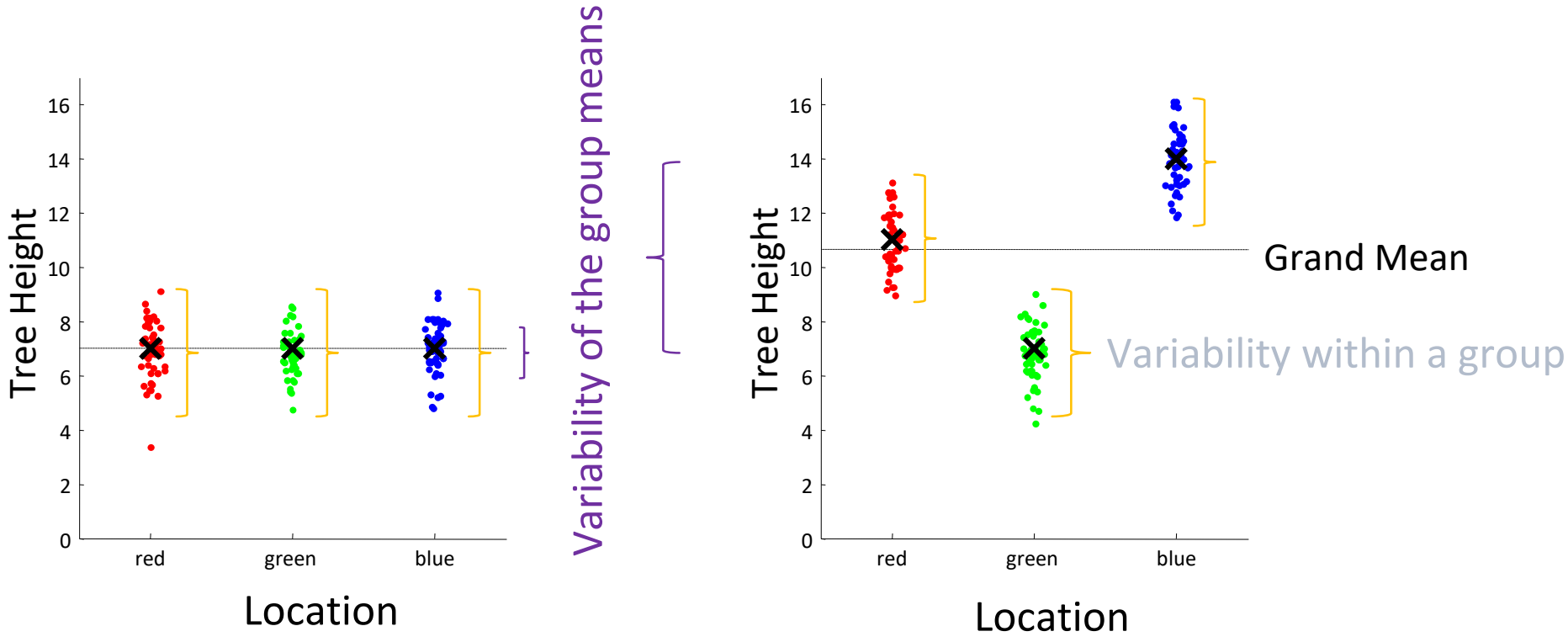


Strategy

$$\frac{\text{Variability between treatments}}{\text{Variability within treatments}}$$

$$\frac{\text{Small number}}{\text{Large number}}$$

$$\frac{\text{Large number}}{\text{Small number}}$$

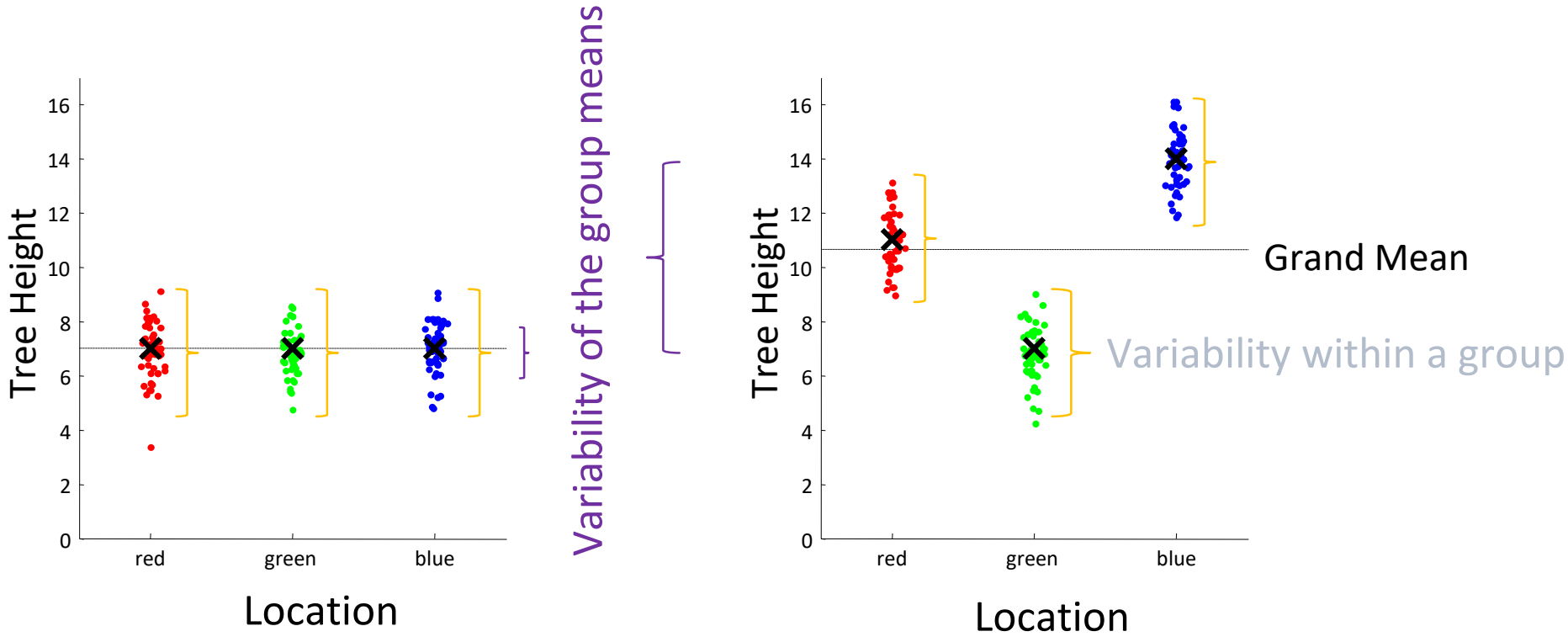


Strategy

If F is ***small*** then it indicates support for the null hypothesis.
If F is ***large*** then it indicates support for the alternative hypothesis.

$$F = \frac{\text{Small number}}{\text{Large number}} = \text{Small number}$$

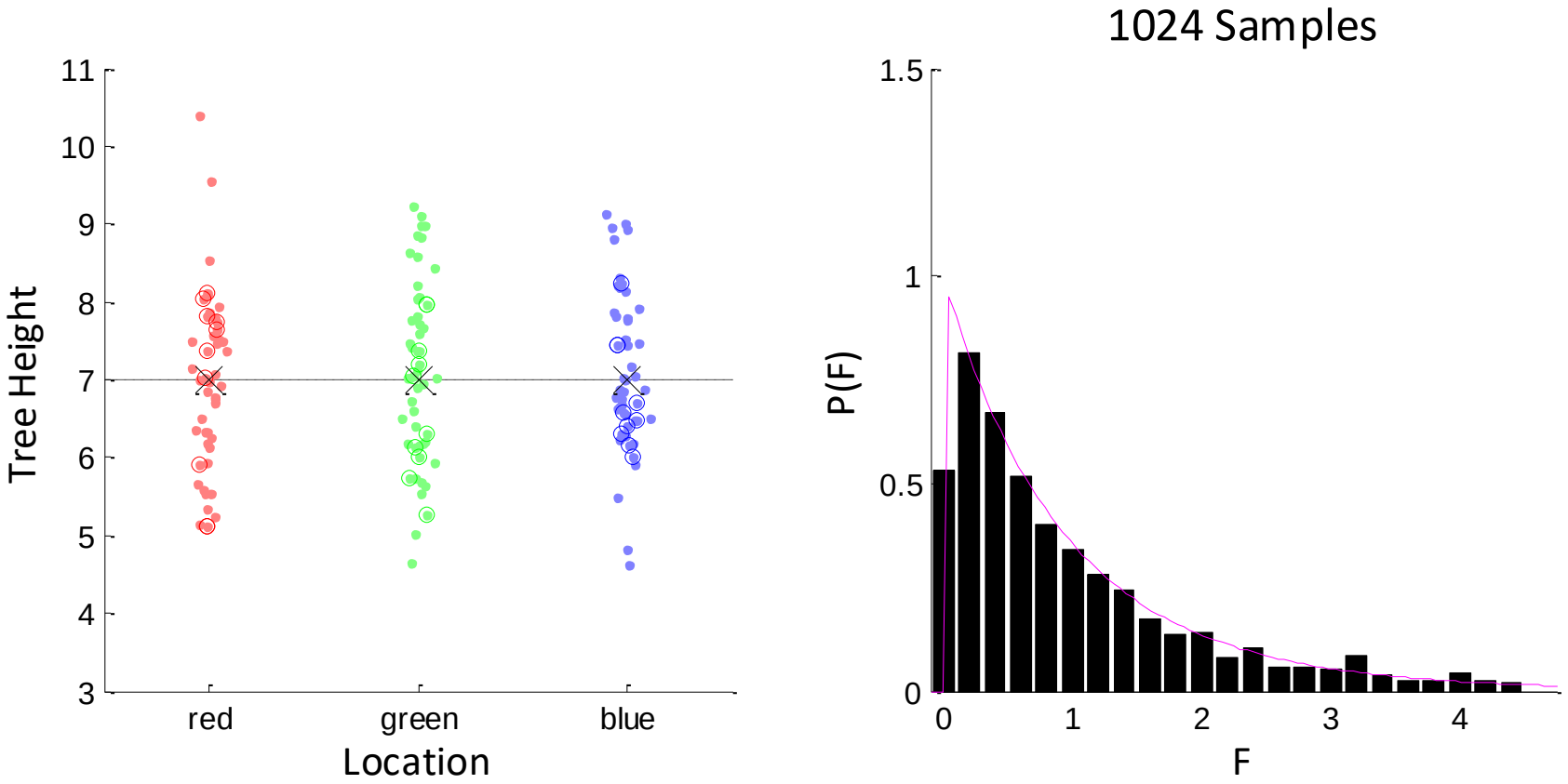
$$F = \frac{\text{Large number}}{\text{Smallnumber}} = \text{Large number}$$



Strategy

This way we can perform one test with a = 0.05

How is *F* distributed under the null Hypothesis?

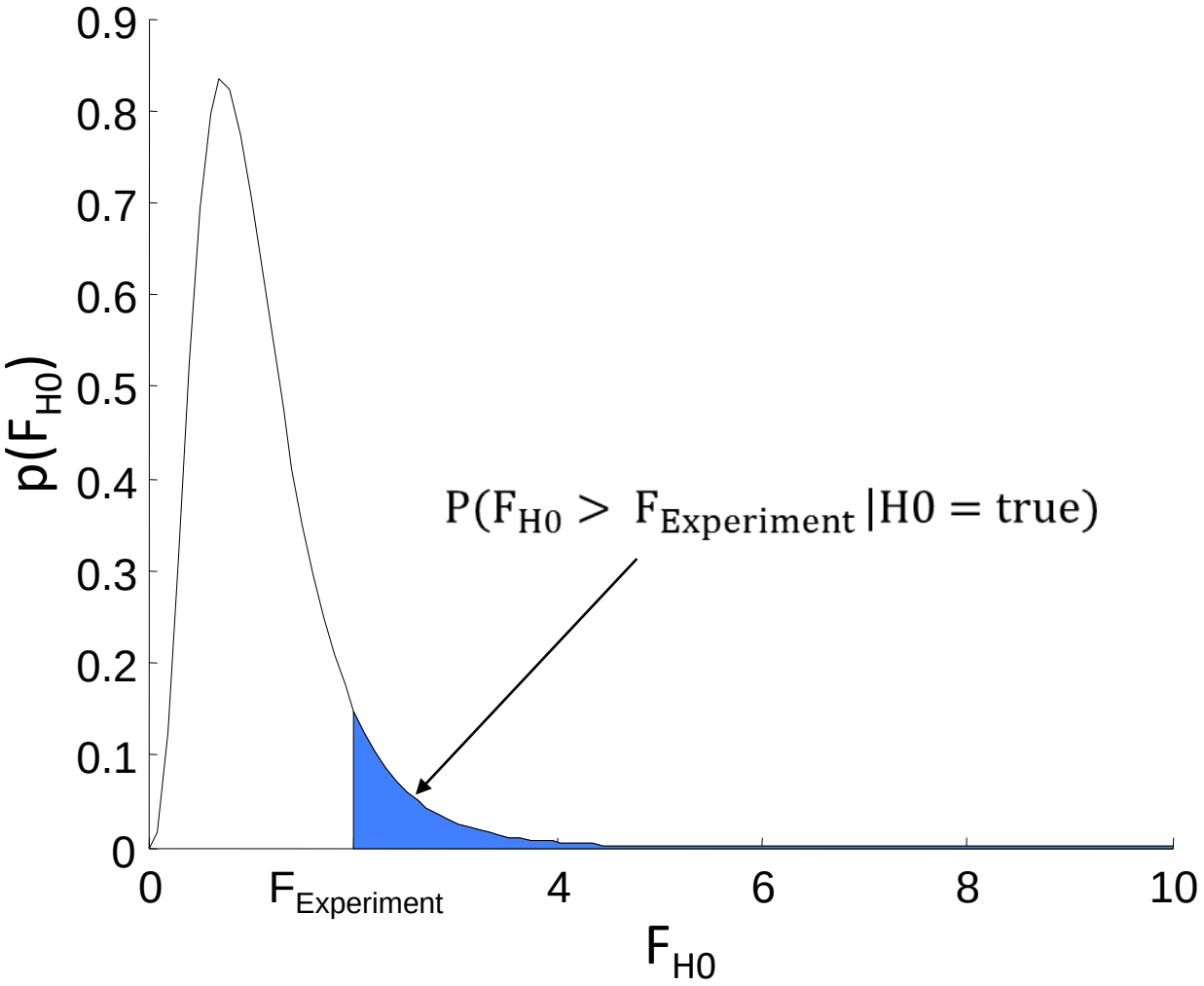


Logic of the ANOVA

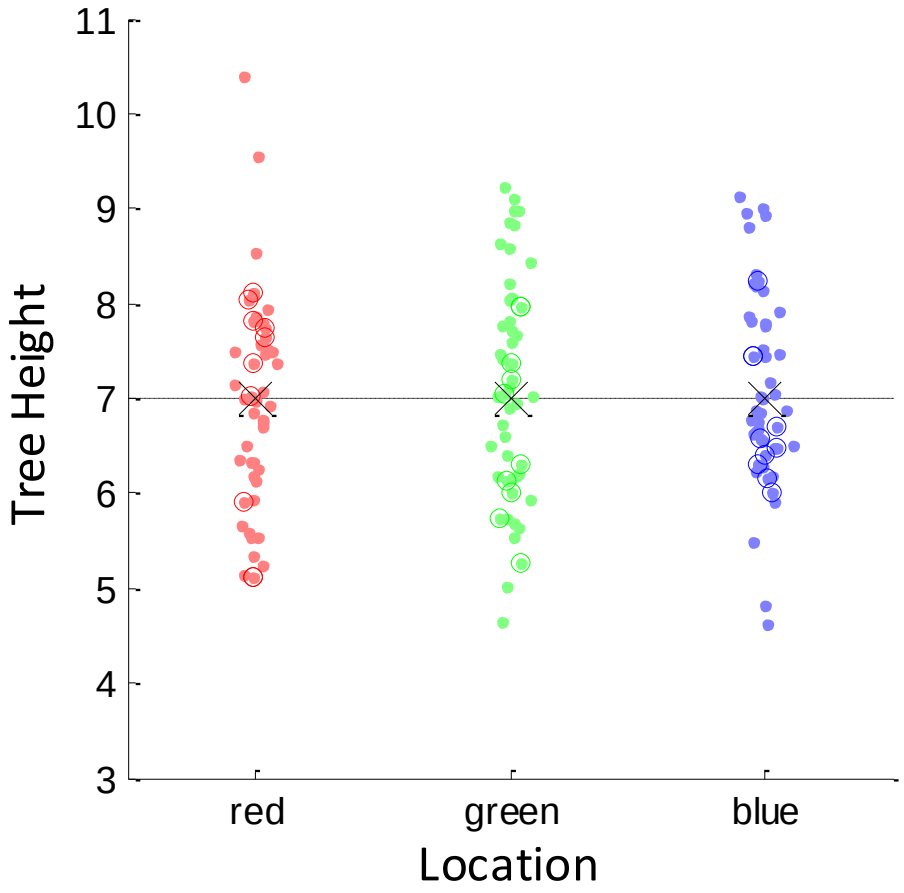
- If the null hypothesis is true then F is expected to be around 1, or less, on average
- If an ANOVA we compute the probability that a value equal to or greater than the F-value from our experiment would be observed if the null hypothesis was true, i.e,

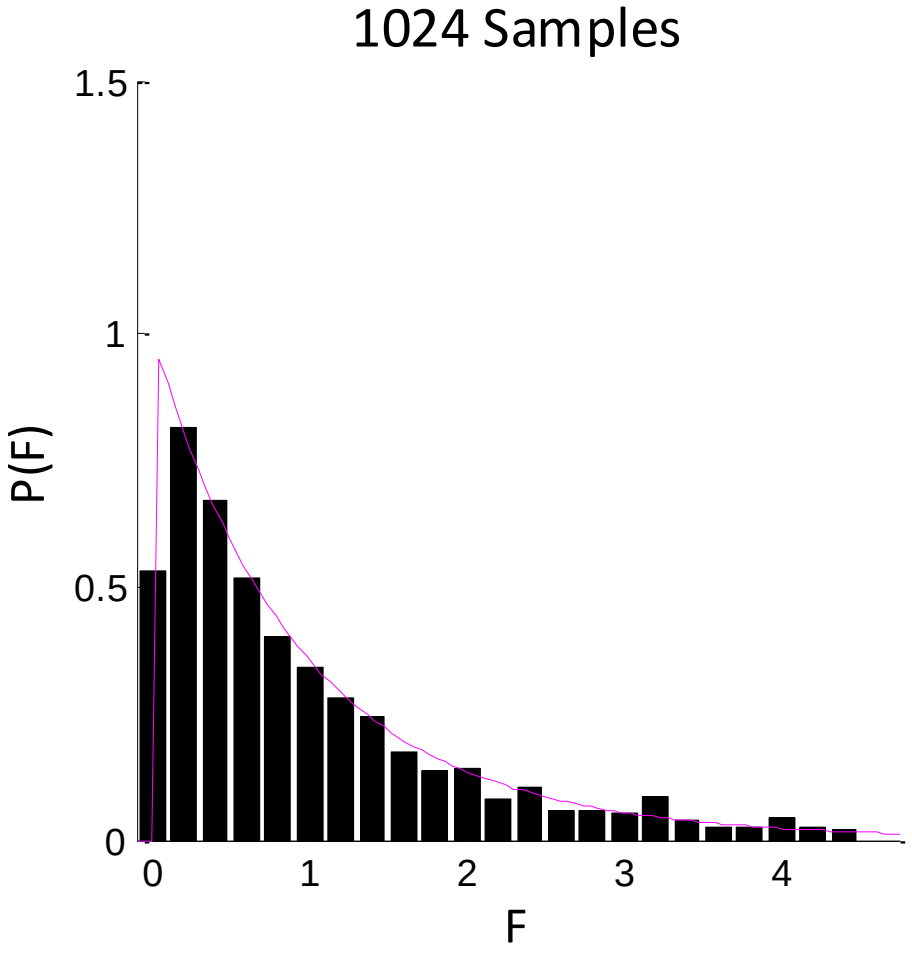
$$P(F_{H0} > F_{Experiment} | HO = true))$$

Logic of the ANOVA



How Is *F* Distributed Under the Null Hypothesis?





Distribution of *F* Under the Null Hypothesis

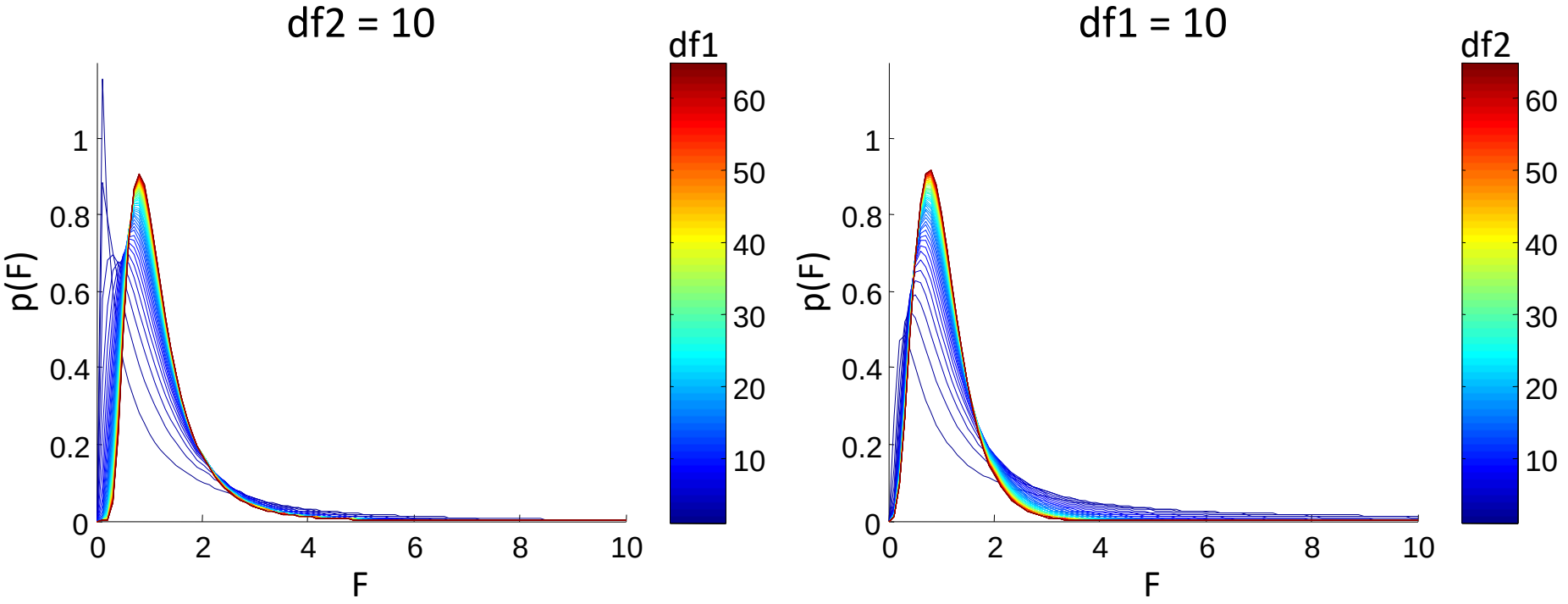
- In general, the distribution of *F* obtained from all possible samples under the null hypothesis follows the form:

$$p(F;d_1,d_2)=\frac{\sqrt{\frac{(d_1F)^{d_1}d_2^{d_2}}{(d_1F+d_2)^{d_1+d_2}}}}{FB\left(\frac{d_1}{2},\frac{d_2}{2}\right)}$$

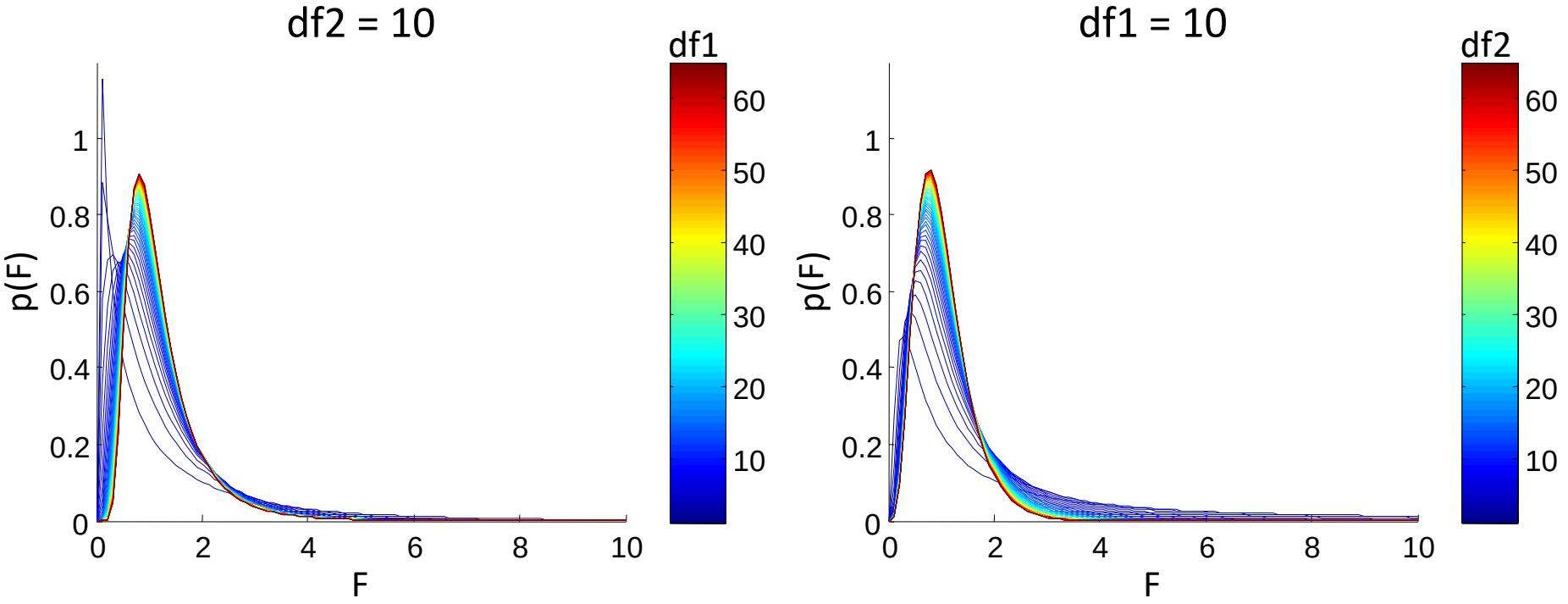
- Where:
  - d_1 = number of Groups – 1 = 3 -1
  - d_2 = total number of points – number of groups = 150 – 3

$$B(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}dt$$

Distribution of *F* Under the Null Hypothesis



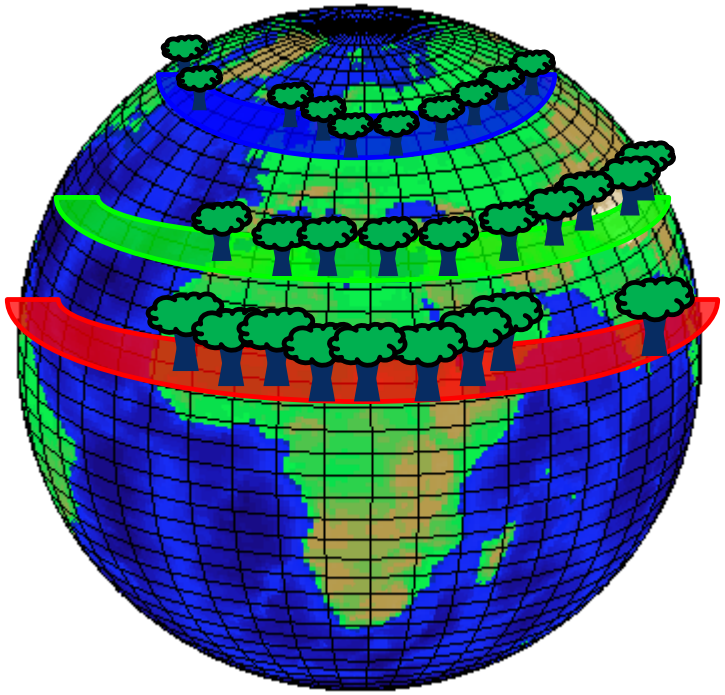
Distribution of *F* Under the Null Hypothesis



$$\frac{\text{Variability of the group means}}{\text{Variability within a group}}$$

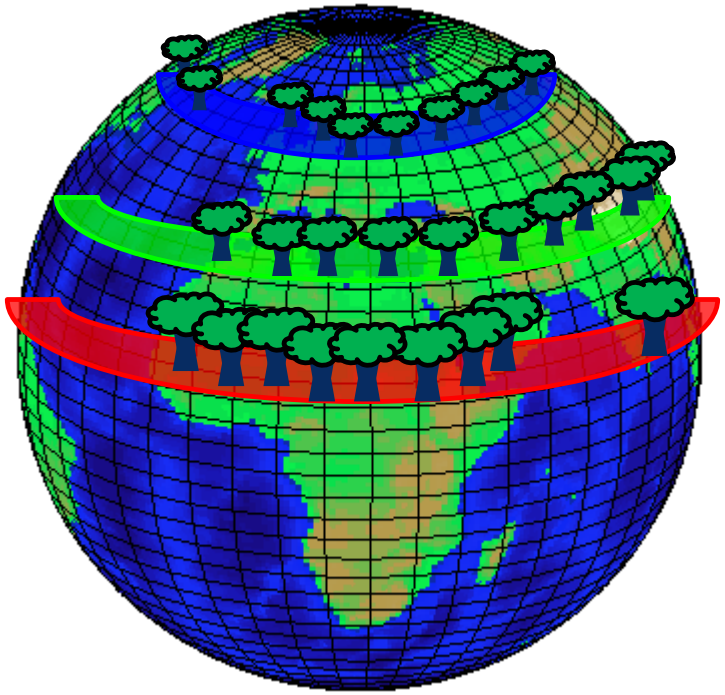
Tree Heights

- Suppose we want to compare tree heights over the three regions shown on the right in the red, green, and blue bands.
- Are the tree heights all the same, or does at least one region contain trees whose heights differ from those in the other regions?



Worked Example

- H0: The tree heights in the three regions are the same.
- H1: The tree heights in at least one of the regions differs from the others.



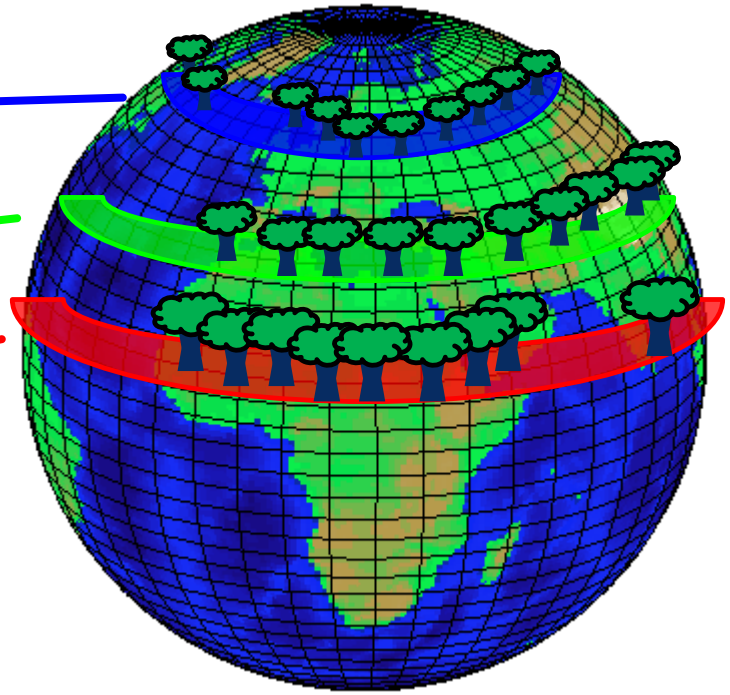
ANOVA Example

- H_0: All three tree groups are equal.
  - $\mu_{North} = \mu_{South} = \mu_{Africa}$
- H_1: At least one tree population is different from the others.
  - $\mu_N = \mu_S \neq \mu_A$
  - $\mu_N \neq \mu_S = \mu_A$
  - $\mu_N \neq \mu_A = \mu_S$
  - $\mu_N \neq \mu_S \neq \mu_A$

Worked Example

Tree Height:

5	4	6
8	7	9
12	11	14

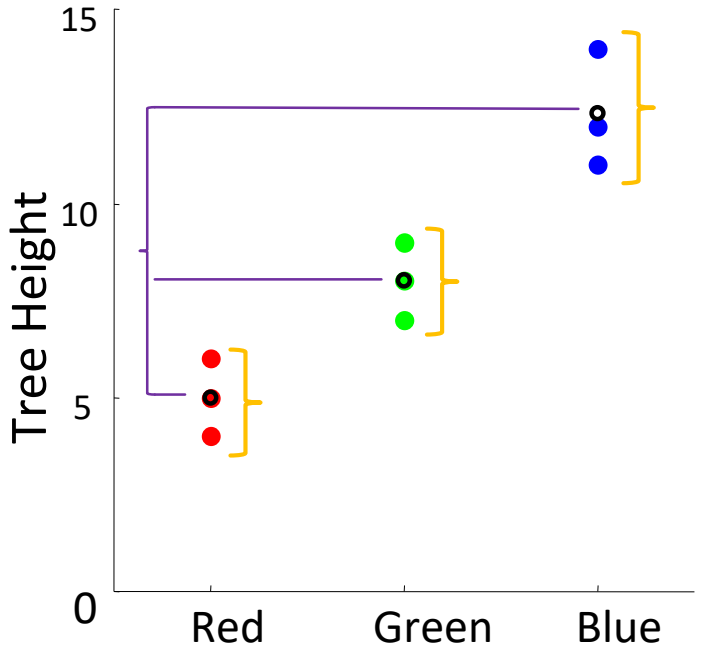


Worked Example

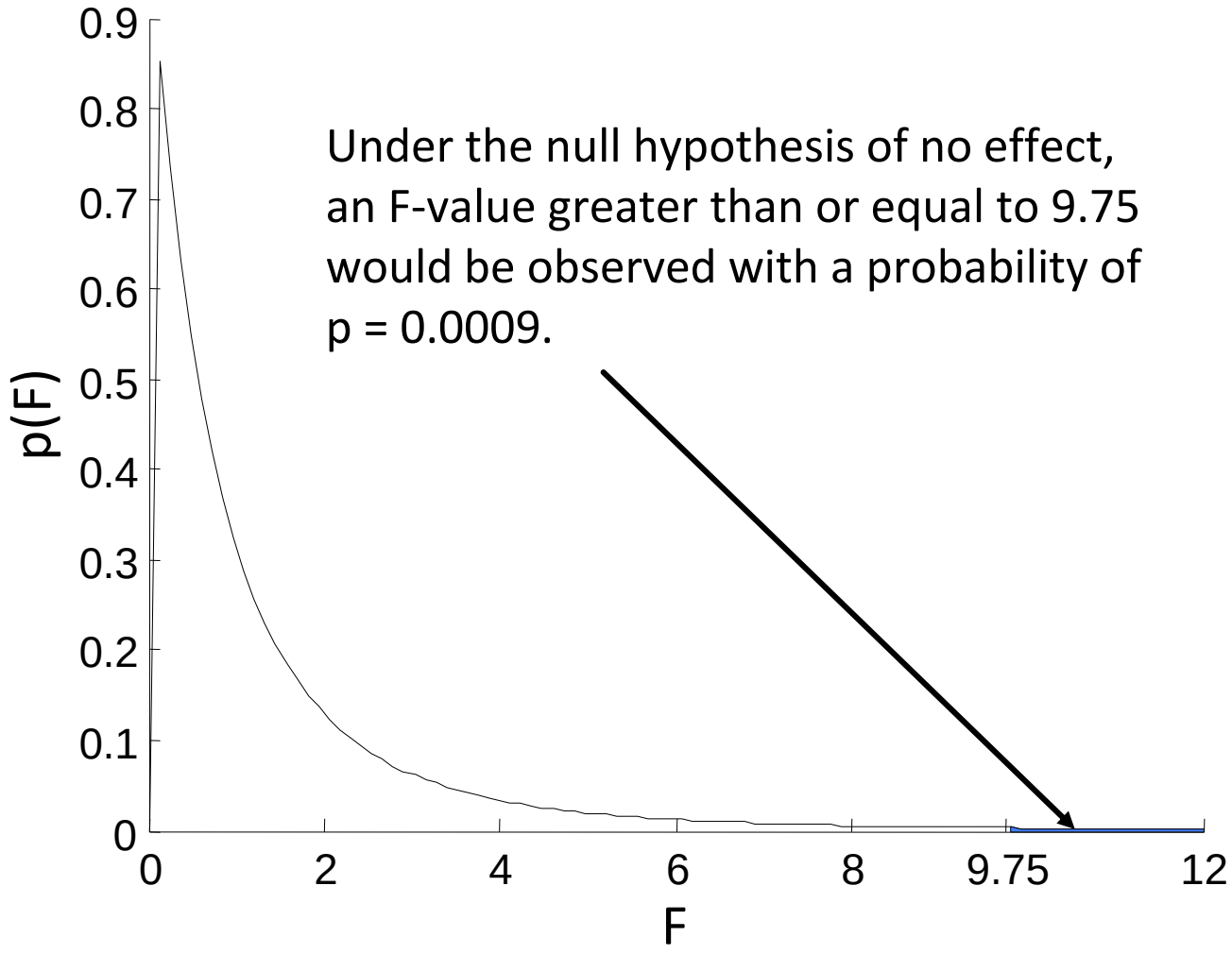
$$MS_{\text{within treatments}} = \frac{SS_{\text{within treatments}}}{df_{\text{within treatments}}} = \frac{8.6667}{6} = 1.4444$$

$$MS_{\text{between treatments}} = \frac{SS_{\text{between treatments}}}{df_{\text{between treatments}}} = \frac{28.16}{2} = 14.08$$

$$F = \frac{MS_{\text{between treatments}}}{MS_{\text{within treatments}}} = \frac{14.08}{1.44} = 9.75$$



Worked Example



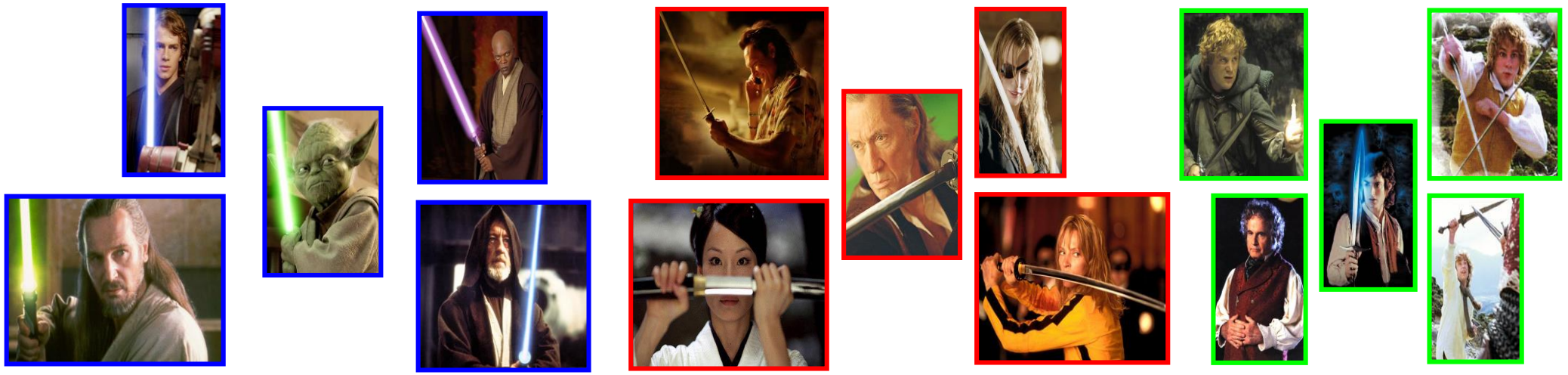
Under the null hypothesis of no effect, an F-value greater than or equal to 9.75 would be observed with a probability of p = 0.0009.

One-Way ANOVA

Three sword fighters are comparing their swords. Obi Wan Kenobi claims that light sabers are the best, Beatrix Kiddo swears by Hattori Hanzo katanas and Frodo Baggins prefers Elvish daggers

One-Way ANOVA

To determine which one is actually the best you recruit 5 Jedi knights, 5 deadly viper assassins and 5 hobbits. Each one is followed through several battles and you count the number of dismemberments/kills each fighter achieves with his/her respective sword



One-Way ANOVA

Light Saber	Katana	Dagger
6	6	0
8	5	4
5	9	0
4	4	1
2	6	0

Step 1: State the hypotheses, and specify the alpha level

- The null hypothesis states that there is no difference among the swords in terms of number of dismemberments/kills committed. In symbols, we would state
  - H_0: $\mu_{\text{Light Saber}} = \mu_{\text{Katana}} = \mu_{\text{Dagger}}$ (the type of sword used has no effect).

- There are a number of possible statements for the alternative hypothesis. The general alternative hypothesis is:
  - H_1: At least one of the mean numbers of dismemberments/kills is different.

  That is, the type of sword has an effect on the number of dismemberments/kills.

- For this test, we'll use $\alpha = 0.05$.

Step 2: Locate the critical region

- For a one-way independent-measures ANOVA the critical region depends on two values: $df_{\text{between treatments}}$ and $df_{\text{within treatments}}$.
  - $df_{\text{between treatments}} = k - 1 = 3 - 1 = 2$
  - $df_{\text{within treatments}} = N - k = 15 - 3 = 12$

- "k" is the number of *k*ategories (number of different swords).

- "N" is the total Number of cells in your table (3 categories x 5 subjects per category = 15).

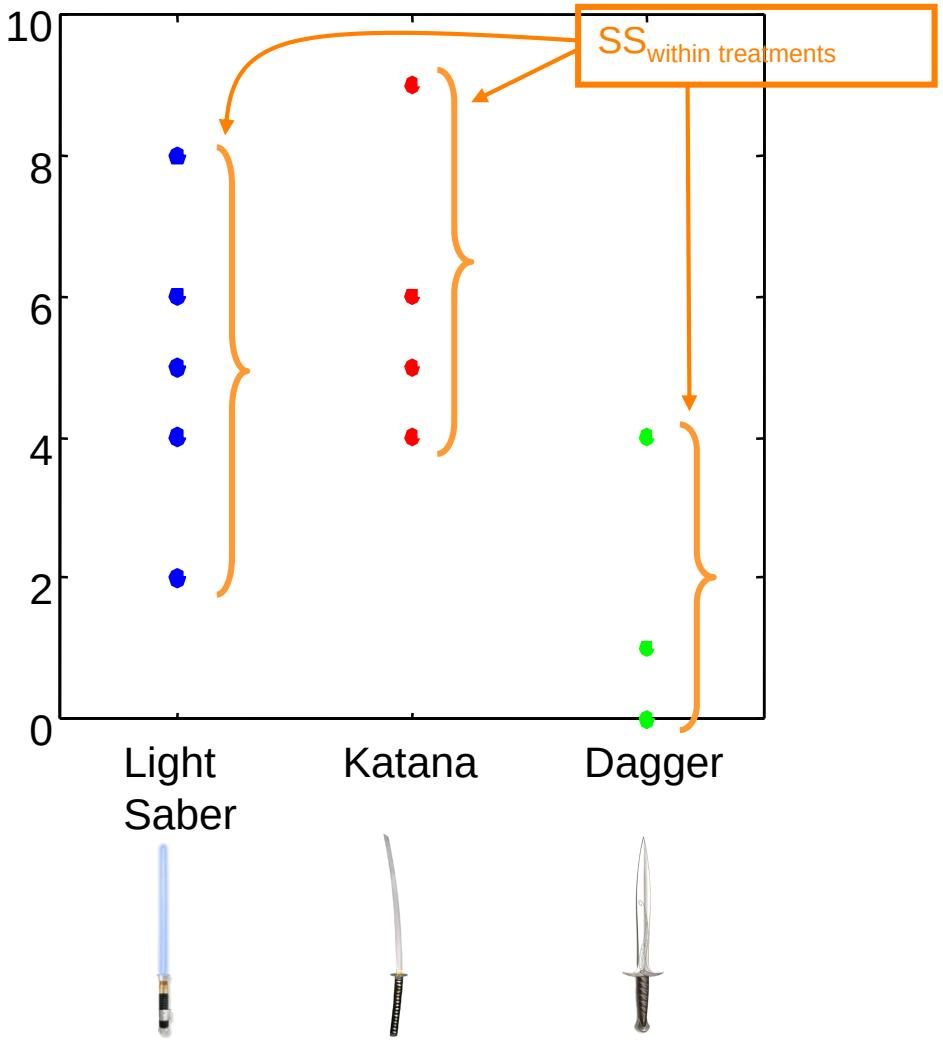
- The F-ratio for this problem will have df = 2, 12. Now we must consult the F-distribution table for df = 2 in the numerator and df = 12 in the denominator.

Step 3: Perform the analysis

The analysis involves the following steps:

1. Perform the analysis of SS
2. Calculate mean squares
3. Calculate the F-ratio

Step 3.1: Perform the analysis of SS



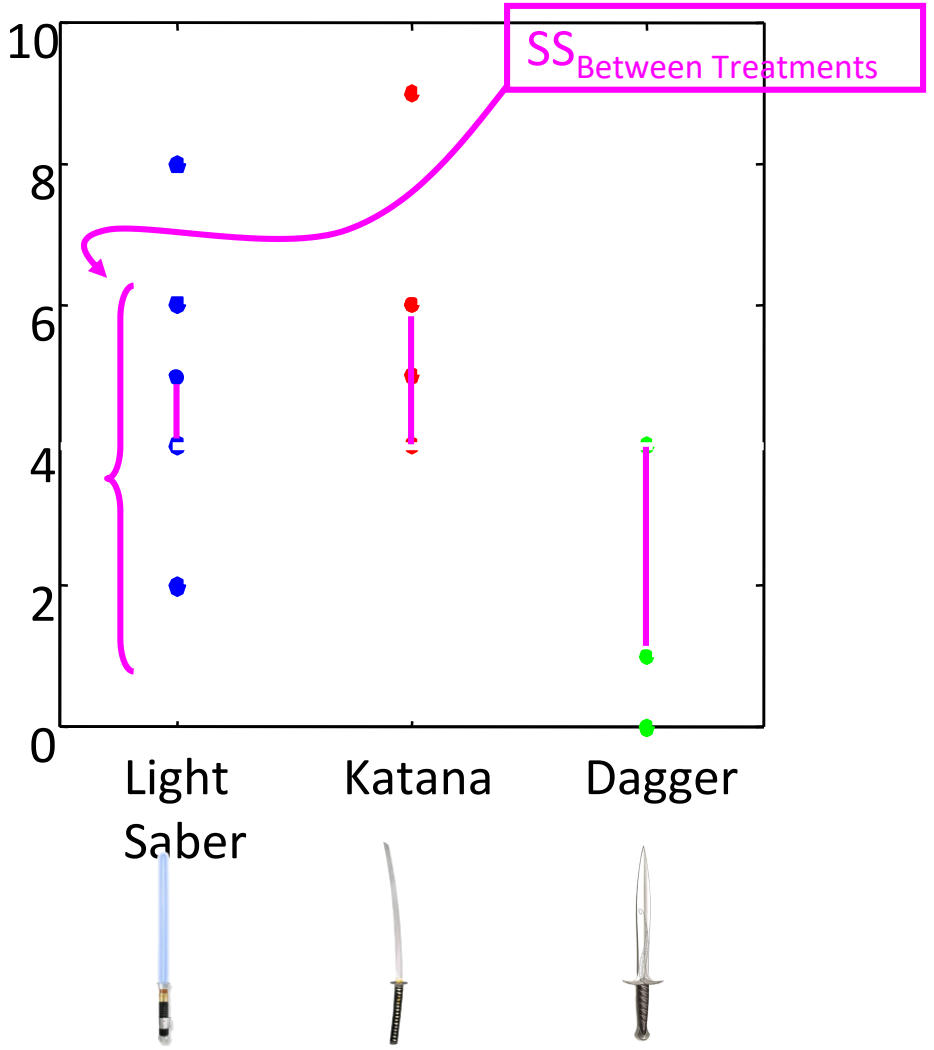
Step 3.1: Perform the analysis of SS

$$SS_{\text{within treatments}} = \sum SS_{\text{inside each treatment}} \qquad SS_{\text{inside each treatment}} = \sum_{i=1}^{n} (x_i - M)^2$$

Light Saber	$(x_i - M)^2$	Katana	$(x_i - M)^2$	Dagger	$(x_i - M)^2$
6	$(6-5)^2=1$	6	0	0	1
8	9	5	1	4	9
5	0	9	9	0	1
4	1	4	4	1	0
2	9	6	0	0	1
M = 5	SS = 20	M = 6	SS = 14	M = 1	SS = 12

$$SS_{\text{within treatments}} = SS_{\text{Light Saber}} + SS_{\text{Katana}} + SS_{\text{Dagger}} = 20 + 14 + 12 = 46$$

Step 3.1: Perform the analysis of SS



Step 3.1: Perform the analysis of SS

Light Saber	Katana	Dagger
6	6	0
8	5	4
5	9	0
4	4	1
2	6	0
M = 5	M = 6	M = 1

$$SS_{\text{between treatments}} = \sum_{i=1}^{k} n_i (M_i - M_G)^2$$

$$= 5 \cdot (5-4)^2 + 5 \cdot (6-4)^2 + 5 \cdot (1-4)^2$$

$$= 70$$

- n_i = number of elements in treatment group "*i*."
- M_i = mean for treatment group "*i*."
- M_G = grand mean.
- k = number of treatment groups.
- N = total number of scores.

$$M_G = \frac{\sum_{i-1}^{N} x_i}{N} = \frac{6+6+0+8+5+4+5+9+0+4+4+1+2+6+0}{15} = 4$$

Step 3.2: Calculate mean squares

$$MS_{\text{between treatments}} = \frac{SS_{\text{between treatments}}}{df_{\text{between treatments}}} = \frac{70}{2} = 35$$

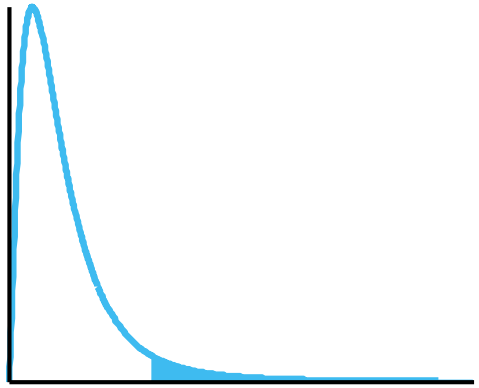
$$MS_{\text{within treatments}} = \frac{SS_{\text{within treatments}}}{df_{\text{within treatments}}} = \frac{46}{12} = 3.83$$

Step 3.3: Calculate the F-ratio

$$F = \frac{MS_{\text{between treatments}}}{MS_{\text{within treatments}}} = \frac{35}{3.83} = 9.14$$

Step 2: Locate the critical region

- Table entries in lightface type are critical values for the 0.05 level of significance.
- Boldface type values are for the 0.01 level of significance.



	Degrees of freedom: numerator				
Degrees of freedom: denominator	1	2	3	4	…
11	4.84 9.65	3.98 7.20	3.59 6.22	3.36 5.67	
12	4.75 9.33	3.88 6.93	3.49 5.95	3.26 5.41	
13	4.67 9.07	3.80 6.70	3.41 5.74	3.18 5.20	
…					

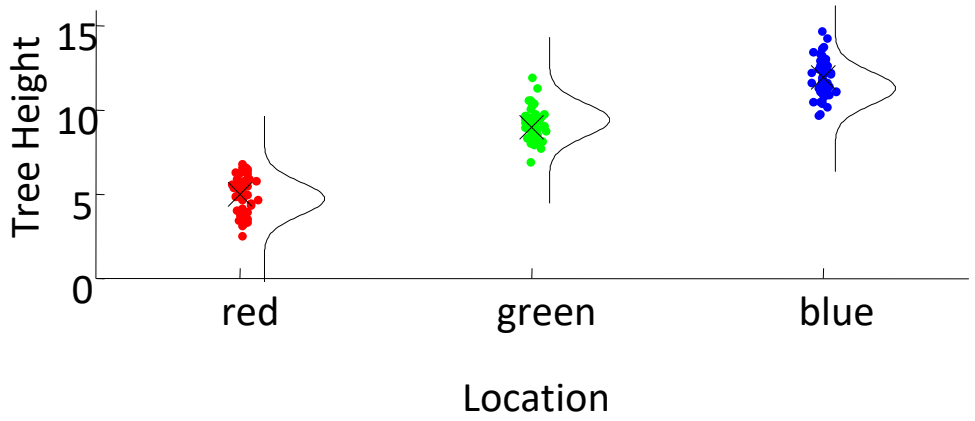
Step 4: Make a decision about H_0, and state a conclusion

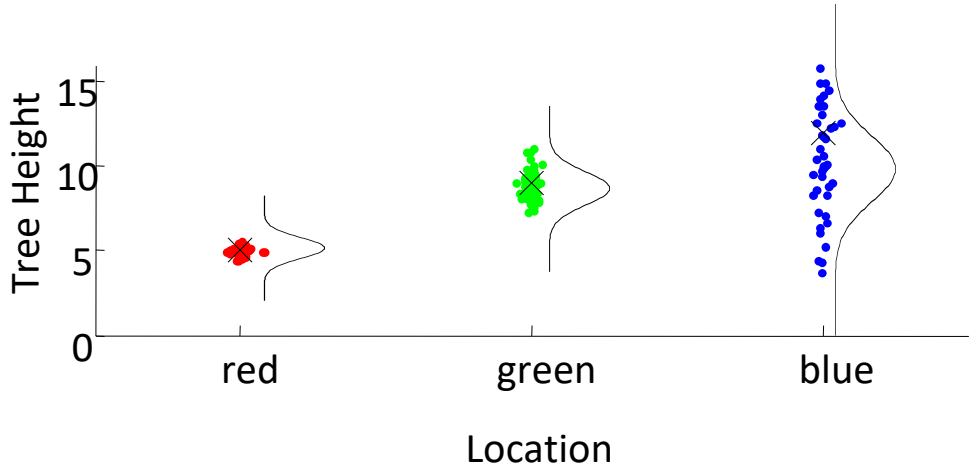
The obtained *F* of 9.14 exceeds the critical value of 3.88. Therefore, we can reject the null hypothesis. The type of sword used has a significant effect on the number of dismemberments/kills achieved, $F(2,12) = 9.14$, $p < 0.05$.

Source	SS	df	MS	F
Between treatments	70	2	35	9.14
Within Treatments	46	12	3.83	

Assumptions of ANOVA

- Each sample is assumed to have been obtained independently from all other samples
- Homogeneity of variance, i.e. homoscedastic and not heteroscedastic
- This assumption may be checked using Hartley's F-max test.

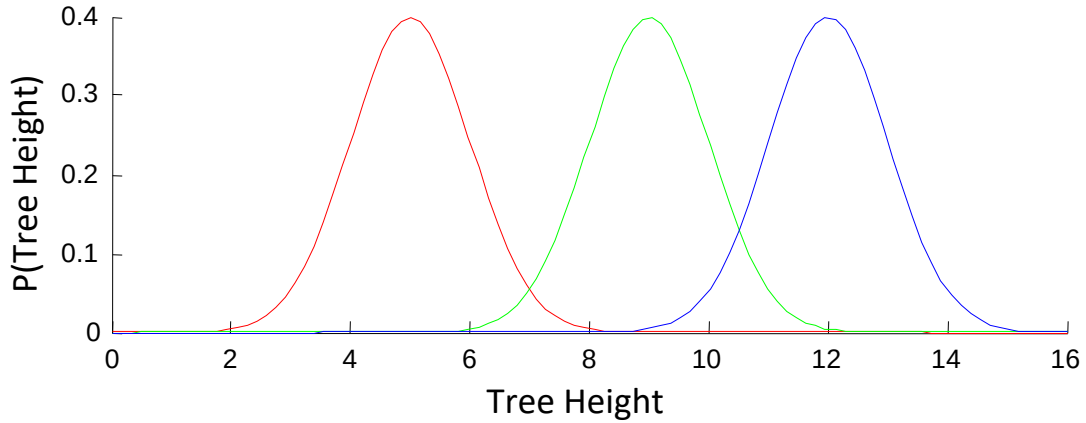


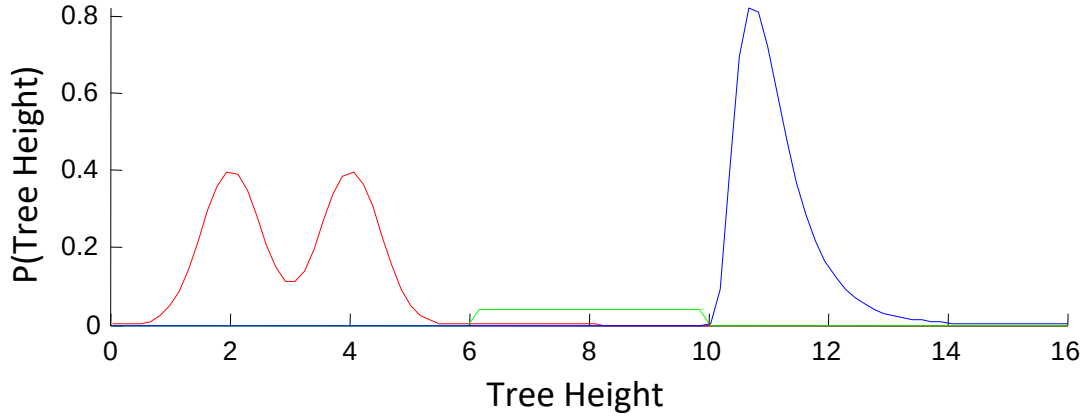


Assumptions of ANOVA

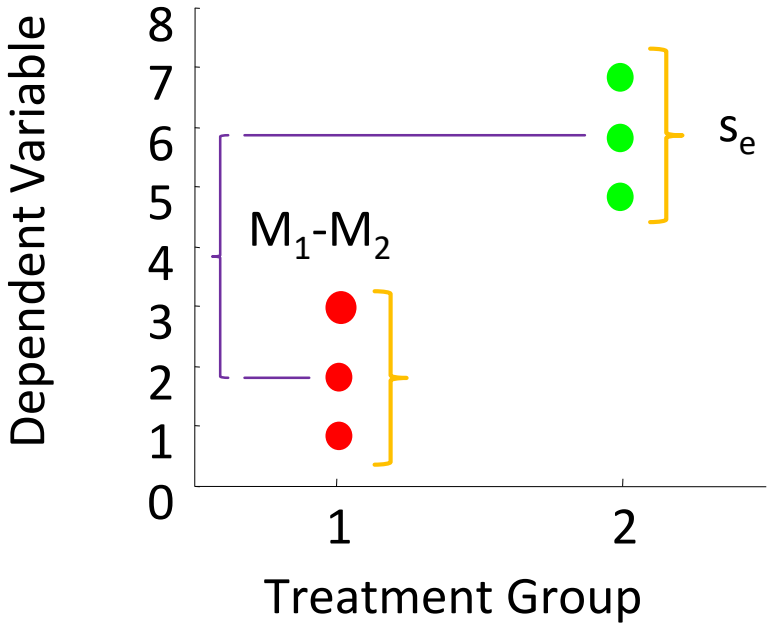
Normally distributed populations:

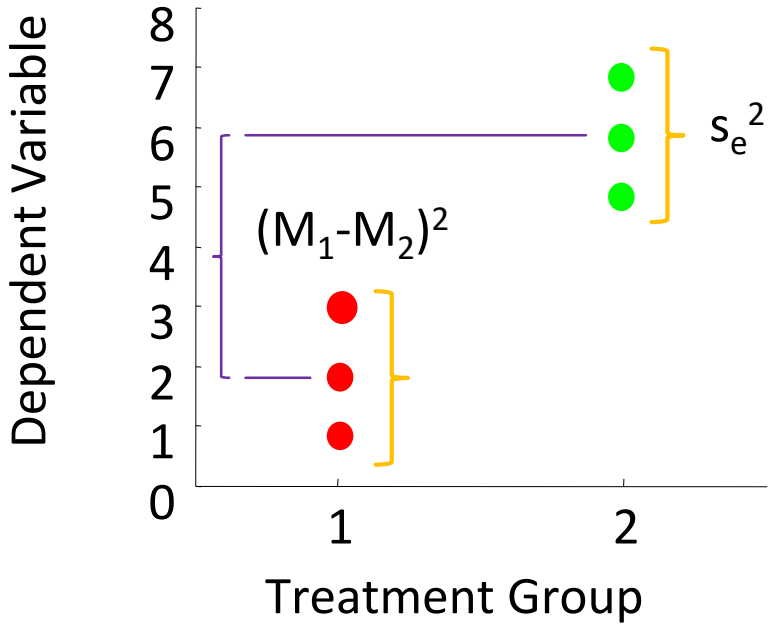
- The populations from which each treatment group is sampled is assumed to follow a normal (Gaussian) distribution.
- This assumption may be checked using the Kolmogorov-Smirnov test.





Relationship Between *t* and *F*





$$F=t^2$$

Worked Example

Effect size:

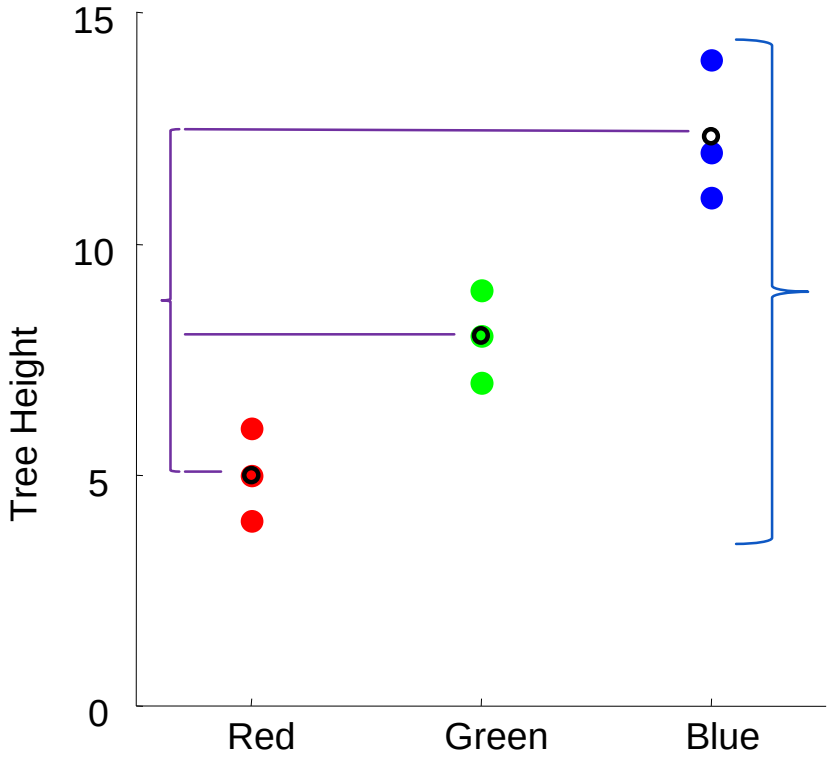
$h^2 = SS_{\text{between treatments}}/SS_{\text{total}}$

$= 216/3108$

$= 0.069$

Thus 6.9% of the total variability in the data is accounted for by the region from which the trees are sampled.

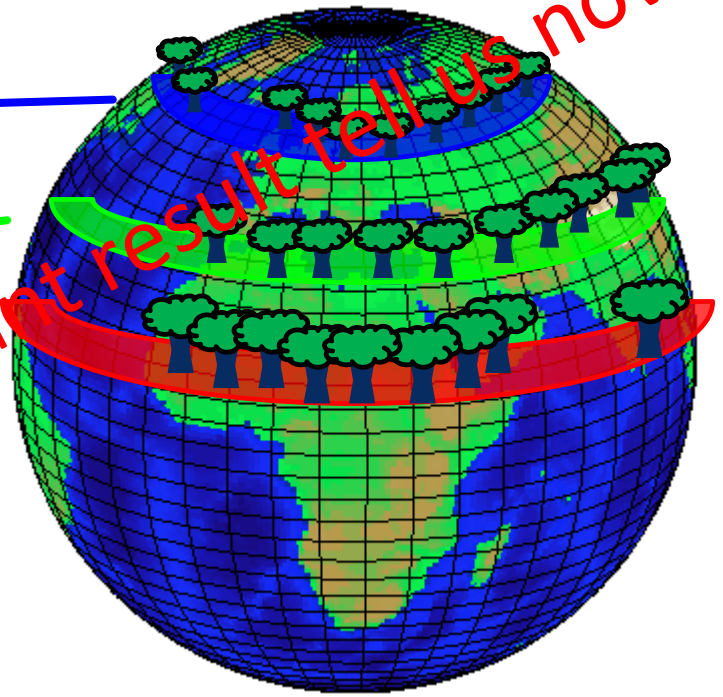
h^2	Cohen's Guideline
0-0.01	Small Effect
0.01-0.09	Medium Effect
0.25-1	Large Effect



Worked Example

Tree Height:

5	4	6
8	7	9
12	11	14



And what does a significant result tell us now?

What the ANOVA Does *Not* Tell You

If the null hypothesis is rejected, the ANOVA does not tell you how many regions contain trees with different heights – it could be that two regions are the same and one region differs, or it could be that all three regions differ. To know this, post-hoc tests are necessary.

Post-Hoc Tests

- Post-hoc tests tell you which regions contain differing tree heights.
- They allow you to compare tree heights in a pair-wise fashion.
- Since we already know that the overall ANOVA is significant we are guaranteed not to exceed a type I error rate of $\alpha = 0.05$.

Post-Hoc Tests

Post-Hoc Comparisons

- Many methods, eg:
  - Tukey's HSD test
  - Scheffé test

Scheffé Test

- Maintains Type I error rate at α across comparisons – for all possible comparisons (not just all pairwise).
- Most conservative post-hoc test – differences must be larger to be significant.
- Cost of cautiousness: reduced power.
- Approach: calculate *F* ratio for every comparison of interest → evaluate against F_{crit} for the original ANOVA.

Scheffé Test: Step 1

Light Saber	Katana	Dagger
6	6	0
8	5	4
5	9	0
4	4	1
2	6	0
T = 25	T = 30	T = 5

Light Sabers versus Katanas:

$$SS_{between} = \sum_{i=1}^{2} \frac{T_i^2}{n_i} - \frac{G_{1,2}^2}{N_{1,2}}$$

$$= \sum_{\text{Light Saber \& Katana}} \frac{T^2}{n} - \frac{G_{\text{Light Saber \& Katana}}^2}{N_{\text{Light Saber \& Katana}}}$$

$$= \left(\frac{25^2}{5} + \frac{30^2}{5} \right) - \frac{(25+30)^2}{(5+5)}$$

$$= 2.5$$

Scheffé Test: Step 1

Light Saber	Katana	Dagger
6	6	0
8	5	4
5	9	0
4	4	1
2	6	0
T = 25	T = 30	T = 5

Light Sabers versus Daggers:

$$SS_{between} = \sum_{i=1}^{2} \frac{T_i^2}{n_i} - \frac{G_{1,2}^2}{N_{1,2}}$$

$$= \sum_{\text{Light Saber \& Dagger}} \frac{T^2}{n} - \frac{G_{\text{Light Saber \& Dagger}}^2}{N_{\text{Light Saber \& Dagger}}}$$

$$= \left(\frac{25^2}{5} + \frac{5^2}{5} \right) - \frac{(25+5)^2}{(5+5)}$$

$$= 40$$

Scheffé Test: Step 1

Light Saber	Katana	Dagger
6	6	0
8	5	4
5	9	0
4	4	1
2	6	0
T = 25	T = 30	T = 5

Katanas versus Daggers:

$$SS_{between} = \sum_{i=1}^{2} \frac{T_i^2}{n_i} - \frac{G_{1,2}^2}{N_{1,2}}$$

$$= \sum_{\text{Katana \& Dagger}} \frac{T^2}{n} - \frac{G_{\text{Katana \& Dagger}}^2}{N_{\text{Katana \& Dagger}}}$$

$$= \left(\frac{30^2}{5} + \frac{5^2}{5} \right) - \frac{(30+5)^2}{(5+5)}$$

$$= 62.5$$

Scheffé Test: Steps

Light saber
6
8
5
4
2
Sum = 25
M = 5

Katana
6
5
9
4
6
Sum = 30
M = 6

Elvish dagger
0
4
0
1
0
Sum = 5
M = 1

$$df_{between} = 2$$

$$MS_{within} = 3.83$$

From the original ANOVA table

$$SS_{between} = \sum_{k \in comp} n_k \left(M_i - G_{comp}\right)^2$$

$$MS_{between} = \frac{SS_{between}}{df_{between}}$$

$$F = \frac{MS_{between}}{MS_{within}}$$

Scheffé Test: Steps

1. Calculate the grand mean for each comparison (G_{Comp})

2. Compute $SS_{between} = \sum_{k \in comp} n_k (M_i - G_{comp})^2$

3. Calculate $MS_{between}$ for the comparison*: $MS_{between} = \frac{SS_{between}}{df_{between}}$

4. Calculate *F* for the comparison: $F = \frac{MS_{between}}{MS_{within}}$

*$df_{between}$ is the $df_{between}$ from the original ANOVA table (ie $df_{between} = 2$)

Scheffé Test

Light saber
6
8
5
4
2
Sum = 25
M = 5

Katana
6
5
9
4
6
Sum = 30
M = 6

Elvish dagger
0
4
0
1
0
Sum = 5
M = 1

Light saber versus Katana

$$G_{comp} = \frac{25+30}{10} = 5.5$$

$$SS_{between} = 5\,(5-5.5)^2 + 5\,(6-5.5)^2 = 2.5$$

$$MS_{between} = \frac{2.5}{2} = 1.25$$

$$F = \frac{1.25}{3.83} = 0.3264$$

Scheffé Test

Light saber
6
8
5
4
2
Sum = 25
M = 5

Katana
6
5
9
4
6
Sum = 30
M = 6

Elvish dagger
0
4
0
1
0
Sum = 5
M = 1

Light saber versus Elvish dagger

$$G_{comp} = \frac{25+5}{10} = 3$$

$$SS_{between} = 5\,(5-3)^2 + 5\,(1-3)^2 = 40.0$$

$$MS_{between} = \frac{40}{2} = 20$$

$$F = \frac{20}{3.83} = 5.2219$$

Scheffé Test

Light saber
6
8
5
4
2
Sum = 25
M = 5

Katana
6
5
9
4
6
Sum = 30
M = 6

Elvish dagger
0
4
0
1
0
Sum = 5
M = 1

Katana versus Elvish dagger

$$G_{comp} = \frac{30+5}{10} = 3.5$$

$$SS_{between} = 5\,(6-3.5)^2 + 5\,(1-3.5)^2 = 62.5$$

$$MS_{between} = \frac{62.5}{2} = 31.25$$

$$F = \frac{31.25}{3.83} = 8.1593$$

Scheffé Test: Step 2

- Calculate $MS_{between}$ for the comparison:

$$MS_{between} = \frac{SS_{between}}{df_{between}}$$

- $df_{between}$ is the $df_{between}$ from the original ANOVA table (ie $df_{between} = 2$)

$$MS_{\text{between: Light Saber vs Katana}} = \frac{SS_{\text{between: Light Saber vs Katana}}}{df_{\text{between}}} = \frac{2.5}{2} = 1.25$$

$$MS_{\text{between: Light Saber vs Dagger}} = \frac{SS_{\text{between: Light Saber vs Dagger}}}{df_{\text{between}}} = \frac{40}{2} = 20$$

$$MS_{\text{between: Katana vs Dagger}} = \frac{SS_{\text{between: Katana vs Dagger}}}{df_{\text{between}}} = \frac{62.5}{2} = 31.25$$

Scheffé Test: Step 3

- Calculate *F* for the comparison.

$$F = \frac{MS_{between}}{MS_{within}}$$

$$F_{\text{Light Saber vs Katana}} = \frac{MS_{\text{Light Saber vs Katana}}}{MS_{\text{within}}} = \frac{1.25}{3.83} = 0.3264$$

$$F_{\text{Light Saber vs Dagger}} = \frac{MS_{\text{Light Saber vs Dagger}}}{MS_{\text{within}}} = \frac{20}{3.83} = 5.2219$$

$$F_{\text{Katana vs Dagger}} = \frac{MS_{\text{Katana vs Dagger}}}{MS_{\text{within}}} = \frac{31.25}{3.83} = 8.1593$$

Scheffé Test: Step 4

- Compare to F_{crit} and state conclusions
  - F_{crit} is the critical F-value from the original ANOVA (i.e. $F_{crit} = 3.88$).
  - $F_{\text{Light Saber vs Katana}} = 0.3624 < 3.88 \therefore$ N.S.
  - $F_{\text{Light Saber vs Dagger}} = 5.2219 > 3.88 \therefore$ Sig.
  - $F_{\text{Katana vs Dagger}} = 8.1593 > 3.88 \therefore$ Sig.

- Significant differences in the number of dismemberments/kills were found between Light Sabers and Daggers and between Katanas and Daggers ($p<0.05$).

Scheffé Test

- **Recap**:
  - Compute ANOVA.
  - If there is a significant overall effect, compute new $SS_{between}$ for each specific comparison.
  - **Step 1**: $SS_{between}$ is $SS_{between} = \sum \frac{T^2}{n} - \frac{G^2}{N}$ where T, n, G, and N are computed only from the groups in the comparison (make believe the other groups do not exist).
  - **Step 2**: Compute new $MS_{between}$ as $SS_{between}$ / $df_{between}$, where $df_{between}$ is the original $df_{between}$ from the ANOVA table.
  - **Step 3**: Compute new *F* as $MS_{between}$ / MS_{within}.
  - **Step 4**: Lookup F_{crit} (same as original ANOVA) and compare to new *F*.

Repeated measures ANOVA

One-Way Repeated Measures ANOVA

Same idea as with repeated measures t-test

Two way ANOVA

The Problem

- After becoming a scientist you are involved in a laboratory accident that results in you acquiring super powers.
- You realize that with great power comes great responsibility so you decide to dedicate your life to fighting crime.
- You have many loved ones who you don’t want your enemies to hurt, so you decide you’re going to need a costume to conceal your true identity.

The Problem

- You examine several other super hero costumes and notice that they generally are made of either spandex, cotton or leather. Each of these materials has different advantages (promotes circulation, breathes, durability) but you want to know which one is conducive to catching the most evil villains.

- Also, you want to find the time of day you can fight crime at that will let you catch the most evil villains.

The Set Up

- You gather 30 of your super hero friends and divide them into 3 costume groups – spandex, cotton and leather. Within each of these groups, half of the subjects fight crime during the day while the other half fight crime during the night. You measure the number of evil villains caught by each super hero over a 1 month period.

The Data

	Spandex	Cotton	Leather
	18	10	3
	10	8	5
	16	12	1
	12	6	7
	14	14	9
	5	6	15
	7	14	13
	3	10	17
	9	8	11
	1	12	19

The Data

	Spandex	Cotton	Leather
Day 	18	10	3
	10	8	5
	16	12	1
	12	6	7
	14	 14	 9
 Night	5	6	15
	7	14	13
	3	10	17
	9	8	11
	1	12	19

Step 1: State the hypotheses and select α

- For the costumes, the null hypothesis states that there is no difference in number of villains caught while wearing costumes made of any of the three materials.

- H_0: $\mu_{spandex} = \mu_{cotton} = \mu_{leather}$

- H_1: At least one of the mean number of villains caught differs for one of the costume types.

Step 1: State the hypotheses and select α

- For the time of day for crime fighting the null hypothesis states that there is no difference in the mean number of villains caught in the day or at night.

- H_0: $\mu_{day} = \mu_{night}$

- The alternative hypothesis states that the mean number of villains caught during the day is different from the mean number of villains caught at night.

- H_1: $\mu_{day} \neq \mu_{night}$

Step 1: State the hypotheses and select α

- For the interaction between costume type and time of day the null hypothesis can be stated in two different ways. First, the difference in number of villains caught between the different costume types will be the same over the different times of day. Second, the difference in number of villains caught between the different times of day will be the same over the different costume types. In more general terms:

- H_0: The effect of costume type does not depend on the time of day (and the effect of time of day does not depend on the costume type).

- H_1: The effect of one factor does depend on the level of the other factor.

Step 1: State the hypotheses and select α

We will adopt alpha value of $\alpha = 0.05$

Step 2: Locate the critical region

- $df_{\text{between treatments}}$ = (number of treatment groups) - 1 = 6 - 1 = 5
- $df_{\text{within treatments}} = \sum_{i=1}^{K}(n_i - 1)$ = (5-1) + (5-1) + (5-1) + (5-1) + (5-1) + (5-1) = 24
- df_{costume} = (number of different costumes) - 1 = 3 - 1 = 2

- df_{time} = (number of times of day) - 1 = 2 - 1 = 1
- $df_{\text{costume x time}} = df_{\text{between treatments}} - df_{\text{costume}} - df_{\text{time}} = 5 - 2 - 1 = 2$
- df_{total} = N - 1 = 30 - 1 = 29

- $F_{\text{crit-costume}}(2, 24) = 3.4028$
- $F_{\text{crit-time}}(1, 24) = 4.2597$
- $F_{\text{crit-costume x time}}(2, 24) = 3.4028$

Step 3: Compute the F-ratios

$$SS_{\text{between treatments}} = \sum_{k=1}^{K}\left(\frac{T_k^2}{n_k}\right) - \frac{G^2}{N}$$

- K = The total number of categories (i.e. spandex & day, spandex & night, cotton & day, cotton & night, leather & day, leather & night) = 6.
- k = The index for the categories (for spandex & day, k = 1, for spandex & night, k = 2, for cotton & day, k = 3, etc.).
- T_k = The total number of villains caught in category k.
- n_k = The number of subjects in category k.
- G = The grand total (i.e. the total number of villains caught over all categories).
- N = The total number of subjects over all categories.

Step 3: Compute the F-ratios

	Spandex	Cotton	Leather
Day	18	10	3
	10	8	5
	16	12	1
	12	6	7
	14	14	9
	$T_1 = 70$	$T_3 = 50$	$T_5 = 25$
Night	5	6	15
	7	14	13
	3	10	17
	9	8	11
	1	12	19
	$T_2 = 25$	$T_4 = 50$	$T_6 = 75$

Step 3: Compute the F-ratios

$$SS_{\text{between treatments}} = \sum_{k=1}^{K}\left(\frac{T_k^2}{n_k}\right) - \frac{G^2}{N}$$

$$= \left(\frac{70^2}{5} + \frac{25^2}{5} + \frac{50^2}{5} + \frac{50^2}{5} + \frac{25^2}{5} + \frac{75^2}{5}\right) - \frac{295^2}{30}$$

$$= 454.1667$$

Step 3: Compute the F-ratios

$$SS_{\text{within tre atments}} = \sum_{k=1}^{K} SS_k$$

- SS_k = The sum of squared deviations from the mean within category *k*.

$$SS_k = \sum_{i=1}^{n_k} X_{i,k}^2 - \frac{\left(\sum_{k=1}^{K} T_k\right)^2}{n_k}$$

- ✓ n_k = The number of subjects in category k.
- ✓ $X_{i,k}$ = Number of villains caught by subject *i* in category k.

Step 3: Compute the F-ratios

	Spandex		Cotton		Leather	
Day	18	324	10	100	3	9
	10	100	8	64	5	25
	16	256	12	144	1	1
	12	144	6	36	7	49
	14	196	14	196	9	81
	$T_1 = 70$	$\Sigma(X^2) = 1020$	$T_3 = 50$	$\Sigma(X^2) = 540$	$T_5 = 25$	$\Sigma(X^2) = 165$
Night	5	25	6	36	15	255
	7	49	14	196	13	169
	3	9	10	100	17	289
	9	81	8	64	11	121
	1	1	12	144	19	361
	$T_2 = 25$	$\Sigma(X^2) = 165$	$T_4 = 50$	$\Sigma(X^2) = 540$	$T_6 = 75$	$\Sigma(X^2) = 1165$

Step 3: Compute the F-ratios

$$SS_k = \sum_{i=1}^{n_k} X_{i,k}^2 - \frac{\left(\sum_{k=1}^{K} T_k\right)^2}{n_k}$$

$$SS_1 = 1020 - \frac{70^2}{5} = 40$$

$$SS_2 = 165 - \frac{25^2}{5} = 40$$

$$SS_3 = 540 - \frac{50^2}{5} = 40$$

$$SS_4 = 540 - \frac{50^2}{5} = 40$$

$$SS_5 = 165 - \frac{25^2}{5} = 40$$

$$SS_6 = 1165 - \frac{75^2}{5} = 40$$

$$SS_{\text{within tre atments}} = \sum_{k=1}^{K} SS_k$$

$$= 40 + 40 + 40 + 40 + 40 + 40$$

$$= 240$$

Step 3: Compute the F-ratios

$$SS_{\text{costume}} = \sum_{k=1}^{K_{\text{costume}}} \left(\frac{\mathrm{T}_k^2}{\mathrm{n}_k} \right) - \frac{G^2}{N}$$

- $K_{costume}$ = The number of different types of costumes we're testing ($K_{costume}$ = 3).
- T_k = The total number of villains caught (over all times of day) using costume k.
- n_k = The number of subjects wearing costume k (over all times of day).
- G = The grand total number of villains caught (over all subjects and all costumes).
- N = The total number of subjects used (over all costume types and times of day).

Step 3: Compute the F-ratios

	Spandex	Cotton	Leather
Day	18	10	3
	10	8	5
	16	12	1
	12	6	7
	14	14	9
Night	5	6	15
	7	14	13
	3	10	17
	9	8	11
	1	12	19
Costume Total	95	100	100

Step 3: Compute the F-ratios

$$SS_{\text{costume}} = \sum_{k=1}^{K_{\text{costume}}} \left(\frac{\mathrm{T}_{\mathrm{k}}^2}{\mathrm{n}_{\mathrm{k}}} \right) - \frac{G^2}{N}$$

$$= \left(\frac{T_{\text{spandex}}^2}{n_{\text{spandex}}} + \frac{T_{\text{cotton}}^2}{n_{\text{cotton}}} + \frac{T_{\text{leather}}^2}{n_{\text{leather}}} \right) - \frac{G^2}{N}$$

$$= \left(\frac{95^2}{10} + \frac{100^2}{10} + \frac{100^2}{10} \right) - \frac{295^2}{30}$$

$$= 1.67$$

Step 3: Compute the F-ratios

$$SS_{\text{time}} = \sum_{k=1}^{K_{\text{time}}} \left(\frac{\mathrm{T}_{\mathrm{k}}^2}{\mathrm{n}_{\mathrm{k}}} \right) - \frac{G^2}{N}$$

- K_{time} = The number of different times of day we're testing (K_{time} = 2).
- T_k = The total number of villains caught (over all costumes) at time of day k.
- n_k = The number of subjects tested at time of day k (over all costumes).
- G = The grand total number of villains caught (over all subjects and all times of day).
- N = The total number of subjects used (over all costume types and times of day).

Step 3: Compute the F-ratios

	Spandex	Cotton	Leather	Time of Day Total
Day	18	10	3	145
	10	8	5	
	16	12	1	
	12	6	7	
	14	14	9	
Night	5	6	15	150
	7	14	13	
	3	10	17	
	9	8	11	
	1	12	19	

Step 3: Compute the F-ratios

$$SS_{\text{time}} = \sum_{k=1}^{K_{\text{time}}} \left(\frac{\mathrm{T}_{\mathrm{k}}^2}{\mathrm{n}_{\mathrm{k}}} \right) - \frac{G^2}{N}$$

$$= \left(\frac{T_{\text{Day}}^2}{n_{\text{Day}}} + \frac{T_{\text{Night}}^2}{n_{\text{Night}}} \right) - \frac{G^2}{N}$$

$$= \left(\frac{145^2}{15} + \frac{150^2}{15} \right) - \frac{295^2}{30}$$

$$= 0.83$$

Step 3: Compute the F-ratios

$$SS_{\text{costume} \times \text{time of day}} = SS_{\text{between treatments}} - SS_{\text{costume}} - SS_{\text{time of day}}$$

$$= 454.17 - 1.67 - 0.83$$

$$= 451.67$$

Step 3: Compute the F-ratios

$$MS_{\text{costume}} = \frac{SS_{\text{costume}}}{df_{\text{costume}}} = \frac{1.67}{2} = 0.83$$

$$MS_{\text{time of day}} = \frac{SS_{\text{time of day}}}{df_{\text{time of day}}} = \frac{0.83}{1} = 0.83$$

$$MS_{\text{costume} \times \text{time of day}} = \frac{SS_{\text{costume} \times \text{time of day}}}{df_{\text{costume} \times \text{time of day}}} = \frac{451.67}{2} = 225.83$$

$$MS_{\text{within treatments}} = \frac{SS_{\text{within treatments}}}{df_{\text{within treatments}}} = \frac{240}{24} = 10$$

$$F_{\text{costume}} = \frac{MS_{\text{costume}}}{MS_{\text{within treatments}}} = \frac{0.83}{10} = 0.083$$

$$F_{\text{time of day}} = \frac{MS_{\text{time of day}}}{MS_{\text{within treatments}}} = \frac{0.83}{10} = 0.083$$

$$F_{\text{costume} \times \text{time of day}} = \frac{MS_{\text{costume} \times \text{time of day}}}{MS_{\text{within treatments}}} = \frac{225.83}{10} = 22.58$$

Step 4: Make a decision about each H_0, and state conclusions

- For the interaction of costume type with time of day, the obtained F-ratio, F = 22.58, exceeds the critical value of F = 3.4028. Therefore, we reject the null hypothesis. We conclude that there is a significant interaction between costume type and time of day in predicting number of evil villains caught.

- Since we found a significant interaction we disregard the main effects (i.e. since the effects of costume type depend on the time of day there's no point in looking at the average effects of costume type averaged over the different times of day because we'll be losing valuable information if we do).

Step 4: Make a decision about each H_0, and state conclusions

Source	SS	df	MS	F	p
Between Treatments	454.17	5			
Costume	1.67	2	0.83	0.08	0.92
Time of Day	0.83	1	0.83	0.08	0.78
Costume x Time of Day Interaction	451.67	2	225.83	22.58	.000003
Within Treatments	240	24	10		
Total	694.17	29			

Effect Size

$$\eta^2_{\text{costume}} = \frac{SS_{\text{costume}}}{SS_{\text{total}} - SS_{\text{time of day}} - SS_{\text{costume} \times \text{time of day}}} = \frac{SS_{\text{costume}}}{SS_{\text{costume}} + SS_{\text{within treatments}}}$$

$$= \frac{1.67}{694.17 - 0.83 - 451.67} = \frac{1.67}{1.67 + 240}$$

$$= 0.007$$

$$\eta^2_{\text{time of day}} = \frac{SS_{\text{time of day}}}{SS_{\text{total}} - SS_{\text{costume}} - SS_{\text{costume} \times \text{time of day}}} = \frac{SS_{\text{time of day}}}{SS_{\text{time of day}} + SS_{\text{within treatments}}}$$

$$= \frac{0.83}{694.17 - 1.67 - 451.67} = \frac{0.83}{0.837 + 240}$$

$$= 0.003$$

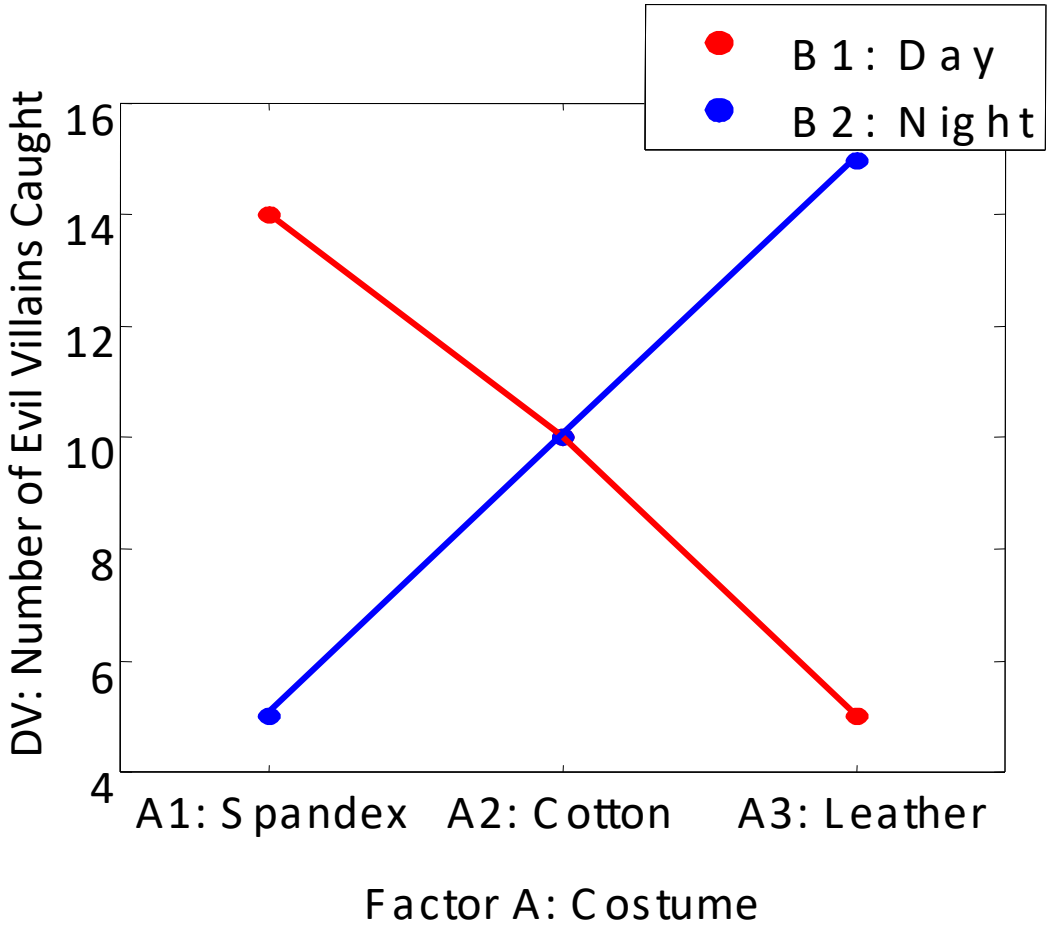
$$\eta^2_{\text{costume} \times \text{time of day}} = \frac{SS_{\text{costume} \times \text{time of day}}}{SS_{\text{total}} - SS_{\text{costume}} - SS_{\text{time of day}}} = \frac{SS_{\text{costume} \times \text{time of day}}}{SS_{\text{costume} \times \text{time of day}} + SS_{\text{within treatments}}}$$

$$= \frac{451.67}{694.17 - 1.67 - 0.83} = \frac{451.67}{451.67 + 240}$$

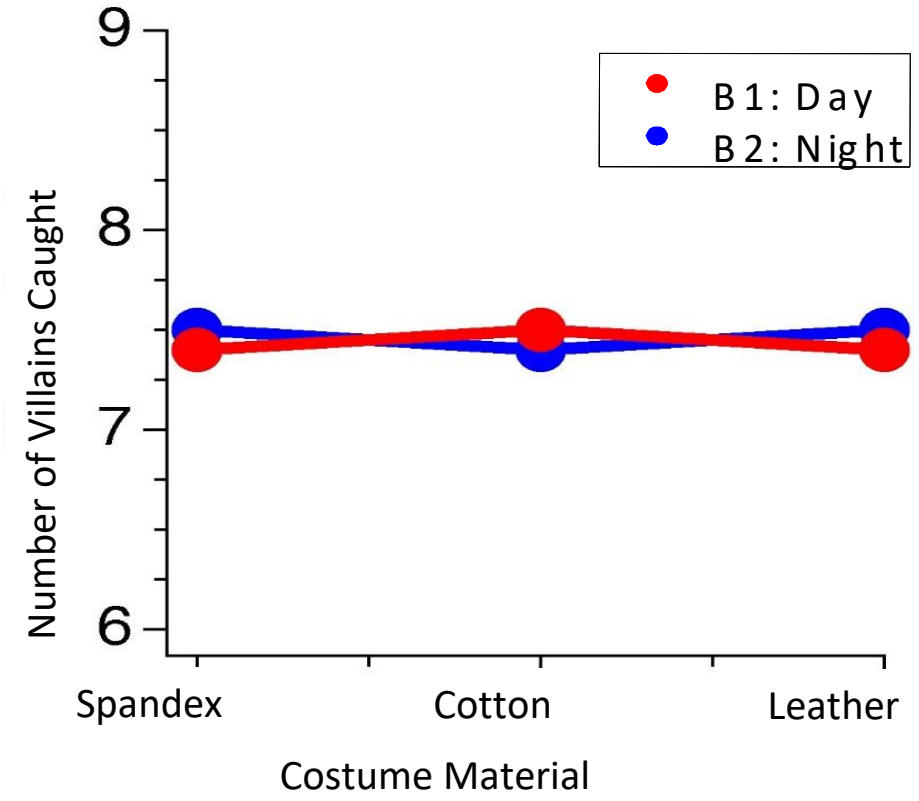
$$= 0.65$$

Graphing 2-way Designs

- Good first step: graph the data
- Infer main effects and interactions from the data
- Still need to run appropriate statistic tests

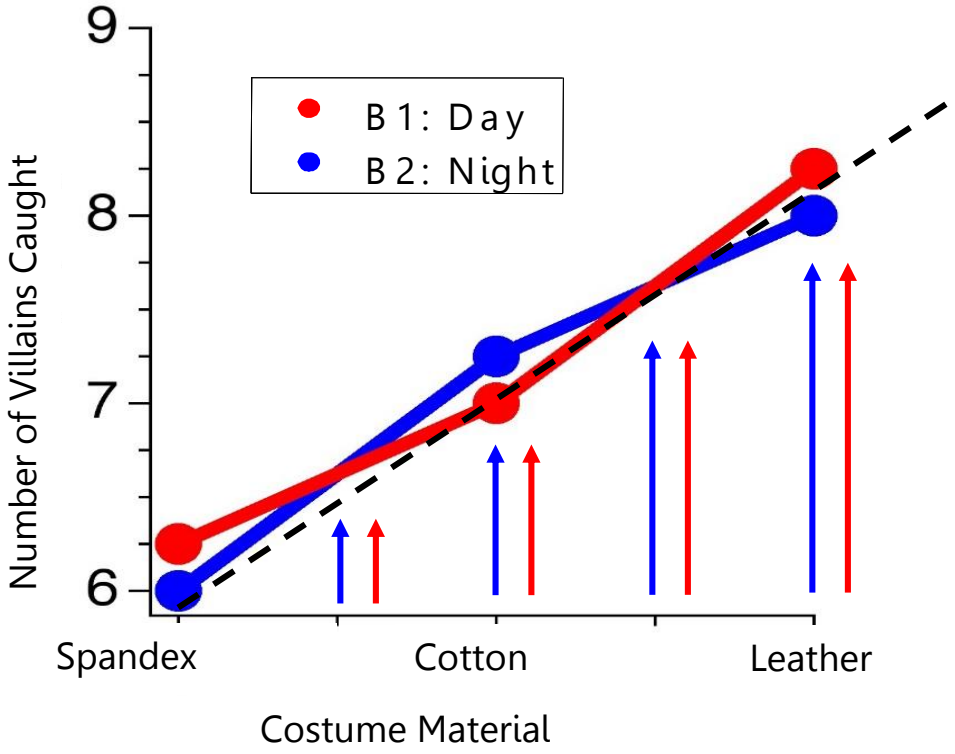


Some Possibilities …



- No main effect of costume material.
- No main effect of time.
- No interactions.

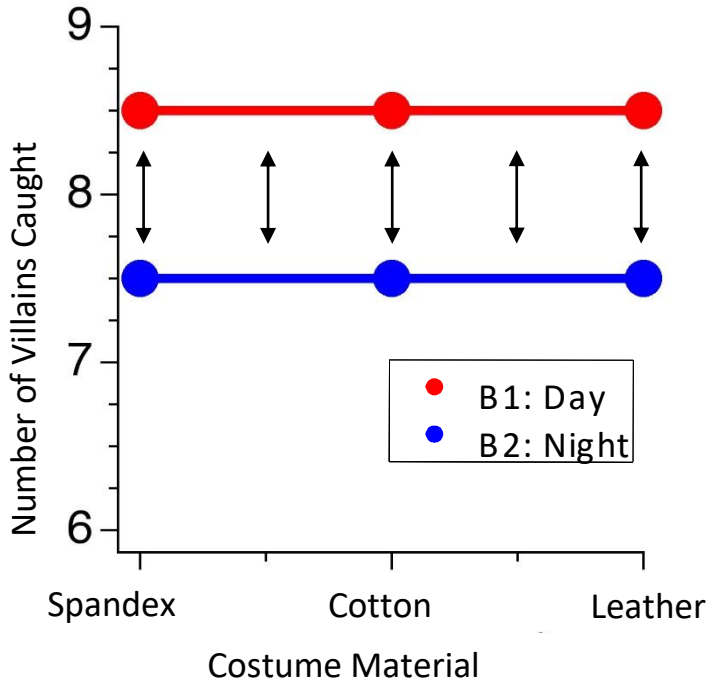
Some Possibilities …



The slope across levels of the first factor (costume material) determine the effect.

- ✓ **Main effect of costume material**
- ✓ No main effect of time.
- ✓ No interactions.

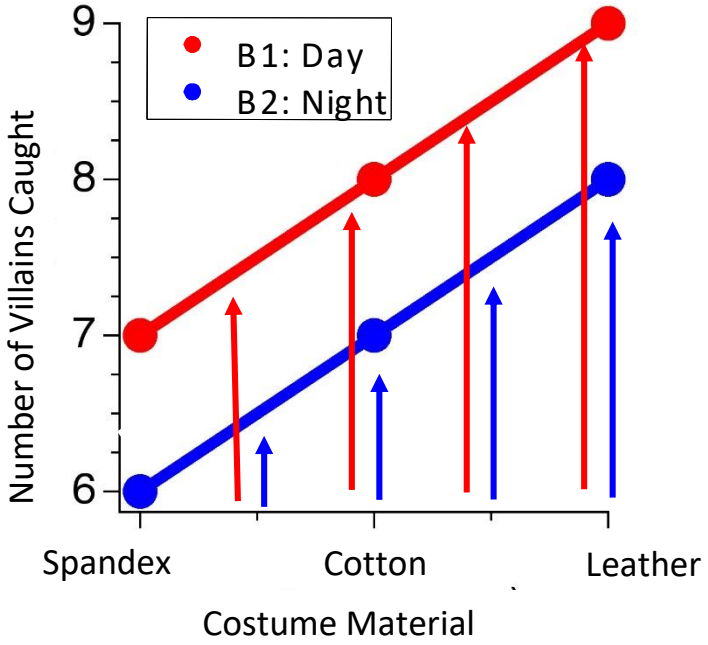
Some Possibilities …

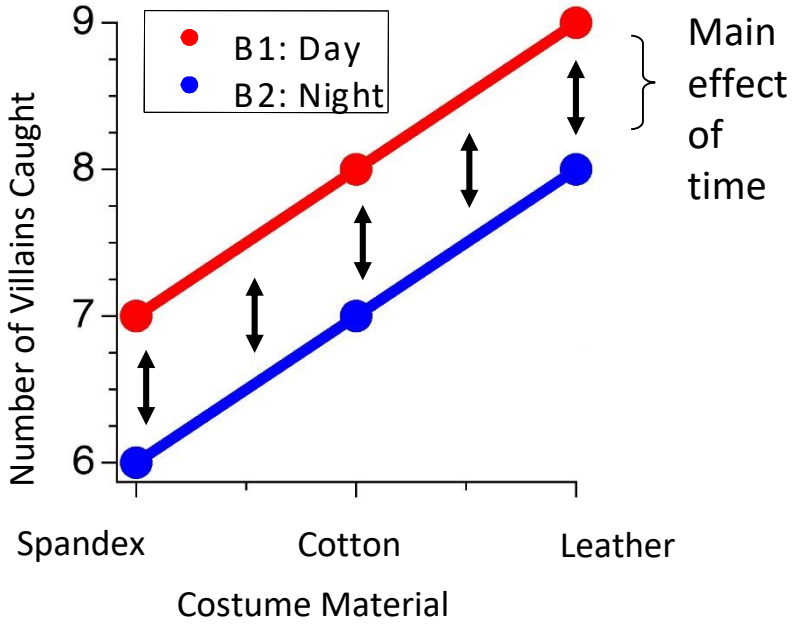


The separation between levels of the second factor (time) determines effect.

- ✓ No main effect of costume material
- ✓ **Main effect of time**.
- ✓ No interactions.

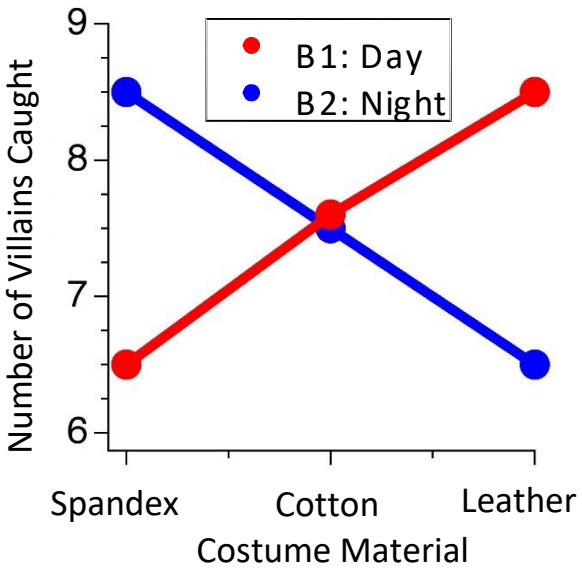
Some Possibilities …





- ✓ **Main effect of costume material.**
- ✓ **Main effect of time**.
- ✓ No interactions.

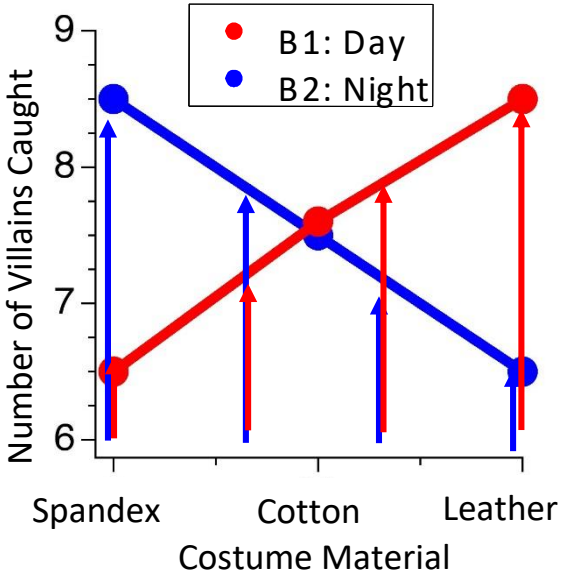
Some Possibilities …



✓ **Interaction!**

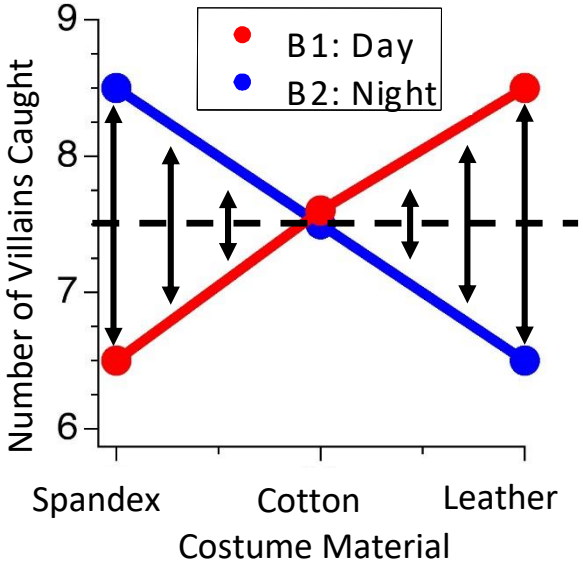
✓ The levels of one factor influence the effect of the other factor.

✓E.g. Time influences whether costume material leads to less or more villains caught.



✓ No main effect of Exercise.

✓ It just happens that, in this example, costume material has exactly opposite effects at day and at night -> **Interaction!**



✓ No main effect of time.

✓ It just happens that, in this example, time has exactly opposite effects across costume types -> **Interaction!**

How badly things can go wrong

- A large variety of people are not able to copy statistical values (t, p, df) correctly from the statistics program to the publication.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	216.425	2	108.212	.449	.649
Within Groups	2891.863	12	240.989		
Total	3108.288	14			

Behavior Research Methods
pp 1-22

First online: 23 October 2015

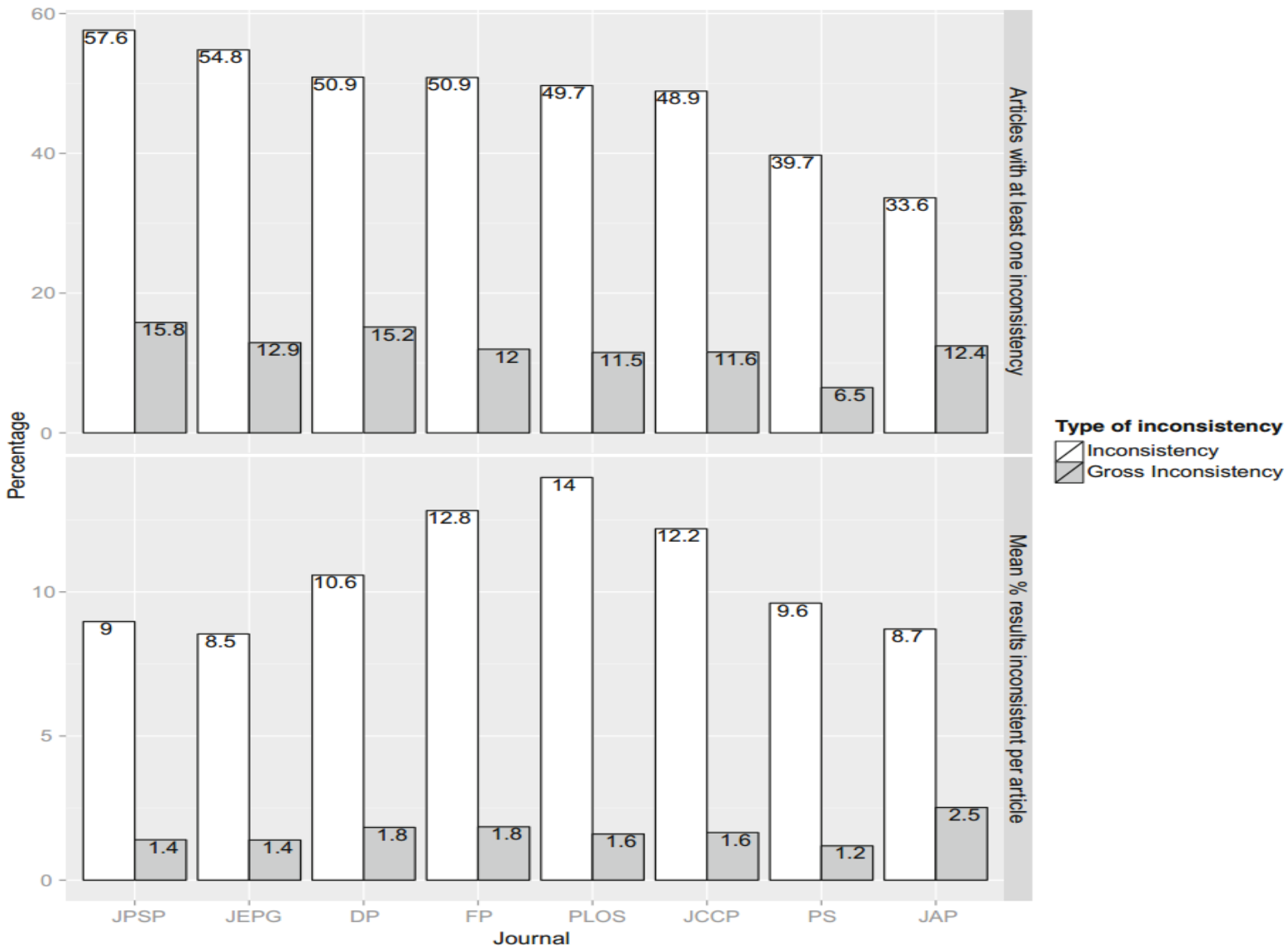
The prevalence of statistical reporting errors in psychology (1985–2013)

Michèle B. Nuijten , Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp, Jelte M. Wicherts

10.3758/s13428-015-0664-2

Abstract

This study documents reporting errors in a sample of over 250,000 *p*-values reported in eight major psychology journals from 1985 until 2013, using the new R package "statcheck." statcheck retrieved null-hypothesis significance testing (NHST) results from over half of the articles from this period. In line with earlier research, we found that half of all published psychology papers that use NHST contained at least one *p*-value that was inconsistent with its test statistic and degrees of freedom. One in eight papers contained a grossly inconsistent *p*-value that may have affected the statistical conclusion. In contrast to earlier findings, we found that the average prevalence of inconsistent *p*-values has been stable over the years or has declined. The prevalence of gross inconsistencies was higher in *p*-values reported as significant than in *p*-values reported as nonsignificant. This could indicate a systematic bias in favor of significant results. Possible solutions for the high prevalence of reporting inconsistencies could be to encourage sharing data, to let co-authors check results in a so-called "co-pilot model," and to use statcheck to flag possible inconsistencies in one's own manuscript or during the review process.


Articles with at least one inconsistency
Mean % results inconsistent per article
Percentage
Journal
JPSP
JEPG
DP
FP
PLOS
JCCP
PS
JAP
57.6
15.8
54.8
12.9
50.9
15.2
50.9
12
49.7
11.5
48.9
11.6
39.7
6.5
33.6
12.4
9
1.4
8.5
1.4
10.6
1.8
12.8
1.8
14
1.6
12.2
1.6
9.6
1.2
8.7
2.5
Type of inconsistency
Inconsistency
Gross Inconsistency

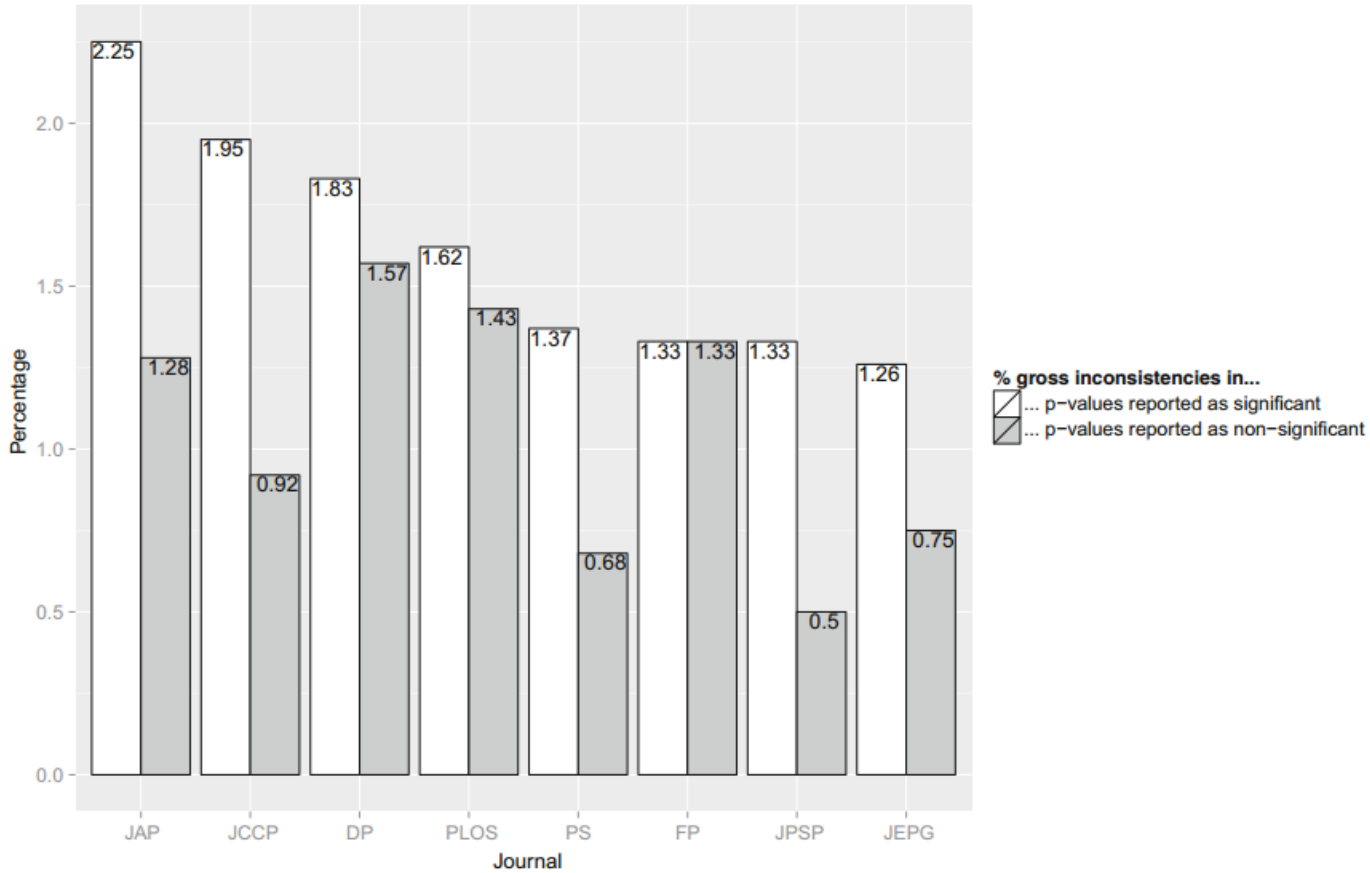


Fig. 6 The percentage of gross inconsistencies in *p*-values reported as significant (white bars) and nonsignificant (gray bars), split up by journal. For the journals *Journal of Applied Psychology* (JAP), *Journal of Consulting and Clinical Psychology* (JCCP), *Developmental Psychology* (DP), *Public Library of Science* (PLoS), *Psychological Science* (PS), *Frontiers in Psychology* (FP), *Journal of Personality and Social Psychology* (JPSP), and *Journal of Experimental Psychology: General* (JEPG), respectively, the total number of significant *p*-values was 11,654, 21,120, 29,962, 22,071, 12,482, 7,377, 78,889, and 14,084, and the total number of nonsignificant *p*-values was 3,119, 5,558, 6,698, 9,134, 2,936, 2,712, 17,868, and 4,407

A paper, which does not exist with > 700 citations

[CITATION] The art of writing a scientific article
J Van der Geer, JAJ Hanraads, RA Lupton - J. Sci. Commun, 2000
Cited by 712 Related articles

Take home messages

Take Home Messages

1. With an ANOVA you can avoid the multiple testing problem—to some extent.
2. More factors may improve or deteriorate power.

END Class 6

Scheffé Test

- **Step 1**: Calculate $SS_{between}$ for the comparison.

$$SS_{between} = \sum \frac{T^2}{n} - \frac{G^2}{N}$$

- **Step 2**: Calculate $MS_{between}$ for the comparison.

$$MS_{between} = \frac{SS_{between}}{df_{between}}$$

- **Step 3**: Calculate *F* for the comparison.

$$F = \frac{MS_{between}}{MS_{within}}$$

- **Step 4**: Compare to F_{crit} and state conclusions.

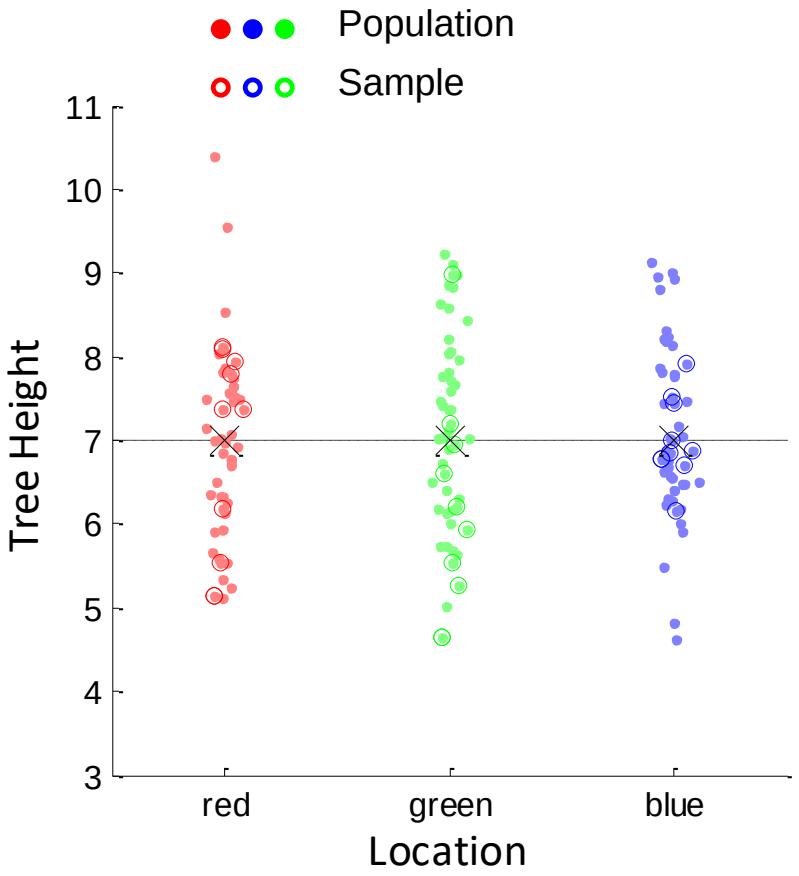
Scheffé Test: Step 1

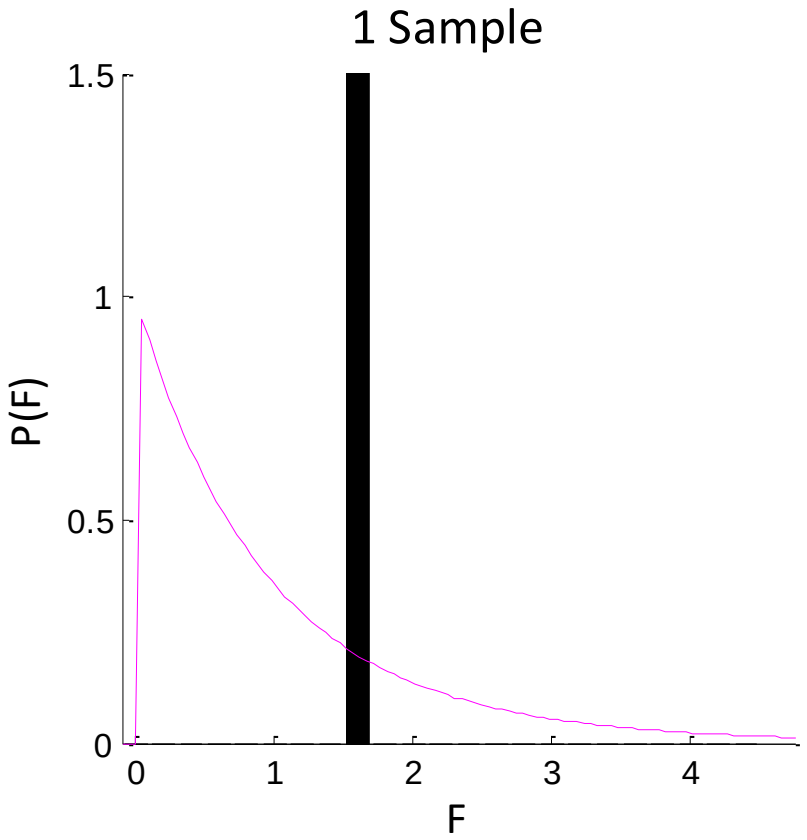
- Calculate $SS_{between}$ for the comparison:

$$SS_{between} = \sum_{i=1}^{2} \frac{T_i^2}{n_i} - \frac{G_{1,2}^2}{N_{1,2}}$$

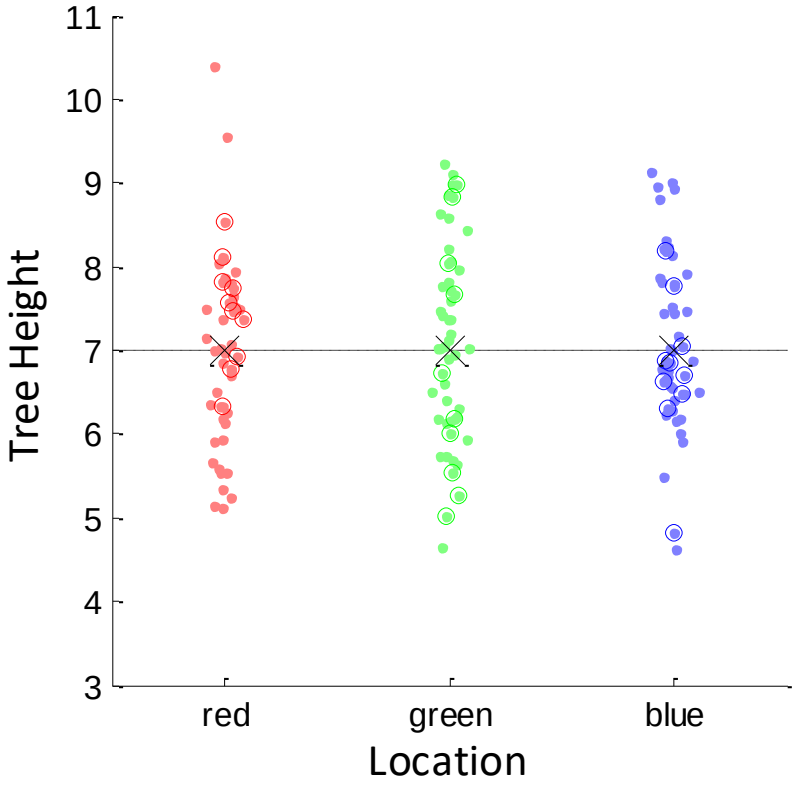
- T_i's are the total numbers of dismemberments/kills for the two sword types being compared
- n_i's are the number of subjects in the two sword types being compared.
- G is the grand total number of dismemberments/kills for only the two swords being compared.
- N is the number of subjects for the two sword types being compared.

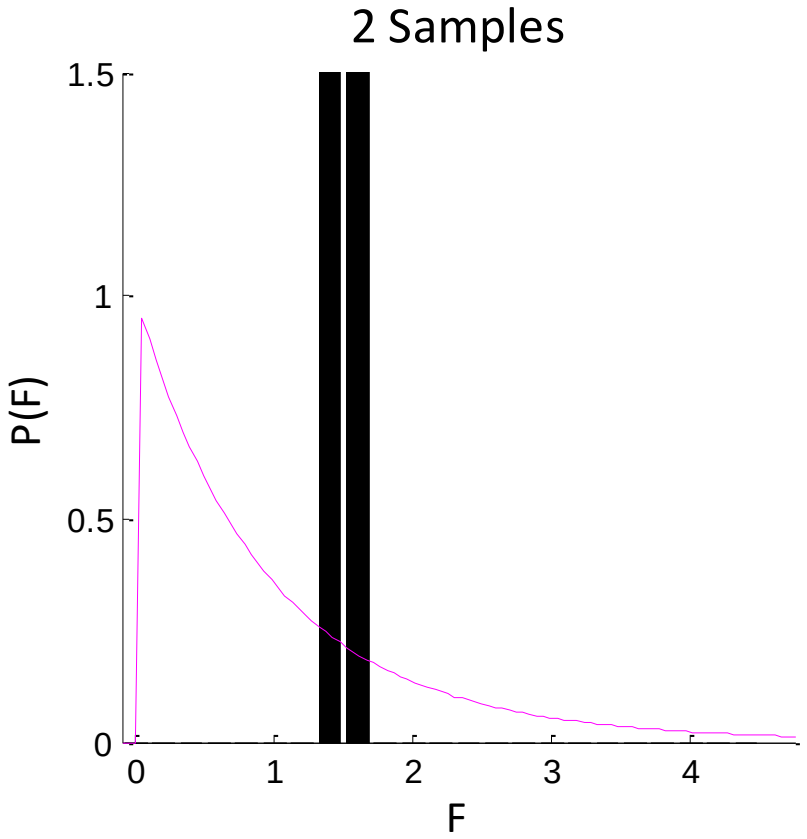
How Is *F* Distributed Under the Null Hypothesis?



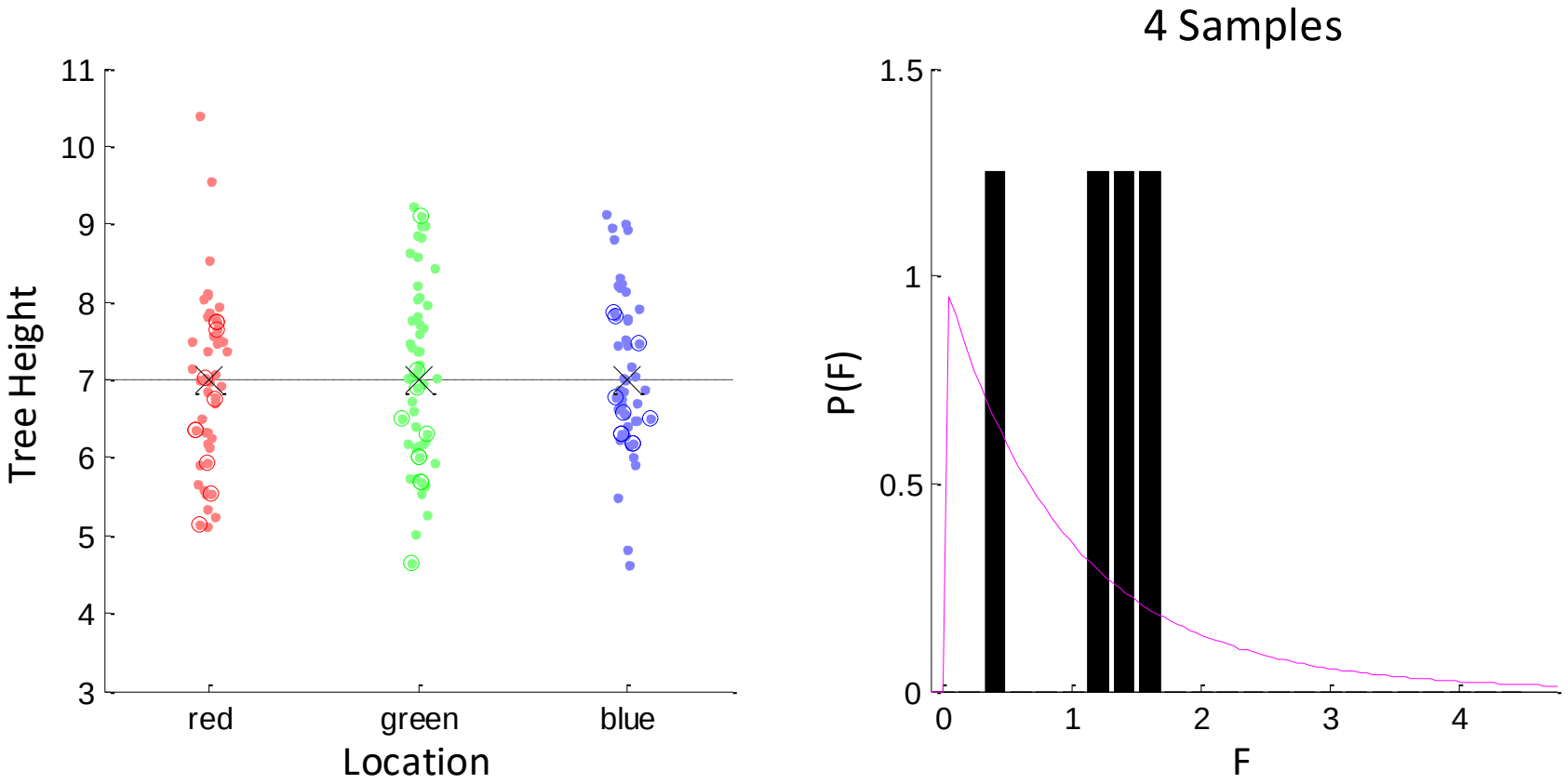


How Is *F* Distributed Under the Null Hypothesis?

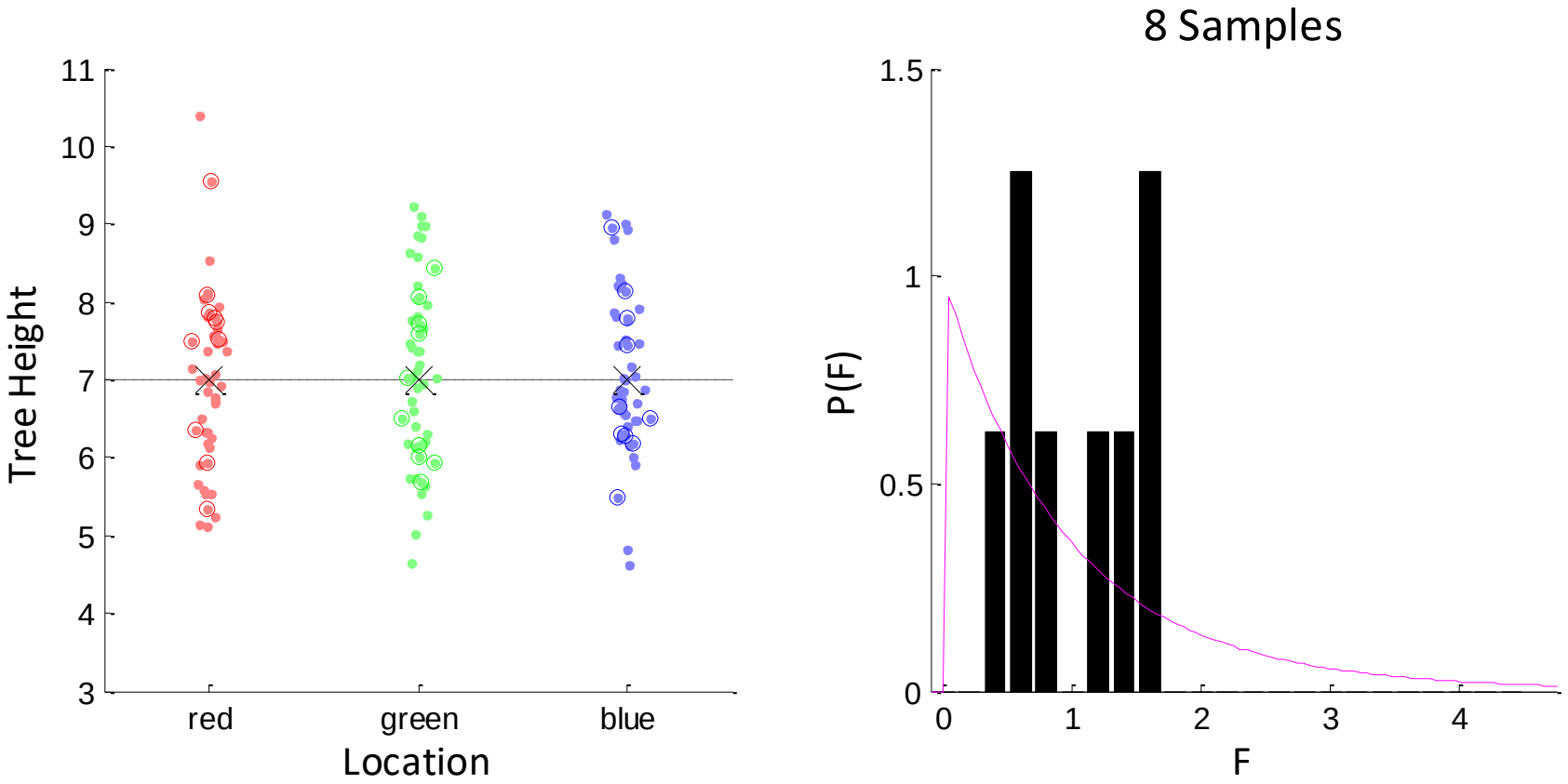




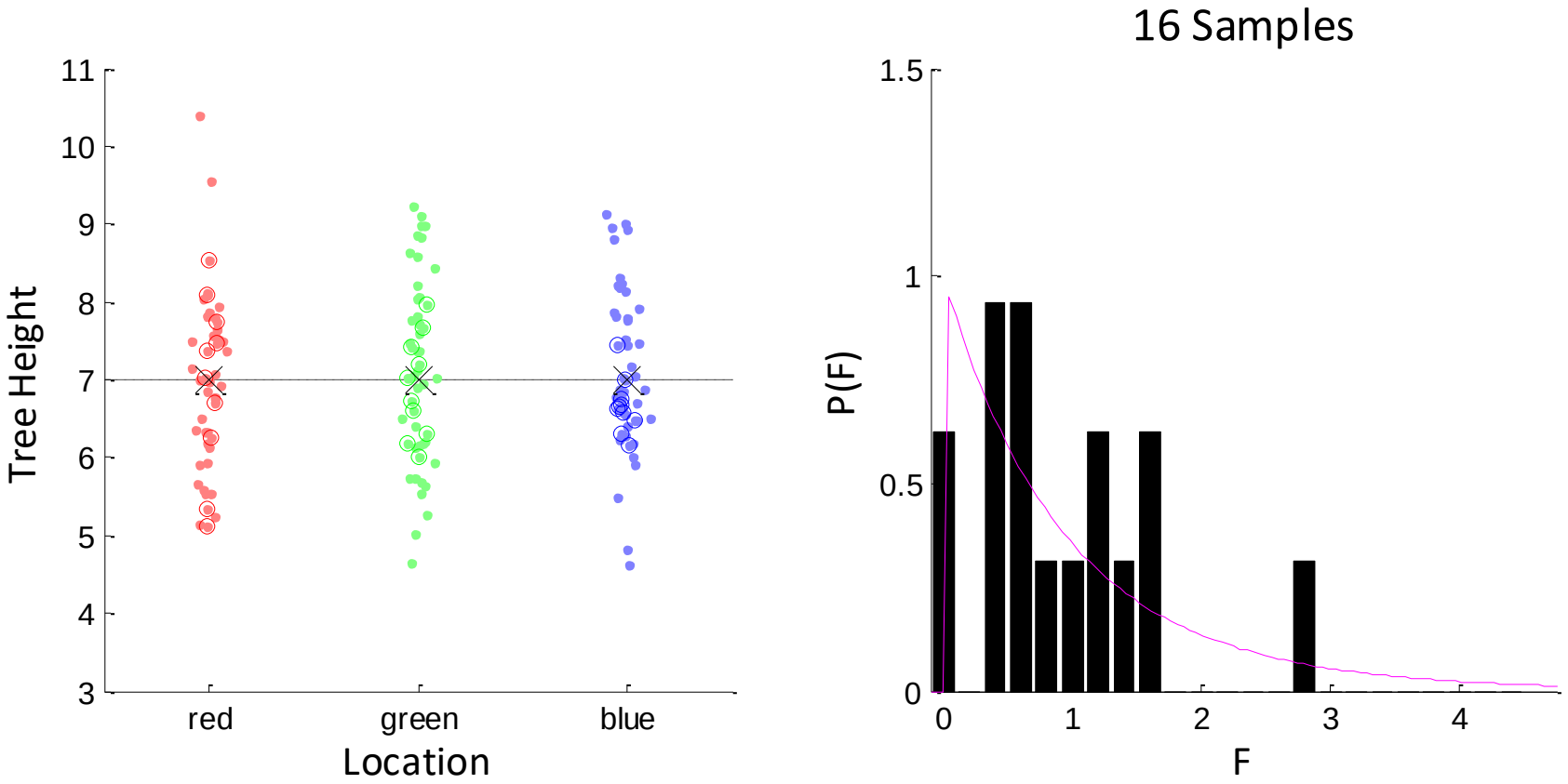
How Is *F* Distributed Under the Null Hypothesis?



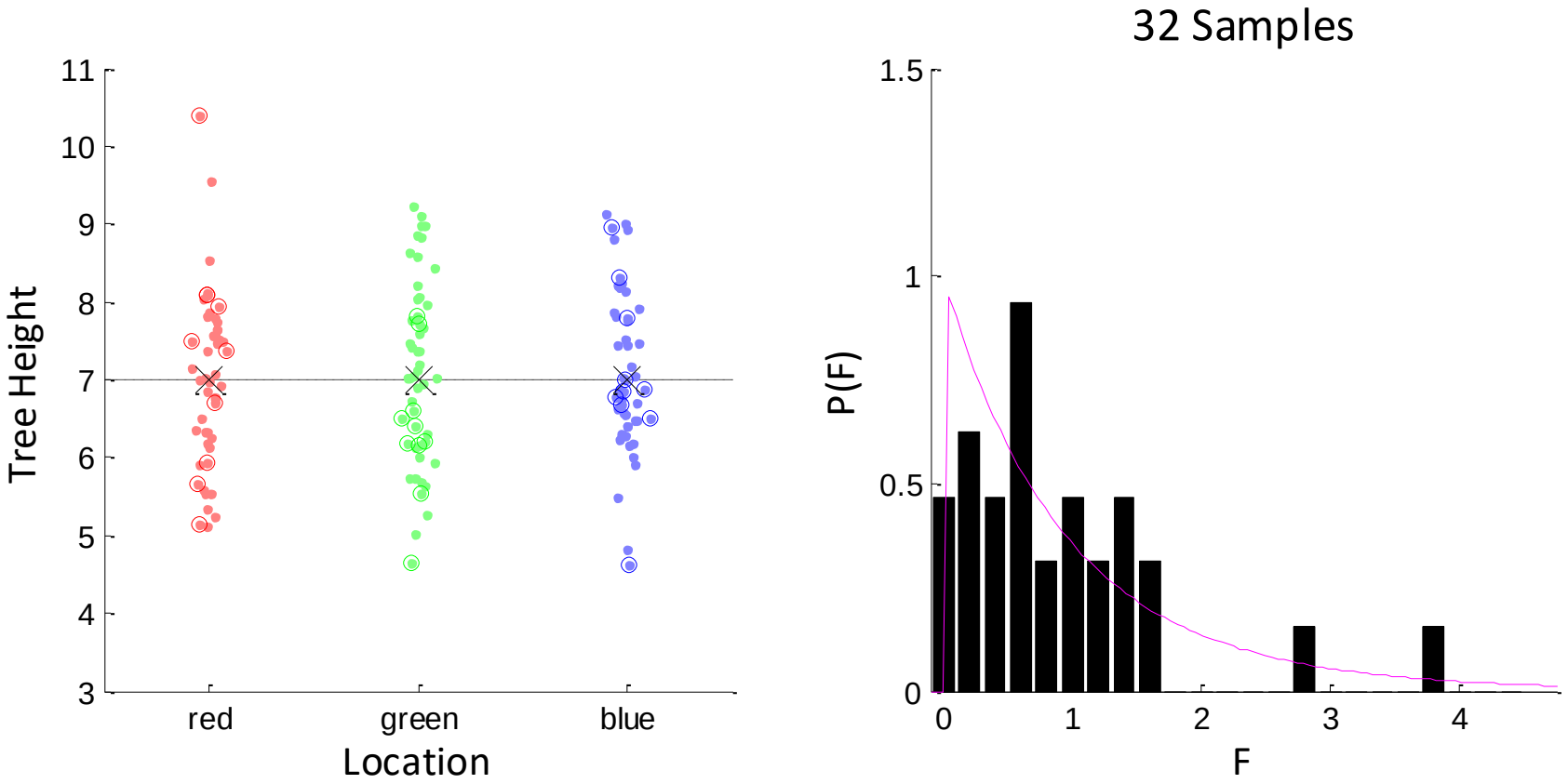
How Is *F* Distributed Under the Null Hypothesis?



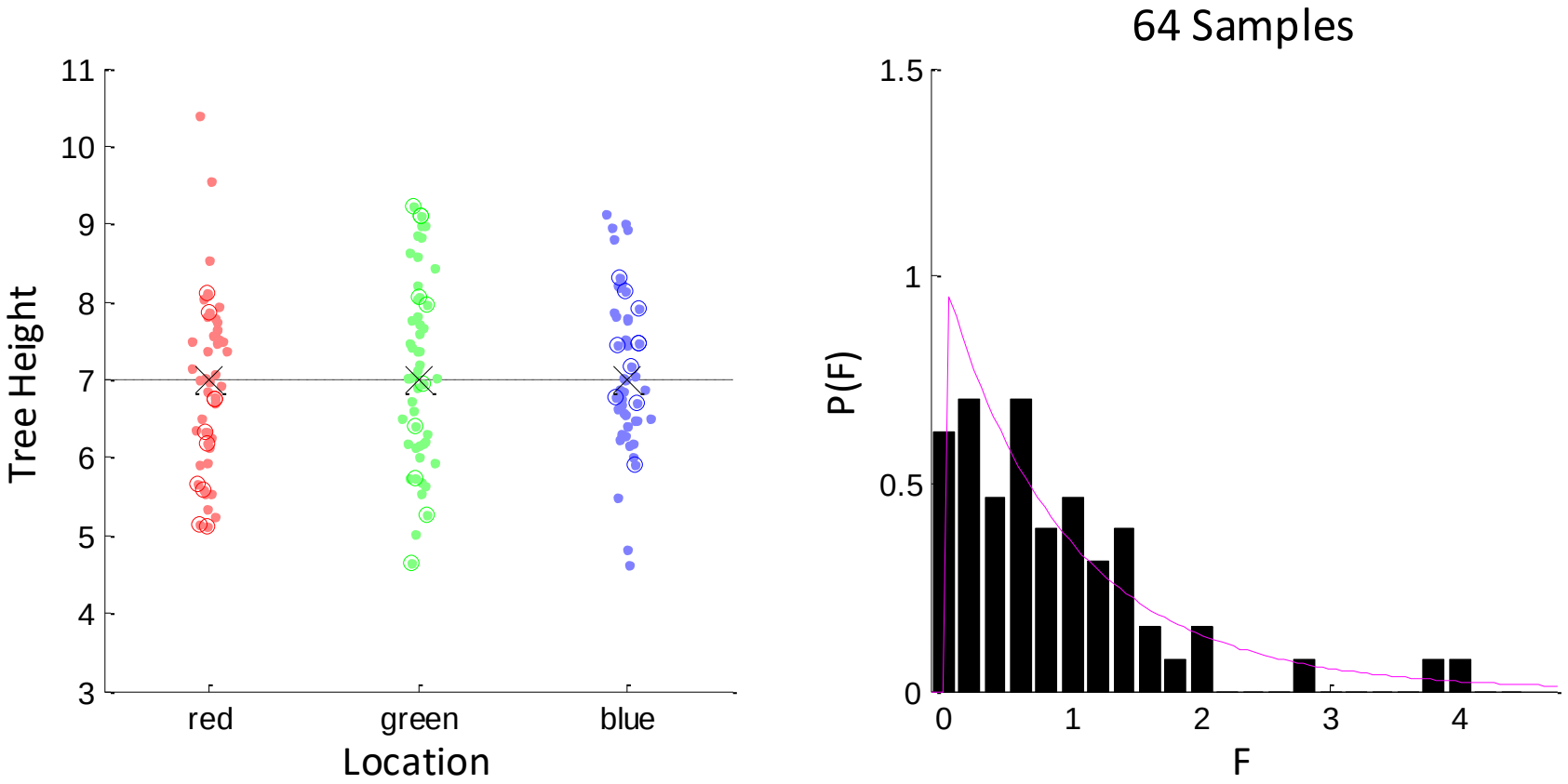
How Is *F* Distributed Under the Null Hypothesis?



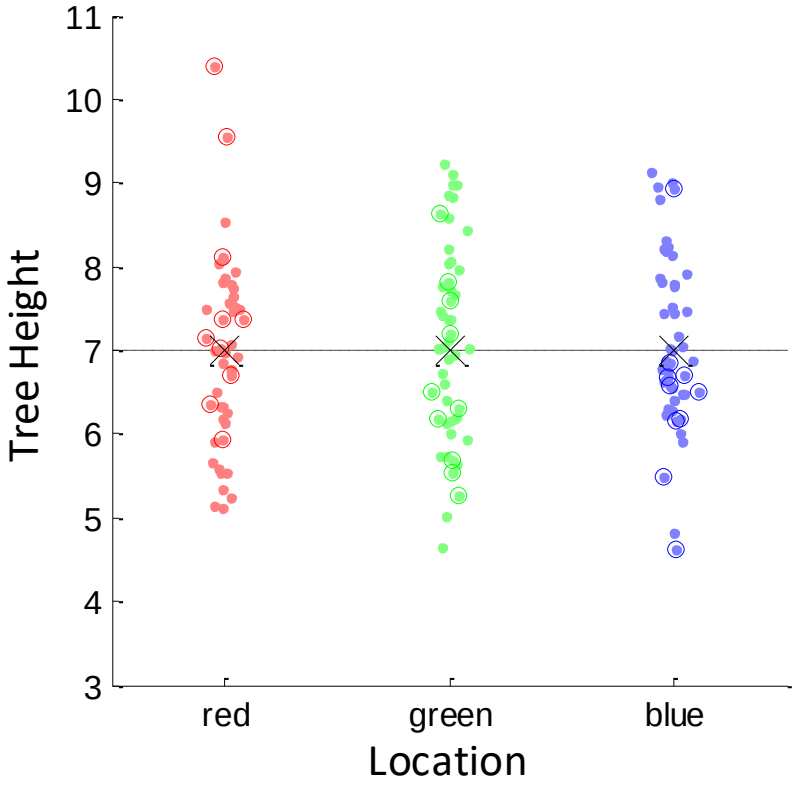
How Is *F* Distributed Under the Null Hypothesis?

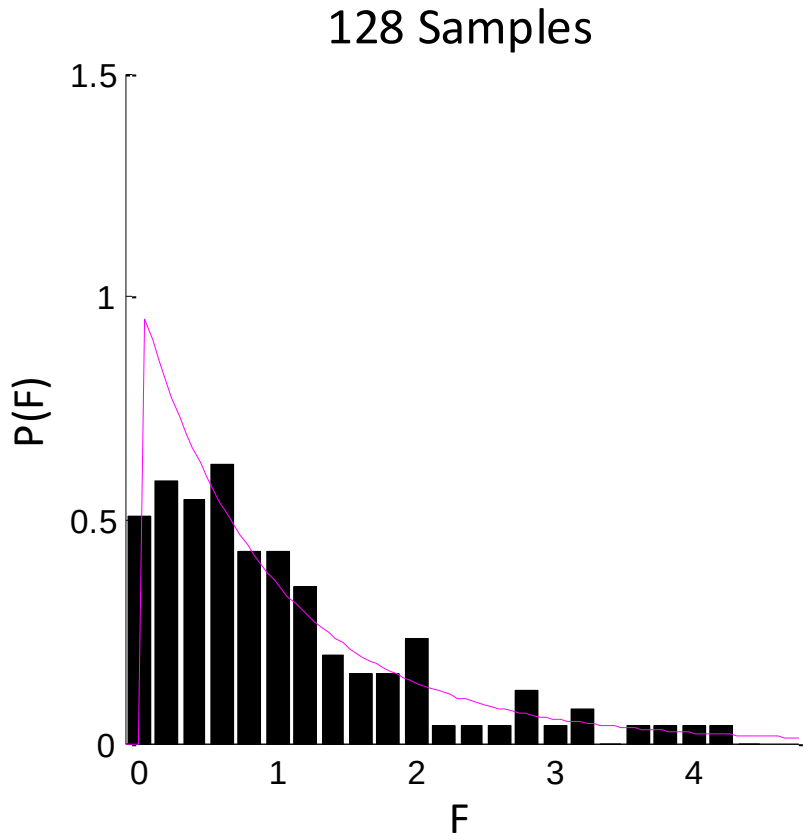


How Is *F* Distributed Under the Null Hypothesis?

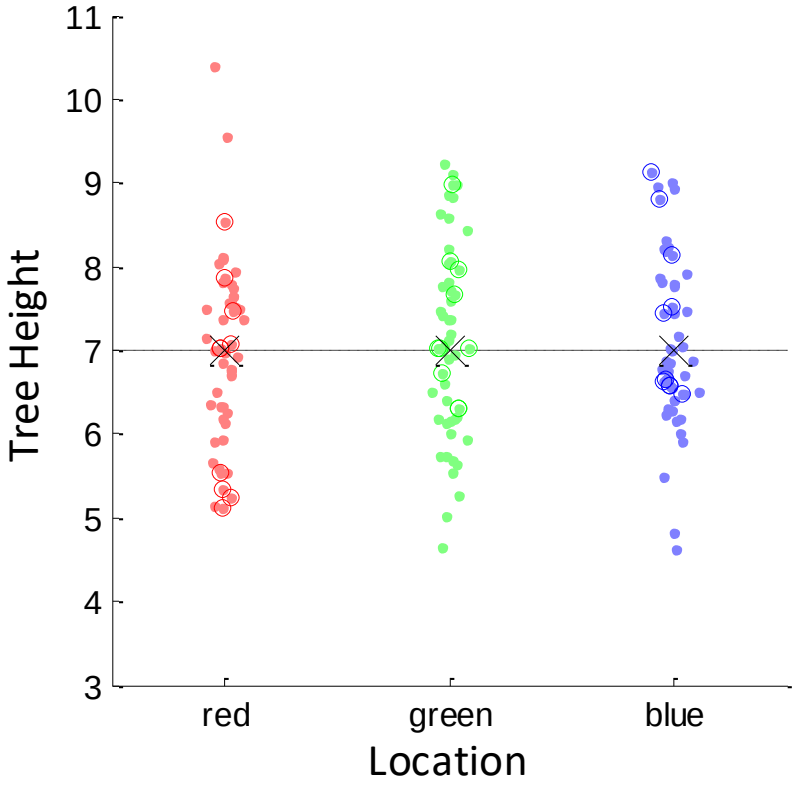


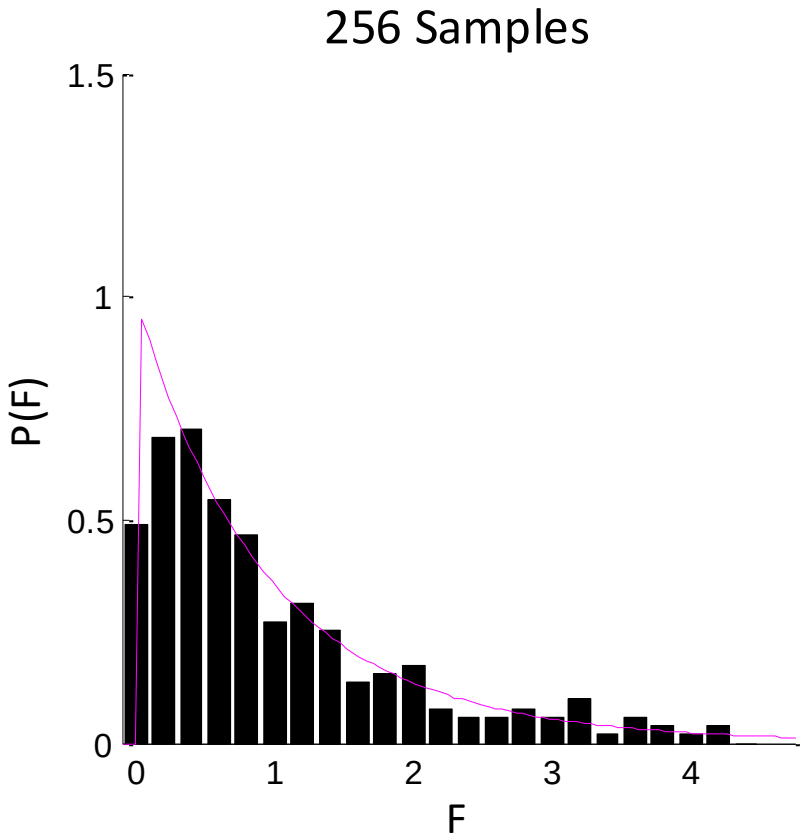
How Is *F* Distributed Under the Null Hypothesis?



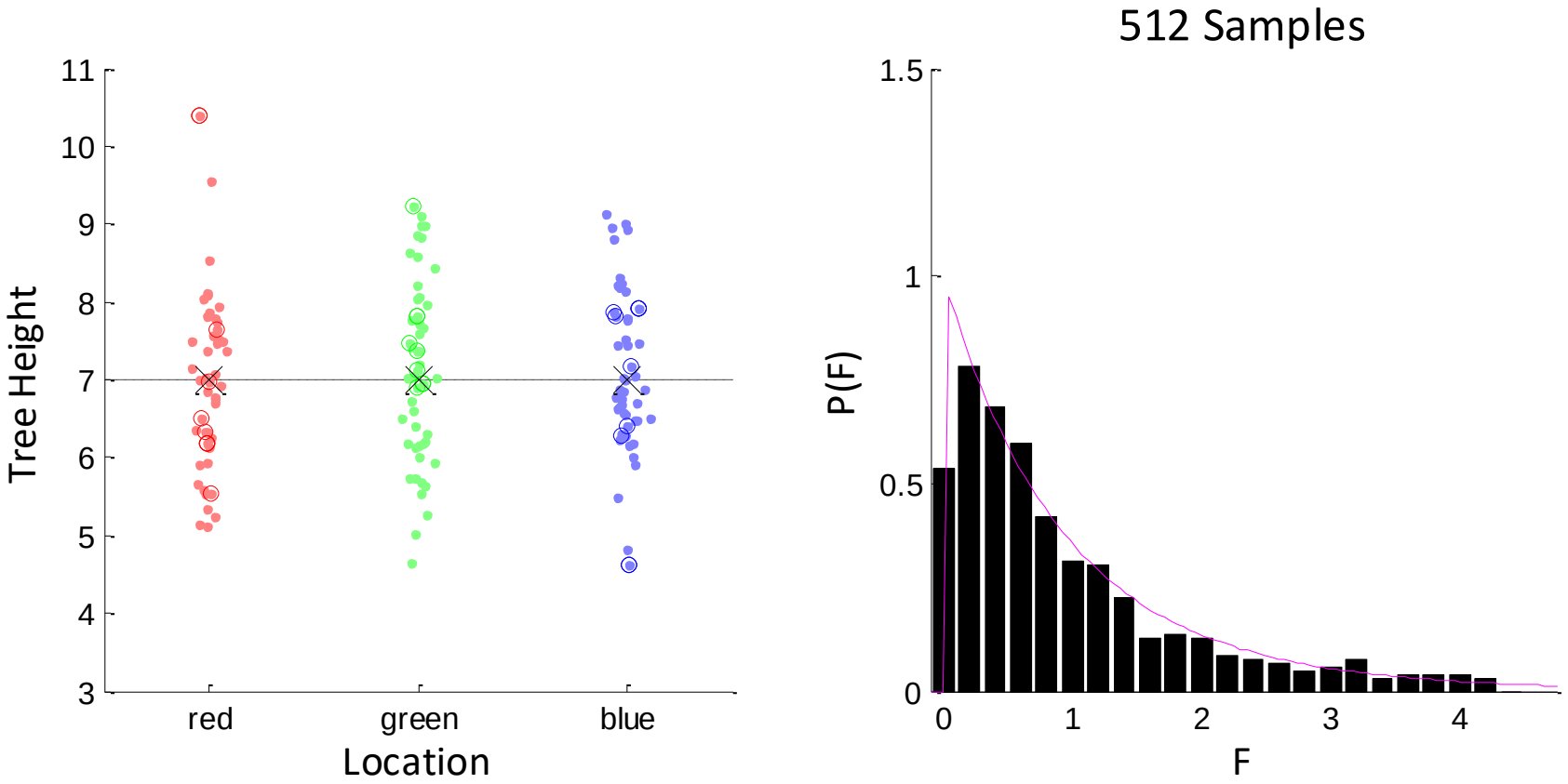


How Is *F* Distributed Under the Null Hypothesis?





How Is *F* Distributed Under the Null Hypothesis?



How Is *F* Distributed Under the Null Hypothesis?