

# **Analysis of Variance**

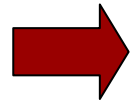
**(and discussion of Bayesian and frequentist statistics)**

**Dan Frey**

**Assistant Professor of Mechanical Engineering and Engineering Systems**



# Plan for Today



Efron, 2004

- Bayesians, Frequentists, and Scientists
- Analysis of Variance (ANOVA)
  - Single factor experiments
  - The model
  - Analysis of the sum of squares
  - Hypothesis testing
  - Confidence intervals

# Brad Efron's Biographical Information

- Professor of Statistics at Stanford University
- Member of the National Academy of Sciences
- President of the American Statistical Association
- Winner of the Wilks Medal
- "... renowned internationally for his pioneering work in computationally intensive statistical methods that substitute computer power for mathematical formulas, particularly the bootstrap method. The goal of this research is to extend statistical methodology in ways that make analysis more realistic and applicable for complicated problems. He consults actively in the application of statistical analyses to a wide array of health care evaluations."

# Bayesians, Frequentists, and Scientists

by Brad Efron

- How does the paper characterize the differences between the two approaches?
- What is currently driving a modern combination of these ideas?
- What lessons did you take away from the examples given?

Efron, B., 2005, "Bayesians, Frequentists, and Scientists," *Journal of the American Statistical Association*, 100, (469):1-5.

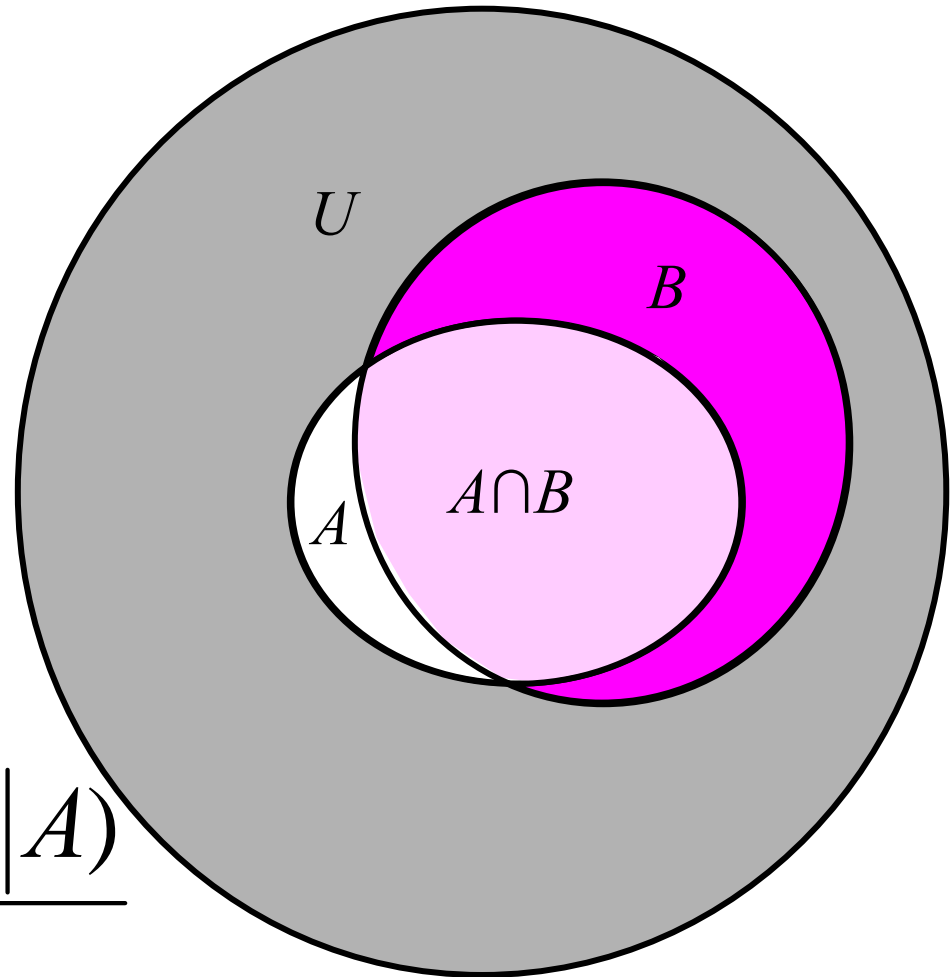
# Bayes' Theorem

$$\Pr(A|B) \equiv \frac{\Pr(A \cap B)}{\Pr(B)}$$

$$\Pr(B|A) \equiv \frac{\Pr(A \cap B)}{\Pr(A)}$$

with a bit of algebra

$$\Pr(A|B) = \frac{\Pr(A) \Pr(B|A)}{\Pr(B)}$$



# Bayes' Theorem and Hypotheses

"prior" probability

this is the  $p$ -value  
statistical tests provide

$$\Pr(H|D) = \frac{\Pr(H) \Pr(D|H)}{\Pr(D)}$$

$\Pr(D) = \sum_{i=1}^n \Pr(D|H_i) \Pr(H_i)$

# Bayes and Breast Cancer

- The probability that a woman randomly selected from all women in the US has breast cancer is 0.8%.
- If a woman has breast cancer, the probability that a mammogram will show a positive result is 90%.
- If a woman does not have breast cancer, the probability of a positive result is 7%.
- Take for example, a woman from the US who has a single positive mamogram. What is the probability that she actually has breast cancer?

- |    |                   |
|----|-------------------|
| 1) | ~90%              |
| 2) | ~70%              |
| 3) | ~9%               |
| 4) | <1%               |
| 5) | none of the above |

# Bayes' Theorem and Hypotheses

"base rate" in the population 0.8% power of the test to detect the disease 90%

$$\Pr(C|P) = \frac{\Pr(C) \Pr(P|C)}{\Pr(P)} = \frac{0.72\%}{6.9\% + 0.72\%} = 9.3\%$$


$$\Pr(P) = \Pr(P|\sim C) \Pr(\sim C) + \Pr(P|C) \Pr(C)$$

"false alarm" rate of the test 7%


1- "base rate" 99.2%



Figure removed due to copyright restrictions.  
Figure 1 in G. Gigerenzer, and A. Edwards. "Simple tools for understanding risks: from innumeracy to insight." *British Medical Journal* 327 (2003), 741-744.



**Probabilistic formulation:** The probability that a woman has breast cancer is 0.8%. If she has breast cancer, the probability that a mammogram will show a positive result is 90%. If a woman does not have breast cancer, the probability of a positive result is 7%. Take for example, a woman who has a positive result. What is the probability that she actually has breast cancer?



**Frequency format:** Eight out of every 1000 women have breast cancer. Of these eight women with breast cancer seven will have a positive result with mammography. Of the 992 women who do not have breast cancer some 70 will have a positive mammogram. Take for example, a sample of women who have positive mammograms. What proportion of these women actually have breast cancer?

# False Discovery Rates

Image removed due to copyright restrictions.

Courtesy of Bradley Efron. Used with permission.  
Source: "Modern Science and the Bayesian-Frequentist Controversy."  
[http://www-stat.stanford.edu/~brad/papers/NEW-ModSci\\_2005.pdf](http://www-stat.stanford.edu/~brad/papers/NEW-ModSci_2005.pdf)

# Wilcoxon Null Distribution

- Assign ranks to two sets of unpaired data
- The probability of occurrence of any total or a lesser total by chance under the assumption that the group means are drawn from the same population:

$$P = 2 \left\{ 1 + \sum_{i=1}^{r-q} \sum_{j=1}^q \Pi_j^i - \sum_{n=1}^{r-q} \left[ (r-q-n+1) \Pi_{q-1}^{q-2+n} \right] \right\} / \frac{\lfloor 2q \rfloor}{\lfloor q \rfloor \times \lfloor q \rfloor}$$

$\Pi_j^i$  represents the number of  $j$ -part partitions of  $i$ ,

$r$  is the serial number of possible rank totals, 0, 1, 2,  $\dots$   $r$ .

$q$  is the number of replicates, and

$n$  is an integer representing the serial number of the term in the series.

*Unequal 5-part partitions  
of 20*

1-2-3-4-10

1-2-3-5-9

1-2-3-6-8

1-2-4-5-8

1-2-4-6-7

1-3-4-5-7

2-3-4-5-6

# Wilcoxon "rank sum" test

(as described in the Matlab "help" system)

`p = ranksum(x,y,'alpha',alpha)`

Description -- performs a two-sided rank sum test of the null hypothesis that data in the vectors `x` and `y` are independent samples from identical continuous distributions with equal medians, against the alternative that they do not have equal medians. `x` and `y` can have different lengths. ...The test is equivalent to a Mann-Whitney U-test.

# Wilcoxon "rank sum" test

(what processing is carried out)

```
BRCA1=[-1.29 -1.41 -0.55 -1.04 1.28 -0.27 -0.57];
```

```
BRCA2=[-0.70 1.33 1.14 4.67 0.21 0.65 1.02 0.16];
```

```
p = ranksum(BRCA1,BRCA2,'alpha',0.05)
```

Form ranks  
of combined  
data sets

-1.41		
-1.29		
-1.04	-0.70	#4
-0.57		
-0.55		
-0.27	0.16	
	0.21	
	0.65	#8 thru
	1.02	#12
	1.14	
1.28		
	1.33	#14 and
	4.67	#15

rank sum of BRCA2 is  
83

largest possible is 92  
smallest possible is 36

network algorithms applied  
to find p-value is ~2%

# Empirical Bayes

- "... the prior quantities are estimated frequentistically in order to carry out Bayesian calculations."
- "... if we had only one gene's data ... we would have to use the Wilcoxon null, but with thousands of genes to consider at once, most of which are probably null, we can empirically estimate the null distribution itself. Doing so gives far fewer significant genes in this case."
- "... Estimating the null hypothesis itself from the data sounds a little crazy, but that's what I mean about huge data sets presenting new opportunities..."

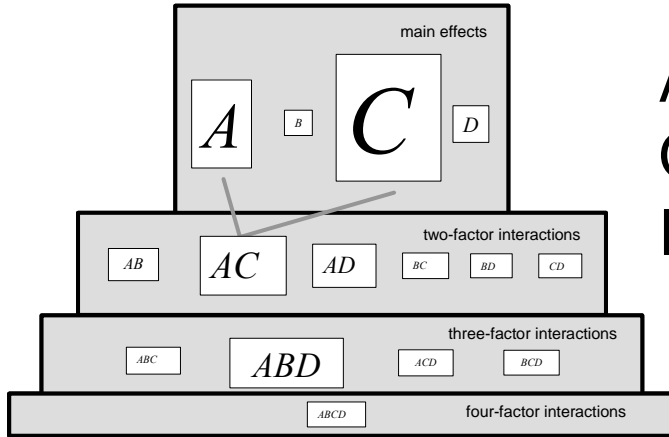
# The Bootstrap

Image removed due to copyright restrictions.

Figures 4 and 5 in Efron, Bradley. "Modern Science and the Bayesian-Frequentist Controversy."

[http://www-stat.stanford.edu/~brad/papers/NEW-ModSci\\_2005.pdf](http://www-stat.stanford.edu/~brad/papers/NEW-ModSci_2005.pdf)

# Same Basic Ideas in My Research

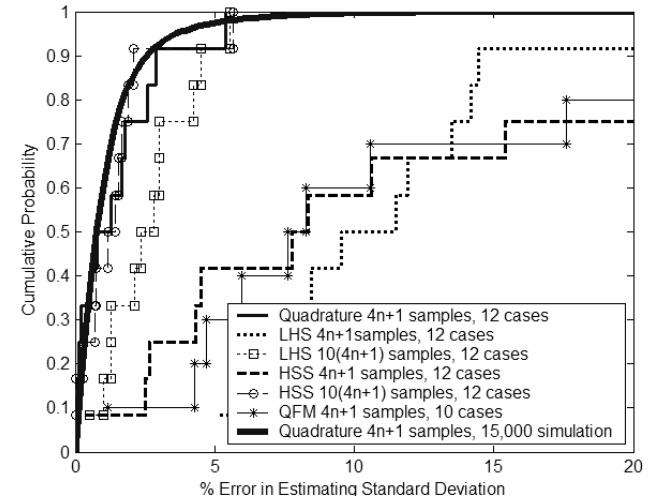


A notion of structure in a problem domain.  
Quantified by a huge body of data.  
Encoded in a probabilistic model.

$$f(\beta_i | \delta_i) = \begin{cases} N(0,1) & \text{if } \delta_i = 0 \\ N(0, c^2) & \text{if } \delta_i = 1 \end{cases}$$

$$\Pr(\delta_{ij} = 1 | \delta_i, \delta_j) = \begin{cases} p_{00} & \text{if } \delta_i + \delta_j = 0 \\ p_{01} & \text{if } \delta_i + \delta_j = 1 \\ p_{11} & \text{if } \delta_i + \delta_j = 2 \end{cases}$$

Used to from "bootstrap" estimates of statistics (in this case, performance of techniques for design of computer experiments).





# Wisdom Regarding the Foundations of Statistics

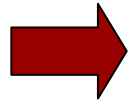
*It is often argued academically that no science can be more secure than its foundations, and that, if there is controversy about the foundations, there must be even more controversy about the higher parts of the science. As a matter of fact, the foundations are the most controversial part of many, if not all, sciences... As in other sciences, controversies about the foundations of statistics reflect themselves to some extent in everyday practice, but not nearly so catastrophically as one might imagine. I believe that here, as elsewhere, catastrophe is avoided, primarily because in practical situations common sense generally saves all but the most pedantic of us from flagrant error... Although study of the foundations of a science does not have the role that would be assigned to it by naïve first-things-firstism, it certainly has a continuing importance as the science develops, influencing, and being influenced by, the more immediately practical parts of the science.*

# Wisdom Regarding the Foundations of Statistics

*It is often argued academically that no science can be more secure than its foundations, and that, if there is controversy about the foundations, there must be even more controversy about the higher parts of the science. As a matter of fact, the foundations are the most controversial part of many, if not all, sciences... As in other sciences, controversies about the foundations of statistics reflect themselves to some extent in everyday practice, but not nearly so catastrophically as one might imagine. I believe that here, as elsewhere, catastrophe is avoided, primarily because in practical situations common sense generally saves all but the most pedantic of us from flagrant error... Although study of the foundations of a science does not have the role that would be assigned to it by naïve first-things-firstism, it certainly has a continuing importance as the science develops, influencing, and being influenced by, the more immediately practical parts of the science.*

# Plan for Today

- Efron, 2004
  - Bayesians, Frequentists, and Scientists



## Analysis of Variance (ANOVA)

- Single factor experiments
- The model
- Analysis of the sum of squares
- Hypothesis testing
- Confidence intervals

# Single Factor Experiments

- A single **experimental factor** is varied
- The parameter takes on various **levels**

Cotton weight percentage	Observations				
	1	2	3	4	5
15	7	7	15	11	9
20	12	17	12	18	18
25	14	18	18	19	19
30	19	25	22	19	23
35	7	10	11	15	11

$a=5$  replicates

Each cell is a  $y_{ij}$

Each row is a treatment  $i$

↑  
experimental factor

Fiber strength in lb/in<sup>2</sup>

# Breakdown of Sum Squares

“Grand Total  
Sum of Squares”

$$GTSS = \sum_{i=1}^a \sum_{j=1}^n y_{ij}^2$$

“Total Sum of  
Squares”

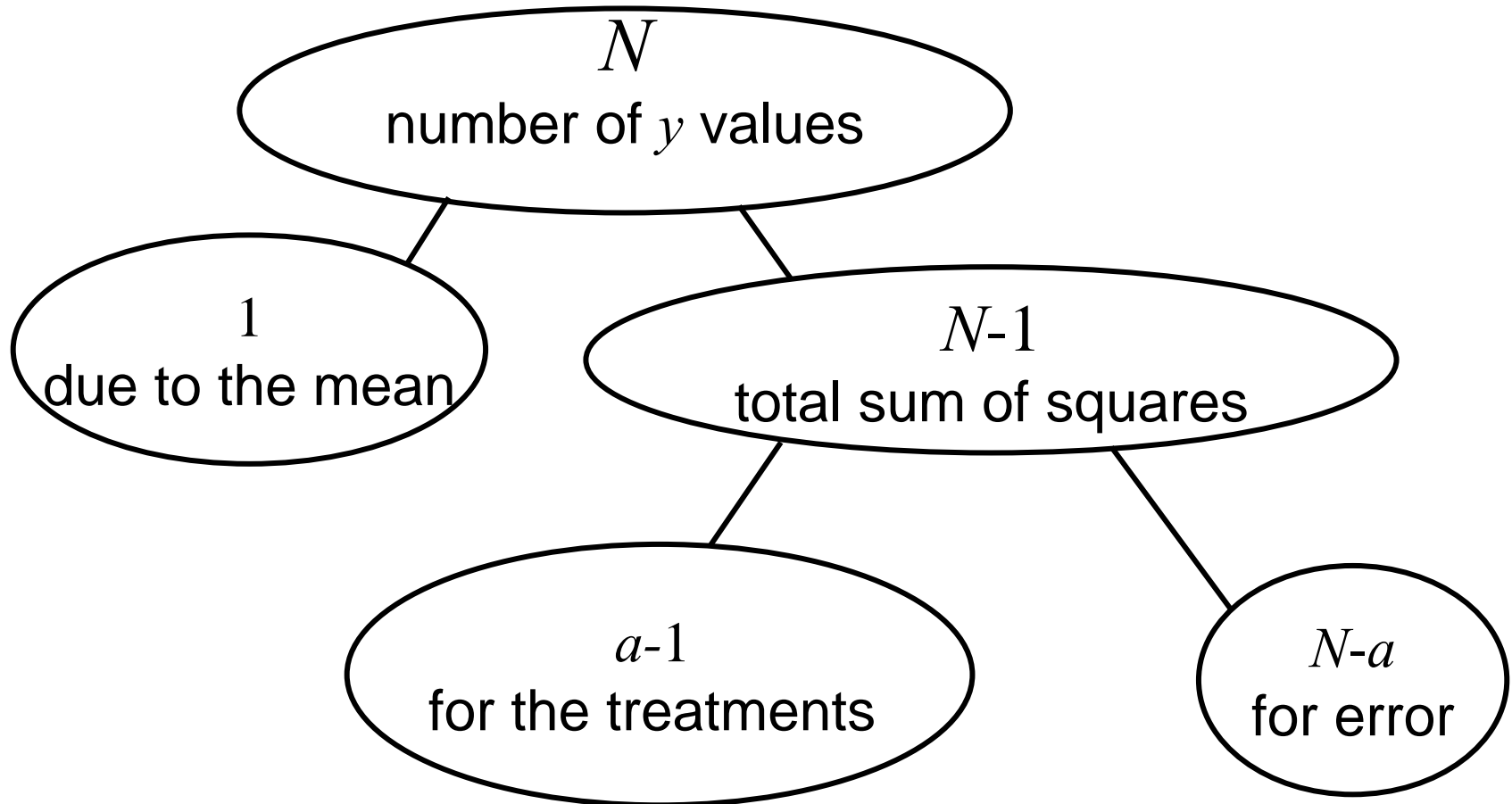
$SS$  due to mean  
 $= n\bar{y}_{..}^2$

$$SS_T = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2$$

$$SS_{Treatments} = n \sum_{i=1}^a (\bar{y}_{i.} - \bar{y}_{..})^2$$

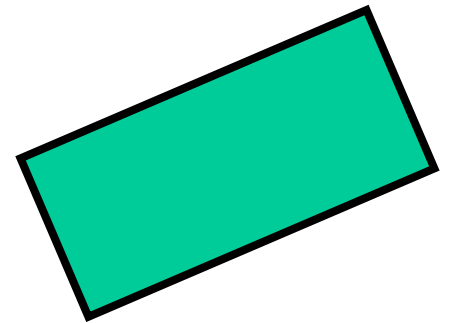
$SS_E$

# Breakdown of DOF

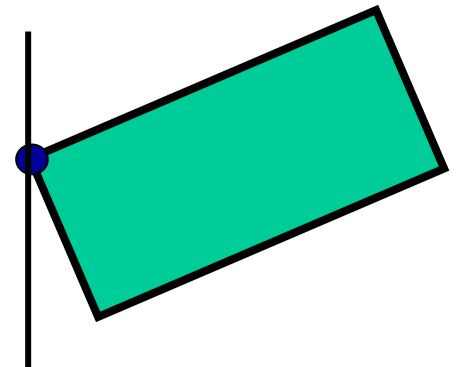


# What is a "Degree of Freedom"?

- How many scalar values are needed to unambiguously describe the state of this object?



- What if I were to fix the  $x$  position of a corner?



# What is a "Degree of Freedom"?

- How many scalar values are needed to unambiguously describe the outcome of this experiment?

Cotton weight percentage	Observations				
	1	2	3	4	5
15	7	7	15	11	9
20	12	17	12	18	18
25	14	18	18	19	19
30	19	25	22	19	23
35	7	10	11	15	11

- What if I were to tell you  $\bar{y}_{..}$  ?
- What if I were to tell you  $\bar{y}_i$   $i = 1...4$  ?



# Example of ANOVA

- What do I get if I compute the mean of these values?
- What is the variance of these values related to?
- If this data were taken in the presence of time trend, how would the tables change if the experimental procedure were altered to eliminate the trend?

	Observations				
Cotton weight percentage	1	2	3	4	5
15	7	7	15	11	9
20	12	17	12	18	18
25	14	18	18	19	19
30	19	25	22	19	23
35	7	10	11	15	11

Anova: Single Factor					
SUMMARY					
Groups	Count	Sum	Average	Variance	
Row 1	5	49	9.8	11.2	
Row 2	5	77	15.4	9.8	
Row 3	5	88	17.6	4.3	
Row 4	5	108	21.6	6.8	
Row 5	5	54	10.8	8.2	

ANOVA						
Source of Variat	SS	df	MS	F	P-value	F crit
Between Gr	475.76	4	118.94	14.75682	9.13E-06	2.866081
Within Group	161.2	20	8.06			
Total	636.96	24				

# Treatment Effects Model

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


Cotton weight percentage	Observations				
	1	2	3	4	5
15	7	7	15	11	9
20	12	17	12	18	18
25	14	18	18	19	19
30	19	25	22	19	23
35	7	10	11	15	11

Each cell is a  $y_{ij}$

Each row is treatment  $i$

Replicates in columns

# Assumptions of the Model

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


- Error is normally distributed
- With equal variance
  - across treatments and
  - over time
- Effects of other factors do not bias the results

# Testing Equality of Treatment Means

- Hypotheses

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_a = 0$$

$$H_1 : \tau_i \neq 0 \text{ for at least one } i$$

- Test statistic

$$F_0 = \frac{MS_{Treatments}}{MS_E}$$

- Criterion for rejecting  $H_0$

$$F_0 > F_{\alpha, a-1, N-a}$$

# Cochran's Theorem

- Let  $Z_i$  be NID(0,1) for  $i=1,2,\dots,\nu$  and

$$\sum_{i=1}^{\nu} Z_i^2 = Q_1 + Q_2 + \dots + Q_s$$

Where  $s < \nu$  and  $Q_i$  has  $\nu_i$  degrees of freedom. Then  $Q_1, Q_2, \dots, Q_s$  are independent chi-square random variables with  $\nu_1, \nu_2, \dots, \nu_s$  degrees of freedom respectively iff  $\nu = \nu_1 + \nu_2 + \dots + \nu_s$

- Implies that  $\frac{MS_{Treatments}}{MS_E}$  is  $F_{a-1, N-a}$

# Model Adequacy Checking

- Normality
  - normal probability plot of residuals
- Independence / constant variance
  - plot residuals versus time sequence
  - plot residuals versus fitted values
  - Bartlett's Test `[ndim, prob] = barttest(x,0.05)`

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_a^2$$

$$H_1 : \sigma_i^2 \neq \sigma_j^2 \text{ for at least one } i, j$$

# Randomization Distribution

- The null hypothesis implies that these observations are not a function of the treatments
- If that were true, the allocation of the data to treatments (rows) shouldn't affect the test statistic
- How likely is the statistic observed under re-ordering?

Cotton weight percentage	Observations				
	1	2	3	4	5
15	7	7	15	11	9
20	12	17	12	18	18
25	14	18	18	19	19
30	19	25	22	19	23
35	7	10	11	15	11

# Randomization Distribution

- This code reorders the data from the cotton experiment
- Does the ANOVA
- Repeats 5000 times
- Plots the pdf of the F ratio

```
trials=5000;bins=trials/100;  
X=[7 7 15 11 9 12 17 12 18 18 14 18 18  
19 19 19 25 22 19 23 7 10 11 15 11];  
group=ceil([1:25]/5);  
  
for i=1:trials  
    r=rand(1,25);  
    [B,INDEX] = sort(r);  
    Xr(1:25)=X(INDEX);  
    [p,table,stats] = anova1(Xr, group,'off');  
    Fratio(i)=cell2mat(table(2,5));  
end  
  
hold off  
[n,x] = hist(Fratio,bins);  
n=n/(trials*(x(2)-x(1)));  
colormap hsv  
bar(x,n)  
hold on  
  
xmax=max(Fratio);  
x=0:(xmax/100):xmax;  
y = fpdf(x,4,20);  
plot(x,y,'LineWidth',2)
```



# The Effect of Heteroscedascity

- This Matlab code generates data with a no treatment effect on location
- But dispersion is affected by group  
 $\sim N(0, \text{group})$
- Type I error rate rises substantially

```
for i=1:1000
    group=ceil([1:50]/10);
    X=group.*random('Normal',0,1,1,50);
    [p,table,stats] = anova1(X, group,'off');
    reject_null(i)=p<0.05;
end
plot(group,X,'+'); mean(reject_null)
```

# How Important Is Normality?

- This code includes **uniformly** distributed variates
- The randomization distribution is computed
- Plots the pdf of the F ratio
- Looks like the F distribution!

```
trials=5000;bins=trials/100;
X=random('Uniform',0,1,1,25);
group=ceil([1:25]/5);

for i=1:trials
    r=rand(1,25);
    [B,INDEX] = sort(r);
    Xr(1:25)=X(INDEX);
    [p,table,stats] = anova1(Xr, group,'off');
    Fratio(i)=cell2mat(table(2,5));
end

hold off
[n,x] = hist(Fratio,bins);
n=n/(trials*(x(2)-x(1)));
colormap hsv
bar(x,n)
hold on

xmax=max(Fratio);
x=0:(xmax/100):xmax;
y = fpdf(x,4,20);
plot(x,y,'LineWidth',2)
```

# Determining Sample Size

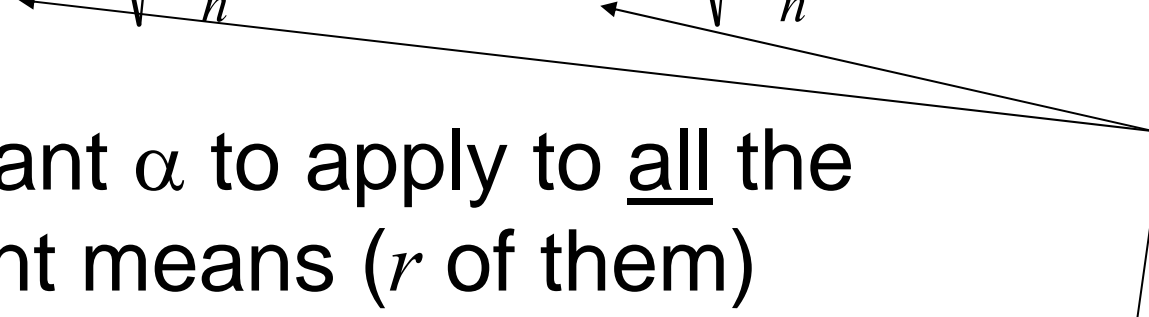
- Can be done on the basis of Type II error probability
  - BUT this requires an estimate of treatment effects compared to error
- OR the experimenter can specify desired width in the confidence interval
  - This only requires an estimate of experimental error

# Balance

- When the number of samples at each treatment is equal across all treatments, the design is said to be balanced
- Unbalance causes no difficulty in the computation of ANOVA tables
- BUT a balanced design provides
  - More robustness to violation of the homoscedascity assumption
  - The greatest possible power of the hypothesis test

# Confidence Intervals

- One-at-a-time confidence intervals apply to each treatment mean

$$\bar{y}_i - t_{\alpha/2, N-a} \sqrt{\frac{MS_E}{n}} \leq \mu_i \leq \bar{y}_i + t_{\alpha/2, N-a} \sqrt{\frac{MS_E}{n}}$$
A thin black line originates from the two square root terms in the equation above. It extends downwards and to the right, then turns left to point towards the text in the next block.

- If you want  $\alpha$  to apply to all the treatment means ( $r$  of them) simultaneously just replace  $\alpha/2$  with  $\alpha/2r$

Bonferroni method

# Confidence Intervals Between Treatment Means

- For a treatment mean

$$\bar{y}_{i.} - t_{\alpha/2, N-a} \sqrt{\frac{MS_E}{n}} \leq \mu_i \leq \bar{y}_{i.} + t_{\alpha/2, N-a} \sqrt{\frac{MS_E}{n}}$$

- For a difference between treatment means

$$\bar{y}_{i.} - \bar{y}_{j.} - t_{\alpha/2, N-a} \sqrt{\frac{2MS_E}{n}} \leq \mu_i - \mu_j \leq \bar{y}_{i.} - \bar{y}_{j.} + t_{\alpha/2, N-a} \sqrt{\frac{2MS_E}{n}}$$

Note the factor of two

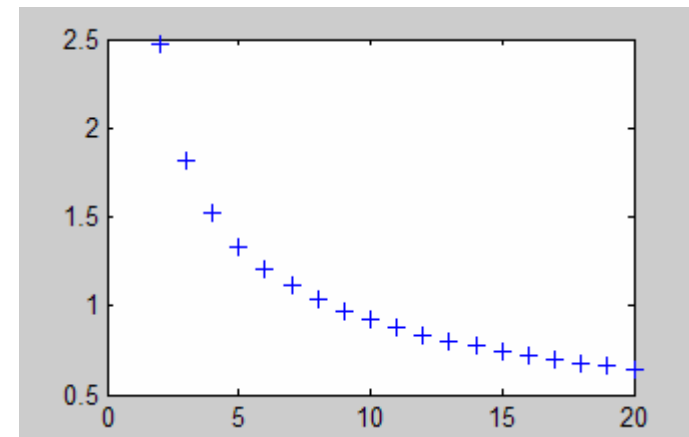


# Determining Sample Size from Confidence Intervals

- State the desired width  $\pm ci$  (e.g. on a difference between treatment means)
- Estimate  $\sigma$  of the experimental error
- Solve for  $n$  such that

$N = n \cdot a$  for a balanced design  $\rightarrow ci \approx t_{\alpha/2, N-a} \sqrt{\frac{2\sigma^2}{n}}$

```
s=3; % estimated experimental error
a=5; %number of treatments
n=2:20;
y=tinv(0.95,n*a-a).*sqrt(s./n);
plot(n,y,'+')
```



# Discovering Dispersion Effects

- Earlier we considered non-constant variance as a difficulty – a violation of our model assumptions
- BUT sometimes we are interested in studying and exploiting these “dispersion effects” (a.k.a. robust design)
- Analysis can proceed as usual (ANOVA, etc)
- Best to use  $\log(s)$  rather than  $s$  as the response

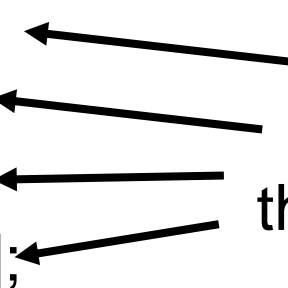


# Example 3-12

## Smelting Experiment

```
X=[0.05 0.04 0.05 0.06 0.03 0.05  
0.04 0.02 0.03 0.05 0.03 0.02  
0.09 0.13 0.11 0.15 0.08 0.12  
0.03 0.04 0.05 0.05 0.03 0.02];  
logX=log(X);  
[p,table,stats] = anova1(logX');
```

Standard deviations of the voltage with 4 control algorithms



# Next Steps

- Friday 20 April, recitation (by Frey)
- Monday 23 April
  - PS#6 due
  - Multiple regression
- Wednesday 25 April
  - Design of Experiments