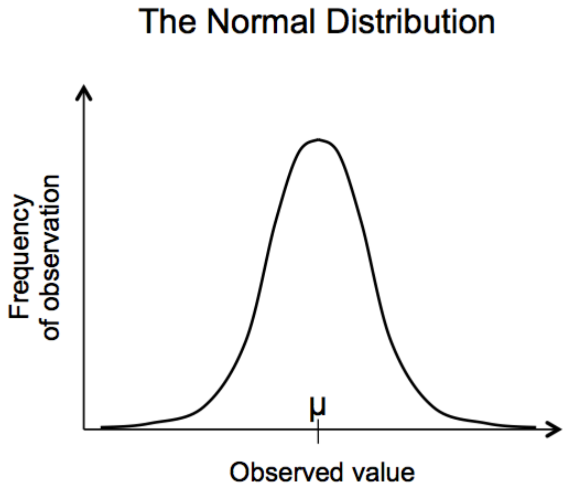
Lecture 4
Topic 3: General linear models (GLMs), the fundamentals of the analysis of variance (ANOVA), and completely randomized designs (CRDs)

The general linear model



1. A purely mathematical entity.

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

2. A distribution of related individuals.
(what we see)

3. A distribution of errors.
(what we think underlies what we see)

One population: An observation is explained as a mean plus a random deviation ε_i (error):

$$\mathbf{Y_i = \mu + \varepsilon_i}$$

The $\boldsymbol{\varepsilon}_i$'s are assumed to be from a population of uncorrelated $\boldsymbol{\varepsilon}$'s with mean zero. Independence among $\boldsymbol{\varepsilon}$'s is assured by random sampling.

Two populations: Each observation is explained as a grand mean *plus* an effect of its group (i.e. treatment) plus a random deviation ε_i (error):

$$\mathbf{Y_{ij} = \mu + \tau_i + \varepsilon_{ij}}$$

$$\mu + \tau_1 = \mu_1 \quad \text{and} \quad \mu + \tau_2 = \mu_2$$
$$\tau_1 + \tau_2 = 0$$

An equivalent expression of this model:

$$Y_{ij} = \bar{Y}_{..} + (\bar{Y}_{i.} - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_{i.})$$

Example: Imagine an experiment with 10 tomato plants (yum!), each in its own pot. 5 of the pots receive fertilizer and the other 5 do not. The total yield of each plant (in kg) is recorded and the data are presented below:

Plant	Fertilized	Not Fertilized
1	2.05	1.21
2	1.65	0.95
3	1.76	1.32
4	2.08	1.33
5	1.84	1.15

Treatment Means	1.876	1.192
General Mean	1.534	
Treatment Effects	0.342	-0.342

Under the assumption of a general linear model, the yield of each tomato plant in the above experiment has the following general form:

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

For Plant 3 receiving fertilizer, the equation looks like this:

$$y_{3,Fert} = \mu + \tau_{Fert} + \varepsilon_{3,Fert}$$
$$y_{3,Fert} = 1.534 + 0.342 - 0.116 = 1.76$$

More than two populations (The Model I or fixed model ANOVA)

1. Treatment effects are additive and fixed by the researcher.

$$\boxed{\sum \tau_i = 0 \quad \rightarrow \quad H_0: \tau_1 = \ldots = \tau_t = 0 \quad H_1: \text{Some } \tau_i \neq 0}$$

2. Errors are random, independent, and normally distributed with a common variance about a zero mean.

3. In the case of a false H_0 (i.e. some $\tau_i \neq 0$), there will be an additional component of variation due to treatment effects equal to:

$$r\sum\frac{\tau_i^2}{t-1}$$

"Significant relative to what?"
Significant relative to **error**.

If the effect due to treatment (i.e. signal) is found to be significantly larger than the fluctuations among observations due to error (i.e noise), the treatment effect is said to be real and significant.

The F distribution

From a normally distributed population (or from two populations with equal variance σ^2):

1. Sample n_1 items and calculate their variance s_1^2
2. Sample n_2 items and calculate their variance s_2^2
3. Construct a ratio of these two sample variances $\left(s_1^2 / s_2^2\right)$

This ratio of this statistic will be close to 1 and its expected distribution is called the **F-distribution**, characterized by two values for df ($df_1 = n_1 - 1$, $df_2 = n_2 - 1$).

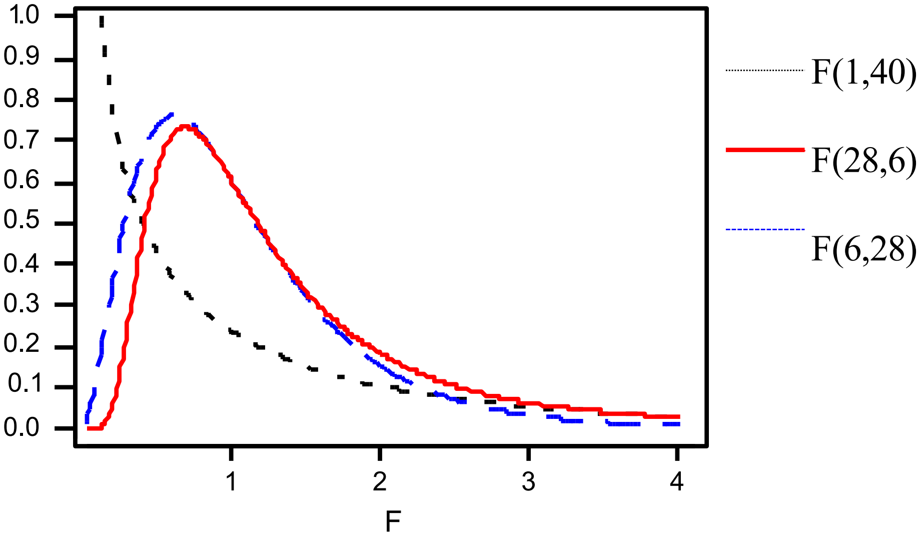


Figure 1 Three example F-distributions

Values in an F table (Table A6) represent the area under the curve to the *right* of the given F-value, with df_1 and df_2.

$$\frac{(n-1)s^2}{\sigma^2} \text{ is distributed according to } \chi^2_{n-1}$$

$$\frac{s_1^2}{s_2^2} \text{ is distributed according to } F_{n_1-1,n_2-1}$$

$F_{(1,9),\,\boldsymbol{\alpha=0.05}} = (t_{9,\,\boldsymbol{\alpha/2}})^2$
$5.12 = 2.262^2$

← Analogous to the relationship between χ^2 and Z.

Example: Pull 10 observations at random from each of two populations. Now test H$_0$: $\sigma_1^2 = \sigma_2^2$ vs. H$_1$: $\sigma_1^2 \neq \sigma_2^2$ (a two-tailed test):

$$F_{\frac{\alpha}{2}=0.025,[df_1=9,df_2=9]} = 4.03$$

Interpretation: The ratio $\left(s_1^2 / s_2^2\right)$, taken from samples of 10 individuals from normally distributed populations with equal variance, is expected to be larger than 4.03 ($F_{\alpha/2=0.025,[9,9]}$) or lower than 0.24 ($F_{1-\alpha/2=0.975,[9,9]}$) *by chance* only 5% of the time.

Testing the hypothesis of equality of two means

The ratio between two estimates of σ^2 can also be used to test differences between *means*:

$$H_0: \mu_1 = \mu_2 \text{ versus } H_1: \mu_1 \neq \mu_2$$

How can we use *variances* to test the differences between *means*? By being creative in how we obtain estimates of σ^2.

$$F = \frac{\text{estimate of } \sigma^2 \text{ from sample means}}{\text{estimate of } \sigma^2 \text{ from individuals}}$$

The denominator is an estimate of σ^2 provided by the individuals *within* a sample. If there are multiple samples, it is a ***weighted average*** of those sample variances.

The numerator is an estimate of σ^2 provided by the means *among* samples. Recall:

$$s_{\bar{Y}(n)}^2 = \frac{s^2}{n}, \text{ so } s^2 = ns_{\bar{Y}(n)}^2$$

$$F = \frac{s_{among}^2}{s_{within}^2} = \frac{ns_{\bar{Y}}^2}{s^2}$$

The fundamental premise underlying ANOVA: When two populations have different *means* (but the same variance), the estimate of σ^2 based on sample means will include a contribution attributable to the difference among population means and F will be higher than expected by chance.

Example: Yields (100 lbs/acre) of wheat varieties 1 and 2 from plots to which the varieties were randomly assigned:

Varieties	**Replications**					$\mathbf{Y_{i.}}$	$\bar{Y}_{\mathbf{i.}}$	$\mathbf{s^2_i}$
1	19	14	15	17	20	85	$\bar{Y}_{1.} = 17$	6.5
2	23	19	19	21	18	100	$\bar{Y}_{2.} = 20$	4.0
						$Y_{..} = 185$	$\bar{Y}_{..} = 18.5$	

Treatments t = 2; Replications n = 5

Begin by assuming that the two populations have the same (unknown) variance σ^2.

1. Estimate the average variance ***within samples*** (the *experimental error*):

$$s_1^2 = \frac{\sum_j (Y_{1j} - \bar{Y}_{1.})^2}{n_1 - 1} \quad , \quad s_2^2 = \frac{\sum_j (Y_{2j} - \bar{Y}_{2.})^2}{n_2 - 1}$$

$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} = 4*6.5 + 4* 4.0 / (4 + 4) = 5.25 \equiv s^2_{within}$$

2. Estimate the variance **between (or among) samples**:

$$s^2_{\bar{Y}} = \frac{\sum_{i=1}^{t} (\bar{Y}_{i.} - \bar{Y}_{..})^2}{t - 1} = [(17 - 18.5)^2 + (20 - 18.5)^2] / (2\text{-}1) = 4.5$$

$$\mathbf{n}\, s^2_{\bar{Y}} = 5 * 4.5 = 22.5 \equiv s^2_{between}$$

To test H_0, we form a ratio of these two estimates:

$$\boldsymbol{F = s_b^2 / s_w^2 = 22.5 / 5.25 = 4.29}$$

Under our assumptions (normality, equal variance), this ratio is distributed according to an $F_{(t-1, t(n-1))} = F_{(1,8)}$ distribution. From Table A.6, we find $F_{0.05,(1,8)} = 5.32$.

$$F_{calc} = 4.29 < 5.32 = F_{crit}$$

SO, we fail to reject H_0 at $\alpha = 0.05$. An F value of 4.29 or larger happens just by chance about 7% of the time for these degrees of freedom.

Modern statistics began in the mind of Ronald Fisher, the first to recognize that variation is not just “noise” drowning “signal,” at best a nuisance to be ignored. Variance itself is a valid object of study, a fingerprint that provides great insight into the mechanisms of natural phenomena. In his words:

> "The populations which are the object of statistical study always display variation in one or more respects. To speak of statistics as the study of variation also serves to emphasize the contrast between the aims of modern statisticians and those of their predecessors. For until comparatively recent times, the vast majority if workers in this field appear to have had no other aim than to ascertain aggregate, or average, values. The variation itself was not an object of study, but was recognized rather as a troublesome circumstance which detracted from the value of the average….From the modern point of view, the study of the causes of variation of any variable phenomenon, from the yield of wheat to the intellect of [people], should be begun by the examination and measurement of the variation which presents itself."
>
> R.A. Fisher
> Statistical Methods for Research Workers (1925)

ANOVA: Single factor designs

The Completely Randomized Design (CRD)

- CRD is the basic ANOVA design
- A single factor is varied to form the different treatments
- These treatments are applied randomly to experimental units
- There are a total of n = rt independent experimental units in the experiment
- H_0: $\mu_1 = \mu_2 = \ldots = \mu_t$ versus H_1: Not all μ_i are equal.

The results of the analysis are usually summarized in an ANOVA table:

Source	df	SS Definition	SS	MS	F
Total	n - 1	$\sum_{i,j}(Y_{ij} - \overline{Y}_{..})^2$	TSS		
Treatments	t – 1	$r\sum_{i}(\overline{Y}_{i.} - \overline{Y}_{..})^2$	SST	SST/(t-1)	MST/MSE
Error	t(r-1) = n - t	$\sum_{i,j}(Y_{ij} - \overline{Y}_{i.})^2$	TSS - SST	SSE/(n-t)	

The mean square for error (MSE): The average dispersion of the observations around their respective group means. It is a valid estimate of a common σ^2, the experimental error, ***if*** the assumption of equal variances is true.

The mean square for treatment (MST): An independent estimate of σ^2, *when the null hypothesis is true*.

The F test: If there are differences among treatment means, there will be an additional source of variation in the experiment due to treatment effects equal to $r\sum\tau_i^2/(t-1)$.

$$\boldsymbol{F = \mathrm{MST/MSE}}$$

$$Expected\ \frac{MST}{MSE} = \frac{\sigma^2 + r\sum\tau_i^2/(t-1)}{\sigma^2}$$

The *F*-test is sensitive to the presence of the added component of variation due to treatment effects. In other words, ANOVA permits us to test whether there are any nonzero treatment effects.

Example

Inoculation of clover with *Rhizobium* strains [ST&D Table 7.1]

Treatments	**3DOK1**	**3DOK5**	**3DOK4**	**3DOK7**	**3DOK13**	**Composite**
Rep 1	19.4	17.7	17.0	20.7	14.3	17.3
Rep 2	32.6	24.8	19.4	21.0	14.4	19.4
Rep 3	27.0	27.9	9.1	20.5	11.8	19.1
Rep 4	32.1	25.2	11.9	18.8	11.6	16.9
Rep 5	33.0	24.3	15.8	18.6	14.2	20.8
Mean	28.8	24.0	14.6	19.9	13.3	18.7
Variance	33.64	14.27	16.94	1.28	2.04	2.56

t = 6, r = 5, overall mean = 19.88

The ANOVA table for this experiment:

Source	**df**	**SS**	**MS**	**F**
Treatment	5	847.05	169.41	14.37**
Error	24	282.93	11.79	
Total	29	1129.98		

1. The mean square error (MSE = 11.79) is just the pooled variance or the average of variances within each treatment (i.e. MSE = $\Sigma s_i^2 / t$).

2. The *F* value (14.37) indicates that the variation among treatments is over 14 times larger than the mean variation within treatments.

$$14.37 > F_{crit} = F_{(5,24),0.05} = 2.62\text{, so we reject } H_0$$

Expected mean squares and F tests

EMS: Algebraic expressions which specify the underlying model parameters estimated by the calculated mean squares and which are used to determine the appropriate error terms for F tests.

EMS table for this one-way (CRD) classification experiment, featuring **t** treatments and **r** replications:

Source	df	MS	EMS
Trtmt	t-1	MST	$\sigma_\varepsilon^2 + r\sum\frac{\tau^2}{t-1}$
Error	t(r-1)	MSE	σ_ε^2

The appropriate test statistic (F) is a ratio of mean squares that is chosen such that the expected value of the *numerator* differs from the expected value of the *denominator* only by the specific factor being tested.

Testing the assumptions associated with ANOVA

1. **Independence of errors:** Guaranteed by the random allocation of experimental units.

2. **Normal distribution of errors:** Shapiro-Wilk test.

3. **Homogeneity of variances:** Several methods are available to test the assumption that variance is the same within each of the groups defined by the independent factor.

Levene's Test: An ANOVA of the absolute values of the residuals.

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}.$$

The residuals (ε_{ij}) are the deviations from the treatment means.

Original data

Treatment	A	B	C
Rep 1	18	17	26
Rep 2	19	15	23
Rep 3	15	13	25
Rep 4	16	15	22
Average	**17**	**15**	**24**

Residuals

Treatment	A	B	C
Rep 1	1	2	2
Rep 2	2	0	-1
Rep 3	-2	-2	1
Rep 4	-1	0	-2
Average	**0**	**0**	**0**

Absolute Values of Residuals

Treatment	A	B	C
Rep 1	1	2	2
Rep 2	2	0	1
Rep 3	2	2	1
Rep 4	1	0	2
Average	**1.5**	**1**	**1.5**

Advantages of the CRD

1. Simple design
2. Can easily accommodate unequal replications per treatment
3. Loss of information due to missing data is small
4. Maximum d.f. for estimating the experimental error
5. Can accommodate unequal variances, using a Welch's variance-weighted ANOVA

The disadvantage

The experimental error includes all the variation in the system except for the component due exclusively to the treatments.

Power

The power of a test is the probability of detecting a nonzero treatment effect. To calculate the power of the F test in an ANOVA, use Pearson and Hartley's power function charts (1953, Biometrika 38:112-130).

To begin, calculate ϕ:

$$\phi = \sqrt{\frac{r}{MSE}\sum\frac{\tau_i^2}{t}}$$

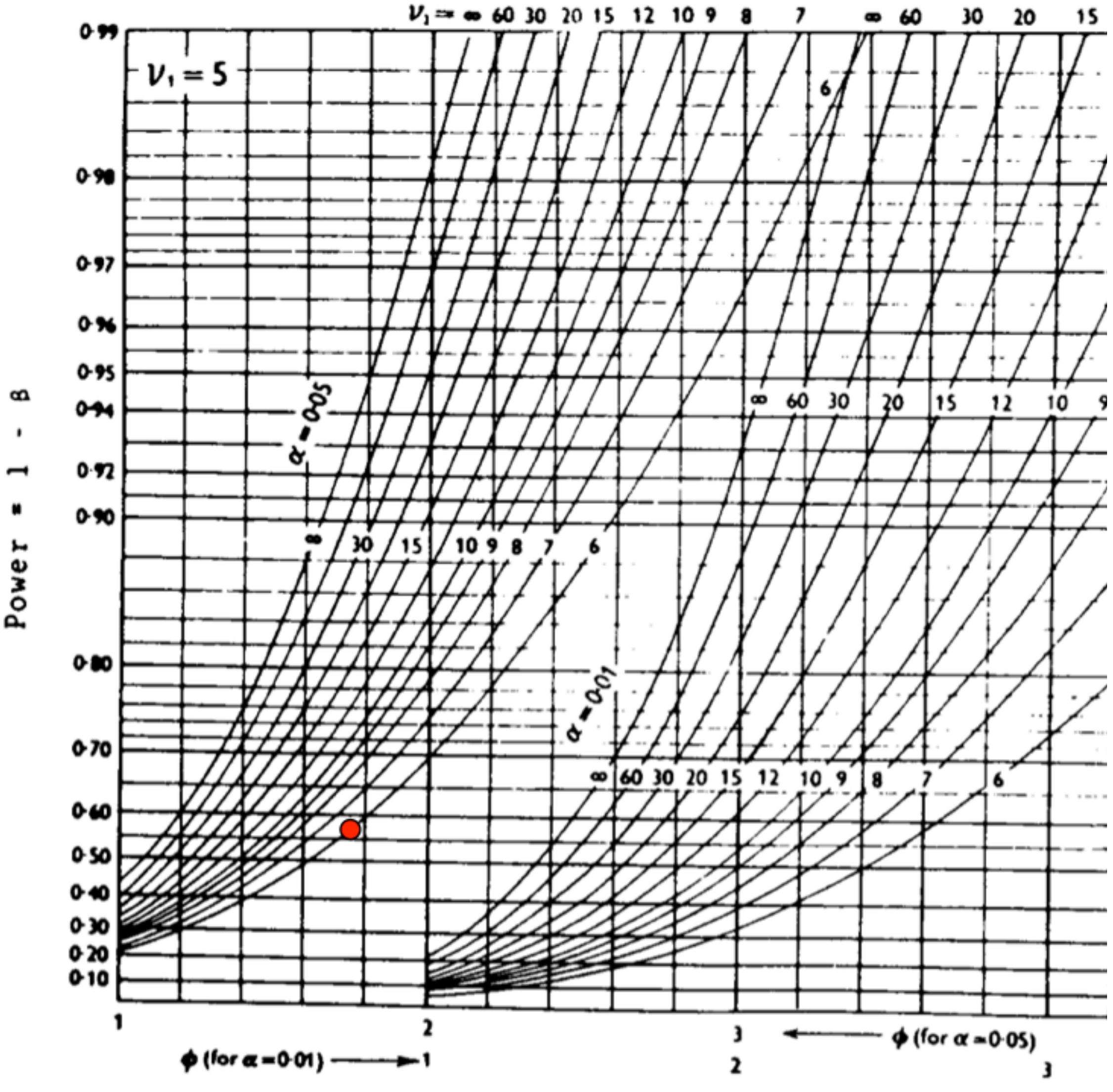
Things to notice:

1. More replications leads to higher ϕ (and higher power).
2. Less error in the model (MSE) leads to higher ϕ.
3. Larger treatment effects (τ_i, our "detection distance") leads to higher ϕ.

Example: Suppose an experiment has t = 6 treatments with r = 2 replications each. Given the MSE and the required α = 5%, you calculate ϕ = 1.75.

$$\phi = \sqrt{\frac{r}{MSE}\sum\frac{\tau_i^2}{t}}$$

To find the power associated with this ϕ, use Chart v_1 = t-1 = 5 and the set of curves corresponding to α = 5%. Select curve v_2 = t(r-1) = 6. The height of this curve corresponding to the abscissa of ϕ = 1.75 is the power of the test.



Power and sample size in analys

In this case, the power is slightly greater than 0.55.

Sample size

To calculate the number of replications for a given α and desired power:

1) Specify the constants
2) Start with an arbitrary r to compute ϕ
3) Use the appropriate chart to find the power
4) Iterate the process until a minimum r value is found which satisfies the required power for a given α level.

We can simplify the general power formula if we assume all $\boldsymbol{\tau_i}$ are zero except the two extreme treatment effects (let's call them τ_K and τ_L , so that **d** = $|\mu_K - \mu_L|$:

$$\phi = \sqrt{\frac{d^2 * r}{2t * MSE}}$$

Example: Suppose that 6 treatments will be involved in a study and the anticipated difference between the extreme means is 15 units. What is the required sample size so that this difference will be detected at α = 1% and power = 90%, knowing that σ^2 = 12? (note, t = 6, α = 0.01, β = 0.10, d = 15, and MSE = 12).

$$\phi = \sqrt{\frac{d^2 * r}{2t * MSE}}$$

r	df	ϕ	(1-β) for α=1%
2	6(2-1)= 6	1.77	0.22
3	6(3-1)= 12	2.17	0.71
4	6(4-1)= 18	2.50	0.93

Thus 4 replications are required for each treatment to satisfy the required conditions.