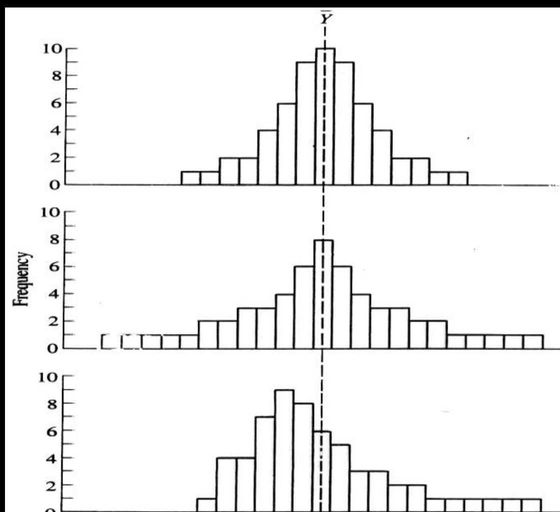
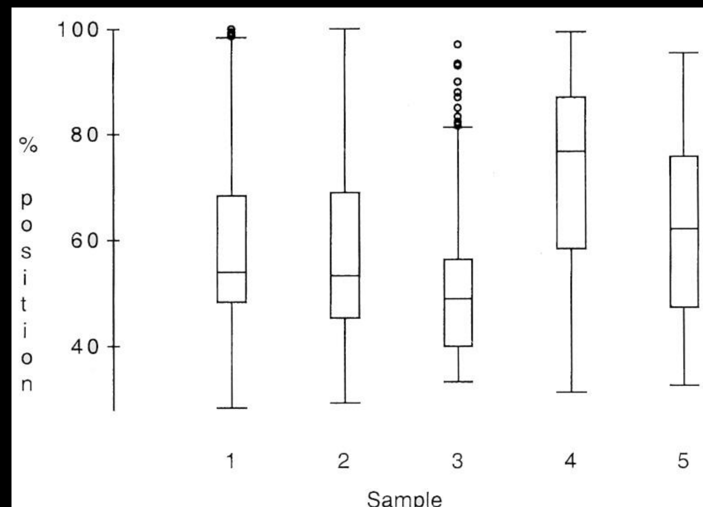
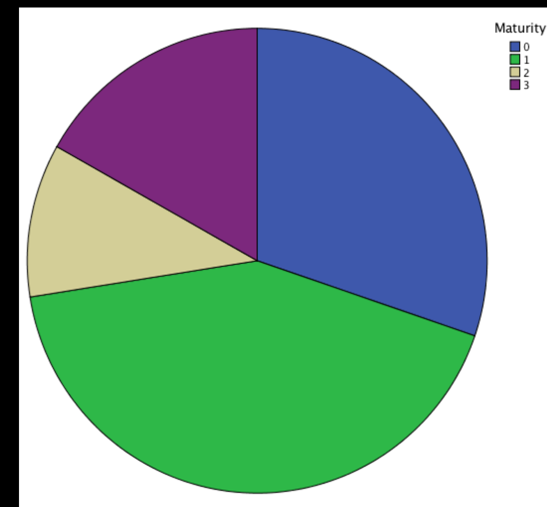
Data Handling :

A practical approach







Lecture 5 ANOVA I

Dr Yu Mo, Zoology

moyu@tcd.ie | https://github.com/github-moyu/Teaching

Summary of lecture 4

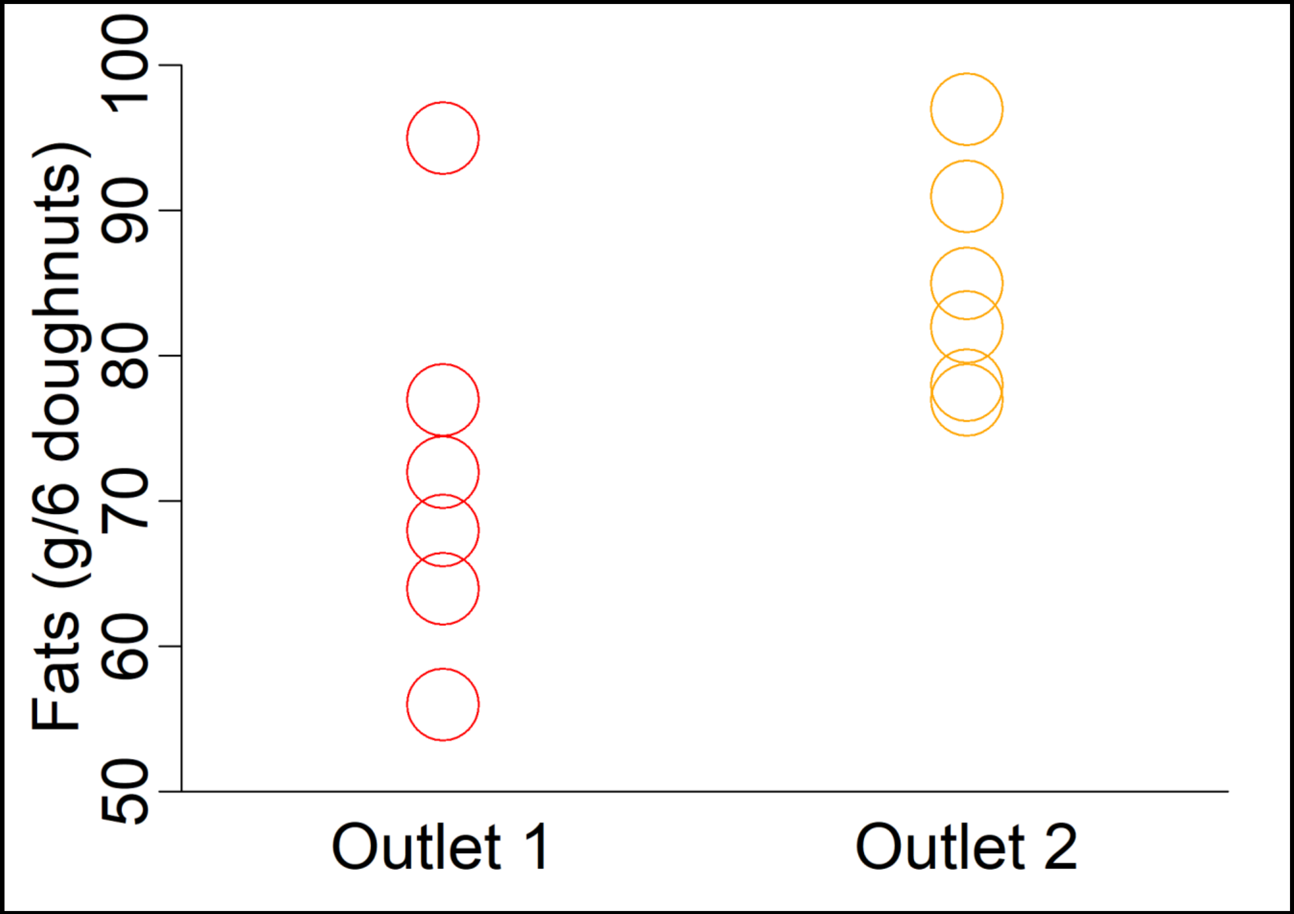
- The importance of **plotting** data prior to analysis
- The need for **summary statistics**
- The central importance of erecting a **null and alternate hypothesis**
- Comparison of two groups using a **t-test**
- Generation of **t** and **p** value, **df**

Fats in doughnuts

Outlet	Fat
1	64
1	72
1	68
1	77
1	56
1	95

Outlet	Fat
2	78
2	91
2	97
2	82
2	85
2	77

Fats in doughnuts

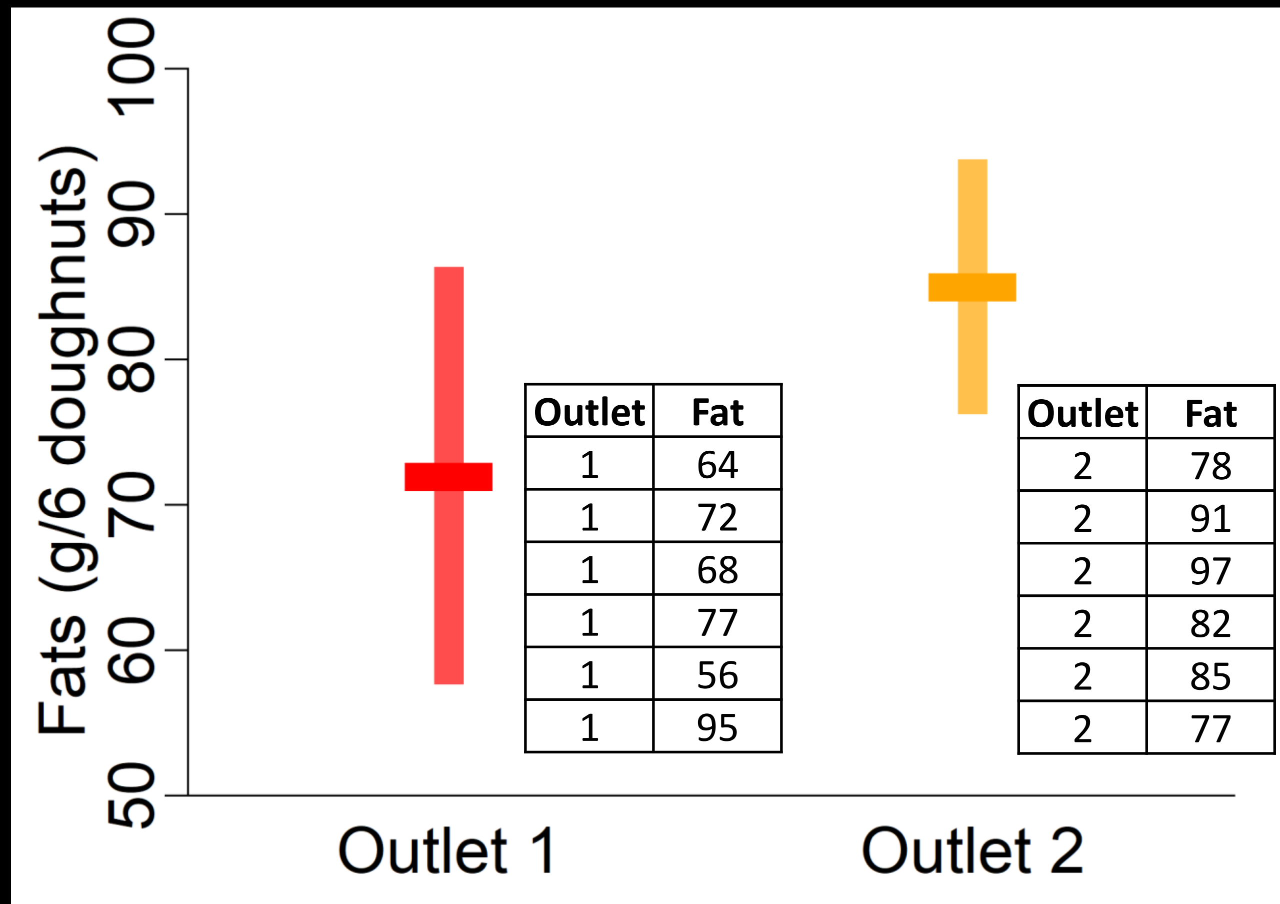


Fats in doughnuts

Mean ± SD

$$\bar{X} = \frac{\sum X}{n}$$

$$SD = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$$



Outlet	Fat
1	64
1	72
1	68
1	77
1	56
1	95

Outlet	Fat
2	78
2	91
2	97
2	82
2	85
2	77

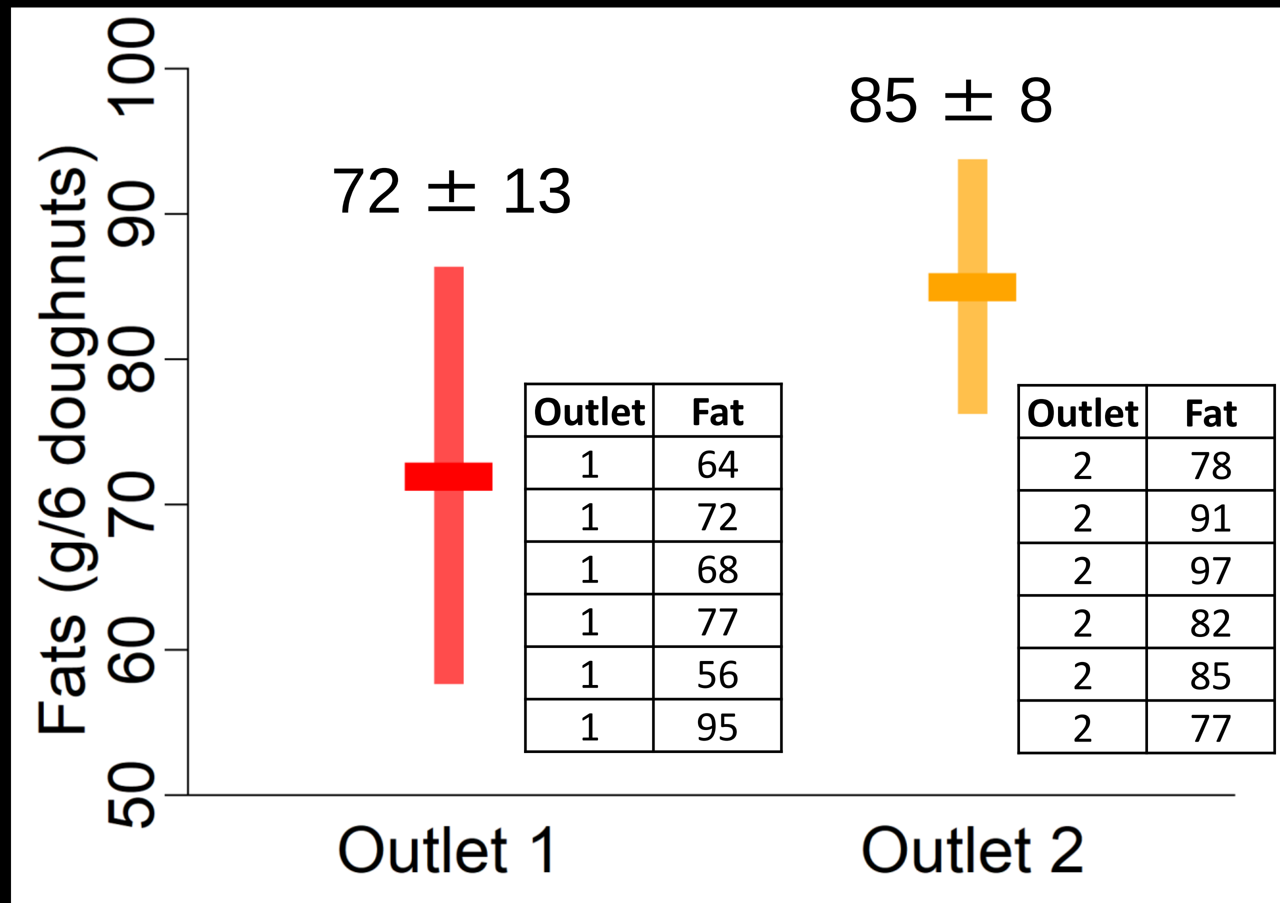
Fats in doughnuts

Independent t-test

t = -2.0624
df = 10
p = 0.06612

If alpha 0.05

Fail <u>to</u> reject H_0



Outlet	Fat
1	64
1	72
1	68
1	77
1	56
1	95

Outlet	Fat
2	78
2	91
2	97
2	82
2	85
2	77

Fats in doughnuts

Analysis of variance (ANOVA)

- Extension of the independent t-test
- Comparison of more than two means
- Can extend to much more complex analyses
  - Assumptions
    - Normality
    - Equality of variances
    - Transformation may help to fulfill these requirements

Null and Alternate hypotheses

- Simply

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1: not all the μ_s are equal

- Model approach

H_0: each obs. $Y_i = \mu + E_i$

μ = overall mean
E_i = a random error term

H_1: each obs. $Y_{ij} = \mu + \alpha_j + E_{ij}$

μ = overall mean
α_j = group/treatment effect
E_i = a random error term

ANOVA

- Between group variability: a measure of the difference between the means for each group and that of the grand mean
- Within group variability: a measure of the difference between each individual value and that of the individual's group mean
- F value: $Var_{between}$ / Var_{within}

ANOVA

Sum of Squares BETWEEN

$$SS_B = \sum n_j(\bar{\boldsymbol{X}}_{,j} - \bar{\bar{X}})^2$$

Mean of Squares BETWEEN

$$MS_B = \frac{SS_B}{j-1}$$

Sum of Squares WITHIN

$$SS_w = \sum_{j=1}^{k} \sum_{i=1}^{n} \left(X_{i,j} - \bar{\boldsymbol{X}}_{,j}\right)^2$$

Mean of Squares WITHIN

$$MS_w = \frac{SS_w}{n-j}$$

$$F = \frac{MS_B}{MS_w}$$

The data set : Doughnuts and fats
4 groups : 6 observations per group

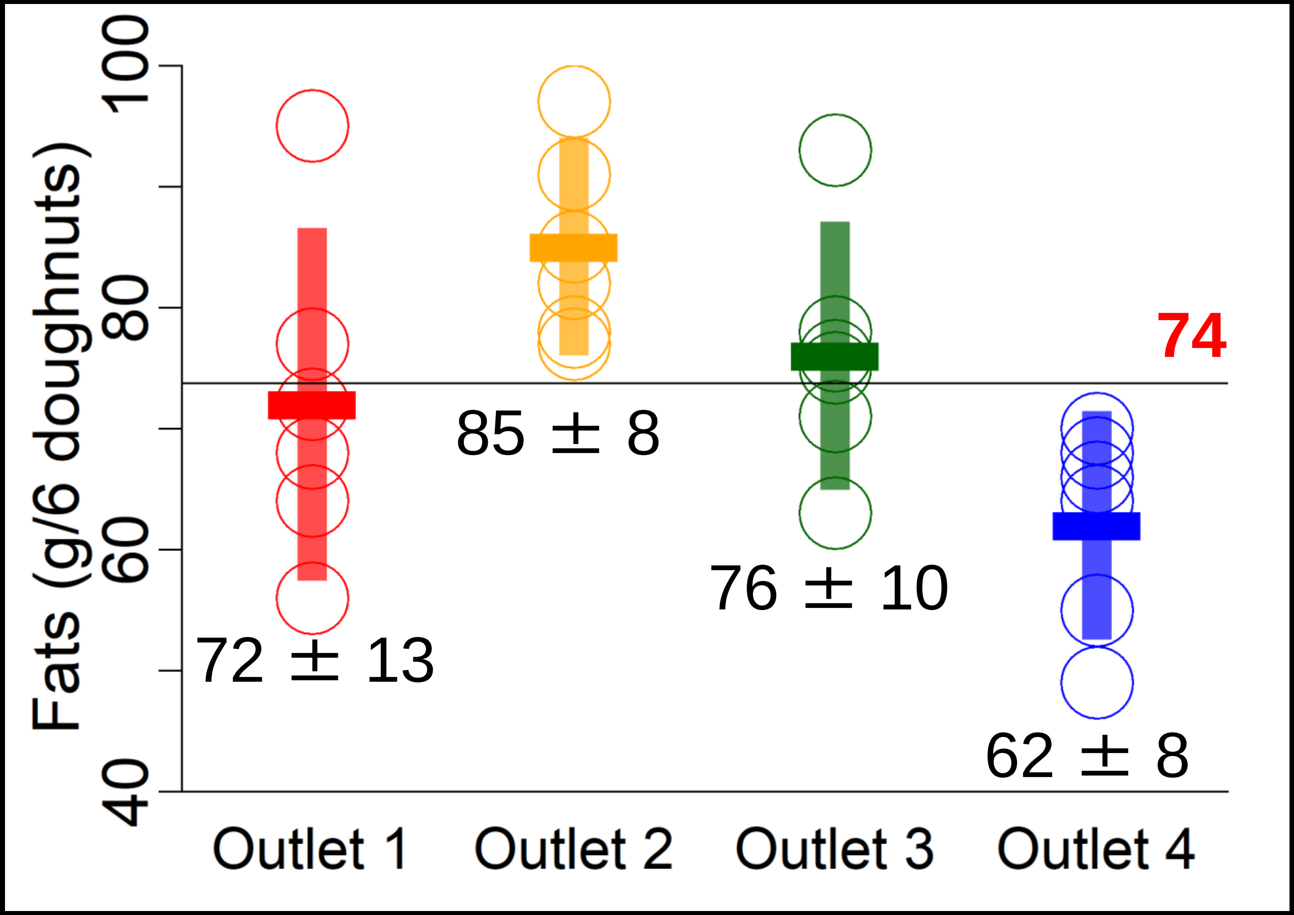
Obs	Outlet 1	Outlet 2	Outlet 3	Outlet 4	
1	64	78	75	55	
2	72	91	93	66	
3	68	97	78	49	
4	77	82	71	64	
5	56	85	63	70	
6	95	77	76	68	
Mean	72	85	76	62	74
S.D.	13.34	7.77	9.88	8.22	

Balanced, fully replicated, one factor design with X levels of factor

Analyse by ONE-WAY ANOVA

Balanced designs have equal numbers of obs. in each factor combination. Statistics are simpler and more powerful

Fats in doughnuts



Within group variance example

$$SS_w = \sum_{j=1}^{k} \boxed{\sum_{i=1}^{n} \left(X_{i,j} - \bar{X}_{,j}\right)^2}$$

Obs	Outlet 1	$\bar{X}_{,1}$	$X_{i,1}$-$\bar{X}_{,1}$	$X_{i,1}$–$\bar{X}_{,1}{}^2$
1	64	72	-8	64
2	72	72	0	0
3	68	72	-4	16
4	77	72	5	25
5	56	72	-16	256
6	95	72	23	529
Sum	-	-	-	890

Calculation of Mean Square Within

Sum of square within group

$$SS_w = SS_{,1} + SS_{,2} + SS_{,3} + SS_{,4}$$
$$= 890 + 302 + 488 + 338$$
$$= 2018$$

$$SS_w = \boxed{\sum_{j=1}^{k}} \sum_{i=1}^{n} \left(X_{i,j} - \bar{\boldsymbol{X}}_{,j}\right)^2$$

Man of square within group

$$MS_w = 2018 \,/\, (24\text{-}4)$$
$$= 100.9$$

$$MS_w = \frac{SS_w}{n - j}$$

Calculation of Mean Square Between

Sum of square between group

SS_B = 6*(72-73.75)^2 + 6*(85-73.75)^2 +
6*(76-73.75)^2 + 6*(62-73.75)^2

=1636

$$SS_B = \sum n_j(\bar{\boldsymbol{X}}_{,\boldsymbol{j}} - \bar{\bar{X}})^2$$

Mean of square between group

MS_B =1636 / (4-1)

= 545.3

$$MS_B = \frac{SS_B}{j-1}$$

Calculation of F ratio

$F = MS_B/MS_W$

With j-1 and n-j degrees of freedom

j-1 = numerator (j number of groups)

n-j = denominator (n number of obs)

df Denominator n-j: 24-4 = 20

	Alpha	1	2	3	4	5	V1
V2							
16							
17							
18							
19							
	0.75						
	0.5						
	0.25						
	0.1						
20	0.05	4.35	3.49	3.1	2.87	2.71	
	0.025			3.86			
	0.01			4.94			
	0.005			5.82			
	0.001			8.10			

df
Numerator
j-1
4-1 = 3

Critical values of the F distribution

Expressed in terms of Null and alternate hypothesis

F < 3.10 accept Ho

F > 3.10 reject Ho

$H_0 = MS_B \leq MS_W$

$H_1 = MS_B > MS_W$

Two important values

Sum of Squares WITHIN	Sum of Squares BETWEEN
$SS_w = 2018$	$SS_B = 1636$
$MS_w = 100.9$	$MS_B = 545.3$

$F = \frac{MS_B}{MS_w} = 545.3/100.9 = 5.40$

F table at 3,20 df(j-1),(n-j)
F < 3.1 accept H0
F > 3.1 reject H0

```
>     #ANOVA
>     oneway.test(Fat~Outlet, data=data,var.equal = TRUE )

        One-way analysis of means

data:  Fat and Outlet
F = 5.4063, num df = 3, denom df = 20, p-value = 0.006876

>     data$Outlet2 <- as.factor(data$Outlet)
>     aov <- aov(Fat~Outlet2, data=data)
>   summary(aov)
            Df Sum Sq Mean Sq F value  Pr(>F)
Outlet2      3   1636   545.5   5.406 0.00688 **
Residuals   20   2018   100.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Why can't we do a series of t-tests ?

	t-test	ANOVA
Ho:	$\mu_1 = \mu_2$ $\mu_1 = \mu_3$ $\mu_1 = \mu_4$ $\mu_2 = \mu_3$ $\mu_2 = \mu_4$ $\mu_3 = \mu_4$	$\mu_1 = \mu_2 = \mu_3 = \mu_4$
Type I error:	$1-(.95)^6 = .26$	0.5

26% chance of suggesting an effect when there isn't one!