Analysis of Variance

The sample mean and variance

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed (iid).

- The sample mean was defined as

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

- The sample variance was defined as

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

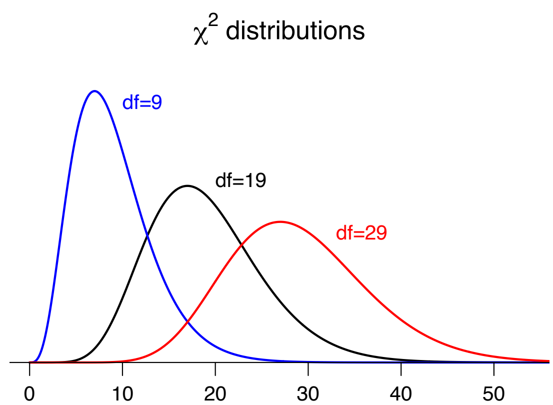
I haven't spoken much about variances (I generally prefer looking at the SD), but we are about to start making use of them!

The distribution of the sample variance

If X_1, X_2, ..., X_n are iid Normal (mean=μ, var=σ^2),

then the sample variance S^2 satisfies $(n-1)\ S^2/\sigma^2 \sim \chi^2_{n-1}$

$\longrightarrow$ When the X_i are not normally distributed, this is not true.



Let $W \sim \chi^2(\text{df} = n - 1)$

E(W) = n–1

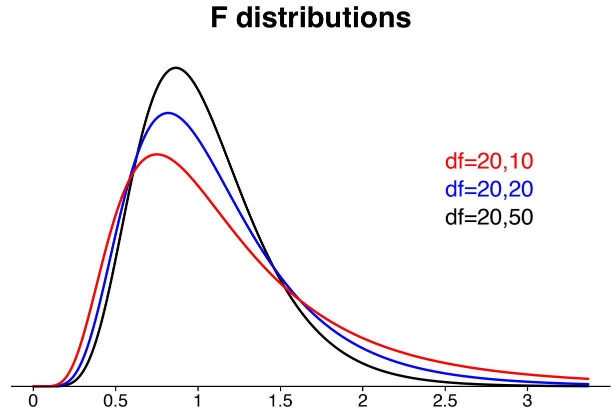
var(W) = 2(n–1)

SD(W) = $\sqrt{2(n-1)}$

The F distribution

Let $Z_1 \sim \chi^2_m$, and $Z_2 \sim \chi^2_n$. Assume Z_1 and Z_2 are independent.

$\longrightarrow$ Then $\dfrac{Z_1/m}{Z_2/n} \sim F_{m,n}$



The distribution of the sample variance ratio

Let $X_1, X_2, \ldots, X_m$ be iid Normal (μ_x, σ_x^2).

Let $Y_1, Y_2, \ldots, Y_n$ be iid Normal (μ_y, σ_y^2).

Then $(m-1) \times S_x^2/\sigma_x^2 \sim \chi^2_{m-1}$ and $(n-1) \times S_y^2/\sigma_y^2 \sim \chi^2_{n-1}$.

Hence

$$\frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} \sim F_{m-1,n-1}$$

or equivalently

$$\frac{S_x^2}{S_y^2} \sim \frac{\sigma_x^2}{\sigma_y^2} \times F_{m-1,n-1}$$

Hypothesis testing

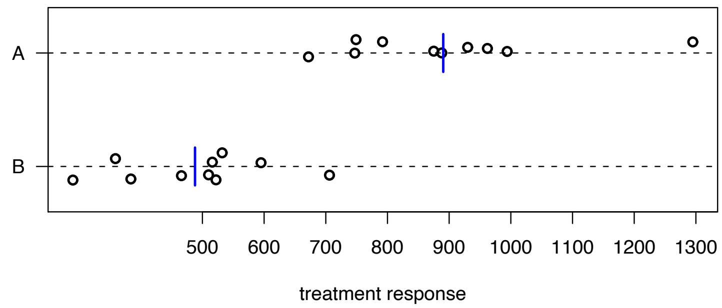
Let $X_1, X_2, \ldots, X_m$ be iid Normal (μ_x, σ_x^2).

Let $Y_1, Y_2, \ldots, Y_n$ be iid Normal (μ_y, σ_y^2).

We want to test H_0: $\sigma_x^2 = \sigma_y^2$ versus H_a: $\sigma_x^2 \neq \sigma_y^2$

$\longrightarrow$ Under the null hypothesis $S_x^2/S_y^2 \sim F_{m-1,n-1}$

Example



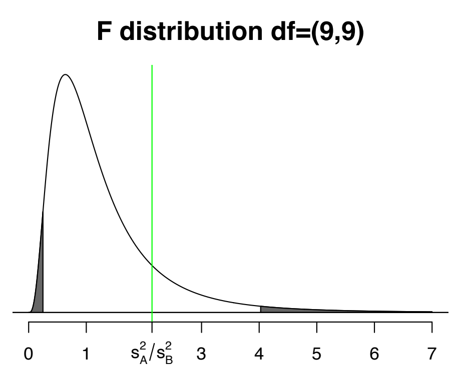
Are the variances the same in the two groups?

Example

We want to test H_0: $\sigma_A^2 = \sigma_B^2$ versus H_a: $\sigma_A^2 \neq \sigma_B^2$

$\longrightarrow$ At the 5% level, we reject the null hypothesis if our test statistic, the ratio of the sample variances (treatment group A versus B), is below 0.25 or above 4.03.

The ratio of the sample variances in our example is 2.14. We therefore do not reject the null hypothesis.



Confidence interval for the variance ratio

Let $X_1, X_2, \ldots, X_m$ be iid Normal (μ_x, σ_x^2).

Let $Y_1, Y_2, \ldots, Y_n$ be iid Normal (μ_y, σ_y^2). *X,Y* independent.

$$\frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} \sim F_{m-1,n-1}$$

Let L be the 2.5th and U be the 97.5th percentile of F(m–1, n–1).

$\longrightarrow$ $\Pr\{L < (S_x^2/\sigma_x^2)/(S_y^2/\sigma_y^2) < U\} = 95\%$.

$\longrightarrow$ $\Pr\{(S_x^2/S_y^2)/U < \sigma_x^2/\sigma_y^2 < (S_x^2/S_y^2)/L\} = 95\%$.

Thus, the interval $\{\ (S_x^2/S_y^2)/\ U\ ,\ \ (S_x^2/S_y^2)/\ L\ \}$
is a 95% confidence interval for σ_x^2/σ_y^2.

Example

m = 10; n = 10.

2.5th and 97.5th percentiles of F(9,9) are 0.248 and 4.026.

Note that, since m = n, L = 1/U.

$s_x^2/s_y^2 = 2.14$

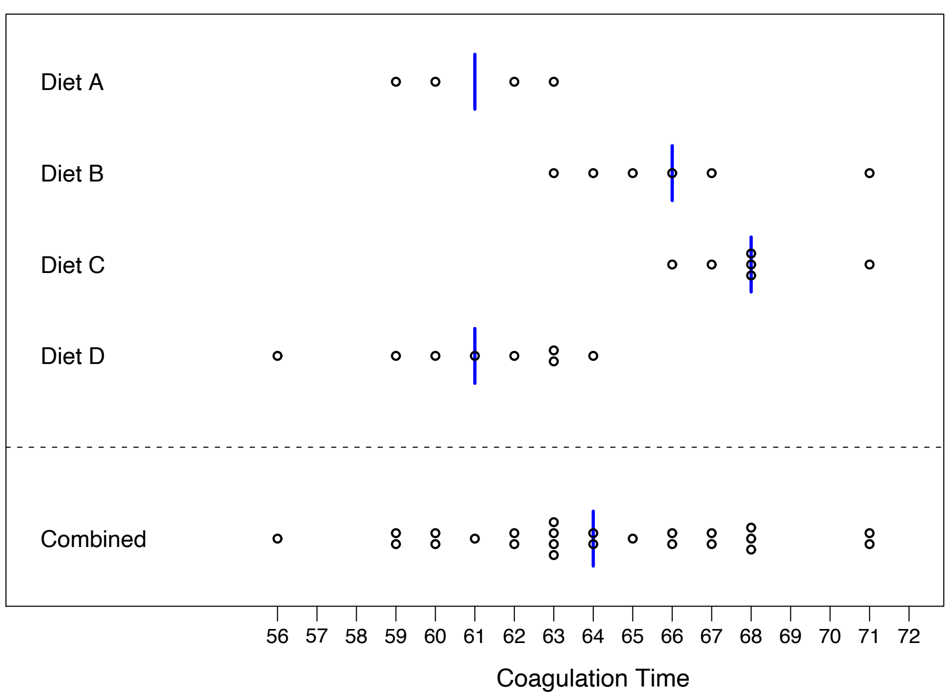
$\longrightarrow$ The 95% confidence interval for σ_x^2/σ_y^2 is
(2.14 / 4.026, 2.14 / 0.248) = (0.53, 8.6)

How about a 95% confidence interval for σ_x/σ_y?

Blood coagulation time

T		avg
A	62 60 63 59	61
B	63 67 71 64 65 66	66
C	68 66 71 67 68 68	68
D	56 62 60 61 63 64 63 59	61
		64

Blood coagulation time



Notation

Assume we have k treatment groups.

n_t	number of cases in treatment group t
N	number of cases (overall)
Y_{ti}	response i in treatment group t
$\bar{Y}_{t\cdot}$	average response in treatment group t
$\bar{Y}_{\cdot\cdot}$	average response (overall)

Estimating the variability

We assume that the data are random samples from four normal distributions having the same variance σ^2, differing only (if at all) in their means.

We can estimate the variance σ^2 for each treatment t, using the sum of squared differences from the averages within each group.

Define, for treatment group t,

$$S_t = \sum_{i=1}^{n_t} (Y_{ti} - \bar{Y}_{t\cdot})^2.$$

Then

$$E(S_t) = (n_t - 1) \times \sigma^2.$$

Within group variability

The within-group sum of squares is the sum of all treatment sum of squares:

$$S_W = S_1 + \cdots + S_k = \sum_t \sum_i (Y_{ti} - \bar{Y}_{t\cdot})^2$$

The within-group mean square is defined as

$$M_W = \frac{S_1 + \cdots + S_k}{(n_1 - 1) + \cdots + (n_k - 1)} = \frac{S_W}{N - k} = \frac{\sum_t \sum_i (Y_{ti} - \bar{Y}_{t\cdot})^2}{N - k}$$

It is our first estimator of σ^2.

Between group variability

The between-group sum of squares is

$$S_B = \sum_{t=1}^{k} n_t (\bar{Y}_{t\cdot} - \bar{Y}_{\cdot\cdot})^2$$

The between-group mean square is defined as

$$M_B = \frac{S_B}{k - 1} = \frac{\sum_t n_t (\bar{Y}_{t\cdot} - \bar{Y}_{\cdot\cdot})^2}{k - 1}$$

It is our second estimator of σ^2.

That is, if there is no treatment effect!

Important facts

The following are facts that we will exploit later for some formal hypothesis testing:

- The distribution of S_W/σ^2 is χ^2(df=N-k)
- The distribution of S_B/σ^2 is χ^2(df=k-1)   if there is no treatment effect!
- S_W and S_B are independent

Variance contributions

$$\sum_t \sum_i (Y_{ti} - \bar{Y}_{..})^2 = \sum_t n_t(\bar{Y}_{t.} - \bar{Y}_{..})^2 + \sum_t \sum_i (Y_{ti} - \bar{Y}_{t.})^2$$

$$S_T = S_B + S_W$$

$$N - 1 = k - 1 + N - k$$

ANOVA table

source	sum of squares	df	mean square
between treatments	$S_B = \sum_t n_t(\bar{Y}_{t\cdot} - \bar{Y}_{\cdot\cdot})^2$	$k-1$	$M_B = S_B/(k-1)$
within treatments	$S_W = \sum_t \sum_i (Y_{ti} - \bar{Y}_{t\cdot})^2$	$N-k$	$M_W = S_W/(N-k)$
total	$S_T = \sum_t \sum_i (Y_{ti} - \bar{Y}_{\cdot\cdot})^2$	$N-1$	

Example

source	sum of squares	df	mean square
between treatments	228	3	76.0
within treatments	112	20	5.6
total	340	23	

The ANOVA model

We write $Y_{ti} = \mu_t + \epsilon_{ti}$ with $\epsilon_{ti} \sim$ iid N(0,σ^2).

Using $\tau_t = \mu_t - \mu$ we can also write

$$Y_{ti} = \mu + \tau_t + \epsilon_{ti}.$$

The corresponding analysis of the data is

$$y_{ti} = \bar{y}_{..} + (\bar{y}_{t.} - \bar{y}_{..}) + (y_{ti} - \bar{y}_{t.})$$

The ANOVA model

Three different ways to describe the model:

A. Y_{ti} independent with $Y_{ti} \sim N(\mu_t, \sigma^2)$

B. $Y_{ti} = \mu_t + \epsilon_{ti}$ where $\epsilon_{ti} \sim$ iid $N(0, \sigma^2)$

C. $Y_{ti} = \mu + \tau_t + \epsilon_{ti}$ where $\epsilon_{ti} \sim$ iid $N(0, \sigma^2)$ and $\sum_t \tau_t = 0$

Hypothesis testing

We assume

$$Y_{ti} = \mu + \tau_t + \epsilon_{ti} \qquad \text{with} \qquad \epsilon_{ti} \sim \text{iid N}(0,\sigma^2).$$

Equivalently, $Y_{ti} \sim$ independent $N(\mu_t, \sigma^2)$

We want to test

$$H_0 : \tau_1 = \cdots = \tau_k = 0 \qquad \text{versus} \qquad H_a : H_0 \text{ is false.}$$

Equivalently, $H_0 : \mu_1 = \ldots = \mu_k$

For this, we use a one-sided F test.

Another fact

It can be shown that

$$E(M_B) = \sigma^2 + \frac{\sum_t n_t \tau_t^2}{k-1}$$

Therefore

$$E(M_B) = \sigma^2 \quad \text{if } H_0 \text{ is true}$$

$$E(M_B) > \sigma^2 \quad \text{if } H_0 \text{ is false}$$

Recipe for the hypothesis test

Under H_0 we have

$$\frac{M_B}{M_W} \sim F_{k-1,\, N-k}.$$

Therefore

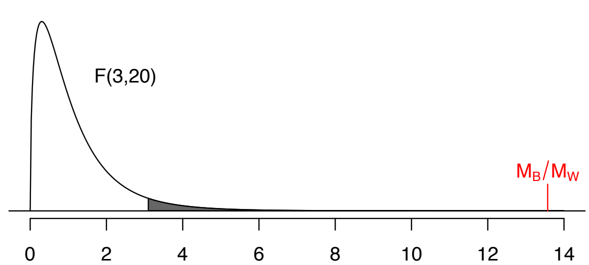
- Calculate M_B and M_W.
- Calculate M_B/M_W.
- Calculate a p-value using M_B/M_W as test statistic, using the right tail of an F distribution with $k-1$ and $N-k$ degrees of freedom.

Example (cont)

$H_0 : \tau_1{=}\tau_2{=}\tau_3{=}\tau_4{=}0$ versus $H_a : H_0$ is false.

$M_B = 76$, $M_W = 5.6$, therefore $M_B/M_W = 13.57$.

Using an F distribution with 3 and 20 degrees of freedom, we get a pretty darn low p-value. Therefore, we reject the null hypothesis.



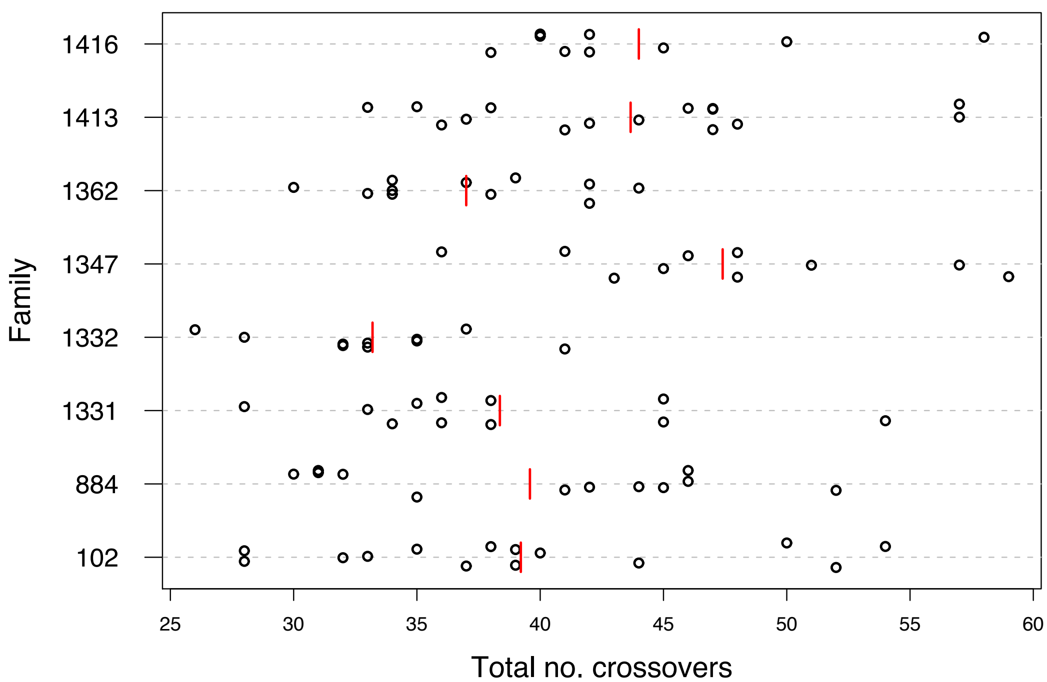
The R function aov() does all these calculations for you!

Example

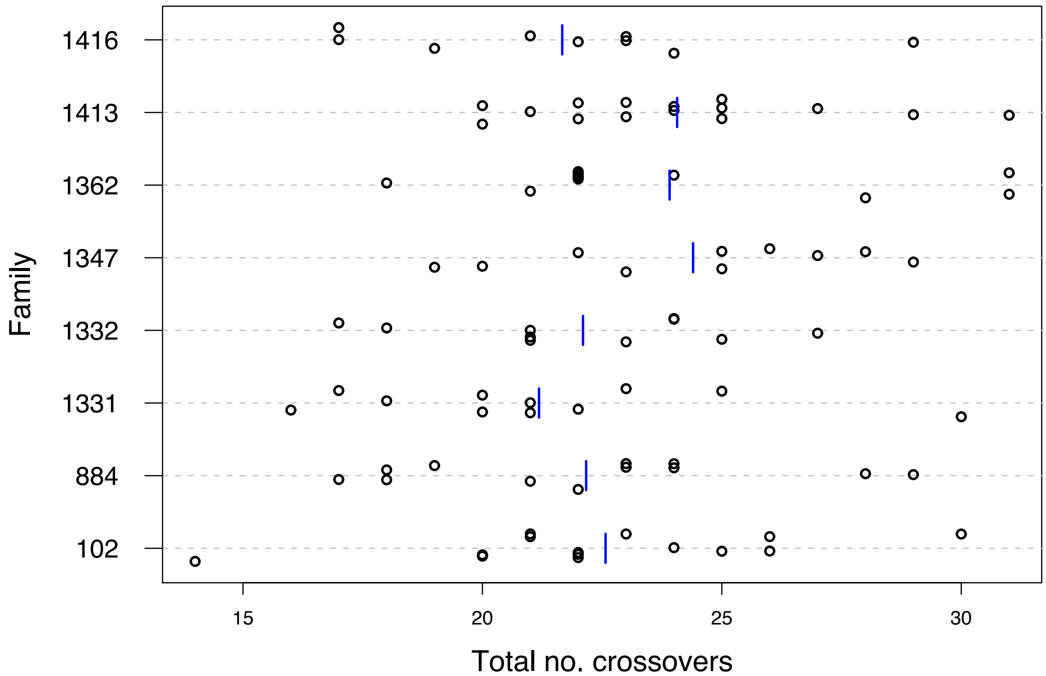
For each of 8 mothers and 8 fathers, we observe (estimates of) the number of crossovers, genome-wide, in a set of independent meiotic products.

⟶ Do the fathers (or mothers) vary in the number of crossovers they deliver?

Female meioses



Male meioses



ANOVA tables

Female meioses:

source	SS	df	MS	F	P-value
between families	1485	7	212.2	4.60	0.0002
within families	3873	84	46.1		
total	5358	91			

Male meioses:

source	SS	df	MS	F	P-value
between families	114	7	16.3	1.23	0.30
within families	1112	84	13.2		
total	1226	91			

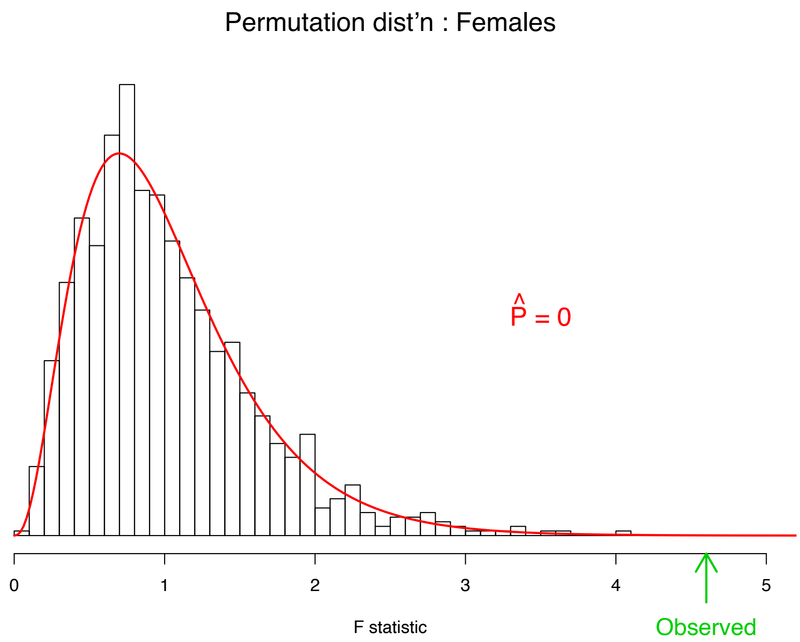
Permutation test

The P-values calculated above are based on the assumption that the measurements in the underlying populations are normally distributed.

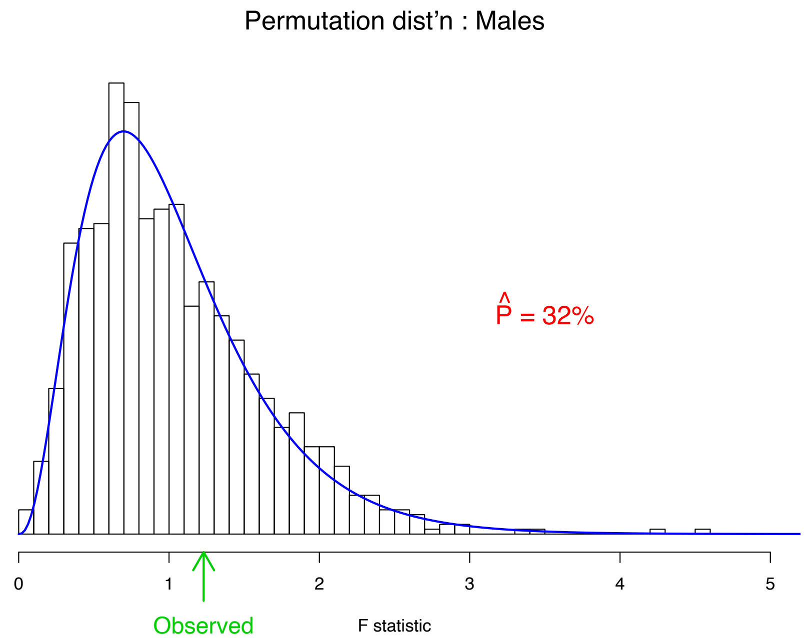
Alternatively, one may use a permutation test to obtain P-values:

1. Permute (shuffle) the XO counts relative to the family IDs.
2. Re-calculate the F statistic.
3. Repeat (1) and (2) many times (1000 or 10,000 times, say).
4. Estimate the P-value as the proportion of the F statistics from permuted data that are bigger or equal to the observed F statistic.

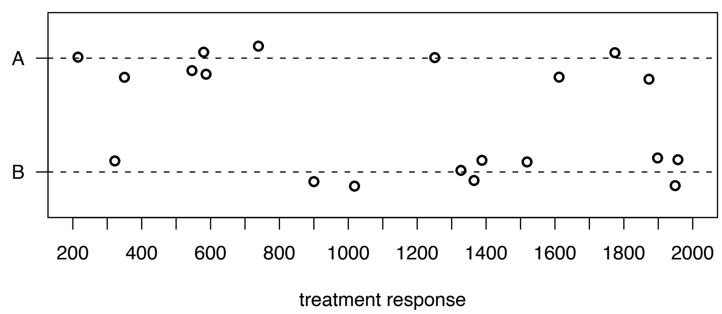
Female meioses



Male meioses



Another example



Are the population means the same?

By now, we know two ways of testing that:

Two-sample t-test, and ANOVA with two treatments.

$\longrightarrow$ But do they give similar results?

ANOVA table

source	sum of squares	df	mean square
between treatments	$S_B = \sum_t n_t(\bar{Y}_{t\cdot} - \bar{Y}_{\cdot\cdot})^2$	$k - 1$	$M_B = S_B/(k-1)$
within treatments	$S_W = \sum_t \sum_i (Y_{ti} - \bar{Y}_{t\cdot})^2$	$N - k$	$M_W = S_W/(N-k)$
total	$S_T = \sum_t \sum_i (Y_{ti} - \bar{Y}_{\cdot\cdot})^2$	$(N - 1)$	

ANOVA for two groups

The ANOVA test statistic is M_B/M_W, with

$$M_B = n_1(\bar{Y}_1 - \bar{Y}_{\cdot\cdot})^2 + n_2(\bar{Y}_2 - \bar{Y}_{\cdot\cdot})^2$$

and

$$M_W = \frac{\sum_{i=1}^{n_1}(Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2}(Y_{2i} - \bar{Y}_2)^2}{n_1 + n_2 - 2}$$

Two-sample t-test

The test statistic for the two sample t-test is

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s\sqrt{1/n_1 + 1/n_2}}$$

with

$$s^2 = \frac{\sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)^2}{n_1 + n_2 - 2}$$

This also assumes equal variance within the groups!

Result

$$\frac{M_B}{M_W} = t^2$$

Reference distributions

If there was no difference in means, then

$$\frac{M_B}{M_W} \sim F_{1,n_1+n_2-2}$$

$$t \sim t_{n_1+n_2-2}$$

Now does this mean $F_{1,n_1+n_2-2} = (t_{n_1+n_2-2})^2$?

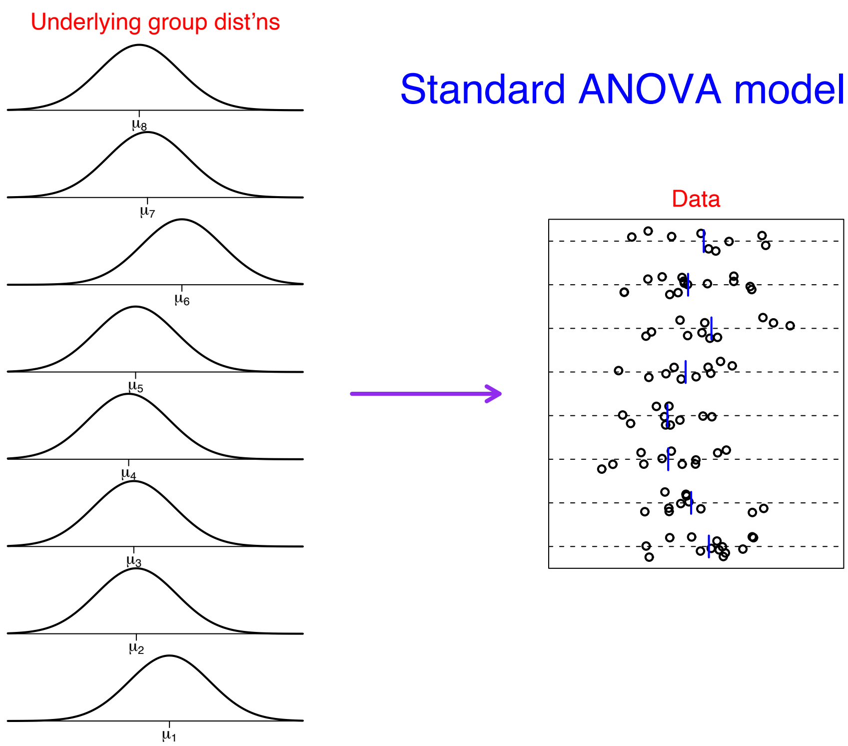
A few facts

$$F_{1,k} = t_k^2$$

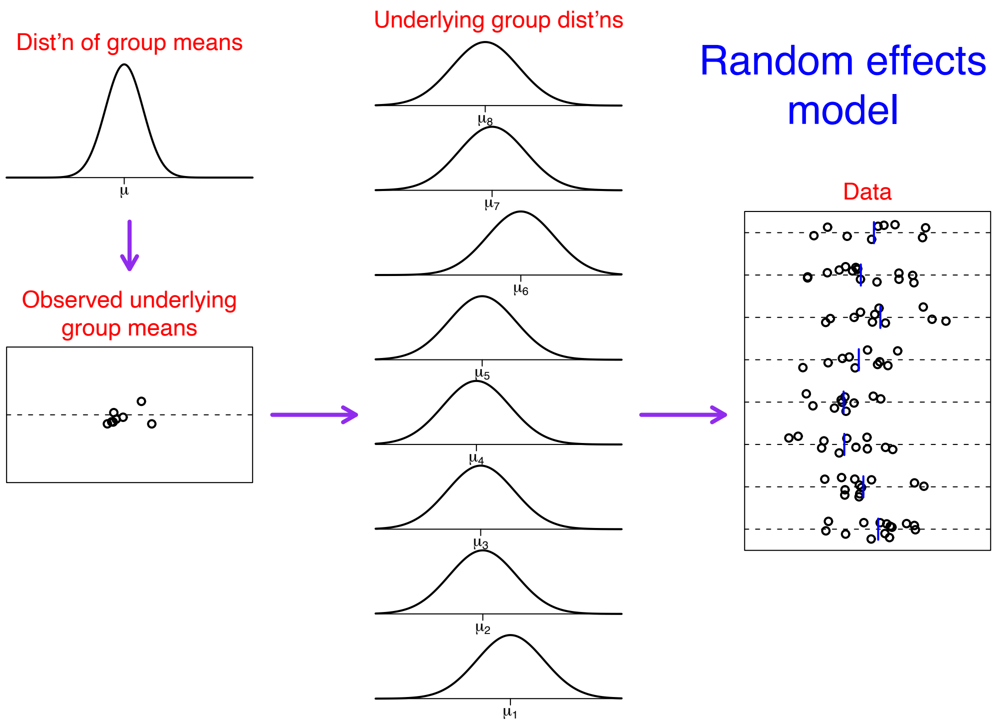
$$F_{k,\infty} = \frac{\chi_k^2}{k}$$

$$N(0,1)^2 = \chi_1^2 = F_{1,\infty} = t_\infty^2$$

Fixed effects



Random effects



The random effects model

Two different ways to describe the model:

A. $\mu_t \sim$ iid $N(\mu, \sigma_A^2)$

$$Y_{ti} = \mu_t + \epsilon_{ti} \text{ where } \epsilon_{ti} \sim \text{ iid } N(0, \sigma^2)$$

B. $\tau_t \sim$ iid $N(0, \sigma_A^2)$

$$Y_{ti} = \mu + \tau_t + \epsilon_{ti} \text{ where } \epsilon_{ti} \sim \text{ iid } N(0, \sigma^2)$$

$\longrightarrow$ We add another layer of sampling.

Hypothesis testing

$\rightarrow$ In the standard ANOVA model, we considered the μ_t as fixed but unknown quantities.

We test the hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ (versus H_0 is false) using the statistic M_B/M_W from the ANOVA table and the comparing this to an $F(k-1, N-k)$ distribution.

$\rightarrow$ In the random effects model, we consider the μ_t as random draws from a normal distribution with mean μ and variance σ_A^2.

We seek to test the hypothesis $H_0 : \sigma_A^2 = 0$ versus $H_a : \sigma_A^2 > 0$.

As it turns out, we end up with the same test statistic and same null distribution. For one-way ANOVA, that is!

Estimation

For the random effects model it can be shown that

$$\mathrm{E}(\mathrm{M_B}) = \sigma^2 + \mathrm{n_0} \times \sigma_A^2$$

where

$$\mathrm{n_0} = \frac{1}{k-1}\left(N - \frac{\sum_t n_t^2}{\sum_t n_t}\right)$$

Recall also that $\mathrm{E}(\mathrm{M_W}) = \sigma^2$.

Thus, we may estimate σ^2 by $\hat{\sigma}^2 = \mathrm{M_W}$.

And we may estimate σ_A^2 by $\hat{\sigma}_A^2 = (\mathrm{M_B} - \mathrm{M_W})/\mathrm{n_0}$

(provided that this is ≥ 0).

The first example

The samples sizes for the 8 families were (14, 12, 11, 10, 10, 11, 15, 9), for a total sample size of 92.

Thus, $\mathrm{n_0} \approx 11.45$.

For the female meioses, $\mathrm{M_B} = 212$ and $\mathrm{M_W} = 46$. Thus

$\hat{\sigma} = \sqrt{46} = 6.8$ $\longrightarrow$ overall sample mean = 40.3

$\hat{\sigma}_A = \sqrt{(212 - 46)/11.45} = 3.81.$

For the male meioses, $\mathrm{M_B} = 16.3$ and $\mathrm{M_W} = 13.2$. Thus

$\hat{\sigma} = \sqrt{13.2} = 3.6$ $\longrightarrow$ overall sample mean = 22.8

$\hat{\sigma}_A = \sqrt{(16.3 - 13.2)/11.45} = 0.52.$