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

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

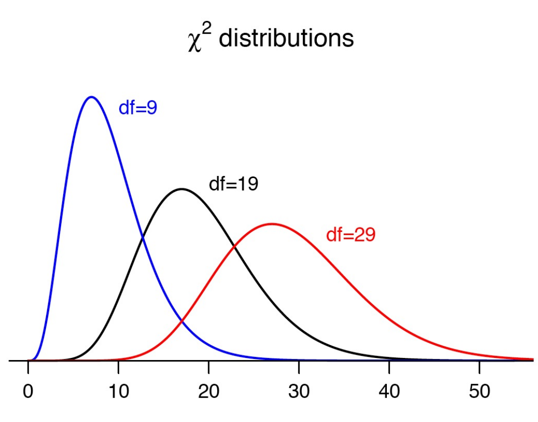
I haven't spoken much about variances (I generally prefer looking at the SD), but we are about to start making use of them.

The distribution of the sample variance

If $X_1, X_2, \ldots, X_n$ are iid normal(μ, σ^2)

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

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



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

$E(W) = n-1$

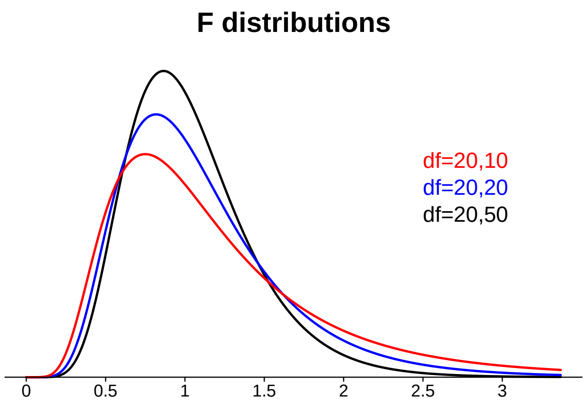
$\text{var}(W) = 2(n-1)$

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

The F distribution

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

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, ..., X_m be iid normal(μ_x, σ_x^2).

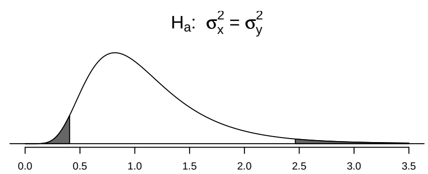
Let Y_1, Y_2, ..., Y_n be iid normal(μ_y, σ_y^2).

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

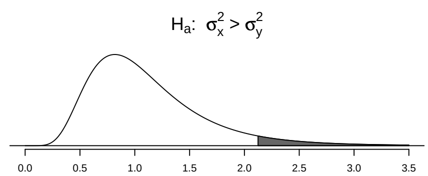
Under the null hypothesis $s_x^2/s_y^2 \sim F_{m-1,n-1}$

Critical regions

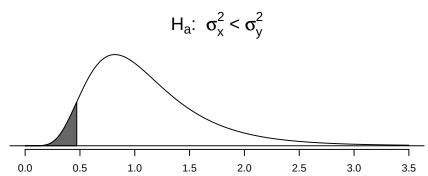
- If the alternative is $\sigma_x^2 \neq \sigma_y^2$, we reject if the ratio of the sample variances is unusually large or unusually small.



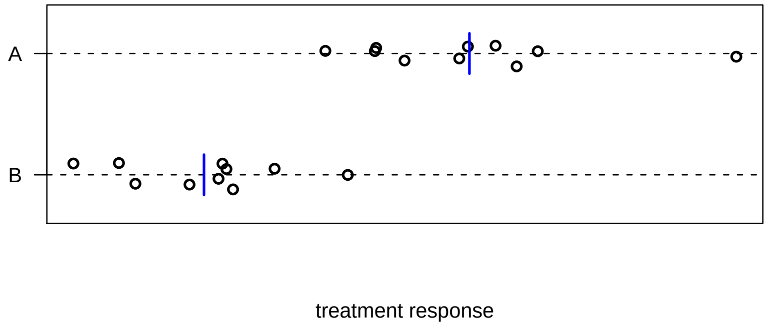
- If the alternative is $\sigma_x^2 > \sigma_y^2$, we reject if the ratio of the sample variances is unusually large.



- If the alternative is $\sigma_x^2 < \sigma_y^2$, we reject if the ratio of the sample variances is unusually small.



Example



Are the variances the same in the two groups?

Example

We want to test H_0: $\sigma^2_A = \sigma^2_B$ versus H_a: $\sigma^2_A \neq \sigma^2_B$

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 σ_x^2/σ_y^2

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).

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

Let L = 2.5th %ile and U = 97.5th %ile of F(m–1, n–1).

Then Pr[$L < (s_x^2/\sigma_x^2)/(s_y^2/\sigma_y^2) < U$] = 95%.

Thus 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

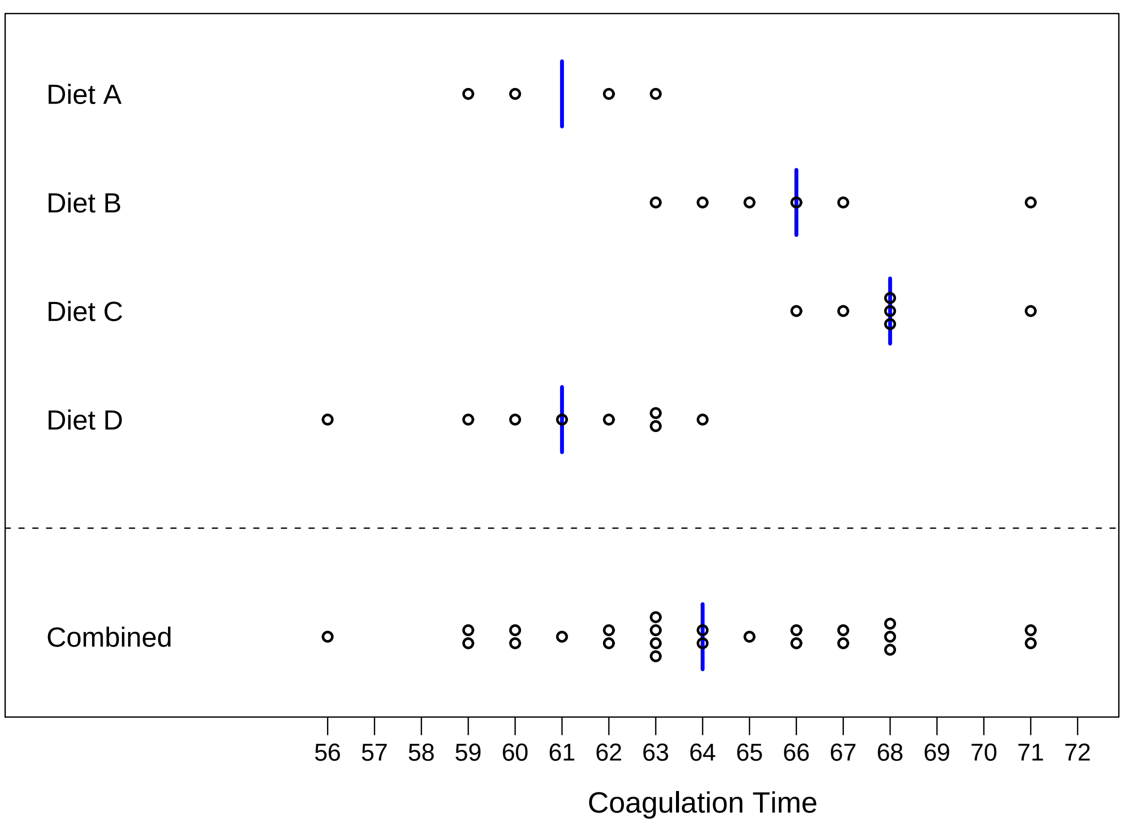
The 95% confidence interval for σ_x^2/σ_y^2 is

(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



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 estimate 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 estimate 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

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

Hypothesis testing

We assume

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

[equivalently, $Y_{ti} \sim$ independent N(μ_t, σ^2)]

We want to test

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

[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)

$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$

$H_a : H_0$ 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!

Now what did we do...?

$$\begin{pmatrix} 62 & 63 & 68 & 56 \\ 60 & 67 & 66 & 62 \\ 63 & 71 & 71 & 60 \\ 59 & 64 & 67 & 61 \\ & 65 & 68 & 63 \\ & 66 & 68 & 64 \\ & & & 63 \\ & & & 59 \end{pmatrix} = \begin{pmatrix} 64 & 64 & 64 & 64 \\ 64 & 64 & 64 & 64 \\ 64 & 64 & 64 & 64 \\ 64 & 64 & 64 & 64 \\ & 64 & 64 & 64 \\ & 64 & 64 & 64 \\ & & & 64 \\ & & & 64 \end{pmatrix} + \begin{pmatrix} -3 & 2 & 4 & -3 \\ -3 & 2 & 4 & -3 \\ -3 & 2 & 4 & -3 \\ -3 & 2 & 4 & -3 \\ & 2 & 4 & -3 \\ & 2 & 4 & -3 \\ & & & -3 \\ & & & -3 \end{pmatrix} + \begin{pmatrix} 1 & -3 & 0 & -5 \\ -1 & 1 & -2 & 1 \\ 2 & 5 & 3 & -1 \\ -2 & -2 & -1 & 0 \\ & -1 & 0 & 2 \\ & 0 & 0 & 3 \\ & & & 2 \\ & & & -2 \end{pmatrix}$$

	observations		grand average		treatment deviations		residuals
	y_{ti}	=	$\bar{y}_{..}$	+	$\bar{y}_{t.} - \bar{y}_{..}$	+	$y_{ti} - \bar{y}_{t.}$
Vector	**Y**	=	**A**	+	**T**	+	**R**
Sum of Squares	98,644	=	98,304	+	228	+	112
D's of Freedom	24	=	1	+	3	+	20