# ANOVA, F test

Joe Felsenstein

Department of Genome Sciences and Department of Biology

# Analysis of variance

If we have a number $p$ of groups, with sample sizes $n$, and we take as the null hypothesis that they come from the same normal distribution, we can make two estimates of the standard deviation:

- From the variance of the means around the overall mean:

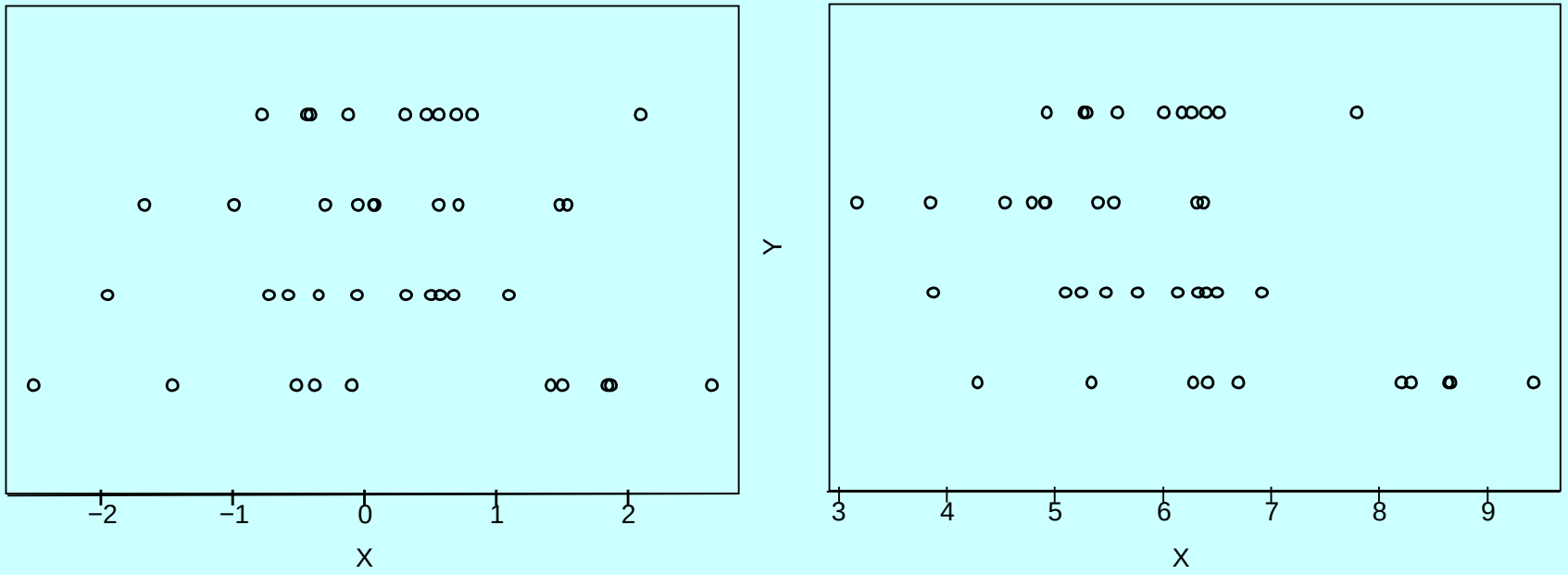
$$\frac{\sum_i \left(\bar{X}_i - \bar{X}\right)^2}{p-1} = \frac{\hat{s^2}}{n}$$

- From the within-group variances. If all groups have the same variance *(assumption!)* then we can also estimate it by pooling the within group sum of squares:

$$\frac{\sum_i \sum_j \left(X_{ij} - \bar{X}_i\right)^2}{(n-1)p} = \hat{s^2}$$

# ANOVA and F

True differences between groups inflate the estimate that comes from variances of between-group means:




These two estimates of the variance are independent (under the assumptions).

# R. A. Fisher (1890-1962) in about 1930

Inventor of about half of modern mathematical statistics (maximum likelihood, likelihood ratio test, analysis of variance, F distribution, consistency, sufficiency, efficiency, P values, etc. etc.)

Also one of the two greatest of the three great founders of theoretical population genetics, and thus a father of the modern evolutionary synthesis, as well as the primary developer of the theory of quantitative characters. He has been called the greatest mathematical scientist of the 20th century.

# The F distrbution

- An estimated variance is a sum of squares of variable around their mean, divided by the number of degrees of freedom.

# The F distrbution

- An estimated variance is a sum of squares of variable around their mean, divided by the number of degrees of freedom.
- The sum of squares (around their mean) of $n$ variables from a normal distribution with mean 0 and variance 1 is drawn from a Chi-square ($\chi^2$) distribution with $n$ degrees of freedom (it's a particular Gamma distribution, a scaling of the waiting time until the $n/2$-th phone call).

# The F distrbution

- An estimated variance is a sum of squares of variable around their mean, divided by the number of degrees of freedom.
- The sum of squares (around their mean) of $n$ variables from a normal distribution with mean 0 and variance 1 is drawn from a Chi-square ($\chi^2$) distribution with $n$ degrees of freedom (it's a particular Gamma distribution, a scaling of the waiting time until the $n/2$-th phone call).
- When the mean is estimated from the sample, the sum of squares has a $\chi^2$ distribution with $n-1$ degrees of freedom.

# The F distrbution

- An estimated variance is a sum of squares of variable around their mean, divided by the number of degrees of freedom.
- The sum of squares (around their mean) of $n$ variables from a normal distribution with mean 0 and variance 1 is drawn from a Chi-square ($\chi^2$) distribution with $n$ degrees of freedom (it's a particular Gamma distribution, a scaling of the waiting time until the $n/2$-th phone call).
- When the mean is estimated from the sample, the sum of squares has a $\chi^2$ distribution with $n-1$ degrees of freedom.
- The variance in that case is distributed as a $\chi^2$ value divided by its degrees of freedom.

# The F distrbution

- An estimated variance is a sum of squares of variable around their mean, divided by the number of degrees of freedom.
- The sum of squares (around their mean) of $n$ variables from a normal distribution with mean 0 and variance 1 is drawn from a Chi-square ($\chi^2$) distribution with $n$ degrees of freedom (it's a particular Gamma distribution, a scaling of the waiting time until the $n/2$-th phone call).
- When the mean is estimated from the sample, the sum of squares has a $\chi^2$ distribution with $n-1$ degrees of freedom.
- The variance in that case is distributed as a $\chi^2$ value divided by its degrees of freedom.
- When the variance of the true distribution of values is $\sigma^2$ and not 1, the estimated variance is distributed as $\sigma^2\chi^2_d/d$, where $d$ is the degrees of freedom.

# The F distrbution

- An estimated variance is a sum of squares of variable around their mean, divided by the number of degrees of freedom.
- The sum of squares (around their mean) of $n$ variables from a normal distribution with mean 0 and variance 1 is drawn from a Chi-square ($\chi^2$) distribution with $n$ degrees of freedom (it's a particular Gamma distribution, a scaling of the waiting time until the $n/2$-th phone call).
- When the mean is estimated from the sample, the sum of squares has a $\chi^2$ distribution with $n-1$ degrees of freedom.
- The variance in that case is distributed as a $\chi^2$ value divided by its degrees of freedom.
- When the variance of the true distribution of values is $\sigma^2$ and not 1, the estimated variance is distributed as $\sigma^2\chi^2_d/d$, where $d$ is the degrees of freedom.
- In a ratio of two independent variances from the same distribution, the $\sigma^2$ disappears. This is the F distribution, with degrees of freedom $d_1$ and $d_2$.

# The F distrbution

- An estimated variance is a sum of squares of variable around their mean, divided by the number of degrees of freedom.
- The sum of squares (around their mean) of $n$ variables from a normal distribution with mean 0 and variance 1 is drawn from a Chi-square ($\chi^2$) distribution with $n$ degrees of freedom (it's a particular Gamma distribution, a scaling of the waiting time until the $n/2$-th phone call).
- When the mean is estimated from the sample, the sum of squares has a $\chi^2$ distribution with $n-1$ degrees of freedom.
- The variance in that case is distributed as a $\chi^2$ value divided by its degrees of freedom.
- When the variance of the true distribution of values is $\sigma^2$ and not 1, the estimated variance is distributed as $\sigma^2\chi^2_d/d$, where $d$ is the degrees of freedom.
- In a ratio of two independent variances from the same distribution, the $\sigma^2$ disappears. This is the F distribution, with degrees of freedom $d_1$ and $d_2$.
- R. A. Fisher calculated the density function of this distribution, and with colleagues calculated its tail probabilities for reasonable values of $d_1$ and $d_2$.

## The table of data looks like this

| Groups | | | | |
|---|---|---|---|---|
| 2.428 | 3.171 | 2.976 | . . . | 3.092 |
| 2.741 | 3.421 | 3.042 | . . . | 2.997 |
| 2.898 | | 3.172 | . . . | 3.113 |
| | | 3.232 | . . . | 2.960 |

Note that the number of values in each cell can differ, and there is no meaning to what order you write them down in the cell (the horizontal rows here are meaningless).

## How to do a one-way ANOVA in R?

If the data is in a “data frame” with a column for the number of the group, and a column for the value (and each being for one observation), it looks like this:

| | grp | x |
|---|---|---|
| 1 | 1 | 29.11601 |
| 2 | 1 | 29.00317 |
| 3 | 1 | 28.75443 |
| 4 | 1 | 29.00025 |
| 5 | 1 | 28.56627 |
| 6 | 1 | 28.84333 |
| 7 | 1 | 28.96895 |
| 8 | 1 | 28.56424 |
| 9 | 1 | 28.93370 |
| 10 | 1 | 28.66744 |
| 11 | 2 | 31.00009 |
| 12 | 2 | 30.93960 |
| ... | | |

## How to do a one-way ANOVA in R?

Use a “linear model” explaining variable `x` by variable `grp`, where the latter is considered as a “factor” which means its numerical size isn’t important, just the different values as markers of groups.
The R notation for this model is somewhat wierd: `x~factor(grp)`

If the data frame is called `mm` you do

```
aa <- lm(x ~ factor(grp), data=mm)
```

and then after that to get the ANOVA:

```
anova(aa)
```

and the resulting printout is the analysis of variance table:

```
Analysis of Variance Table

Response:  x
               Df  Sum Sq  Mean Sq  F value  Pr(>F)
 factor(grp)    3  43.774   14.591   215.33  < 2.2e-16 ***
 Residuals     36   2.439    0.068
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

## Two-way ANOVA

In addition to groups you can have much more elaborate designs involving rows and columns, etc. Mathematically, you can write the statistical model as something like this:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

so that the $k$th observation in the cell in row $i$ and column $j$ is the sum of an overall mean (a constant) plus a random effect specific to the row, a random effect specific to the column, a random effect specific to that cell, plus an "error" specific to that observation.

If each row of the data frame has an observation number in column 1, a row number (R), a column number (C), and the variable (Y), then in the linear model command, we do:

```
aa <- lm(Y ~ factor(R)+factor(C)+factor(R:C), data=mm)
```

This works!

# A two-way table that could be analyzed this way

| | | | | |
|---|---|---|---|---|
| 4.289 | 4.078 | 4.819 | 4.402 | 4.640 |
| 4.104 | 3.769 | 4.231 | 3.882 | 4.299 |
| 5.003 | 3.786 | 3.646 | 3.869 | 3.888 |
| 4.338 | 4.735 | 3.999 | 4.452 | 3.840 |
| 4.034 | 4.790 | 4.558 | 4.431 | 4.400 |
| 4.098 | 4.085 | 3.617 | 4.114 | 4.694 |

Again, within each cell we assume the order of the 3 items is meaningless.

But what if it isn't meaningless. Suppose the rows are populations, the columns are genes, and in each population the same three individuals have their expression measured for all 5 genes?

# The analysis of variance table is like this

Analysis of Variance Table

Response: Y

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| factor(row) | 4 | 24.616 | 6.154 | 2.8950 | 0.02584 * |
| factor(col) | 3 | 8.419 | 2.806 | 1.3202 | 0.27207 |
| factor(row):factor(col) | 12 | 33.487 | 2.791 | 1.3127 | 0.22320 |
| Residuals | 100 | 212.575 | 2.126 | | |

—

Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1