ANALYSIS OF VARIANCE (ANOVA)

ANOVA . . .

. . . is a powerful technique for estimating and comparing the means of two or more populations, and particularly for testing the null hypothesis that the means are the same.
The technique described is frequently called *single-factor experiments* — we study the variability of the response variable X in function of a one-way classification, i.e. classification which can be described in terms of a single *predictory variable* (e.g. the number of the population).

Problem: study the variability of RV X in k populations ($k \geq 3$) – in each of those populations our random variable X (μ, σ) has some expected value $E(X) = \mu_i$, where $i = 1, 2, \ldots, k$ and **the same** (common) variance σ^2.

$E(X)$	population	sample	values
μ_1	1	n_1	$x_{11}, x_{12}, \ldots, x_{1j}, \ldots, x_{1n_1}$
μ_2	2	n_2	$x_{21}, x_{22}, \ldots, x_{2j}, \ldots, x_{2n_2}$
$\ldots$	$\ldots$	$\ldots$	$\ldots, \ldots, \ldots, \ldots, \ldots, \ldots$
μ_i	i	n_i	$x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{in_i}$
$\ldots$	$\ldots$	$\ldots$	$\ldots, \ldots, \ldots, \ldots, \ldots, \ldots$
μ_k	k	n_k	$x_{k1}, x_{k2}, \ldots, x_{kj}, \ldots, x_{kn_k}$

k populations; each of them with n_i values

In total we have $n = \sum_{i=1}^{k} n_i$ values. We calculate
a) group-means (for every population):

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}; \quad i = 1, \ldots, k$$

b) global (or *grand*) mean:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \bar{x}_i$$

and the (total) sum of squares $q = SST$:

$$\begin{aligned} q &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[(x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x}) \right]^2 \\ &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + 0 + \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 \equiv q_R + q_G \end{aligned}$$

three sums: $q,\ q_R,\ q_G$

q (SST) — is the Σ of deviations of x_{ij} with respect to the grand mean; the number of DF associated with SST equals to $n-1$ since SST contains n summands and one constraint $\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})=0\right)$.

q_R (SSR) — is the Σ of deviations within the groups (with respect to the group means); the number of DF associated with SSR equals to $n-k$ since SST contains n summands and k constraints $\left(\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)=0;\ \ i=1,\ldots,k\right)$.

q_G (SSG) — Σ of deviations of the group means with respect to the grand mean; the number of DF associated with SSG equals to $k-1$ since SST contains k summands and one constraint $\left(\sum_{i=1}^{k}(\bar{x}_i-\bar{x})=0\right)$.

if the means – within different groups – are equal
all the three statistics are "equally good" estimators of σ^2:

$$\frac{Q}{n-1} \qquad \frac{Q_R}{n-k} \qquad \frac{Q_G}{k-1}$$

.

All the three statistics:

$$\frac{Q}{n-1} \qquad \frac{Q_R}{n-k} \qquad \frac{Q_G}{k-1}$$

are "equally good" estimators of σ^2
if the means – within different groups – are equal.

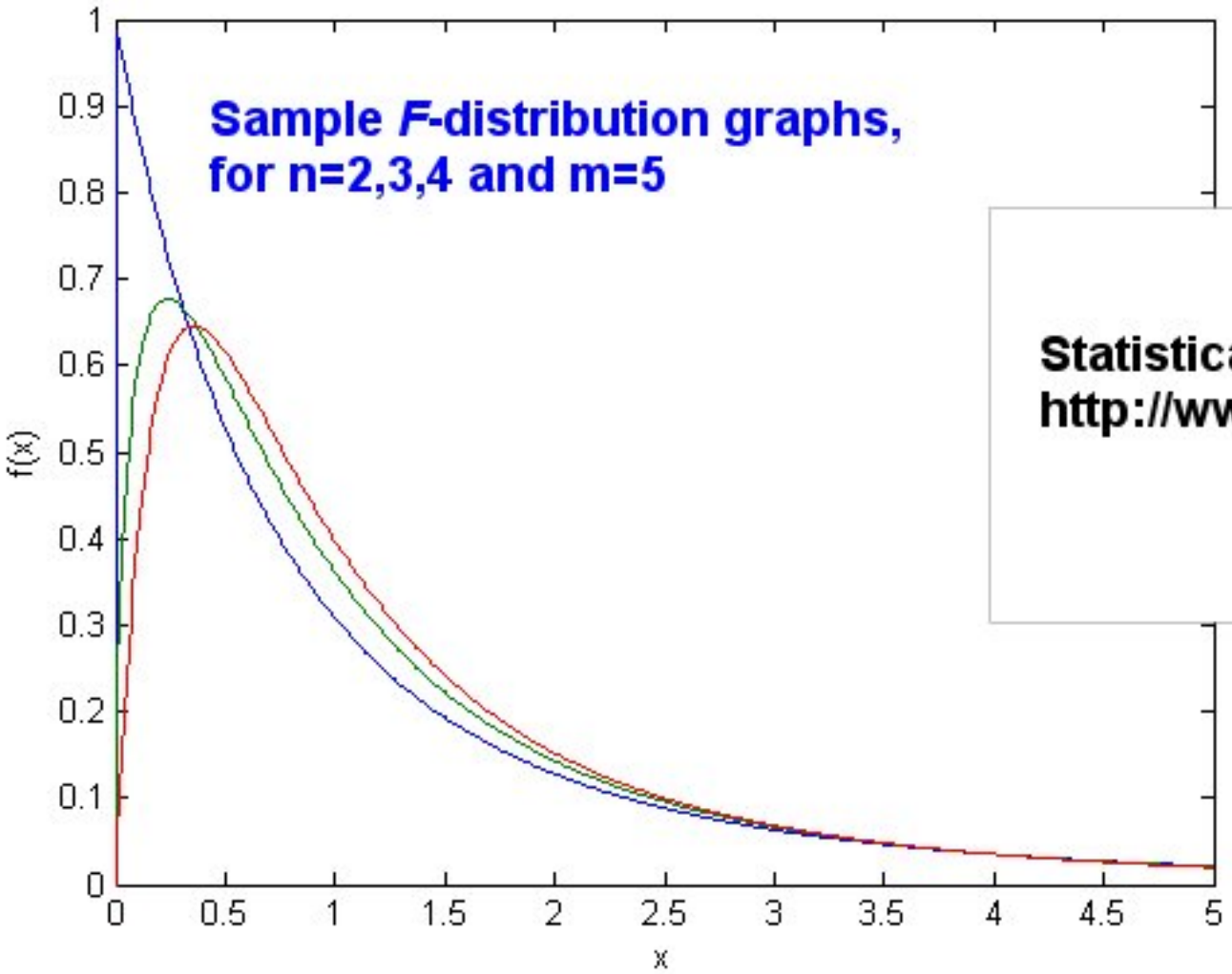
In other words – we are testing:

$H_0: \quad \mu_1 = \mu_2 = \ldots \mu_k,$ against
$H_1: \quad$ At least two of the means are not equal.

The statistic:

$$\mathcal{F} = \frac{\frac{Q_G}{k-1}}{\frac{Q_R}{n-k}}$$

has the Snedecor $\mathcal{F}$ distribution with $(k-1, n-k)$ degrees of freedom.



Statistical Analysis Handbook
http://www.statsref.com/

Example: R. E. Walpole, R. H. Myers ,,Probability and statistics for engineers and scientists"

Table: Absorption of Moisture in Concrete Aggregates

	Aggregate					
	1	2	3	4	5	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16 854
Mean	553.3	569.3	610.5	465.2	610.7	561.8

Example – solution

In other words – we are testing:

H_0 : $\mu_1 = \mu_2 = \ldots \mu_k$ $k = 5$, against

H_1 : At least two of the means are not equal.

Critical region is: $f > 2.76$ with $\nu_1 = 4$ and $\nu_2 = 25$ degrees of freedom.

Computations:

$SST = (551 - 561.8)^2 + (457 - 561.8)^2 + \ldots + (679 - 561.8)^2 = 209.38$

$SSG = 6\left[(553.3 - 561.8)^2 + \ldots + (610.7 - 561.8)^2\right] = 85.36$

$SSR = 209.38 - 85.36 = 124.02$

$$\mathcal{F} = \frac{\frac{Q_G}{k-1}}{\frac{Q_R}{n-k}} = \frac{\frac{85.36}{4}}{\frac{124.02}{25}} = \boxed{4.30 > 2.76}.$$

Decision: reject H_0 and conclude that the aggregates do not have the same absorption.