Lecture 18: Analysis of Variance (ANOVA)

GENOME 560

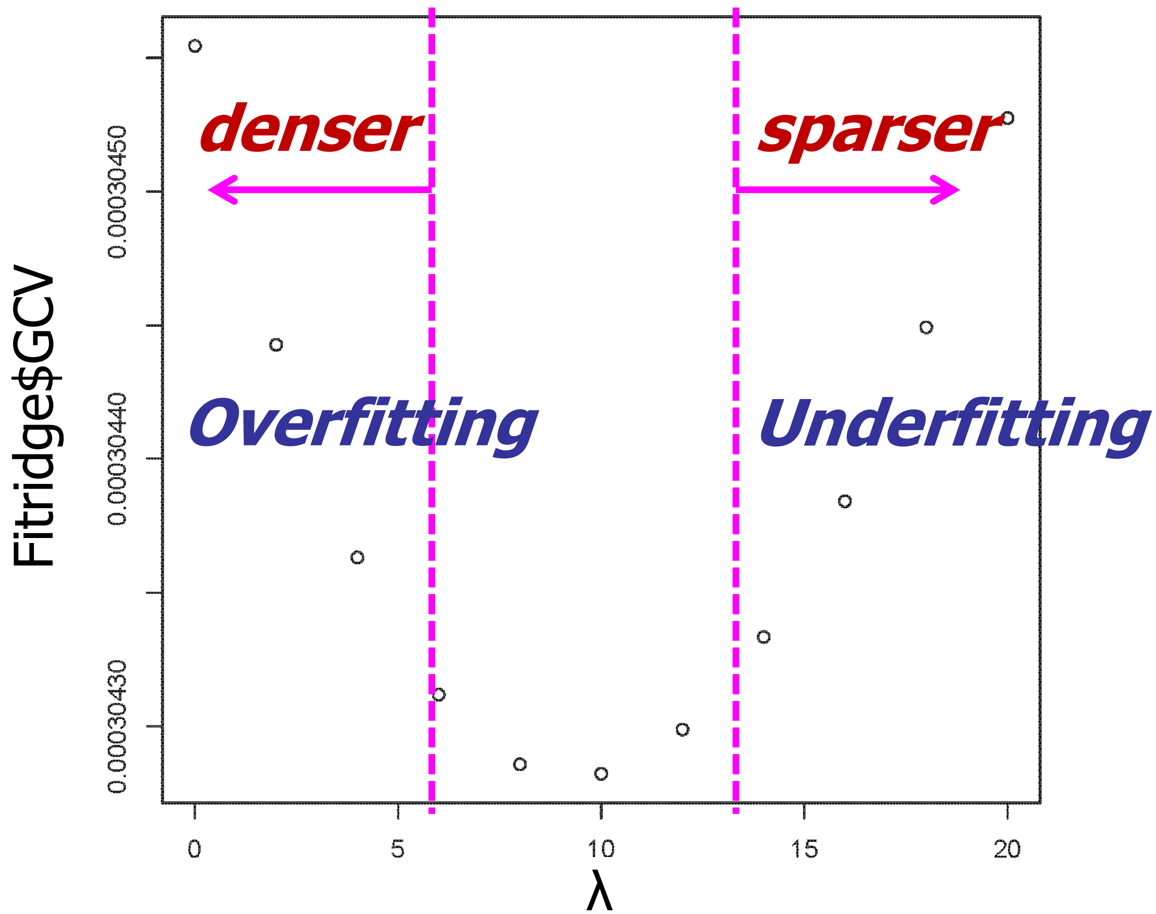
Su-In Lee, CSE & GS (suinlee@uw.edu)

Review of Last Lecture

- How to determine the value of λ ?
- Cross-validation
- Overfitting vs. underfitting

Review: Overfitting vs. Underfitting

$$f(\beta_0, \beta_1, \cdots, \beta_p) = \boxed{\sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}) \right]^2} + \boxed{\lambda \sum_{j=1}^{p} \beta_j^2}$$



Analysis of variance (ANOVA)

- A collection of statistical models used to analyze the *differences between group means*

- In its simplest form, ANOVA tests *whether or not the means of several groups are equal*

Motivating example

- A random sample of some quantitative trait was measured in individuals randomly sampled from population

- Let's test *whether or not the trait means of different genotype groups are equal*

- Genotype of a certain SNP
  - AA: 82, 83, 97
  - AG: 83, 78, 68
  - GG: 38, 59, 55

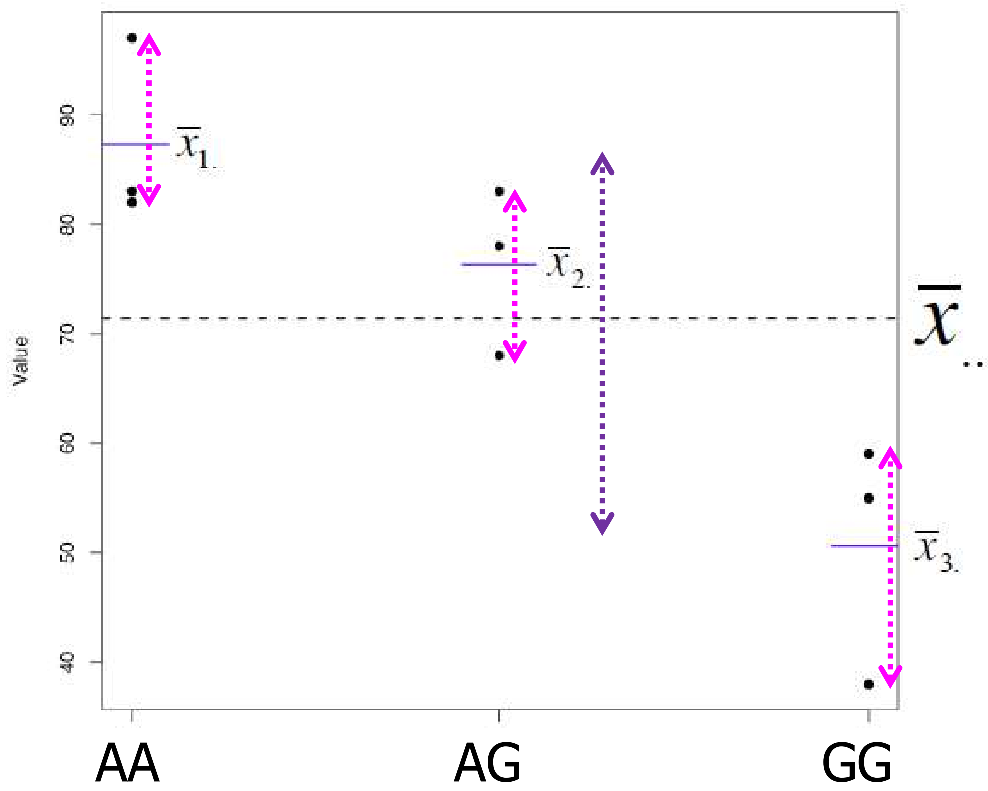
There are N (= 9) individuals and K (= 3) groups …

Basic idea of ANOVA

- The *observed variance* in a particular variable is partitioned into *components attributable to different sources of variation*

Rationale of ANOVA

- Partition total variation of the data into two sources
  - 1. Variation *within* genotype groups
  - 2. Variation *between* genotype groups



x_{ij} **:** *j*th observation in the *i*th genotype group

x_{11} x_{12} x_{13}

AA: 82, 83, 97 $\xrightarrow{\textit{average}}$ $\bar{x}_{1.}$

AG: 83, 78, 68 $\xrightarrow{\textit{average}}$ $\bar{x}_{2.}$

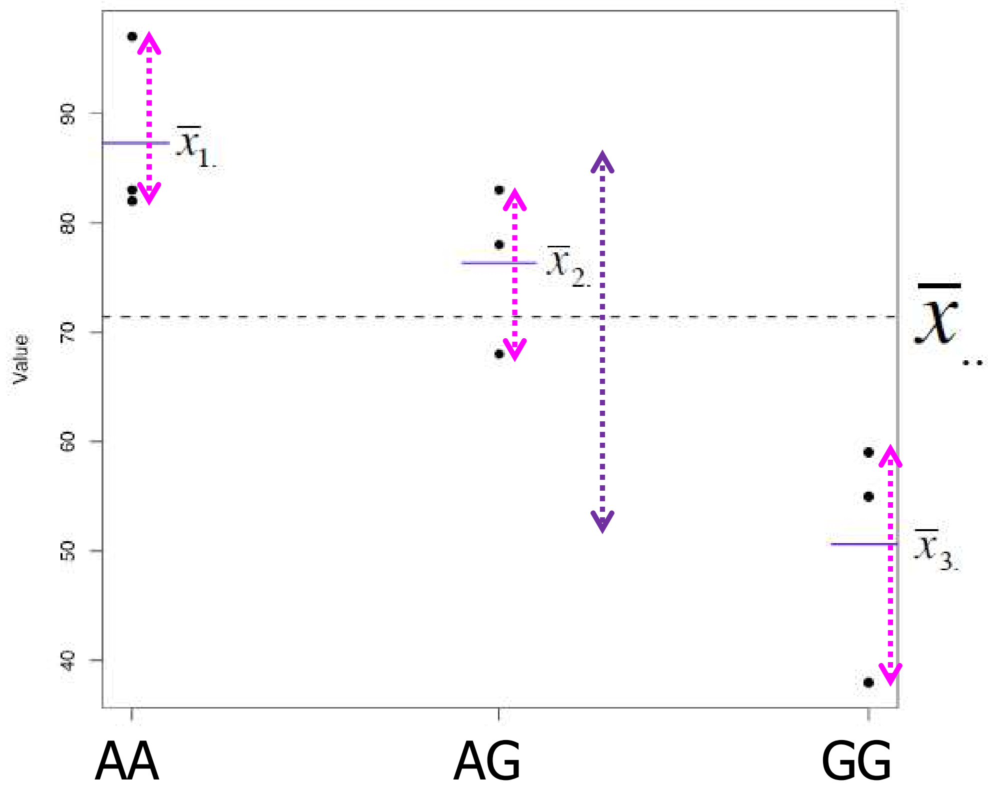
GG: 38, 59, 55 $\xrightarrow{\textit{average}}$ $\bar{x}_{3.}$

Grand mean $\bar{x}_{..}$

Rational of ANOVA

- Comparing between *variation within genotype groups* and *variation between genotype groups*

- If H_0 is true, the *standardized variances* are equal to one another



The Details

- Our Data:
  - AA: 82, 83, 97 $\quad \bar{x}_{1.} = (82+83+97)/3 = 87.3$
  - AG: 83, 78, 68 $\quad \bar{x}_{2.} = (83+78+68)/3 = 76.3$
  - GG: 38, 59, 55 $\quad \bar{x}_{3.} = (38+59+55)/3 = 50.6$

- Let x_{ij} denote the data from the ith group and jth observation

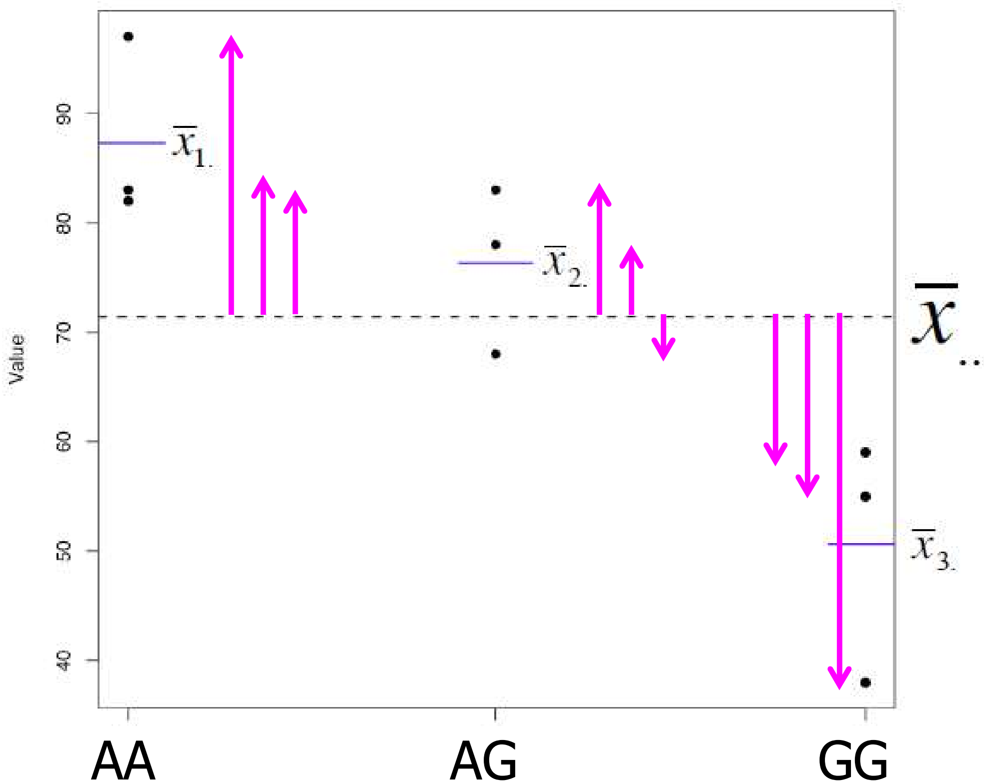
- Overall, or *grand mean*, is:

$$\bar{x}_{..} = \sum_{i=1}^{K} \sum_{j=1}^{J} \frac{x_{ij}}{N}$$

$$\bar{x}_{..} = \frac{82+83+97+83+78+68+38+59+55}{9} = 71.4$$

Total variation

- **SST:** Sum of squared deviations *about the grand mean across all N observations*

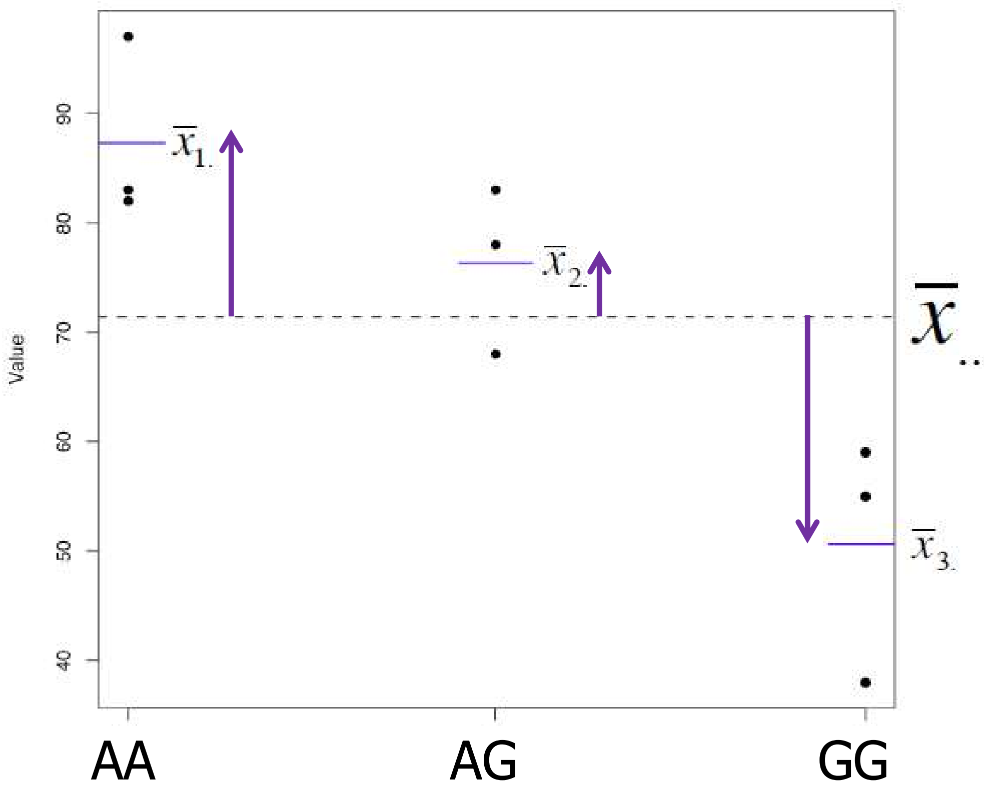


SST

$$\sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{..})^2$$

Between group variation

- **SST$_G$:** Sum of squared deviations *for each group mean about the grand mean*

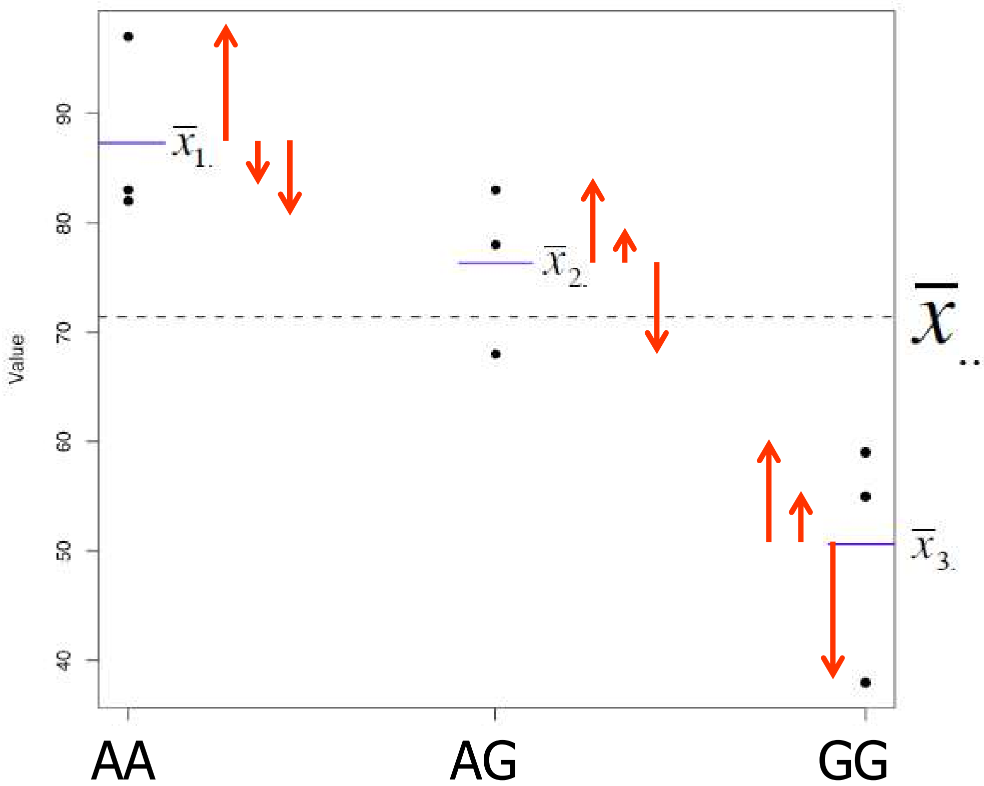


SST$_G$

$$\sum_{i=1}^{K} n_i \bullet (\bar{x}_{i.} - \bar{x}_{..})^2$$

Within group variation

- **SST$_E$:** Sum of squared deviations *for all observations within each group from that group mean, summed across all groups*



SST$_E$

$$\sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{i.})^2$$

Partitioning Total Variation

- Variation is simply average squared deviations from the mean

$$SST = SST_G + SST_E$$

$$\sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij} - \bar{x}_{..})^2 \qquad \sum_{i=1}^{K} n_i \bullet (\bar{x}_{i.} - \bar{x}_{..})^2 \qquad \sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij} - \bar{x}_{i.})^2$$

Sum of squared deviations about the grand mean across all N observations

Sum of squared deviations for each group mean about the grand mean

Sum of squared deviations for all observations within each group from that group mean, summed across all groups

In our example

$$SST = SST_G + SST_E$$

$$\sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{..})^2 \qquad \sum_{i=1}^{K} n_i \bullet (\bar{x}_{i.}-\bar{x}_{..})^2 \qquad \sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{i.})^2$$

$$\begin{aligned}&(82-71.4)^2+(83-71.4)^2+(97-71.4)^2+\\&(83-71.4)^2+(78-71.4)^2+(68-71.4)^2+\\&(38-71.4)^2+(59-71.4)^2+(55-71.4)^2=\end{aligned}$$

$$\begin{aligned}&3\bullet(87.3-71.4)^2+\\&3\bullet(76.3-71.4)^2+\\&3\bullet(50.6-71.4)^2=\end{aligned}$$

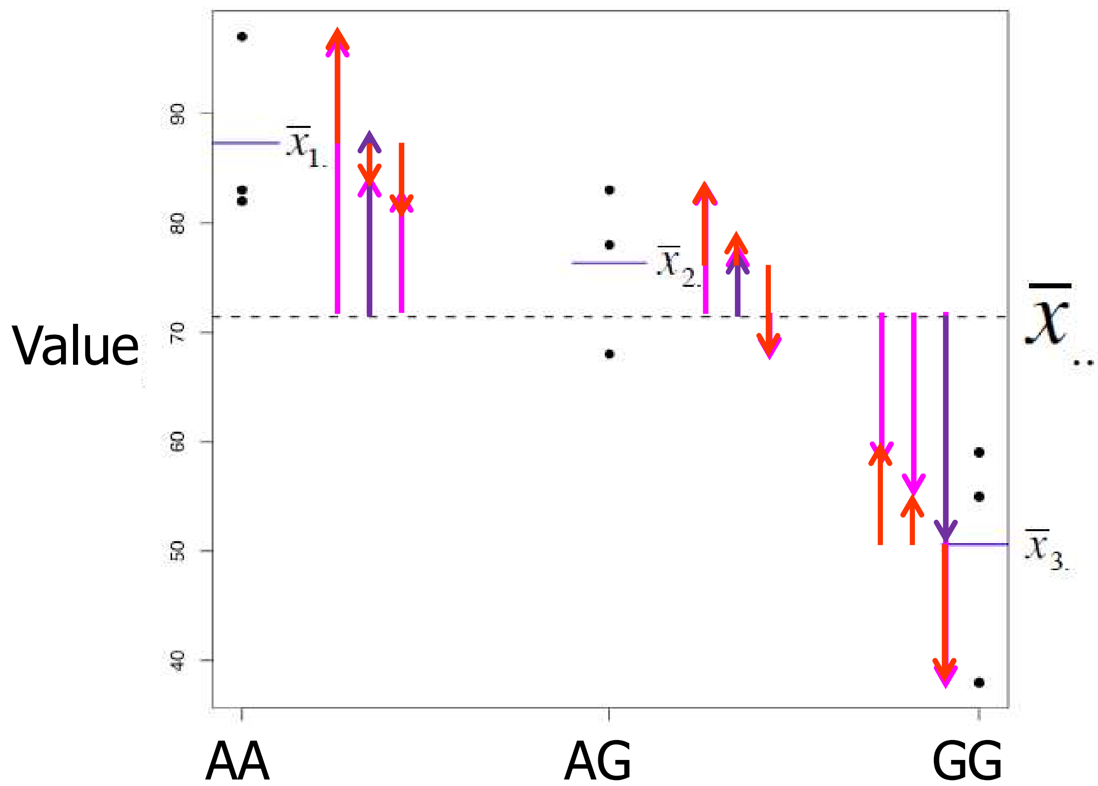
$$\begin{aligned}&(82-87.3)^2+(83-87.3)^2+(97-87.3)^2+\\&(83-76.3)^2+(78-76.3)^2+(68-76.3)^2+\\&(38-50.6)^2+(59-50.6)^2+(55-50.6)^2=\end{aligned}$$

2630.2   2124.2   506

In our example

$$\text{SST} = \text{SST}_G + \text{SST}_E$$

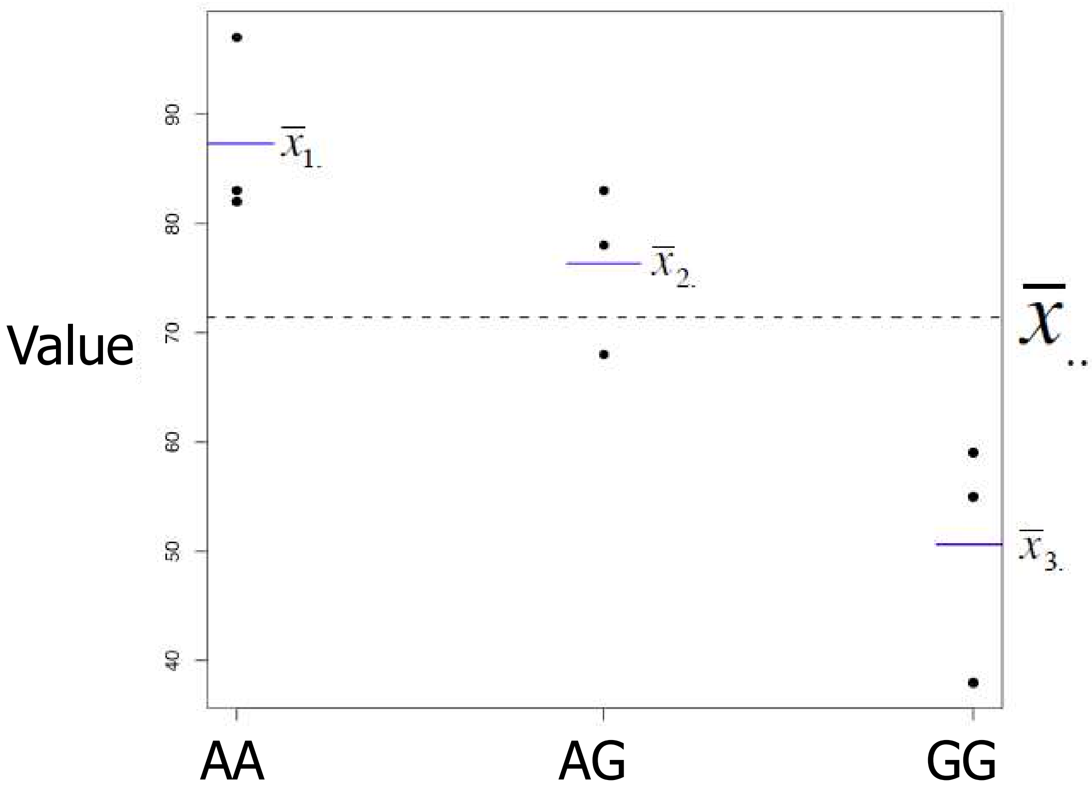
$$\sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{..})^2 \qquad \sum_{i=1}^{K} n_i \bullet (\bar{x}_{i.}-\bar{x}_{..})^2 \qquad \sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{i.})^2$$



Basic questions

- Are the trait means of different groups are equal?

$$SST \quad = \quad SST_G \quad + \quad SST_E$$



The F-Test

- Is the difference in the means of the groups more than background noise (=variability within groups) ?

Summarizes the mean differences between all groups at once.

$$\boxed{F = \frac{\text{Variability between groups}}{\text{Variability within groups}}}$$

Analogous to pooled variance from a t-test.

The F-Test

- The test statistic has an *F-distribution under the null hypothesis*

Summarizes the mean differences between all groups at once.

$$\boxed{F = \frac{\text{Variability between groups}}{\text{Variability within groups}}}$$

Analogous to pooled variance from a t-test.

Outline

- Basics on analysis of variance (ANOVA)
- One-way ANOVA
- R session
- Next time: Two-way ANOVA

Basic Framework of ANOVA

- Want to study the effect of one or more ***qualitative (discrete)*** variables on a ***quantitative (continuous)*** outcome variable

- Qualitative variables are referred to as ***factors***
  - e.g., Genotype of a certain SNP

- Characteristics that differentiates factors are referred to as ***levels***
  - e.g., three genotypes of a SNP

One-Way ANOVA

- Simplest case, also called *single factor ANOVA*
  - The *outcome* variable is the variable you're comparing
  - The *factor* variable is the categorical variable being used to define the groups
  - The *one-way* is because each value is classified in exactly one way

- ANOVA easily generalizes to more factors

Assumptions of ANOVA

- Samples are independent
- Responses for a given group are independently and identically distributed normal random variables
- Variances of populations are equal

One-Way ANOVA: Null Hypothesis

- The null hypothesis is that the means of K groups are all equal

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_K$$

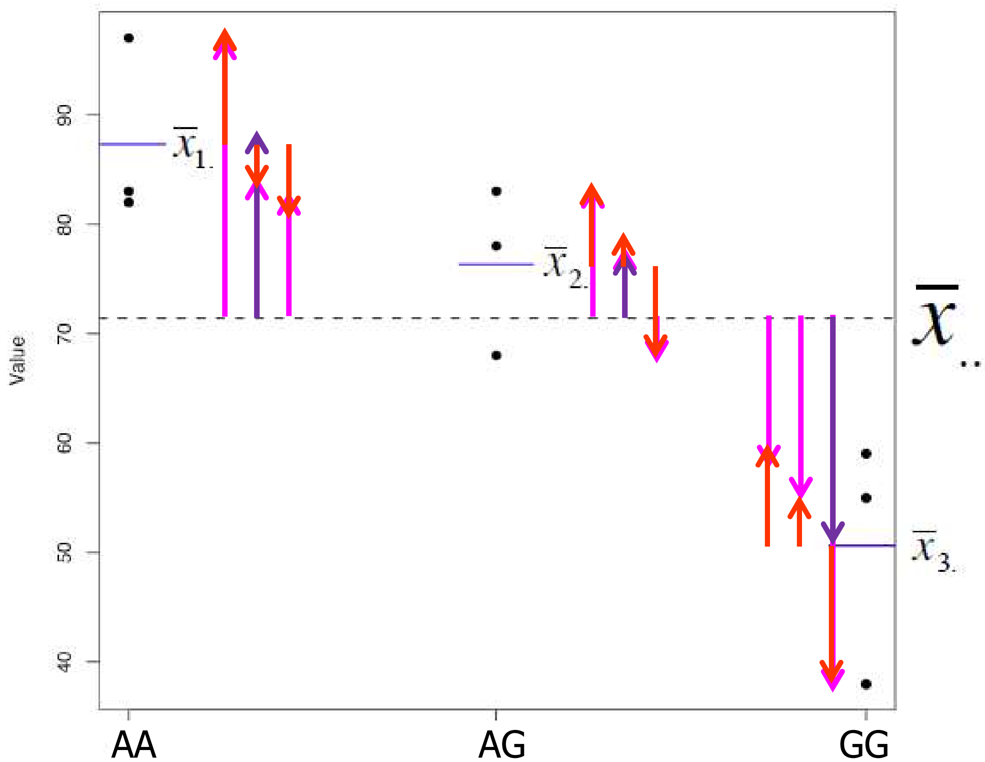
- The alternative hypothesis is that at least one of the means is different

Revisiting the genotype group example

- Total variation can be partitioned into between-group variation and within-group variation

$$\text{SST} = \text{SST}_G + \text{SST}_E$$

$$\sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{..})^2 \qquad \sum_{i=1}^{K} n_i \bullet (\bar{x}_{i.}-\bar{x}_{..})^2 \qquad \sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{i.})^2$$



x_{ij} **:** *j*th observation in the *i*th genotype group

x_{11} x_{12} x_{13}

AA: 82, 83, 97 —*average*→ $\bar{x}_{1.}$

AG: 83, 78, 68 —*average*→ $\bar{x}_{2.}$

GG: 38, 59, 55 —*average*→ $\bar{x}_{3.}$

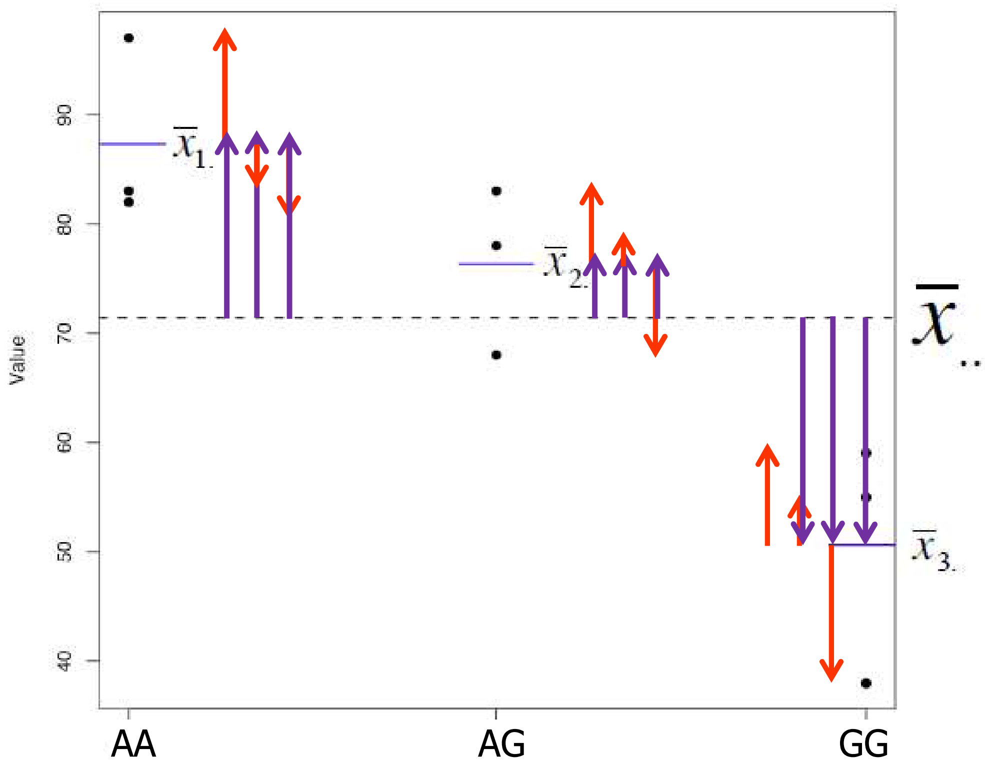
Grand mean $\bar{x}_{..}$

ANOVA: comparing variances

- Compare between-group variation with within-group variation

$$\text{SST} = \boxed{\text{SST}_G} + \boxed{\text{SST}_E}$$

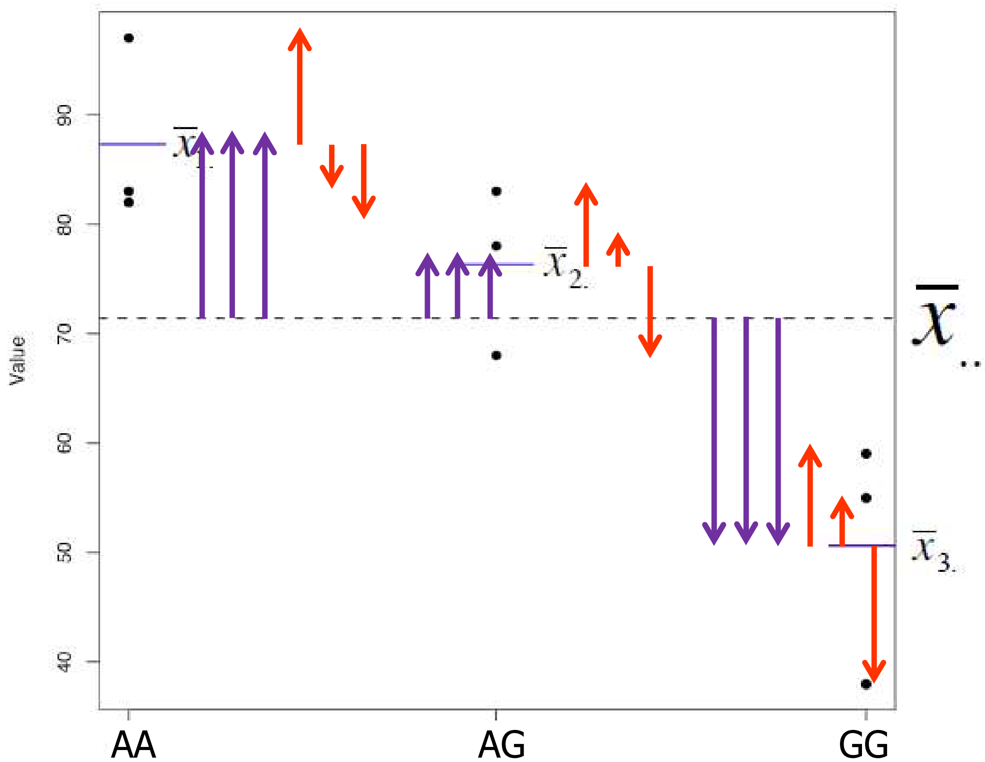
$$\sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij} - \bar{x}_{..})^2 = \boxed{\sum_{i=1}^{K} n_i \bullet (\bar{x}_{i.} - \bar{x}_{..})^2} + \boxed{\sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij} - \bar{x}_{i.})^2}$$



ANOVA: comparing variances

- Compare between-group variation with within-group variation

$$\underset{\sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{..})^2}{\text{SST}} = \boxed{\underset{\sum_{i=1}^{K} n_i \bullet (\bar{x}_{i.}-\bar{x}_{..})^2}{\text{SST}_G}} + \boxed{\underset{\sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{i.})^2}{\text{SST}_E}}$$



- Are SST_G and SST_E comparable?
- In a null model, are they expected to be equal?
  - Which one is more likely to be larger in a null model?
- They need to be standardized.

Calculating the variance

- Population variance (σ^2) measures the deviation among individual measurements from the population mean (μ) for the entire population.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

  - Degrees of freedom (df): the number of independent pieces of information that go into the estimate of a parameter.
- Calculate the sum of the squared deviations from the mean and then divide it by the df.

Calculating the variance

- The variance is a measure of how spread a set of data is.
- Given, N data points $x_1, \dots, x_N$, the sum of the squared deviations from the population mean (μ) measures the spreadness.

$$\sum_{i=1}^{N} (x_i - \mu)^2$$

  - The larger the N is, the larger the sum is.
- Degrees of freedom (df): the number of independent pieces of information that go into the estimate of a parameter.
- Population variance (σ^2) is defined as the average squared deviation from the sample mean (μ):

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

Divide by the degrees of freedom

Calculating the variance

- Population variance (σ^2) is defined as the average squared deviation from the sample mean (μ):

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

- In many cases, we do not know what the mean μ is.
- Instead, we can use the sample mean $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.
- The degrees of freedom of an estimate of a parameter is equal to the number of independent scores that go into the estimate minus the number of parameters used as intermediate steps in the estimation of the parameter itself.
- Sample variance (s^2) can be computed as:

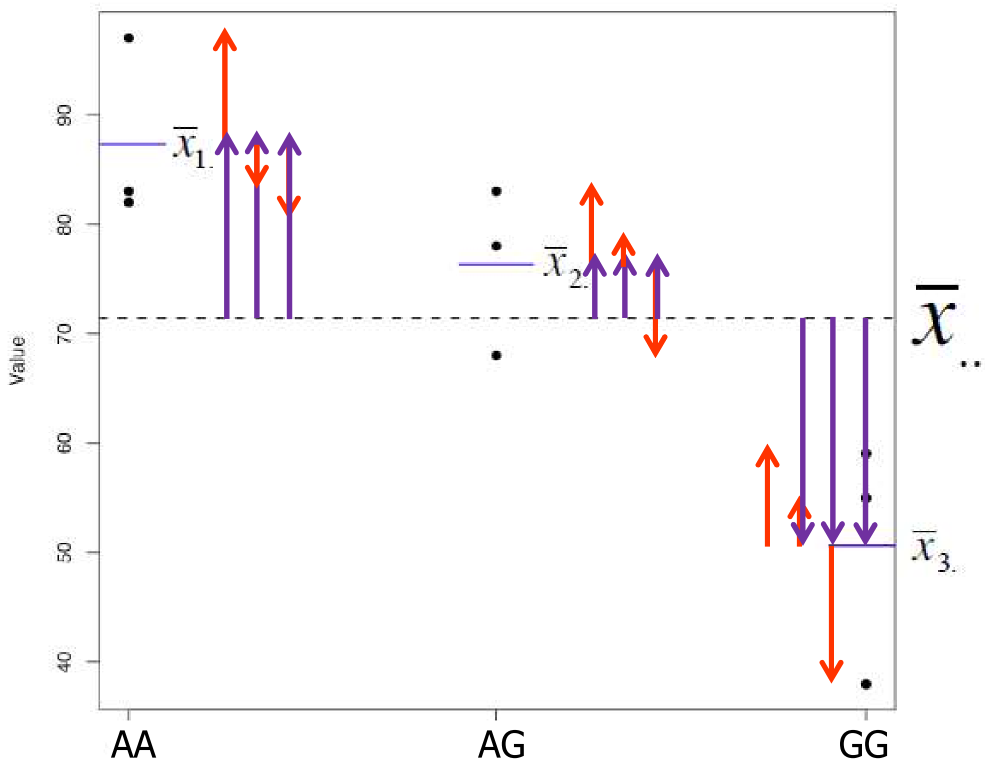
$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Divide by the degrees of freedom

Degrees of freedom in ANOVA

- Compare between-group variation with within-group variation

$$\underset{\sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{..})^2}{\text{SST}} = \boxed{\underset{\sum_{i=1}^{K} n_i \bullet (\bar{x}_{i.}-\bar{x}_{..})^2}{\text{SST}_G}} + \boxed{\underset{\sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{i.})^2}{\text{SST}_E}}$$



- There are N data points and K groups
- Df: (# independent scores) – (# intermediate scores)
- Between-group variance
  - Df: (K-1)
- Within-group variance
  - Df: (N-K)

Standardized Variances

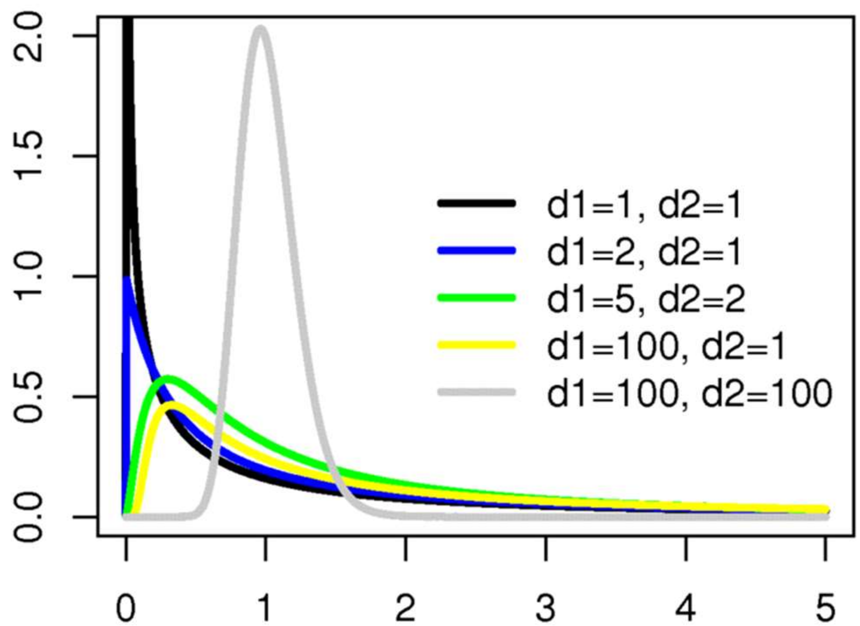
- To make the sum of squares comparable, we divide each one by their associated *degrees of freedom*
  - SST : $N - 1$ $(9 - 1 = 8)$
  - SST_G : $K - 1$ $(3 - 1 = 2)$
  - SST_E : $N - K$ $(9 - 3 = 6)$
- $MST_G = 2142.2 / 2 = 1062.1$
- $MST_E = 506 / 6 = 84.3$

Calculating F Statistics

- The test statistic is the ratio of group and error mean squares

$$F = \frac{MST_G}{MST_E} = \frac{1062.2}{84.3} = 12.59$$

- If H_0 is true, MST_G and MST_E are similary
  - More formally, if H_0 is true, the F ratio has an *F-distribution*



Calculating F Statistics

- The test statistic is the ratio of group and error mean squares

$$F = \frac{MST_G}{MST_E} = \frac{1062.2}{84.3} = 12.59$$

- If H_0 is true MST_G and MST_E are equal
- Critical value for rejection region is $F_{\alpha, K-1, N-K}$
- If we define $\alpha = 0.05$, then $F_{0.05, 2, 6} = 5.14$

ANOVA Table

Source of Variation	df	Sum of Squares	MS	F
Group	k-1	SST_G	$\frac{SST_G}{k-1}$	$\frac{SST_G}{k-1} \Big/ \frac{SST_E}{N-k}$
Error	N-k	SST_E	$\frac{SST_E}{N-k}$	
Total	N-1	SST		

Outline

- Basics on analysis of variance (ANOVA)
- One-way ANOVA
- R session
- Next time: Two-way ANOVA