ANOVA

Young W. Lim

2019-09-06 Fri

Outline

1. Based on

2. ANOVA
   - One-way ANOVA
   - one-way ANOVA Model
   - Two-way ANOVA
   - Within Groups Variance Estimate S_W^2

Based on

"Understanding Statistics in the Behavioral Sciences" R. R. Pagano

ANOVA (1)

- Analysis of variance (ANOVA) is
  a collection of statistical models and
  their associated estimation procedures
  (such as the "variation" among and between groups)
  used to analyze the differences
  among group means in a sample.

ANOVA (2)

- The ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation.

ANOVA (3)

- In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the t-test beyond two means.

Exmple (1)

- The analysis of variance can be used
  as an exploratory tool to explain observations.
  A dog show provides an example.

Exmple (2)

- A dog show is not a random sampling of the breed:
  it is typically limited to dogs
  that are adult, pure-bred, and exemplary.
- A histogram of dog weights from a show
  might plausibly be rather complex,
  like the yellow-orange distribution shown in the illustrations.
- Suppose we wanted to predict the weight of a dog
  based on a certain set of characteristics of eachdog.

One-way ANOVA (1)

- one-way analysis of variance (one-way ANOVA) is a technique that can be used to compare means of two or more samples (using the F distribution).

One-way ANOVA (2)

- can be used only for numerical response data,
  the "Y", usually one variable, and
  numerical or (usually) categorical input data,
  the "X", always one variable, hence "one-way"

One-way ANOVA (3)

- The ANOVA tests the null hypothesis
  that samples in all groups are drawn
  from populations with the same mean values.
- To do this, two estimates are made of the population variance.
- These estimates rely on various assumptions

One-way ANOVA (4)

- The ANOVA produces an F-statistic,
  the ratio of the variance calculated among the means
  to the variance within the samples.

One-way ANOVA (5)

- If the group means are drawn
  from populations with the same mean values,
  the variance between the group means
  should be lower than the variance of the samples,
  following the central limit theorem.
- A higher ratio therefore implies that
  the samples were drawn from
  populations with different mean values.

One-way ANOVA Model (1)

- The normal linear model describes treatment groups with probability distributions which are identically bell-shaped (normal) curves with different means.

One-way ANOVA Model (2)

- Thus fitting the models requires
  only the means of each treatment group
  and a variance calculation
  (an average variance within the treatment groups is used).

- Calculations of the means and the variance
  are performed as part of the hypothesis test.

One-way ANOVA Model (3)

- The commonly used normal linear models
  for a completely randomized experiment are:
- $i = 1, \dots, I$ is an index over experimental units
- $j = 1, \dots, J$ is an index over treatment groups
- I_j is the number of experimental units in the j-th treatment group
- $I = \sum_j I_j$ is the total number of experimental units

One-way ANOVA Model (4)

- $y_{i,j}$ are observations
- μ_j is the mean of the observations for the jth treatment group
- μ is the grand mean of the observations
- τ_j is the jth treatment effect, a deviation from the grand mean
- $\sum \tau_j = 0$
- $\mu_j = \mu + \tau_j$
- $\varepsilon \sim N(0, \sigma^2)$,
- $\varepsilon_{i,j}$ are normally distributed zero-mean random errors.

two-way ANOVA Model (1)

- the two-way analysis of variance (ANOVA) is an extension of the one-way ANOVA that examines the influence of <u>two</u> <u>different</u> <u>categorical</u> <u>independent</u> <u>variables</u> on <u>one</u> <u>continuous</u> <u>dependent</u> <u>variable</u>.
- The two-way ANOVA not only aims at assessing the main <u>effect</u> of each <u>independent variable</u> but also if there is any <u>interaction</u> between them.

two-way ANOVA Model (2)

- Suppose a data set
  for which a dependent variable may be influenced
  by two factors which are potential sources of variation.
- The first factor has I levels $i \in \{1, \ldots, I\})$
  and the second has J levels $j \in \{1, \ldots, J\})$

two-way ANOVA Model (3)

- Each combination (i, j) defines a treatment, for a total of $I \times J$ treatments.
- We represent the number of replicates for treatment (i, j) by n_{ij},
- and let k be the index of the replicate in this treatment $(k \in \{1, \ldots, n_{ij}\})$.

- From these data, we can build a contingency table, where
- $n_{i+} = \sum_{j=1}^{J} n_{ij}$
- $n_{+j} = \sum_{i=1}^{I} n_{ij}$
- $n = \sum_{i,j} n_{ij} = \sum_{i} n_{i+} = \sum_{j} n_{+j}$

  the total number of replicates

two-way ANOVA Model (5)

- The experimental design is balanced
  if each treatment has the same number of replicates, K.
- the design is also said to be orthogonal
  allowing to fully distinguish the effects of both factors.
- We hence can write $\forall i, j \; n_{ij} = K$, and
  $\forall i, j \; n_{ij} = \dfrac{n_{i+} \cdot n_{+j}}{n}$

two-way ANOVA Model (6)

- In the classical approach,
  testing null hypotheses
  (that the factors have no effect)
  is achieved via their significance
  which requires calculating sums of squares.
- Testing if the interaction term is significant
  can be difficult because of
  the potentially large number of degrees of freedom.

two-way ANOVA Model (7)

- Following Gelman and Hill,
  the assumptions of the ANOVA,
  and more generally the general linear model,
  are, in decreasing order of importance:
  1. the data points are relevant
     with respect to the scientific question under investigation;
  2. the mean of the response variable is influenced additively
     (if not interaction term) and linearly by the factors;
  3. the errors are independent;
  4. the errors have the same variance;
  5. the errors are normally distributed.

Within Groups Variance Estimate S_W^2 (1)

- weighted estimate of H_0 population variance, σ^2
  weighted average of s_1^2 and s_2^2

$$s_W^2 = \frac{SS_1 + SS_2}{(n_1 - 1) + (n_2 - 1)} = \frac{SS_1 + SS_2}{N - 2}$$

Within Groups Variance Estimate S_W^2 (2)

- weighted estimate of H_0 population variance, σ^2
  weighted average of \$s_1^2, s_2^2, . . . \$ and s_k^2

$$s_W^2 = \frac{SS_1 + SS_2 + ... + SS_k}{(n_1 - 1) + (n_2 - 1) + ... + (n_k - 1)}$$

$$= \frac{SS_1 + SS_2 + ... + SS_k}{N - k}$$