Why Anova Works

Yakov Shklarov / 2020-01-01

In this expository note, I'd like to explain the theoretical foundations of Anova from the viewpoint of modern linear algebra. We'll take an abstract perspective that ignores the distinctions between Anova, Ancova, and regression analysis: so, for present purposes,[1]

$$\text{Anova} = \text{multiple linear regression} + \text{hypothesis testing} + \text{a way of tabulating the results.}$$

When Anova was introduced in the early 1900s, linear algebra was viewed in terms of matrices and coordinate vectors—abstract vector spaces and operators didn't start to grow popular until the 1920s–1930s. In my opinion, the newer, coordinate-free approach is the best way to understand Anova. But most students of statistics don't know very much linear algebra, so the subject is often presented in an opaque, convoluted way.

If you understand linear algebra and basic probability theory but don't know much about Anova then I hope that the explanation here will be more efficient (and more interesting!) than a lecture course on regression analysis or a standard stats textbook. If you're like me, you'll have an easier time if you understand the theory first, and only later dive into the specifics of experimental design such as contrasts, factorial designs, and blocking.

If you already understand Anova on a practical level but are mystified by the magical formulas for F statistics, hat matrices, and sums-of-squares, then this note might help you pull things together into a coherent framework.

Contents

[1] Depending on who you ask, Anova also entails a certain batching or structuring of coefficients, and/or requires all coefficients to be discrete. But our characterization captures the general mathematical spirit. For a rundown of the differences between Anova, Ancova, and so on, see
`https://www.quora.com/What-are-the-differences-between-ANOVA-ANCOVA-MANCOVA-etc/answer/Justin-Rising`.

1 The theory

Anova relies on four basic concepts, each of which we'll explore in some depth:

1. The spherical multivariate normal distribution,
2. Linear regression via orthogonal projection,
3. The Pythagorean Theorem, and
4. The Snedecor F distribution.

Allow me to give a brief high-level overview. The normal distribution has some very special mathematical features, which provide an interface to linear algebra that doesn't exist for any other probability distribution. Specifically, orthogonal decompositions of Euclidean space correspond to orthogonal decompositions of the (spherical) multivariate normal distribution. It's fortunate that the very same distribution that often arises naturally (by virtue of its role in the Central Limit Theorem) is precisely the one that has this interplay with linear algebra. If it weren't for this coincidence then our entire framework would be far less valuable.

The Pythagorean Theorem finds use here in several ways, but the most salient is that it lets us express the variance of a sample as (loosely speaking) the sum of variances along several orthogonal directions. The result is a rough heuristic for measuring how well a model fits. This is what practitioners mean when they say things like "so-and-so explains 84% of the variance in such-and-such".

The Snedecor F distribution serves as a linchpin to join all this geometry and probability theory with the statistical methodology of hypothesis testing.

1.1 The spherical multivariate normal distribution

The *spherical* or *standard multivariate normal distribution* (henceforth, SMVN) is the continuous probability distribution on $\mathbb{R}^n$ with density

$$f_n(\mathbf{x}) \;=\; \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(x_1^2 + \cdots + x_n^2)\right), \qquad \mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n,$$

where $n \geqslant 0$ and $\sigma > 0$. We'll use the (nonstandard) notation $\mathcal{SN}(\mathbb{R}^n, \sigma^2)$ for this distribution.[2]

The SMVN is special in that it enjoys two very nice properties:

(a) First, it's the product of its marginals, i.e., its density satisfies

$$f_n(\mathbf{x}) \;=\; f(x_1)f(x_2)\cdots f(x_n), \quad \mathbf{x} \in \mathbb{R}^n$$

where

$$f(x) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

Recall that this is just the density of the univariate normal distribution $\mathcal{SN}(\mathbb{R}, \sigma^2)$ ($\equiv N(0, \sigma^2)$).

In intuitive terms, picking a random vector in $\mathbb{R}^n$ according to $\mathcal{SN}(\mathbb{R}^n, \sigma^2)$ is equivalent to picking each of its n coordinates independently according to $\mathcal{SN}(\mathbb{R}, \sigma^2)$.

(b) Second, it's spherically symmetric about the origin. More precisely, it's invariant under orthogonal transformations: If A is an $n \times n$ real orthogonal matrix (i.e., one that represents a rotation or reflection), then applying A to the entire distribution leaves it unchanged. That is, if $X \sim \mathcal{SN}(\mathbb{R}^n, \sigma^2)$ then also $AX \sim \mathcal{SN}(\mathbb{R}^n, \sigma^2)$.

A remark on notation: If V is an n-dimensional real inner product space, then the distribution $\mathcal{SN}(V, \sigma^2)$ is well-defined because by property (b) it doesn't matter which orthonormal basis we use to specify the density. In traditional notation for the multivariate normal, $\mathcal{SN}(V, \sigma^2)$ is $\mathcal{N}_n(0, \sigma^2 I)$ (after picking an orthonormal

[2] The parameter for normal distributions is always the variance σ^2 rather than the standard deviation σ—this looks weird but is consistent with the notation $\mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$ for the general multivariate normal.

ordered basis for V and thus identifying V with $\mathbb{R}^n$). And if W is subspace of V then $\mathcal{SN}(W, \sigma^2)$ is $\mathcal{N}_n(0, \sigma^2 P)$, where P is the matrix that projects $\mathbb{R}^n$ orthogonally onto W (the distribution $\mathcal{N}_n(0, \sigma^2 P)$ is a so-called *degenerate* multivariate normal distribution on $\mathbb{R}^n$ whenever $\dim(W) < n$.)

To see why $\mathcal{SN}(W, \sigma^2) = \mathcal{N}_n(0, \sigma^2 P)$, notice that after a suitable orthogonal change of basis the matrix P becomes $\operatorname{diag}(1, 1, \ldots, 1, 0, 0 \ldots, 0)$ (with $\dim(W)$ 1's), and recall that for any matrix $R \in \mathcal{M}_{n \times n}(\mathbb{R})$ we have

$$X \sim \mathcal{N}_n(0, \Sigma) \implies RX \sim \mathcal{N}_n(0, R\Sigma R^T).$$

An immediate consequence of (a) and (b), which we'll need later, is that if we project the SMVN onto any subspace then we get the SMVN on that subspace (with the same parameter σ^2). To put it formally, take an n-dimensional real inner product space V with a k-dimensional subspace W, where $0 \leqslant k \leqslant n$. Let $\pi : V \to W$ be the orthogonal projection. If $X \sim \mathcal{SN}(V, \sigma^2)$ then $\pi(X) \sim \mathcal{SN}(W, \sigma^2)$. The reason this follows from properties (a) and (b) is that we can first rotate the distribution to line up the first k coordinate axes with W, and then integrate out the other $n - k$ coordinates along the orthogonal complement $W^\perp$. (By the way, a slight modification of this argument explains why *conditioning* on $X \in W$ also yields a multivariate normal. In fact, we can condition on any translate of W (i.e, any affine subspace of V). And this "slicing" property is also valid for non-spherical multivariate normal distributions, because those can be defined as affine transformations of a spherical normal distribution.)

The two properties (a) and (b) of the SMVN—expressibility as the product of marginals, and invariance under orthogonal transformations—are essential to Anova, as we'll see in the next two sections. What's more, no other distributions will work, because in dimension $n \geqslant 2$ these two properties uniquely characterize the SMVN!

Proof. Any probability density f_n on $\mathbb{R}^n$ that satisfies both (a) and (b) must satisfy

$$f(x_1)f(x_2)\cdots f(x_n) \;=\; f_n(\mathbf{x}) \;=\; f_n\left(|\mathbf{x}|\frac{\mathbf{x}}{|\mathbf{x}|}\right) \;=\; f(|\mathbf{x}|)f(0)^{n-1}, \quad \mathbf{x} \in \mathbb{R}^n \quad (\text{where } |\mathbf{x}| = \|\mathbf{x}\|_2).$$

Assume that f is continuous.[3] It's not too hard to show that this implies $f(x) > 0$ for all $x \in \mathbb{R}$, so we can rearrange and take logarithms:

$$\begin{aligned} \frac{f(x_1)}{f(0)}\frac{f(x_2)}{f(0)}\cdots\frac{f(x_n)}{f(0)} \;&=\; \frac{f(r)}{f(0)} \qquad (\text{where } r = \|\mathbf{x}\|) \\ \sum_{i=1}^{n}\big(\log f(x_i) - \log f(0)\big) \;&=\; \log f(r) - \log f(0). \end{aligned}$$

Since f must be an even function, we have $f(x) = f(\sqrt{x^2})$ for all $x \in \mathbb{R}$. Let $g(t) = \log f(\sqrt{t}) - \log f(0)$ for $t \geqslant 0$. The preceding equality becomes

$$\sum_{i=1}^{n} g(x_i^2) \;=\; g(r^2),$$

or (by the Pythagorean Theorem)

$$\sum_{i=1}^{n} g(x_i^2) \;=\; g\left(\sum_{i=1}^{n} x_i^2\right).$$

But g is continuous, and every additive continuous function is linear,[4] so $g(t) = at$ for some $a \in \mathbb{R}$. Thus,

$$\begin{aligned} \log f(x) \;&=\; ax^2 + \log f(0), \quad \text{or} \\ f(x) \;&=\; f(0)e^{ax^2}, \quad x \in \mathbb{R}. \end{aligned}$$

And we're done. We must have $a < 0$ because f is a probability density; $f(0)$ is just a normalizing constant. The usual convention is to put $a = -\frac{1}{2\sigma^2}$, then $f(0) = (2\pi\sigma^2)^{-1/2}$. □

[3] Actually, I believe that the assumption of continuity is unnecessary, though I haven't worked out the details.

[4] The equation $f(x+y) = f(x) + f(y)$ is called *Cauchy's functional equation*. To prove that implies linearity, first show by induction that $f(r) = rf(1)$ for every $r \in \mathbb{Q}$.

Incidentally, there's a well-known trick for evaluating the Gaussian integral $\int_{-\infty}^{\infty} e^{-x^2}\,dx$: Square the integral and convert to polar coordinates. I have a professor who likes to say that a trick, when used more then once, ceases to be a trick and becomes a device. But this is a "true trick": As we've just seen, *it's a theorem* that this method of integration only works for integrands of the form ke^{-cx^2}!

1.2 Linear regression via orthogonal projection

1.2.1 The linear model

I assume that the reader has seen linear regression before, so I'll only give a concise overview.

The (univariate) *linear model* used in Anova is

$$y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where X is an $n \times (p+1)$ real matrix called the *design matrix* (by virtue of its role in the design of experiments); $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ is called the *parameter*; and $\boldsymbol{\varepsilon}$, a random element of $\mathbb{R}^n$ with distribution $\mathcal{SN}(\mathbb{R}^n, \sigma^2)$, is called the *error*. So $y \in \mathbb{R}^n$ is also a random element of $\mathbb{R}^n$ (whereas X and $\boldsymbol{\beta}$ are not random). The terminology I'm using is slightly unorthodox: Statisticians prefer the plural forms *vector of parameters* and *vector of errors*.

For example, if

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

then eq. (1) becomes the so-called *simple linear regression* model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2), \qquad 1 \leqslant i \leqslant n.$$

Simple linear regression is a special case of *multiple linear regression*:

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix},$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2), \qquad 1 \leqslant i \leqslant n.$$

Multiple linear regression should not be confused with *multivariate linear regression*, where in eq. (1) y and $\boldsymbol{\beta}$ are replaced by matrices. I won't discuss multivariate regression further.

We call the model of eq. (1) linear because the *mean response* $\boldsymbol{\mu} := X\boldsymbol{\beta}$ is a linear function of $\boldsymbol{\beta}$. It's not necessary that $\boldsymbol{\mu}$ be linear in the regressors $x_1, \dots, x_q$: in fact, in simple and multiple linear regression $\boldsymbol{\mu}$ is not linear but merely affine in the regressors, unless $\beta_0 = 0$. As another example, the *quadratic polynomial model* is a special case of the linear model:

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix},$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2), \qquad 1 \leqslant i \leqslant n.$$

We can take polynomials of arbitrary degree, so in principle the linear model is flexible enough to fit any sample perfectly (i.e., with zero residual), as long as there are no replicates—no pairs of indices i_1, i_2 such that the i_1th and i_2th row of X are equal, but $y_{i_1} \neq y_{i_2}$.

Each tuple $(x_{i1}, x_{i2}, \ldots, x_{iq}, y_i)$ (associated with the ith row of eq. (1)) is called an *observation* or *sampling unit*. The symbols x_{ij} are called *explanatory variables* or *independent variables* or *regressors*; the symbols y_i are called *predicted variables* or *dependent variables* or *response variables* (actually, there exist at least a dozen synonyms). The sequence of all n observations is called the *sample*. So it's conventional to speak of a "sample of size n", but not of "n samples".

Note that the number of independent variables (q) is often much smaller than the width of the design matrix ($p + 1$).

Once the design matrix is chosen and the sample collected, the immediate goal is to obtain an estimate $\widehat{\boldsymbol{\beta}}$ for the parameter. We'll call the resulting vector $\widehat{\boldsymbol{\mu}} := X\widehat{\boldsymbol{\beta}}$ the *fitted vector*. The fitted vector is an estimate of the mean response $\boldsymbol{\mu}$. We'll call $\boldsymbol{e} := \boldsymbol{y} - X\widehat{\boldsymbol{\beta}}$ the *residual*. (The more common names are *vector of fitted values* and *vector of residuals* respectively.)

It's vital to understand the distinction between the error $\boldsymbol{\varepsilon}$ and the residual $\boldsymbol{e}$. The error is part of the model but cannot be determined from the sample unless the parameter $\boldsymbol{\beta}$ is known. The residual is not part of the model, but is determined from the sample, because $\widehat{\boldsymbol{\beta}}$ is a function of X and $\boldsymbol{y}$. You may wish to peek ahead to fig. 2.

I haven't yet discussed this function $(X, \boldsymbol{y}) \mapsto \widehat{\boldsymbol{\beta}}$. Choosing such a function is a major question in statistics and in machine learning: How do we fit the model? In the language of statistics, the function is called an *estimator* for $\boldsymbol{\beta}$. The output of the function—which is a random variable because $\boldsymbol{y}$ is a random variable—is also called an *estimator*. That is, statisticians often don't draw the conceptual distinction between the function and its output. For a given sample, the particular value that the estimator takes on (that is, the random variate) is called an *estimate* of $\boldsymbol{\beta}$. The symbol $\widehat{\boldsymbol{\beta}}$ denotes either the estimator or the estimate, depending on the context.

A schematic of the linear model is shown in fig. 1. On the left are the vectorized inputs (each is an element of $\mathbb{R}^n$), on the right is the vectorized output (also an element of $\mathbb{R}^n$). The model can be thought of as a machine that takes in data on the left and outputs randomized data on the right. Although we refer to $\boldsymbol{\beta}$ as the "parameter", we can see that there are actually three parameters: $\boldsymbol{\beta}$, σ, and the *functional form* of the model (the way in which the inputs are combined to form the design matrix X.) The number of inputs q and the sample size n are also parameters, but after the data are collected they cannot be changed.

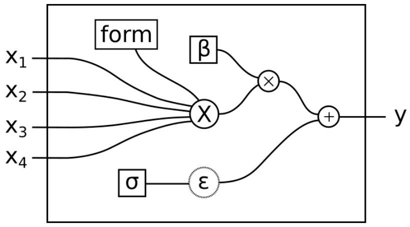


Figure 1: The linear model

A more common viewpoint is to not consider the functional form to be a parameter, but instead to consider each functional form to have its own associated model with parameters $\boldsymbol{\beta}$ and σ. There isn't really one right way to think about things here: The question of whether you're deciding on a model or deciding on a parameter is a matter of semantics.

There are a couple of ways in which the linear can be used. First, a statistician might start with an existing sample, and then try to find a configuration of parameters that is best matches the sample without overfitting. For certain applications it's desirable to have a *parsimonious* model: one where the functional form is simple and there are not too many variables or coefficients β_i. (There's a large assortment of classical methods for choosing a parsimonious model: Mallows's C_p, AIC, added variable plots, variance inflation factors, and so forth.) Typically an iterative optimization method is used: A simple functional form is chosen, then the best possible $\boldsymbol{\beta}$ is found for that functional form (we'll see momentarily how this is done), and the model is evaluated for how well it explains the sample. If the fit is inadequate, then a slightly more sophisticated functional form is chosen, and the process is repeated.

A second way to use the linear model is to decide on the functional form ahead of time, before data are available. Then the data can be collected in a specific way so as to best answer a given question. These sorts of procedures fall under the purview of the subject called the *design of experiments*.

1.2.2 Least squares linear regression

Suppose we've decided on a functional form and collected a sample, so we have a matrix X and a vector $\mathbf{y}$. It's time to fit the model: to find an estimate for $\boldsymbol{\beta}$ and σ^2. Computationally, the easiest way to get an estimate $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is to minimize the Euclidean norm of the residual $\mathbf{e}$ $(= \mathbf{y} - X\widehat{\boldsymbol{\beta}})$:

$$\widehat{\boldsymbol{\beta}} \;=\; \operatorname*{argmin}_{\mathbf{b}\,\in\,\mathbb{R}^{p+1}} |\mathbf{y} - X\mathbf{b}|.$$

For example, in the case of multiple linear regression, this corresponds to finding the coefficients $\widehat{\beta}_0, \ldots, \widehat{\beta}_p$ that minimize the so-called *sum of squares of residuals*

$$SS_{\text{res}} \;:=\; \sum_{i=1}^{n} e_i^2 \;=\; \sum_{i=1}^{n} \bigl(y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip})\bigr)^2.$$

More generally,

$$SS_{\text{res}} \;:=\; \sum_{i=1}^{n} e_i^2 \;=\; \sum_{i=1}^{n} \bigl(y_i - (X\widehat{\boldsymbol{\beta}})_i\bigr)^2.$$

Minimizing $|\mathbf{e}|$ amounts to making $\mathbf{e}$ orthogonal to the column space $C(X) \leqslant \mathbb{R}^n$. For $n = 2$ and $n = 3$, this is visually intuitive (see fig. 2.) To prove the general case, we appeal to the Pythagorean Theorem: Assume that $\widehat{\boldsymbol{\beta}}$ satisfies $(\mathbf{y} - X\widehat{\boldsymbol{\beta}}) \perp C(X)$. Such a $\widehat{\boldsymbol{\beta}}$ necessarily exists by orthogonal decomposition of $\mathbb{R}^n$.[5] And for any other point $\widehat{\boldsymbol{\mu}}' \in C(X)$ we have

$$|\mathbf{y} - X\widehat{\boldsymbol{\beta}}|^2 \;=\; |\mathbf{y} - \widehat{\boldsymbol{\mu}}'|^2 - |X\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\mu}}'|^2 \;\leqslant\; |\mathbf{y} - \widehat{\boldsymbol{\mu}}'|^2,$$

so $X\widehat{\boldsymbol{\beta}}$ is indeed the closest point to $\mathbf{y}$ in $C(X)$.

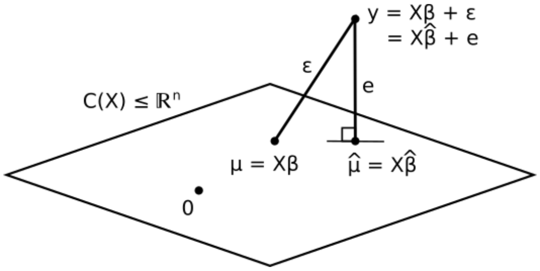


Figure 2: The geometry of least squares

If X has full column rank then we can give a simple explicit formula for the least squares estimate $\widehat{\boldsymbol{\beta}}$. First observe that X^TX is invertible:

$$\mathbf{v} \in \ker(X^TX) \;\Longrightarrow\; (X^TX)\mathbf{v} = \mathbf{0} \;\Longrightarrow\; \mathbf{v}^TX^TX\mathbf{v} = 0 \;\Longrightarrow\; (X\mathbf{v})^T(X\mathbf{v}) = 0 \;\Longrightarrow\; X\mathbf{v} = \mathbf{0} \;\Longrightarrow\; \mathbf{v} = \mathbf{0}.$$

Let $\widehat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{y}$. Then

$$X^T(\mathbf{y} - X\widehat{\boldsymbol{\beta}}) \;=\; X^T\bigl(\mathbf{y} - X(X^TX)^{-1}X^T\mathbf{y}\bigr) \;=\; X^T\mathbf{y} - X^T\mathbf{y} \;=\; \mathbf{0},$$

so $(\mathbf{y} - X\widehat{\boldsymbol{\beta}}) \perp C(X)$ as required.[6]

[5] There's a more general result: For every Hilbert space H, every nonempty closed convex set $E \subseteq H$ has the property that every point in H has a unique best approximation in E. For finite-dimensional Hilbert spaces, the converse is true: If a set E has the property "every point in H has a unique best approximation in E", then E is convex. But it's unknown whether this converse holds for infinite-dimensional Hilbert spaces.

[6] When X doesn't have full column rank, there's more than one $\widehat{\boldsymbol{\beta}}$ that satisfies the normality condition. The typical thing to do then is to minimize $|\widehat{\boldsymbol{\beta}}|$. For an in-depth study of similar problems, see: Wang, Wei, Qiao, *Generalized Inverses: Theory and Computations, Second Edition*, Springer, 2018.

To summarize, the estimate $\widehat{\beta}$ that minimizes the norm of the residual is the solution (not necessarily unique) to

$$X^T(y - X\widehat{\beta}) = 0.$$

This equation is called the *normal equation* because it's satisfied by $\widehat{\beta}$ if and only if $C(X)$ is normal to e.

The matrix H that orthogonally projects $\mathbb{R}^n$ onto $C(X)$ is called the *hat matrix*, because it transforms 'y' into '$\widehat{y}$' (a common synonym for '$\widehat{\mu}$'). If X has full column rank, then the hat matrix is

$$H = X(X^TX)^{-1}X^T.$$

The formula for H should look plausible because we want $Hy = X\widehat{\beta}$ where $\widehat{\beta}$ is the least squares estimate. Here's a proof of the formula. Let $v \in \mathbb{R}^n$, and decompose v as the sum of a vector orthogonal to $C(X)$ and a vector parallel to $C(X)$:

$$v = v^{\perp} + v^{||}.$$

Pick α such that $v^{||} = X\alpha$. Since $X^Tv^{\perp} = 0$, we have

$$\big(X(X^TX)^{-1}X^T\big)v = \big(X(X^TX)^{-1}X^T\big)X\alpha = X(X^TX)^{-1}(X^TX)\alpha = X\alpha = v^{||}.$$

The projection matrix H (which sends y to $\widehat{\mu}$) and its complementary partner $I - H$ (which sends y to e) will play a major role in what is to come.

1.2.3 Least squares and the SMVN

Several important results follow immediately. If these results aren't obvious, try to build a geometric understanding of the discussion in section 1.1.

First, the random variables

$$\widehat{\mu} = \mu + H\varepsilon \qquad \text{and} \qquad e = (I - H)\varepsilon$$

(the least squares fitted vector and residual) are independent. Incidentally, since $\widehat{\beta}$ is a function of $\widehat{\mu}$, this implies that $\widehat{\beta}$ and e are independent.

Second, since $y \sim \mu + \mathcal{SN}(\mathbb{R}^n, \sigma^2)$,

$$\begin{aligned} \widehat{\mu} &\sim \mu + \mathcal{SN}(C(X), \sigma^2) \quad \text{and} \\ e &\sim \mathcal{SN}(C(X)^{\perp}, \sigma^2) \end{aligned}$$

where $C(X)$ is the column space of X and $C(X)^{\perp}$ is its orthogonal complement.

Finally, the covariance matrices of $\widehat{\mu}$ and e are

$$\begin{aligned} \mathrm{Cov}(\widehat{\mu}) &= \sigma^2 H, \\ \mathrm{Cov}(e) &= \sigma^2 (I - H). \end{aligned}$$

1.2.4 Why use least squares (and when not to)

As we've shown, least squares regression is computationally convenient (only matrix arithmetic is required) and gives a clean decomposition of the SMVN. Least squares has another pleasant characteristic: It yields a maximum-likelihood estimate for $\widehat{\beta}$, because the density of the SMVN is monotonically decreasing with distance from the origin, and the least squares estimate of $\widehat{\beta}$ is precisely the one that minimizes $|\widehat{\mu} - y|$.

Unfortunately, in practice, errors are rarely normally distributed. For any real-world system the error will have a skewed and/or fat-tailed distribution. In a sample from a fat-tailed distribution extreme points will look like "outliers" relative to a normal distribution. If you happen to know the true distribution of the error then you can get a maximum likelihood estimate by minimizing a cost function other than sum of squares of residuals. For details on how to find the correct cost function, see

– Boyd, Stephen and Vandenberghe, Lieven. Maximum likelihood estimation. *Convex Optimization*. Cambridge University Press, 2004. pp. 351–357.

The requirement for errors to be normal is a fundamental limitation of Anova. Of course we can still do a least squares fit even if the true error is not normal, but then all F statistics will be garbage. In some scenarios it's possible to get away without normality by appealing to the Central Limit Theorem. But this needs to be done with care: Convergence to the normal can be slower than you might think, especially if the underlying distribution is heavily skewed.

1.3 The Pythagorean Theorem

The usual Pythagorean Theorem for $\mathbb{R}^n$ states that for every vector $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$,

$$|y|^2 = \sum_{i=1}^{n} |y_i|^2.$$

It will be convenient to work at a slightly higher level of abstraction: If $y^{[1]}, y^{[2]}, \ldots, y^{[k]}$ are pairwise orthogonal vectors in $\mathbb{R}^n$ (actually, in any real inner product space) then

$$\left|\sum_{i=1}^{k} y^{[i]}\right|^2 = \sum_{i=1}^{k} \left|y^{[i]}\right|^2.$$

In Anova, the response vector y is decomposed into pairwise orthogonal components by successive projections onto linear subspaces, and then the magnitudes of these components are compared against one another as part of a test of hypotheses. In typical applications the various subspaces are nested within one another, and we'll focus on that case exclusively.

To that end, a *flag* in $\mathbb{R}^n$ is a chain of subspaces that includes $\{0\}$ and $\mathbb{R}^n$. That is, a flag is a sequence $V_0, V_1, \ldots, V_k$ of subspaces with

$$\{0\} = V_0 \subsetneq V_1 \subsetneq V_2 \subsetneq \cdots \subsetneq V_k = \mathbb{R}^n.$$

The word "flag" is meant to evoke an actual flag flying in the wind: The flag spans a subspace of dimension 2, the flagpole spans a subspace of dimension 1, the pointy end of the flagpole is a subspace of dimension 0. Note that we can have $k < n$ (when $k = n$, the flag is said to be *complete*: it contains a subspace of every dimension between 0 and n.)

Every flag gives rise to a sequence of orthogonal projection operators $\pi_i : \mathbb{R}^n \to V_i$, $0 \leqslant i \leqslant k$ (these correspond to successive "hat matrices".) The projection operators $\pi_0, \ldots, \pi_k$ can be used to decompose the vector y as follows:

$$\begin{aligned} y &= y^{[1]} + y^{[2]} + \cdots + y^{[k]} \\ \text{where} \quad y^{[i]} &:= \pi_i(y) - \pi_{i-1}(y) \in V_i \cap V_{i-1}^{\perp} \quad \text{for } 1 \leqslant i \leqslant k. \end{aligned}$$

The components $y^{[i]}$ are pairwise orthogonal because

$$\begin{aligned} y^{[j]} &\in V_j, \\ y^{[i]} &\in V_{i-1}^{\perp} \subseteq V_j^{\perp} \quad \text{for } 1 \leqslant j < i \leqslant n. \end{aligned}$$

It's convenient to also introduce the notation

$$y^{[a,b]} := \sum_{a \leqslant i \leqslant b} y^{[i]} \qquad \text{for } 1 \leqslant a \leqslant b \leqslant k,$$

so that for example

$$y = y^{[1]} + y^{[2,4]} + y^{[5]} + \cdots + y^{[k]}.$$

The identity

$$|y|^2 = |y^{[1]}|^2 + |y^{[2]}|^2 + \cdots + |y^{[k]}|^2 \tag{2}$$

(and variations involving components of the form $\mathbf{y}^{[a,b]}$) is referred to by statisticians as the *Additional Sum of Squares Principle*. The statisticians' terminology and conventions are explained further in section 1.3.1.

The k pairwise orthogonal subspaces $V_i \cap V_{i-1}^{\perp}$, $1 \leqslant i \leqslant k$ decompose the SMVN into k independent components. More generally, we can say something similar even if the mean of the normal distribution is not $\mathbf{0}$. Assume that $\mathbf{y}$ is the response vector in the linear model (eq. (1)), and $\boldsymbol{\mu} \in V_j$ for some fixed $0 \leqslant j < k$. Then, by the properties of the SMVN discussed in section 1.1,

$$\mathbf{y}^{[i]} \;\sim\; \mathcal{SN}(V_i \cap V_{i-1}^{\perp},\, \sigma^2) \quad \text{for } j+1 \leqslant i \leqslant k$$

and, moreover, the random elements $\mathbf{y}^{[j+1]}, \ldots, \mathbf{y}^{[k]}$ are mutually independent.

We'll see in section 1.4 how Anova compares the lengths of the components $\mathbf{y}^{[j+1]}, \ldots, \mathbf{y}^{[k]}$: Loosely speaking, we should expect their lengths to be commensurate; a lack of commensurability is evidence against the hypothesis $\boldsymbol{\mu} \in V_j$. It might not be obvious at this point that the lengths should be commensurate: the effect only gets pronounced when the dimension of the space is large (i.e., for a large sample), because of the phenomenon of concentration of measure in high dimensions (or, to put it another way, because of the Law of Large Numbers.)

1.3.1 The Additional Sum of Squares Principle

The *Additional Sum of Squares Principle* (or *Extra Sum of Squares Principle*) is simply another name for the Pythagorean Theorem in the context of Anova.

Here's the best-known scenario. Suppose that the first column of X is all 1s, for example as in simple or multiple linear regression. Let

$$V_0 = \{\mathbf{0}\}, \quad V_1 = \operatorname{span}\{\mathbf{1}\}, \quad V_2 = C(X), \quad V_3 = \mathbb{R}^n.$$

where $\mathbf{1} = (1, 1, \ldots, 1)$. Then

$$|\mathbf{y}^{[2,3]}|^2 \;=\; |\mathbf{y}^{[2]}|^2 + |\mathbf{y}^{[3]}|^2. \tag{3}$$

This identity is often written as

$$\mathrm{SS_{tot}} = \mathrm{SS_{reg}} + \mathrm{SS_{res}}$$

where

$$\mathrm{SS_{tot}} \;:=\; |\mathbf{y}^{[2,3]}|^2 \;=\; |\mathbf{y} - \mathbf{y}^{[1]}|^2 \;=\; \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{is called the } \textit{total sum of squares},$$

$$\mathrm{SS_{reg}} \;:=\; |\mathbf{y}^{[2]}|^2 \;=\; \sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2 \quad \text{is called the } \textit{regression sum of squares},$$

$$\mathrm{SS_{res}} \;:=\; |\mathbf{y}^{[3]}|^2 \;=\; \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 \quad \text{is called the } \textit{residual sum of squares}.$$

Here $\bar{y}$ is the *sample mean* $\frac{1}{n}\sum_{i=1}^{n} y_i$, and $\hat{\mu}_i$ is the ith fitted value. The total sum of squares is more properly called the *corrected* total sum of squares (it's been corrected for the sample mean $\bar{y}$.)

A number of other conventions are in use:

Quantity	Notation
Total sum of squares	$\mathrm{SS_{tot}}$, SST, TSS
Regression sum of squares	$\mathrm{SS_{reg}}$, SSR, RSS, ESS
Residual sum of squares	$\mathrm{SS_{res}}$, $\mathrm{SS_{err}}$, SSR, RSS, SSE

The "ESS" in the second row of the table stands for *explained sum of squares*. The "SSE" in the third row stands for "sum of squares of the error", a common but incorrect phrase. Some people try to salvage the abbreviation SSE by claiming that it stands for something like "sum of squared estimate of the error", but arguably that's even worse (the residual is not a great estimator for the error, since it's always smaller.)

There are countless variations on the theme. For example, sometimes there are *replicates* in the sample—sampling units with the same list of inputs $(x_1, \ldots, x_q)$ but with different response y. In that case it's customary to put $V_1 = C(X)$ and $V_2 = \text{span}\{(1, 1, \ldots, 1, 0, \ldots, 0),\ (0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0),\ \ldots,\ (0, \ldots, 0, 1, \ldots, 1)\}$ so that projection onto V_2 amounts to computing the mean response within each group. Equation (3) is now written as

$$\text{SS}_{\text{tot}} \;=\; \text{LFSS} + \text{PESS}$$

where LFSS stands for *lack of fit sum of squares* and PESS stands for *pure error sum of squares*. This setup is useful because it often has high statistical power for detecting a lack of fit, i.e., for rejecting the null hypothesis that the data came from the model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. See section 2 for an example of LFSS/PESS.

Another standard scenario is to compare the fit of two different functional forms with corresponding design matrices X_A and X, where $C(X_A) \subseteq C(X)$. (Usually X_A is not described explicitly, but instead $C(X_A)$ is simulated by placing a linear restriction on β.) The flag used here is

$$V_0 = \{\mathbf{0}\}, \quad V_1 = \text{span}\{\mathbf{1}\}, \quad V_2 = C(X_A), \quad V_3 = C(X), \quad V_4 = \mathbb{R}^n.$$

The functional form of X is called the *full model*, the functional form of X_A is called the *restricted model*. The idea is to decide whether the restricted model is adequate to explain the data, or whether the full model is necessary.

1.3.2 Degrees of freedom in Anova

The random variable $\mathbf{y}$ lives in $\mathbb{R}^n$, so it's said to have n *degrees of freedom*.[7] Its component $\mathbf{y}^{[i]}$ lives in $V_i \cap V_{i-1}^{\perp}$ so it's said to have

$$\text{df}_i \;:=\; \dim(V_i \cap V_{i-1}^{\perp}) = \dim(V_i) - \dim(V_{i-1})$$

degrees of freedom. For example, in the first scenario of section 1.3.1 (multiple linear regression), if X has full column rank then

$$\begin{aligned} \text{df}_3 \;&=\; \dim(\mathbb{R}^n) - \dim(C(X)) \;=\; n - p - 1, \\ \text{df}_2 \;&=\; \dim(C(X)) - 1 \;=\; p. \end{aligned}$$

Since $\mathbf{y}^{[3]}$ is simply $\boldsymbol{e}$, the quantity df_3 is called the *degrees of freedom of residual* and is written as df_{res}. And $\mathbf{y}^{[2]} = \pi_2(\mathbf{y}) - \pi_1(\mathbf{y}) = \hat{\boldsymbol{\mu}} - \bar{y}\mathbf{1}$, so the quantity df_2 is called the (corrected) *degrees of freedom of regression* and is written as df_{reg}. Their sum

$$\text{df}_{\text{tot}} \;:=\; \text{df}_{\text{res}} + \text{df}_{\text{res}} \;=\; (n - p - 1) + p \;=\; n - 1$$

is called the (corrected) *total degrees of freedom*. It's equal to the number of degrees of freedom of the component $\mathbf{y}^{[2,3]} := \mathbf{y}^{[2]} + \mathbf{y}^{[3]}$.

The number of degrees of freedom is used for computing the F statistic. An intermediate step is to compute the *sum of squares* of $\mathbf{y}^{[i]}$

$$\text{SS}_i \;:=\; |\mathbf{y}^{[i]}|^2,$$

and *mean square* of $\mathbf{y}^{[i]}$

$$\text{MS}_i \;:=\; \frac{\text{SS}_i}{\text{df}_i}.$$

The mean square should be thought of as the average square deviation from $\mathbf{0}$ along the one-dimensional components of $\mathbf{y}^{[i]}$.

This is all we'll need, though I should mention that the general concept of degrees of freedom in statistics is quite a bit more complicated.

[7] Perhaps generalizing the notion will help clarify it: If $\boldsymbol{w}$ is a random element whose support $\text{supp}\,\boldsymbol{w}$ is a topological manifold, then the *number of degrees of freedom* of $\boldsymbol{w}$ is $\text{df}_{\boldsymbol{w}} := \dim(\text{supp}\,\boldsymbol{w})$.

1.3.3 Bessel's correction and its generalization

We're now in a position to understand intuitively the notorious "$n - 1$" in the formula for the sample variance,

$$s^2 \ := \ \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

This formula gives an unbiased estimate of the variance σ^2 when the sample is drawn independently from $N(\mu, \sigma^2)$ with unknown population mean μ. In terms of our linear model (eq. (1)), X is a single column of 1s and $\beta_0 = \mu$; there are no regressors. Notice that $\boldsymbol{\mu} = \mu\mathbf{1}$. Take the flag

$$V_0 = \{0\}, \quad V_1 = C(X) = \mathrm{span}(\mathbf{1}), \quad V_2 = \mathbb{R}^n.$$

Decompose the response as $\mathbf{y} = \mathbf{y}^{[1]} + \mathbf{y}^{[2]}$. We have

$$\begin{aligned} \mathbf{y} &\sim \boldsymbol{\mu} + \mathcal{SN}(\mathbb{R}^n, \sigma^2), \\ \mathbf{y}^{[1]} &\sim \boldsymbol{\mu} + \mathcal{SN}(\mathrm{span}(\mathbf{1}), \sigma^2), \\ \mathbf{y}^{[2]} &\sim \mathcal{SN}(\mathrm{span}(\mathbf{1})^{\perp}, \sigma^2). \end{aligned}$$

But $\mathbf{y}^{[2]} = (y_1 - \bar{y}, \ldots, y_n - \bar{y})$ and $\mathrm{df}_2 = n - 1$, so the expression for s^2 is precisely the mean square of $\mathbf{y}^{[2]}$. The expectation of s^2 is σ^2 because the expectation of the square of a single one-dimensional component of the SMVN is $\mathbf{E}(x^2) = \mathbf{E}(x^2) - \mathbf{E}(x)^2 = \sigma^2$. Furthermore, it's hopeless to somehow extract information about σ^2 from the remaining component $\mathbf{y}^{[1]}$, because we don't know how far $\mathbf{y}$ is from $\boldsymbol{\mu}$ along the axis $\mathrm{span}(\mathbf{1})$.

More generally, suppose that the sample came from a linear model with arbitrary (but known) functional form. As long as $C(X) \subsetneq \mathbb{R}^n$, we can still learn about σ^2 by examining the component of $\mathbf{y}$ orthogonal to $C(X)$. If X has full column rank $p + 1$, then (by reasoning just as above) an unbiased estimate for σ^2 is

$$s^2 \ = \ \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2.$$

For example, in simple linear regression we get

$$s^2 \ = \ \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

An enticing picture begins to come into focus: The output $\mathbf{y}$ from the model depicted in fig. 1 can be thought of as joint information about the mean $\boldsymbol{\mu}$ (equivalently, $\boldsymbol{\beta}$) and the variance σ^2. If we know a priori that $\boldsymbol{\mu}$ lies within a subspace $C(X) \leqslant \mathbb{R}^n$ then the information decomposes neatly into two orthogonal components $\mathbf{y}^{[1]} \in C(X)$ and $\mathbf{y}^{[2]} \in C(X)^{\perp}$, the former being purely information about the mean and the latter being purely information about the variance. Making this idea precise will have to wait for some other day.

1.3.4 * The Parallel Axis Theorem

The Parallel Axis Theorem is a computational tool for working with second moments, purloined from classical mechanics. While we won't need it for what we're doing, it's nice to know that it's really just the Pythagorean Theorem in disguise.

Theorem (Parallel Axis). *Let X be a real-valued random variable with mean μ and finite variance. Then, for every $c \in \mathbb{R}$,*

$$\mathbf{E}((X - c)^2) \ = \ \mathrm{Var}(X) + (\mu - c)^2.$$

Proof. The random variable X is a point in the Hilbert space $L_2(\xi)$ where ξ is the distribution of X. If we also consider c and μ as constant functions in $L_2(\xi)$, the identity may be rewritten as

$$\|X - c\|^2 \ = \ \|X - \mu\|^2 + \|\mu - c\|^2.$$

But this is a special case of the Pythagorean Theorem because

$$\begin{aligned} \langle X-\mu,\ \mu-c\rangle &= (\mu-c)\langle X-\mu,\ 1\rangle \\ &= (\mu-c)\big(\langle X,1\rangle-\langle\mu,1\rangle\big) \\ &= (\mu-c)(\mu-\mu) \\ &= 0. \end{aligned} \qquad \square$$

1.4 The Snedecor F distribution

The most natural way to define the Snedecor F distribution is in terms of normal random variables. Textbooks often take the backwards approach: They define the Snedecor F distribution via its PDF and then present its characterization in terms of normal variables as a consequence, as though it were just a happy coincidence. We won't bother with the PDF here.

Take $\sigma > 0$ and $d_1, d_2 \in \mathbb{N}_{>0}$. Let $U_1, \ldots, U_{d_1}, W_1, \ldots, W_{d_2}$ be mutually independent random variables with distribution $N(0, \sigma^2)$, and define

$$F_{d_1,d_2} := \text{ the distribution of } \frac{(U_1^2+\cdots+U_{d_1}^2)/d_1}{(W_1^2+\cdots+W_{d_2}^2)/d_2}.$$

The distribution F_{d_1,d_2} depends on d_1 and d_2, but it doesn't depend on σ^2 because rescaling both numerator and denominator by the same quantity leaves F_{d_1,d_2} unchanged. We call F_{d_1,d_2} the *Snedecor F distribution* or simply the *F-distribution* with parameters d_1 and d_2 (sometimes called the *numerator* and *denominator degrees of freedom*, respectively.)

Equivalently, F_{d_1,d_2} is the distribution of

$$\frac{X_{d_1}/d_1}{X_{d_2}/d_2}$$

for independent $X_{d_1} \sim \chi^2(d_1)$ and $X_{d_2} \sim \chi^2(d_2)$, because the chi-square distribution on k degrees of freedom is precisely the distribution of the sum of squares of k independent $N(0,1)$ variables. Similarly, it's the distribution of

$$\frac{|S_{d_1}|^2/d_1}{|S_{d_2}|^2/d_2}$$

for independent $S_{d_1} \sim \mathcal{SN}(\mathbb{R}^{d_1}, \sigma^2)$ and $S_{d_2} \sim \mathcal{SN}(\mathbb{R}^{d_2}, \sigma^2)$. Notice that this last characterization is essentially as a ratio of mean squares (as defined in section 1.3.2.)

Let $X_k \sim \chi^2(k)$. As k grows, the distribution of X_k/k concentrates near 1 (fig. 3). This follows from the Law of Large Numbers because for $U \sim N(0,\sigma^2)$ we have $\mathbf{E}[U^2] = \mathbf{E}[(U-0)^2] = \mathrm{Var}[U] = \sigma^2$. Likewise, $|S_{d_1}|^2/d_1$ converges in probability to σ^2 as $d_1 \to \infty$. When both d_1 and d_2 are large, both numerator and denominator are concentrated near σ^2, and thus F_{d_1,d_2} is concentrated near 1 (fig. 4).

Now on to null hypothesis significance testing. Suppose that we have a flag

$$\{0\} = V_0 \subsetneq V_1 \subsetneq V_2 \subsetneq \cdots \subsetneq V_k = \mathbb{R}^n,$$

and suppose also that $\mu \in V_j$ for some $0 \leqslant j < k$. Let $y \sim \mu + \mathcal{SN}(\mathbb{R}^n, \sigma^2)$, and write

$$y = y^{[1]} + \cdots + y^{[j]} + y^{[j+1]} + y^{[j+2,k]}$$

as in section 1.3. Then evidently

$$\frac{|y^{[j+1]}|^2/\mathrm{df}_{j+1}}{|y^{[j+2,k]}|^2/\mathrm{df}_{j+2,k}} \sim F_{\mathrm{df}_{j+1},\mathrm{df}_{j+2,k}} \tag{4}$$

where $\mathrm{df}_{j+1} = \dim(V_{j+1} \cap V_j^{\perp})$ and $\mathrm{df}_{j+2,k} = \dim(V_{j+1}^{\perp})$.

In this scenario V_j corresponds to the restricted model and V_{j+1} corresponds to the full model. The null hypothesis is that μ belongs to the restricted model. If the observed value of the statistic in eq. (4) is large

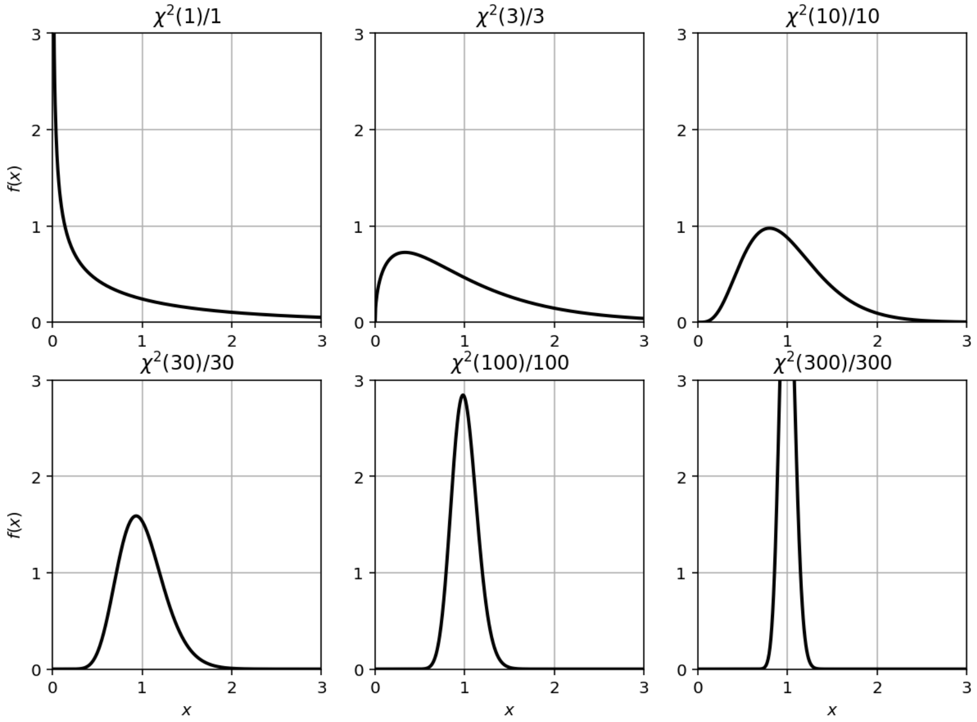


Figure 3: Density of $\chi^2(k)/k$ for $k = 1, 3, 10, 30, 100, 300$

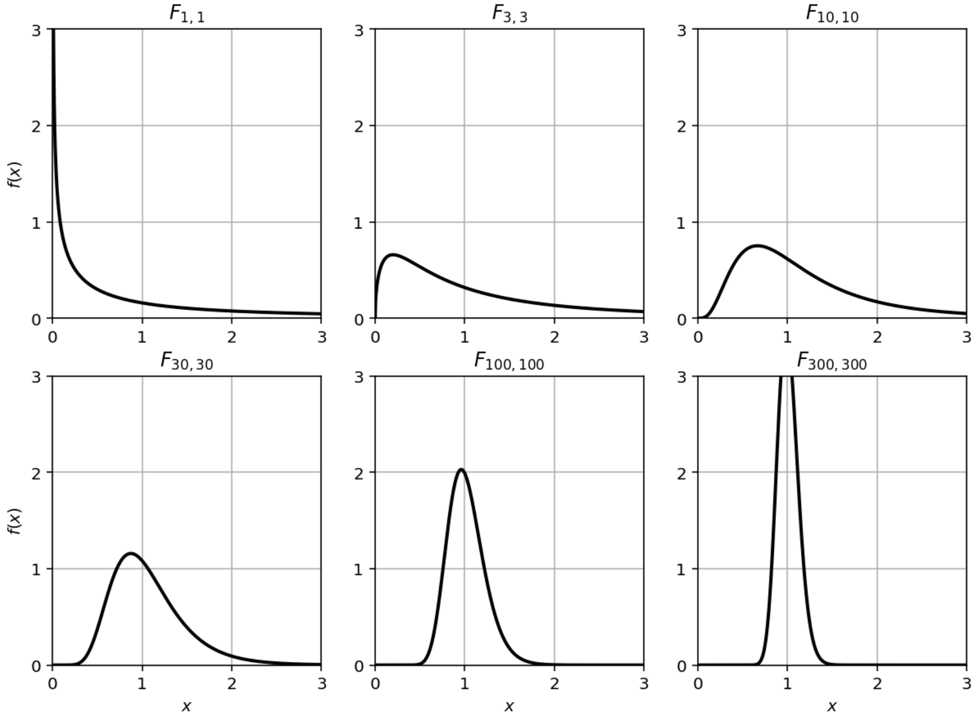


Figure 4: Density of $F_{k,k} = \frac{\chi^2(k)/k}{\chi^2(k)/k}$ for $k = 1, 3, 10, 30, 100, 300$

relative to the mode of the F-distribution (we do a one-tailed test here) then we have evidence against the null hypothesis.

Note that the choice of full model doesn't affect the null hypothesis: it only affects the power of the test. So a high F statistic is evidence against the restricted model but is *not* evidence for the full model; at least, not within the framework of null hypothesis significance testing.

Equation (4) is usually written in a form closer to

$$F_{\mathrm{obs}} = \frac{(\mathrm{SS}'_{\mathrm{res}} - \mathrm{SS}_{\mathrm{res}})/(\mathrm{df}' - \mathrm{df})}{\mathrm{SS}_{\mathrm{res}}/\mathrm{df}} \sim F(\mathrm{df}' - \mathrm{df}, \mathrm{df}),$$

where $\mathrm{SS}_{\mathrm{res}}$ is the sum of squares of residual of the full model, $\mathrm{SS}'_{\mathrm{res}}$ is the sum of squares of residual of the restricted model, df is the number of degrees of freedom of the residual of the full model, and df' is the number of degrees of freedom of the residual of the restricted model.

As a side note, it should now be clear from section 1.3.3 why the F-distribution can be used for comparing the standard deviations of two populations: The *F-test* uses the fact that if the standard deviations are equal then

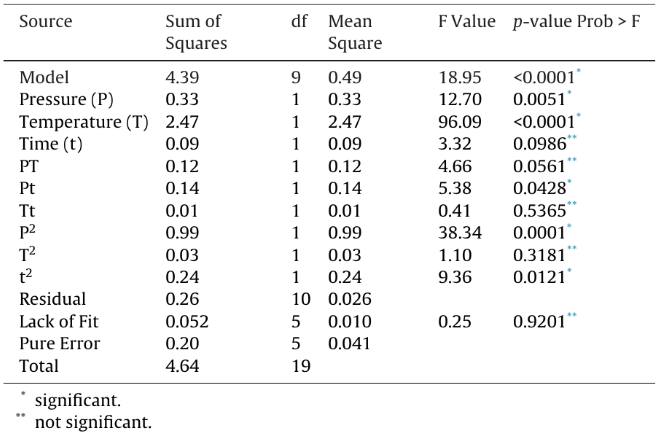
$$\frac{s_1^2}{s_2^2} \sim F_{n_1-1,\, n_2-1},$$

where $s_1^2,\ s_2^2$ are the sample variances and $n_1,\ n_2$ are the sample sizes.

2 Anova tables

The Anova table is a more-or-less standardized format for displaying F tests and their associated p-values for a sequence of linear models. These tables regularly crop up in scientific papers. In this section we'll briefly examine a typical Anova table (fig. 5), filched from a random paper.[8]

We've already covered the meaning behind all the tests, so this section serves only to establish a correspondence between the structure of the table and the development above.



Source	Sum of Squares	df	Mean Square	F Value	*p*-value Prob > F
Model	4.39	9	0.49	18.95	<0.0001*
Pressure (P)	0.33	1	0.33	12.70	0.0051*
Temperature (T)	2.47	1	2.47	96.09	<0.0001*
Time (t)	0.09	1	0.09	3.32	0.0986**
PT	0.12	1	0.12	4.66	0.0561**
Pt	0.14	1	0.14	5.38	0.0428*
Tt	0.01	1	0.01	0.41	0.5365**
P^2	0.99	1	0.99	38.34	0.0001*
T^2	0.03	1	0.03	1.10	0.3181**
t^2	0.24	1	0.24	9.36	0.0121*
Residual	0.26	10	0.026		
Lack of Fit	0.052	5	0.010	0.25	0.9201**
Pure Error	0.20	5	0.041		
Total	4.64	19			

* significant.
** not significant.

Figure 5: An Anova table from the wild

The table was constructed from an analysis of a sample of size 20. The label "Total" in the bottom row should technically be "Corrected Total" (the data were corrected for the sample mean), but it's common to just write "Total". The sum of squares for the corrected total is $|y - \mu\mathbf{1}|^2 = 4.64$ on $20 - 1 = 19$ degrees of freedom, where μ is the overall sample mean $\frac{1}{n}\sum_{i=1}^{20} y_i$. The other rows show a breakdown of the corrected total sum of squares into its components.

[8] The table was found in: Basegmez et al. Biorening of blackcurrant pomace into high value functional ingredients using supercritical CO_2, pressurized liquid and enzyme assisted extractions. Journal of Supercritical Fluids, 2017.

At the highest level, the total sum of squares is broken down into two orthogonal components: the "Model" (or regression) sum of squares, and the "Residual" sum of squares. Notice that $4.64 = 4.39 + 0.26$. Occasionally, the remaining rows are indented to make the hierarchical structure of the table more transparent.

The (full) model used was

$$y \;=\; \beta_0 + \beta_1 P + \beta_2 T + \beta_3 t + \beta_4 PT + \beta_5 Pt + \beta_6 Tt + \beta_7 P^2 + \beta_8 T^2 + \beta_9 t^2 + \varepsilon.$$

Let X be the design matrix of the full model; the size of X is 20×10. The top-level breakdown, in our language, corresponds to the flag

$$V_0 = \{0\}, \quad V_1 = \text{span}(\mathbf{1}), \quad V_2 = C(X), \quad V_3 = \mathbb{R}^{20}.$$

The corrected total sum of squares is $|\mathbf{y}^{[2,3]}|^2 = |\mathbf{y} - \mathbf{y}^{[1]}|^2$; the model sum of squares is $|\mathbf{y}^{[2]}|^2$; the residual sum of squares is $|\mathbf{y}^{[3]}|^2$. Note that $\mathbf{y}^{[2]}$ has $10 - 1 = 9$ degrees of freedom and $\mathbf{y}^{[3]}$ has $20 - 10 = 10$ degrees of freedom, as is displayed in the corresponding rows of the table.

The mean square is simply the ratio of sum of squares to degrees of freedom. The F Value (or F statistic) in the Model row is $18.95 = 0.49/0.026$: the ratio of mean square of the model to the mean square of the residual. The p-value is $\mathbf{P}(F > 18.95)$ where $F \sim F_{9,10}$. This p-value is quite low, which indicates strong evidence against the hypothesis that $\boldsymbol{\mu} \in V_1$, that is, that the data came from a model of the form $y = \mu + \varepsilon$.

Next we move on to the breakdown of the residual into "Lack of Fit" and "Pure Error". These two phrases always indicate that there were replicates in the sample—often an experiment will be designed this way on purpose. Let k be the number of groups (within each group, every sampling unit has the same P, T, and t), and let W be the design matrix of the model that can assign an arbitrary mean to each group:

$$y_i = \mu_1 g_1(i) + \cdots + \mu_k g_k(i) + \varepsilon_i,$$

where $g_j(i) = 1$ if the ith unit is in the jth group, and $g_j(i) = 0$ otherwise. Notice that $C(X) \subseteq C(W)$, so we can extend the flag to

$$V_0 = \{0\}, \quad V_1 = \text{span}(\mathbf{1}), \quad V_2 = C(X), \quad V_3 = C(W), \quad V_4 = \mathbb{R}^{20}.$$

The "Pure Error" sum of squares is $|\mathbf{y}^{[4]}|^2$; the "Lack of Fit" sum of squares is $|\mathbf{y}^{[3]}|^2$. We can deduce from the Pure Error degrees of freedom that there are $k = 20 - 5 = 15$ groups. If the full model is valid then $\boldsymbol{\mu} \in C(X)$, so the ratio $(|\mathbf{y}^{[3]}|^2/5)/(|\mathbf{y}^{[4]}|^2/5)$ is distributed according to $F_{5,5}$. The observed F value 0.25 is not significantly higher than 1, so this test doesn't give evidence against the full model.

The Model sum of squares is broken down into 9 components (rows P through t^2 in the table). Here there are two conventions for displaying the sum of squares: The *sequential sum of squares* (often explicitly indicated in the table header as Seq. SS) and *adjusted sum of squares* (often indicated as Adj. SS).[9] Sequential sum of squares means that the flag is generated by adding one component at a time, in the same order as they appear in the table: For example, the PT row's sum of squares would be computed by comparing the two models

$$\begin{aligned} y &= \beta_0 + \beta_1 P + \beta_2 T + \beta_3 t + \varepsilon, \\ y &= \beta_0 + \beta_1 P + \beta_2 T + \beta_3 t + \beta_4 PT + \varepsilon. \end{aligned}$$

Adjusted sum of squares would mean that all coefficients except β_4 are present in the model, that is, the adjusted sum of squares for the PT row is computed by comparing the two models

$$\begin{aligned} y &= \beta_0 + \beta_1 P + \beta_2 T + \beta_3 t + \beta_5 Pt + \beta_6 Tt + \beta_7 P^2 + \beta_8 T^2 + \beta_9 t^2 + \varepsilon, \\ y &= \beta_0 + \beta_1 P + \beta_2 T + \beta_3 t + \beta_4 PT + \beta_5 Pt + \beta_6 Tt + \beta_7 P^2 + \beta_8 T^2 + \beta_9 t^2 + \varepsilon. \end{aligned}$$

So for example, in the case of adjusted sum of squares, the F statistic is the ratio of the row's mean square to the residual mean square.

[9] A more comprehensive taxonomy is Type I, II, III, and IV sums of squares, but never mind...

One other statistic that is often present at the bottom of an Anova table is the so-called *coefficient of determination* R^2 (pronounced "ar squared"). The coefficient of determination is

$$R^2 \;:=\; \frac{SS_{reg}}{SS_{tot}} \;=\; 1 - \frac{SS_{res}}{SS_{tot}}.$$

It's meant to encode the "fraction of variance explained by the model". A related statistic is the *fraction of variance unexplained*

$$\mathrm{FVU} \;:=\; 1 - R^2 \;=\; \frac{SS_{res}}{SS_{tot}}.$$

Another way to measure the fraction of variance explained is to compare the mean squares instead of the sums of squares. This gives the *adjusted* R^2,

$$R^2_{adj} := 1 - (1 - R^2)\frac{n-1}{n-p-1} \;=\; 1 - \frac{SS_{res}/df_{res}}{SS_{tot}/df_{tot}}.$$

3 Odds and ends

3.1 An inequality involving the correlation coefficient

In an analysis of variance between groups, we naturally expect the in-group variance to be less than the total variance. If the correlation between the group index and the response is close to ± 1 then the in-group variance should be significantly less than the total variance.[10] The following theorem formalizes this idea, and generalizes it to arbitrary bivariate distributions.

Theorem. *Let* X *and* Y *be real-valued random variables with finite positive variance, and* ρ *their correlation:*

$$\rho \equiv \rho(X, Y) \;:=\; \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

Then

$$\mathbf{E}(\mathrm{Var}(Y \mid X)) \;\leqslant\; (1 - \rho^2)\,\mathrm{Var}(Y).$$

Remark. In the case of group-means Anova, X is a discrete variable that specifies the group, and Y is the response. The reason we only have an upper bound is that it's possible that X predicts Y very well but nonlinearly, so we might have $\rho \approx 0$ but still $\mathbf{E}(\mathrm{Var}(Y \mid X)) = 0$.

Proof. We will use the covariance-variance form of the Cauchy–Schwartz Inequality,

$$\begin{aligned} \big|\langle U - \mathbf{E}U,\; V - \mathbf{E}V\rangle\big|^2 \;&\leqslant\; \|U - \mathbf{E}U\|^2 \|V - \mathbf{E}V\|^2 \\ \mathrm{Cov}(U, V)^2 \;&\leqslant\; \mathrm{Var}(U)\,\mathrm{Var}(V). \end{aligned}$$

With $U = \mathbf{E}(Y \mid X)$ and $V = X$, we get

$$\begin{aligned} \mathrm{Var}(\mathbf{E}(Y \mid X))\,\mathrm{Var}(X) \;&\geqslant\; \mathrm{Cov}(\mathbf{E}(Y \mid X),\, X)^2 \\ &=\; \big(\mathbf{E}(\mathbf{E}(Y \mid X)X) - \mathbf{E}(\mathbf{E}(Y \mid X))\mathbf{E}(X)\big)^2 \quad \text{by definition of covariance} \\ &=\; \big(\mathbf{E}(\mathbf{E}(XY \mid X)) - \mathbf{E}(Y)\mathbf{E}(X)\big)^2 \quad \text{by pull-out property and tower rule} \\ &=\; \big(\mathbf{E}(XY) - \mathbf{E}(Y)\mathbf{E}(X)\big)^2 \quad \text{by tower rule} \\ &=\; \big(\mathrm{Cov}(X, Y)\big)^2 \\ &=\; \rho^2\,\mathrm{Var}(X)\,\mathrm{Var}(Y) \quad \text{by definition of } \rho. \end{aligned}$$

[10] There are a few pretty graphs of this phenomenon at `https://en.wikipedia.org/wiki/Analysis_of_variance#Example`.

Dividing through by Var(X):

$$\begin{aligned}
\mathrm{Var}(\mathbf{E}(Y \mid X)) &\geqslant \rho^2\,\mathrm{Var}(Y) \\
\mathrm{Var}(Y) - \mathbf{E}(\mathrm{Var}(Y \mid X)) &\geqslant \rho^2\,\mathrm{Var}(Y) \quad \text{by law of total variance for } Y \\
\mathbf{E}(\mathrm{Var}(Y \mid X)) &\leqslant (1-\rho^2)\,\mathrm{Var}(Y).
\end{aligned}$$

□

4 Should you be using Anova?

The most accessible statistical methods are also the most misused, and Anova is no exception. The regression model used by Anova relies on some very strong assumptions that almost never hold in practice:

Assumption 1: A linear parametric model is appropriate for the system being studied. In particular, the predictors must directly cause the response, i.e., the causal network must look like this:

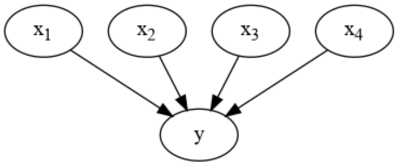


Figure 6: A flat causal network

And it's not enough for the predictors $x_1, x_2, \ldots, x_n$ to be under control merely for the duration of the experiment: It's necessary that the actual real-world system you're studying behaves this way. This is true in industrial settings, where the predictors represent inputs to a production process. And this is the classical application of Anova: Determining crop yield as a function of various growing conditions. But in natural systems, causal networks tend to be much more messy (when they exist at all.)

For further discussion along these lines, see the book by Pearl and Mackenzie in the references below.

Assumption 2: The predictors are determined exactly, with no measurement error or any other kind of error. Error exists only in the response variable y (see section 1.2 for the functional form of the model).

Assumption 3: The errors in the response are independent and identically distributed according to a normal distribution with mean 0.

Proponents claim that Anova is robust against some violations of these assumptions (esp. normality of errors), but I have yet to see a convincing argument.

Remember also that there are many other tools available. Consider your goal:

- Are you trying develop a conceptual model to understand some real-world system? Then you should do more theoretical work to come up with a realistic model instead of shoehorning everything into a linear setting. For example, you could create a causal Bayesian network, or another kind of probabilistic graphical model.
- Are you trying to create an empirical predictive model? Then (depending on what you plan to do with it later) you might not care how simple your model is. Don't use polynomial regression unless you have good reason to believe that a polynomial model is suitable. Better to fit a spline or a Gaussian process,[11] or use some other nonparametric model.

[11] For a guide to fitting Gaussian processes, see `https://yugeten.github.io/posts/2019/09/GP/`.

Anova was developed by Ronald Fisher the early 1900s. Back then, it would have been computationally prohibitive to fit a nonlinear model. Today we can do much better. If your goal is to build a predictive model, look into machine learning. For more on this topic, see

  - Breiman, Leo (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, Vol. 16, No. 3, 199–231.

- If you're setting up a controlled experiment, Anova might be acceptable, but make sure that the assumptions above are satisfied.

Despite its drawbacks, Anova remains very popular, so it's worth taking the time to learn how it works. At the very least, you'll learn to better appreciate its limitations, so you'll be able to identify where it's misused.

As for null significance hypothesis testing, enough has been written on the topic. Let me just point to two excellent papers:

- Cohen, J. (1994). The earth is round (p $<$ .05). *American Psychologist*, 49(12), 997–1003.
- Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert & Jennifer L. Tackett. (2019). Abandon Statistical Significance. *The American Statistician*, 73:sup1, 235–245.

Disclaimer: I'm not an expert, so don't take any of my opinions here too seriously. This is a work in progress! According to statistician Andrew Gelman, Anova is "more important than ever."[12]

5 Acknowledgments

I'd like to thank 8.5tails for giving feedback on the writing. Thank you also to Xuchen for helping with the proof in section 3.1.

6 References

For further reading, see

- Saville, David and Wood, Graham. *Statistical Methods: The Geometric Approach*. Springer, 1991.

The geometric treatment in this note grew (mostly) out of Chapter 4 of Saville and Wood's book. The book is written at a very elementary level, but its coverage is more broad: Seven chapters are devoted to the design of experiments.

For an in-depth critique of traditional statistical modeling—including Anova—see

- Pearl, Judea and Mackenzie, Dana. *The Book of Why*. Basic Books, 2018.

[12] https://statmodeling.stat.columbia.edu/2019/03/28/understanding-how-anova-relates-to-regression/