

1 Degrees of Freedom

Copyright © 2004 Paul Mathews

Rev. 1.1 20Dec04

Mathews Malnar and Bailey, Inc.

217 Third Street, Fairport Harbor, OH 44077

Phone: 440-350-0911

Fax: 440-350-7210

E-mail: pmathews@apk.net, paul@mmbstatistical.com

Web: www.mmbstatistical.com

1.1 Introduction

- Information or data comes in discrete units called *degrees of freedom*.
- In any data set there are as many degrees of freedom as there are observations.
- The purpose of collecting data from a process is to learn about that process, but all data sets contain both signal and noise.
- The purpose of data analysis is to separate the signal from the noise.
- The signal extracted from the data is called a *model* which should capture the systematic or assignable cause variation in the data set. Whatever variation is left over is attributed to noise or common cause variation.
- The relationship between the data, model, and noise can be written:

$$Data \rightarrow Model + ErrorStatement \quad ((1))$$

or for individual observations:

$$y_i = \hat{y}_i + \epsilon_i$$

where the error statement specifies the shape of the ϵ_i distribution and its standard deviation.

- If the model is accurate, then it will approximate the population from which the sample was drawn.
- The complexity of terms that can be included in the model is limited by the magnitude of the noise.
- Just as the variation in the data gets partitioned into two sources, the model and the error, the degrees of freedom get partitioned into those categories, too.

- When we extract a model from data, the model consumes some information or degrees of freedom. Whatever information or degrees of freedom are left over after the model has been constructed are available to form the error statement.
- The standard deviation of the residuals ϵ_i , usually called the standard error of the model, is always given by:

$$s_\epsilon = \sqrt{\frac{\sum \epsilon_i^2}{df_\epsilon}}$$

- Information or data accounting is done by keeping track of the available degrees of freedom and how they are partitioned between the model and the error statement.

1.2 One Sample

Suppose that a sample of size n is drawn from a single stable population and that the measured response is quantitative. The first statistic extracted from the sample data is the mean:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

In the case of the sample mean, we divide $\sum y_i$ by n because there are n pieces of information or degrees of freedom available to determine the sample mean.

After the sample mean has been determined, we can calculate the discrepancies between each y_i and \bar{y} :

$$y_i = \bar{y} + \epsilon_i$$

where the ϵ_i are called the residuals. Notice that this equation has the same form as Equation 1. We must form the error statement from the residuals. We do this by: 1) constructing a histogram of the residuals to determine their shape and 2) calculating their standard deviation:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n \epsilon_i^2}$$

where

$$\epsilon_i = y_i - \bar{y}$$

Although there are n ϵ_i s, we must divide by $n - 1$ because only the first $n - 1$ of the n ϵ_i s are free to vary. To understand this, suppose that a sample of size $n = 10$ has a sample mean of $\bar{y} = 3$ and that the first nine of the y_i are all

$y_i = 3$. Since there are ten degrees of freedom in the data set, by specifying the first nine observations and the sample mean, the tenth observation is not free to vary - in this case it is forced to take on the value $y_{10} = 3$.

These arguments show that whereas the sample mean is calculated from all n degrees of freedom, it consumes one degree of freedom leaving $n - 1$ degrees of freedom to estimate the noise. That is:

$$\begin{aligned} df_{model} &= 1 \\ df_{\epsilon} &= n - 1 \\ df_{total} &= n \end{aligned}$$

1.3 Confidence Interval for μ

A classic application of the sample mean and standard deviation is the calculation of the confidence interval for the true but unknown population mean. The confidence interval is given by:

$$P\left(\bar{y} - t_{\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{y} + t_{\alpha/2} \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

where the t distribution's degrees of freedom are determined by the degrees of freedom available to determine s , i.e. $df_{\epsilon} = n - 1$. (Sometimes the t values are written $t_{\alpha/2, df_{\epsilon}}$ to explicitly indicate the degrees of freedom associated with the t distribution.)

1.4 Two Independent Samples

When two independent samples are drawn from their respective populations for the purpose of comparing their means, and it is likely that the two populations have equal standard deviations, then the sample means are:

$$\bar{y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} y_{1i}$$

and

$$\bar{y}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} y_{2j}$$

Since the standard deviations of the two populations are expected to be the same, we can combine the information about the noise from the two independent samples into one improved estimate of the common population standard deviation:

$$s_{\epsilon} = \sqrt{\frac{\sum_{i=1}^{n_1} \epsilon_{1i}^2 + \sum_{j=1}^{n_2} \epsilon_{2j}^2}{n_1 + n_2 - 2}}$$

where

$$\begin{aligned}\epsilon_{1i} &= y_{1i} - \bar{y}_1 \\ \epsilon_{2j} &= y_{2j} - \bar{y}_2\end{aligned}$$

and we must divide by $df_\epsilon = n_1 + n_2 - 2$ to account for the two degrees of freedom consumed by the prior calculations of the statistics \bar{y}_1 and \bar{y}_2 . That is:

$$\begin{aligned}df_{model} &= 2 \\ df_\epsilon &= n_1 + n_2 - 2 \\ df_{total} &= n_1 + n_2\end{aligned}$$

1.5 Confidence Interval for $\Delta\mu$

The confidence interval for the difference between the two population means would be:

$$P\left(\Delta\bar{y} - t_{\alpha/2}s_\epsilon\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \Delta\mu < \Delta\bar{y} + t_{\alpha/2}s_\epsilon\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = 1 - \alpha$$

where $\Delta\bar{y} = \bar{y}_1 - \bar{y}_2$ and the t distribution's degrees of freedom are $df_\epsilon = n_1 + n_2 - 2$.

1.5.1 Two-Sample t Test with Unequal Variances

When the two populations have variances that are different from each other (heteroscedastic), the noise information from the two samples cannot be combined as they were in the homoscedastic case. There are two different but related methods of determining the degrees of freedom for the heteroscedastic case:

- Hsu's method: $df_\epsilon = \min(n_1 - 1, n_2 - 1)$
- Welch's method: $df_\epsilon = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$

Notice that:

1. Welch's method is an interpolation of the other two methods:

$$\min(n_1 - 1, n_2 - 1) < df_\epsilon (Welch) < n_1 + n_2 - 2$$

2. Hsu's method is the most conservative but least sensitive to small differences between the treatment means.

1.5.2 Degrees of Freedom in ANOVA and Regression

In the model-building methods ANOVA and regression, the total variation in the measured response is always considered relative to the grand mean of the response. The necessary calculation of the grand mean as the first step in these analyses is accounted for by setting the total degrees of freedom associated with a data set equal to the number of observations minus one. In the subsequent models, although a model might contain k coefficients, only the first $k - 1$ are free to vary because the last coefficient must be consistent with the grand mean.

1.6 One-Way ANOVA

The purpose of one-way ANOVA is to determine if there is evidence of differences between the populations means of $k \geq 3$ treatments. That is, we want to test the hypotheses:

$$\begin{aligned} H_0 &: \mu_i = \mu_j \text{ for all possible } i, j \text{ pairs} \\ H_A &: \mu_i \neq \mu_j \text{ for one or more } i, j \text{ pairs} \end{aligned}$$

The model will consist of the k treatment means. If there are n observations in each of the k treatments, then there are nk observations in the data set. If the treatment errors are homoscedastic (i.e. have the same constant variance), which is one of the usual assumptions required of ANOVA, then as in the case of two-sample means, the common treatment variance can be estimated by combining information about the noise from all of the treatments so there will be $df_\epsilon = nk - k$ error degrees of freedom.

The total variation in the data set is measured relative to the grand mean of the data set by:

$$SS_{total} = \sum_{j=1}^k \sum_{i=1}^n (y_{ij} - \bar{y})^2$$

where the calculation of \bar{y} consumes one degree of freedom. It can be shown that SS_{total} can be partitioned into two parts, one associated with differences between treatments:

$$SS_{treatments} = n \sum_{j=1}^k (\bar{y}_j - \bar{y})^2$$

and another associated with variation within treatments:

$$SS_\epsilon = \sum_{j=1}^k \sum_{i=1}^n (y_{ij} - \bar{y}_j)^2$$

where

$$SS_{total} = SS_{treatments} + SS_\epsilon$$

The corresponding degrees of freedom are:

$$\begin{aligned} df_{total} &= nk - 1 \\ df_{model} &= k - 1 \\ df_{\epsilon} &= (nk - 1) - (k - 1) = k(n - 1) \end{aligned}$$

where

$$df_{total} = df_{model} + df_{\epsilon}$$

Although there are k treatment means in the model, only the first $k - 1$ of them are free to vary because the k th one must be consistent with \bar{y} .

The ANOVA calculations are summarized in an ANOVA table:

Source	df	SS	MS	F
Treatments	$k - 1$	$SS_{treatments}$	$\frac{SS_{treatments}}{df_{treatments}}$	$F = \frac{MS_{treatments}}{MS_{\epsilon}}$
Error	$k(n - 1)$	SS_{ϵ}	$\frac{SS_{\epsilon}}{df_{\epsilon}}$	
Total	$nk - 1$	SS_{total}		

where the F distribution has $df_{treatments} = k - 1$ numerator degrees of freedom and $df_{\epsilon} = k(n - 1)$ denominator degrees of freedom. If there are no differences between any of the treatment means, then we expect that $F \approx 1$, but if there are differences then we expect that $F \gg 1$.

Since the F distribution's shape depends on $df_{treatments}$ and df_{ϵ} , it's not sufficient to just observe if F is large compared to one or not. To be more rigorous, we usually calculate the p value corresponding to the F statistic from the area under the F distribution to the right of the ANOVA's F statistic:

$$p = \int_F^{\infty} pdf(F; df_{treatments}, df_{\epsilon}) dF$$

where $pdf()$ indicates the F distribution *probability density function*. When p is small (typically $p < 0.05$) then we reject H_0 and conclude that there are statistically significant differences between the treatment means.

MINITAB and most other statistical software packages automatically calculate and report the p value for the ANOVA.

1.7 One-Way ANOVA: Example

A one-way classification experiment was performed to test five different treatments for possible differences between their population means. Samples of size $n = 8$ were drawn from each treatment. Determine how the ANOVA degrees of freedom are partitioned and complete the ANOVA calculations:

Source	df	SS	MS	F	p
Treatment		360			
Error		700			
Total					

Solution:

Source	df	SS	MS	F	p
Treatment	4	360	90	4.5	0.005
Error	35	700	20		
Total	39	1060			

where $F_{0.995,4,35} = 4.5$ and the model can be summarized with the following statistics:

$$s_\epsilon = \sqrt{MS_\epsilon} = \sqrt{20} = 4.47$$

$$r^2 = 1 - \frac{SS_\epsilon}{SS_{total}} = 1 - \frac{700}{1060} = 0.340$$

$$r_{adj}^2 = 1 - \frac{df_{total}}{df_\epsilon} \frac{SS_\epsilon}{SS_{total}} = 1 - \frac{39}{35} \frac{700}{1060} = 0.264$$

1.8 Linear Regression

When we fit a line to the data in a scatter plot with n points, the line will have the form:

$$(y_i - \bar{y}) = b_1 (x_i - \bar{x}) + \epsilon_i$$

where the line passes through the point (\bar{x}, \bar{y}) and has slope $\frac{\Delta y}{\Delta x} = b_1$. The point (\bar{x}, \bar{y}) and the slope, which comprise the model, each consume one degree of freedom so there will be $df_\epsilon = n - 2$ degrees of freedom to estimate the variation in the observations about the fitted line.

The equation for the line is usually written in an equivalent but alternative form:

$$\begin{aligned} (y_i - \bar{y}) &= b_1 (x_i - \bar{x}) + \epsilon_i \\ y_i &= b_1 (x_i - \bar{x}) + \bar{y} + \epsilon_i \\ &= b_1 x_i + (-b_1 \bar{x} + \bar{y}) + \epsilon_i \\ &= (b_1 x_i + b_0) + \epsilon_i \\ &= \hat{y}_i + \epsilon_i \end{aligned}$$

where $b_0 = \bar{y} - b_1 \bar{x}$ is the y axis intercept, i.e. another point on the line at $(x, y) = (0, b_0)$. (Notice that this equation has the same structure as Equation 1!)

As in ANOVA, a first degree of freedom is consumed by the point (\bar{x}, \bar{y}) , so although the model consumes two degrees of freedom (b_0 and b_1), only one of them is free to vary so:

$$\begin{aligned} df_{total} &= n - 1 \\ df_{model} &= 1 \\ df_\epsilon &= n - 2 \end{aligned}$$

and the standard deviation of the residuals for the error statement is given by:

$$s_\epsilon = \sqrt{\frac{1}{n-2} \sum_{i=1}^n \epsilon_i^2}$$

where $\epsilon_i = y_i - \hat{y}_i$

1.9 Quadratic Regression

If a quadratic function of the form:

$$y_i = b_0 + b_1x_i + b_2x_i^2 + \epsilon_i$$

is fitted to n points in a scatter plot, then

$$\begin{aligned}df_{total} &= n - 1 \\df_{model} &= 2 \\df_{\epsilon} &= n - 3\end{aligned}$$

The standard error will be:

$$s_{\epsilon} = \sqrt{\frac{1}{n-3} \sum_{i=1}^n \epsilon_i^2}$$

where

$$\epsilon_i = y_i - \hat{y}_i$$

1.10 Polynomial Regression

If a k th order polynomial of the form:

$$y_i = b_0 + b_1x_i + b_2x_i^2 + \cdots + b_kx_i^k + \epsilon_i$$

is fitted to n points in a scatter plot, then

$$\begin{aligned}df_{total} &= n - 1 \\df_{model} &= k \\df_{\epsilon} &= n - k - 1\end{aligned}$$

The standard error will be:

$$s_{\epsilon} = \sqrt{\frac{1}{df_{\epsilon}} \sum_{i=1}^n \epsilon_i^2}$$

where

$$\epsilon_i = y_i - \hat{y}_i$$

1.11 Multi-Way ANOVA

When there are two or more variables in a crossed multi-way classification design:

- The total degrees of freedom is always the number of observations minus one.
- The degrees of freedom for a main effect is given by its number of levels minus one.
- The degrees of freedom for an interaction is given by the product of the degrees of freedom of the relevant main effects.
- The error degrees of freedom will always be whatever's left over after subtracting the model degrees of freedom from total degrees of freedom.

1.12 Multi-Way ANOVA: Example

Determine how the degrees of freedom are partitioned in the $3 \times 2 \times 5$ full factorial experiment with three replicates run in blocks and complete the ANOVA calculations.

Source	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Blocks		26			
A		80			
B		8			
C		2000			
AB		30			
AC		400			
BC		36			
ABC		80			
Error		580			
Total					

Solution:

Source	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Blocks	2	26	13	1.3	0.280
A	2	80	40	4.0	0.024
B	1	8	8	0.8	0.375
C	4	2000	500	50	0.000
AB	2	30	15	1.5	0.232
AC	8	400	50	5	0.000
BC	4	36	9	0.9	0.470
ABC	8	80	10	1.0	0.446
Error	58	580	10		
Total	89	3240			

1.13 Multi-Way ANOVA: Example

Simplify the model from the preceding example by eliminating terms that are not statistically significant and recalculate the ANOVA table.

Solution:

Source	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Blocks		26			
A		80			
B		8			
C		2000			
AB		30			
AC		400			
BC		36			
ABC		36			
Error		80			
Total					

1.14 General Linear Model

When an experiment contains one or more qualitative variables and one or more quantitative variables, we need to do ANOVA on the qualitative variables and regression on the quantitative ones. Interaction terms between qualitative and quantitative terms account for possible slope differences between the levels of the qualitative variables. Calculate degrees of freedom for models of this form using the usual rules for ANOVA, regression, and interactions.

1.15 GLM: Example

An experiment was performed to determine how temperature affects the growth of three different strains of tomatos. Three samples of each strain were evaluated at five different levels of temperature. Determine how the degrees of freedom are partitioned if the model must account for possible slope differences between the strains and include a generic curvature term in the model to check for lack of linear fit.

Solution: The model will have the form:

$$\begin{aligned} y = & b_0 + b_{01}(\text{Strain} = 1) + b_{02}(\text{Strain} = 2) + b_{03}(\text{Strain} = 3) \\ & + \text{Temp} [b_2 + b_{21}(\text{Strain} = 1) + b_{22}(\text{Strain} = 2) + b_{23}(\text{Strain} = 3)] \\ & + b_4 \text{Temp}^2 \end{aligned}$$

where $b_{03} = -(b_{01} + b_{02})$ and $b_{23} = -(b_{21} + b_{22})$. Note that the b_{0i} are bias corrections for the different strains and the b_{2i} are slope corrections for the

different strains.

Source	<i>df</i>
Strain	2
Temp	1
Strain*Temp	2
Temp*Temp	1
Error	38
Total	44

1.16 Designed Experiments

Two-level factorial, two-level fractional factorial, and response surface designs are usually analyzed by multiple regression. Use the usual regression and ANOVA rules to determine how the degrees of freedom are partitioned between the model and error sources.

1.17 Designed Experiments: Example

A Box-Behnken three variable experiment with three replicates in blocks was performed. The model fitted to the data includes main effects, two-factor interactions, quadratic terms, and terms for blocks. Determine how the degrees of freedom are allocated in the experiment.

Solution: The model has the form:

$$\begin{aligned}
 y = & b_0 + b_{01}(\text{Block} = 1) + b_{02}(\text{Block} = 2) + b_{03}(\text{Block} = 3) \\
 & + b_1x_1 + b_2x_2 + b_3x_3 \\
 & + b_{12}x_{12} + b_{13}x_{13} + b_{23}x_{23} \\
 & + b_{11}x_1^2 + b_{22}x_2^2 + b_{33}x_3^2 + \epsilon
 \end{aligned}$$

where $b_{03} = -(b_{01} + b_{02})$.

Source	<i>df</i>
Blocks	2
x_1	1
x_2	1
x_3	1
x_{12}	1
x_{13}	1
x_{23}	1
x_1^2	1
x_2^2	1
x_3^2	1
Error	33
Total	44

1.18 Nested Designs

In a nested experiment design, the levels of one variable are unique within the levels of another. If variable A has a levels and each level of A has b unique levels of a second variable B within it, then the ANOVA degrees of freedom are:

Source	df
A	$a - 1$
B(A)	$a(b - 1)$
Error	$ab(n - 1)$
Total	$abn - 1$

The justification for the value of $df_{B(A)}$ is that there are ab levels of B in the study, but a mean must be calculated for each level of A, so only $b - 1$ of the b means within a level of A are free to vary.

1.19 Nested Design: Example

An experiment was performed to compare the strengths of braided rope manufactured on three different machines. It was impossible to run material from the same lot on each machine, so random samples of four different material lots were used on each machine. Determine how the degrees of freedom are partitioned if two samples of each finished rope were tested.

Solution: With the assignments $A : Machine$ and $B : Material$ we have $a = 3$, $b = 4$, and $n = 2$:

Source	df
Machine	$a - 1 = 2$
Material(Machine)	$a(b - 1) = 9$
Error	$ab(n - 1) = 12$
Total	$abn - 1 = 23$

Although there are $ab = 12$ materials involved in the study, there are $a = 3$ machine means so only $a(b - 1) = 12 - 3 = 9$ of the material means are free to vary.

1.20 References

- Montgomery, Design and Analysis of Experiments
- Box, Hunter, and Hunter, Statistics for Experimenters
- Hicks, Fundamental Concepts in the Design of Experiments
- Mathews, Design of Experiments with MINITAB