

Notes in Statistics 2

Rasmus Jacobsen
7th Semester ELITE in Wireless Communication

January 12, 2009

1 Common Notation

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$S_{xY} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

$$S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - n\bar{Y}^2$$

$$X_{i\bullet} = \frac{1}{n} \sum_{j=1}^n X_{ij}$$

$$X_{\bullet j} = \frac{1}{m} \sum_{i=1}^m X_{ij}$$

$$X_{\bullet\bullet} = \frac{1}{m} \sum_{i=1}^m X_{i\bullet} = \frac{1}{nm} \sum_{i=1}^m \sum_{j=1}^n X_{ij}$$

2 Regression

General linear regression is defined as

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + e,$$

where e is a zero mean random variable. A simple linear regression is ($r = 1$)

$$Y = \alpha + \beta x + e$$

2.1 Least Squares Estimators of the Regression Parameters

(See mm1 Ex. 9.4) Let A and B be estimates of α and β , respectively, then an estimate of Y_i for the input variable x_i , $i = 1, \dots, n$ is

$$\hat{Y}_i = A + Bx_i.$$

The sum of squared differences is

$$SS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - A - Bx_i)^2.$$

The parameters for the straight estimated regression line $Y = A + Bx$ are

$$B = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{S_{xY}}{S_{xx}}$$
$$A = \bar{Y} - B\bar{x}.$$

Remark: If the linear model is $Y = Bx$, then to minimize the mean square error, we say

$$SS = \sum_{i=1}^n (Y_i - Bx_i)^2$$
$$\frac{\partial SS}{\partial B} = -2 \sum_{i=1}^n x_i (Y_i - Bx_i)$$

Setting equal to zero yields

$$\sum_{i=1}^n x_i Y_i = B \sum_{i=1}^n x_i^2$$
$$B = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

2.2 Distribution of the Estimators

(See mm1 Ex. 9.9) It is assumed that the error e is normal distributed as

$$e \sim \mathcal{N}(0, \sigma^2).$$

Furthermore, it is assumed that the samples Y_i , $i = 1, \dots, n$, are independent. This gives

$$Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2).$$

It can be shown that

$$E[B] = \beta, \quad \text{Var}(B) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2},$$

and

$$E[A] = \alpha, \quad \text{Var}(A) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n(\sum_{i=1}^n x_i^2 - n\bar{x}^2)}.$$

The difference $Y_i - \hat{Y}$ is called the residuals. The sum of squares of residuals is

$$SS_R = \sum_{i=1}^n (Y_i - \hat{Y})^2 = \sum_{i=1}^n (Y_i - A - Bx_i)^2 = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}.$$

SS_R can be used to determine the variance of an individual response σ^2 as

$$\frac{SS_R}{\sigma^2} \sim \chi_{n-2}^2, \quad E\left[\frac{SS_R}{n-2}\right] = \sigma^2. \quad (1)$$

We use a degree of freedom in each of the two arithmetic means \bar{x} and \bar{Y} . Therefore there are only $n - 2$ left.

It can be shown that

$$A \sim \mathcal{N}\left(\alpha, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{nS_{xx}}\right), \quad B \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{S_{xx}}\right).$$

2.3 Inferences Concerning β

(See mm1 Ex. 9.12) It is relevant to check whether $\beta = 0$, as if so, then this is equivalent to state that the input does not depend on the output.

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta \neq 0.$$

The test statistic is

$$TS = \sqrt{\frac{(n-2)S_{xx}}{SS_R}}|B|.$$

As this is a two-sided test, the p -value can be found as

$$\begin{aligned} p\text{-value} &= \Pr[|T_{n-2}| > TS] \\ &= 2 \Pr[T_{n-2} > TS] \\ &= 2(1 - \Pr[T_{n-2} < TS]) \\ &= 2(1 - \text{cdf}(TS, n-2)) \end{aligned}$$

If $p\text{-value} < a$, where a is some significance level, then reject H_0 . A $100(1-a)$ percent confidence interval for β is

$$\left(B - \sqrt{\frac{SS_R}{(n-2)S_{xx}}} t_{a/2, n-2}, B + \sqrt{\frac{SS_R}{(n-2)S_{xx}}} t_{a/2, n-2} \right),$$

$$\left(B - \sqrt{\frac{SS_R}{(n-2)S_{xx}}} \text{tin}v(1-a/2, n-2), B + \sqrt{\frac{SS_R}{(n-2)S_{xx}}} \text{tin}v(1-a/2, n-2) \right),$$

where $t_{a/2, n-2} = \text{tin}v(1-a/2, n-2)$. This means that we can be $100(1-a)$ percent confident that β lies in this interval.

2.3.1 Regression to the Mean

We want to test if $0 < \beta < 1$. The hypothesis is

$$H_0 : \beta \geq 1 \quad \text{versus} \quad H_1 : \beta < 1,$$

which is equivalent to

$$H_0 : \beta = 1 \quad \text{versus} \quad H_1 : \beta < 1.$$

The p -value is found for this one-sided test as

$$\begin{aligned} p\text{-value} &= \Pr[T_{n-2} > TS] \\ &= 1 - \Pr[T_{n-2} < TS] \end{aligned}$$

It can be that by mistake, the p -value indicates that there is regression to the mean, where this is due to some outside influence. This is referred to as “regression fallacy”.

2.3.2 Inferences Concerning α

The confidence interval is

$$A \pm \sqrt{\frac{\sum_{i=1}^n x_i^2 SS_R}{n(n-2)S_{xx}}} t_{a/2, n-2}$$

2.3.3 Inferences Concerning the Mean Response $\alpha + \beta x_0$

(See mm1 Ex. 9.21) The input is x_0 , and an estimate \hat{Y} is found as

$$\hat{Y} = A + Bx_0.$$

If x_0 is inside the interval used to find A and B , then the confidence interval of the estimator of $\alpha + \beta x_0$ is

$$A + Bx_0 \pm \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \sqrt{\frac{SS_R}{n-2}} t_{a/2, n-2}.$$

This means, that with $100(1-a)$ percent confidence, $\alpha + \beta x_0$ lies in this interval.

2.3.4 Prediction Interval

If, instead, x_0 is outside, then, with $100(1-a)$ percent confidence, the response Y at the input level x_0 will be contained in the interval

$$A + Bx_0 \pm \sqrt{\frac{n+1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \sqrt{\frac{SS_R}{n-2}} t_{a/2, n-2}.$$

2.4 The Coefficient of Determination and the Sample Correlation Coefficient

- S_{YY} is the variation in the set Y_1, \dots, Y_n .
- SS_R is the error variance.

Therefore, $S_{YY} - SS_R$ represents the variation in the input values. The proportion of variance explained by the different input values is

$$R^2 = \frac{S_{YY} - SS_R}{S_{YY}},$$

called the coefficient of determination.

- If R^2 is close to 0, this indicates that a little variation is explained by the different input values; much noise. The model does not fit the data.
- If R^2 is close to 1, this indicates that most of the variation is explained by the different input values; no noise. The model is a good fit.

$R^2 100$ is the percentage of “explained” variance, where the remaining is due to noise.

The correlation coefficient is

$$r = \frac{S_{xY}}{\sqrt{S_{xx}S_{YY}}}, \quad |r| = \sqrt{R^2}$$

2.5 Analysis of Residuals: Assessing the Model

The residual, $(Y_i - \hat{Y}, i = 1, \dots, n)$, is normalized by dividing each residual with the estimated standard deviation. The estimated variance is found in (1), and we have the “standardized residuals”

$$\frac{Y_i - \hat{Y}}{\sigma} = \frac{Y_i - (A + Bx_i)}{\sqrt{\frac{SS_R}{n-2}}}, \quad i = 1, \dots, n$$

- If these values, and the scatter diagram, does not follow any pattern, then the model fits well.
- If the residuals appear to first decrease and then increase, as the input level increases, this often means that a higher (non-linear) order is needed to describe the relationship.
- If the residuals appear to increase as the level increases, this indicates that the variance of the response is not constant.

2.6 Transforming to Linearity

(See mm2 Ex. 9.35) If a model is known, this can be used to transform values into a linear model. An example is the model

$$W(t) \approx ce^{-dt}$$

By transforming the model, then

$$\ln W(t) \approx \ln c - dt,$$

where

$$Y = \ln W(t)$$

$$\alpha = \ln c$$

$$\beta = -d$$

With the estimates of α and β , the relationship can be predicted as

$$\ln W(t) \approx A + Bt$$

$$W(t) \approx e^A e^{Bt}$$

$$W(t) \approx e^{A+Bt}.$$

2.7 Weighted Least Squares

(See mm2 Ex. 9.40) The following equations minimize the weighted sum of squares:

$$\begin{aligned}\sum_{i=1}^n w_i Y_i &= A \sum_{i=1}^n w_i + B \sum_{i=1}^n w_i x_i \\ \sum_{i=1}^n w_i x_i Y_i &= A \sum_{i=1}^n w_i x_i + B \sum_{i=1}^n w_i x_i^2\end{aligned}$$

The equations can be easily solved with respect to A and B .

If e.g. the values for small x_i s should be weighted greater than greater x_i s, then we could have $w_i = 1/x_i$. If the weight should increase as x_i s increase, then we could have $w_i = x_i$.

In matrix notation, let

$$\begin{aligned}\mathbf{V} &= \left[\sum_{i=1}^n w_i Y_i, \sum_{i=1}^n w_i x_i Y_i \right]^T \\ \mathbf{U} &= \begin{bmatrix} \sum_{i=1}^n w_i & \sum_{i=1}^n w_i x_i \\ \sum_{i=1}^n w_i x_i & \sum_{i=1}^n w_i x_i^2 \end{bmatrix} \\ \mathbf{B} &= [A, B]^T\end{aligned}$$

then

$$\mathbf{B} = \mathbf{U} \setminus \mathbf{V}$$

2.8 Polynomial Regression

Polynomial regression is defined as

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r + e.$$

The estimate of β_i is B_i , and by defining

$$\begin{aligned}\mathbf{X} &= [X_1, \dots, X_n]^T \\ \mathbf{Y} &= [Y_1, \dots, Y_n]^T \\ \mathbf{B} &= [B_r, \dots, B_0]^T \quad (\text{note the order})\end{aligned}$$

then the coefficients can be found as

$$\mathbf{B} = \text{polyfit}(\mathbf{X}, \mathbf{Y}, r)$$

2.9 Multiple Linear Regression

(See mm2 Ex. 9.54) Multiple linear regression is defined as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e,$$

where the output depends on k input variables. There are n observations, and by defining

$$\begin{aligned} \mathbf{Y} &= [Y_1, \dots, Y_n]^T \\ \mathbf{X} &= \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{nk} & \cdots & x_{nk} \end{bmatrix} \\ \mathbf{B} &= [B_0, \dots, B_k]^T \end{aligned}$$

then the least square condition and the estimate of the coefficients are

$$\mathbf{X}^T \mathbf{X} \mathbf{B} = \mathbf{X}^T \mathbf{Y}, \quad \mathbf{B} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

The estimate of the variance can be found using SS_R ,

$$\begin{aligned} SS_R &= \sum_{i=1}^n (Y_i - (B_0 + B_1 x_{i1} + \cdots + B_k x_{ik}))^2 \\ &= \mathbf{Y}^T \mathbf{Y} - \mathbf{B}^T \mathbf{X}^T \mathbf{Y}, \end{aligned}$$

as

$$\frac{SS_R}{\sigma^2} \sim \chi_{n-(k+1)}^2, \quad E \left[\frac{SS_R}{n - (k + 1)} \right] = \sigma^2.$$

The degrees of freedom is $n - (k + 1)$ as we use one for each mean \bar{Y} , $\bar{x}_{i1}, \dots, \bar{x}_{ik}$.

3 Analysis of Variance (ANOVA)

3.1 One-Way Analysis of Variance

(See mm3 Ex. 10.5 and 10.12) Groups are formed of size m in where, for each group n (assumed) independent normal random variables exist with

$$X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2), \quad i = 1, \dots, m; j = 1, \dots, n.$$

It is essential that the groups are formed in a way so it is extremely unlikely that one of the groups are inherently superior. Now, the hypothesis to test is

$$H_0 : \mu_1 = \cdots = \mu_m \quad \text{versus} \quad H_1 : \exists i_1, i_2 \text{ s.t. } \mu_{i_1} \neq \mu_{i_2}$$

The approach to test the hypothesis is based on deriving two estimators for the common variance σ^2 .

1. The first estimator is a valid estimator for σ^2 whether H_0 is true or false.
2. The second estimator provides an estimate of σ^2 close to the one found using the first estimator only if H_0 is true. If H_0 is false, the second estimator's estimate tends to exceed that of the first estimator.

The First Variance Estimator Assume that

$$Z_{ij} = \frac{X_{ij} - \mu_i}{\sigma} \sim \mathcal{N}(0, 1),$$

then

$$\sum_{i=1}^m \sum_{j=1}^n Z_{ij}^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(X_{ij} - \mu_i)^2}{\sigma^2} \sim \chi_{nm}^2.$$

The resulting variable is

$$\sum_{i=1}^m \sum_{j=1}^n \frac{(X_{ij} - X_{i\bullet})^2}{\sigma^2} \sim \chi_{nm-m}^2,$$

as m degrees of freedom are lost, one for each sample mean. It follows that

$$\begin{aligned} SS_W &= \sum_{i=1}^m \sum_{j=1}^n (X_{ij} - X_{i\bullet})^2, \quad \frac{SS_W}{\sigma^2} \sim \chi_{nm-m}^2, \quad E \left[\frac{SS_W}{nm-m} \right] = \sigma^2. \\ &= (n-1) \sum_{i=1}^m S_i^2, \quad S_i^2 = \frac{1}{n-1} \sum_{j=1}^n (X_{ij} - X_{i\bullet})^2 \end{aligned}$$

where S_i^2 is the sample variance.

The Second Variance Estimator This second estimator will only be valid if H_0 is true. Therefore we say that the means are equal, this gives

$$X_{i\bullet} \sim \mathcal{N} \left(\mu, \frac{\sigma^2}{n} \right), \quad i = 1, \dots, m.$$

The sum of squares of the m variables is

$$\sum_{i=1}^m \left(\frac{X_{i\bullet} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \right)^2 = n \sum_{i=1}^m \frac{(X_{i\bullet} - \mu)^2}{\sigma^2} \sim \chi_m^2 \quad (2)$$

The resulting variable is

$$n \sum_{i=1}^m \frac{(X_{i\bullet} - X_{\bullet\bullet})^2}{\sigma^2} \sim \chi_{m-1}^2.$$

It follows that

$$SS_b = n \sum_{i=1}^m (X_{i\bullet} - X_{\bullet\bullet})^2, \quad \frac{SS_b}{\sigma^2} \sim \chi_{m-1}^2, \quad E \left[\frac{SS_b}{m-1} \right] = \sigma^2.$$

Comparison of the Estimators A ratio is found comparing the two estimates, and if this ratio is too large, H_0 is rejected. The test statistic is therefore

$$TS = \frac{\frac{SS_b}{m-1}}{\frac{SS_w}{nm-m}}.$$

The F-distribution is defined as

$$F_{n,m} = \frac{\frac{\chi_n^2}{n}}{\frac{\chi_m^2}{m}},$$

which gives

$$\begin{aligned} p\text{-value} &= \Pr[F_{m-1, nm-m} > TS] \\ &= 1 - \Pr[F_{m-1, nm-m} < TS] \\ &= 1 - f_{cdf}(TS, m-1, nm-m) \end{aligned}$$

Matlab Let

$$\mathbf{X} = \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{bmatrix},$$

then

$$\begin{aligned} [pvalue, dummy, st] &= anova1(\mathbf{X}') \\ multcompare(st) \end{aligned}$$

gives

Source	SS	df	MS	F	Prob > F
Columns	135.762	3	45.2542	7.47	0.0044
Error	72.66	12	6.055		
Total	208.423	15			

where in row

1. SS is SS_b , $df = m - 1$, $MS = \frac{SS_b}{m-1}$, $F = TS$, and $Prob > F = p\text{-value}$.
2. SS is SS_w , $df = nm - m$, $MS = \frac{SS_w}{nm-m}$.

3.1.1 Multiple Comparison of Sample Means

(See mm3 Ex. 10.12) If H_0 is rejected it is of interest to find the mean that differs from the others. For this, the T-method is used, which gives the joint confidence interval, that we, with probability $1 - \alpha$ have

$$X_{i_1\bullet} - X_{i_2\bullet} - W < \mu_{i_1} - \mu_{i_2} < X_{i_1\bullet} - X_{i_2\bullet} + W, \quad i_1 \neq i_2$$

where

$$W = \frac{1}{\sqrt{n}} C(m, nm - m, \alpha) \sqrt{\frac{SS_W}{nm - m}}.$$

The value $C(m, nm - m, \alpha)$ can be found in table A5. A scheme is setup comparing the means, and if the difference differs from zero, then an incorrect group is found. An example for $m = 3$ is

$$\begin{aligned} -0.4 &< \mu_1 - \mu_2 < 0.4 \\ 0.2 &< \mu_1 - \mu_3 < 1 \\ 0.2 &< \mu_2 - \mu_3 < 1 \end{aligned}$$

Here μ_3 differs significantly from that of μ_1 and μ_2 .

3.1.2 One-Way Analysis of Variance with Unequal Sample Sizes

(See mm3 Ex. 10.15) For this case, the sample sizes for each group differs, and we have n_1, \dots, n_m . We therefore find the first variance estimator to be

$$SS_W = \sum_{i=1}^m \sum_{j=1}^{n_i} (X_{ij} - X_{i\bullet})^2, \quad \frac{SS_W}{\sigma^2} \sim \chi_{\sum_{i=1}^m n_i - m}^2, \quad E \left[\frac{SS_W}{\sum_{i=1}^m n_i - m} \right] = \sigma^2,$$

and the second variance estimator to be

$$SS_b = \sum_{i=1}^m n_i (X_{i\bullet} - X_{\bullet\bullet})^2, \quad \frac{SS_b}{\sigma^2} \sim \chi_{m-1}^2, \quad E \left[\frac{SS_b}{m-1} \right] = \sigma^2,$$

where, in this case

$$X_{\bullet\bullet} = \frac{1}{m} \sum_{i=1}^m \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}.$$

We have

$$\begin{aligned} TS &= \frac{\frac{SS_b}{m-1}}{\frac{SS_W}{\sum_{i=1}^m n_i - m}}, \\ p\text{-value} &= \Pr[F_{m-1, \sum_{i=1}^m n_i - m} > TS] \\ &= 1 - \Pr[F_{m-1, \sum_{i=1}^m n_i - m} < TS] \\ &= 1 - \text{cdf}(TS, m-1, \sum_{i=1}^m n_i - m). \end{aligned}$$

3.2 Two-Way Analysis of Variance

The observed values of the random variables X_{ij} can be defined as a $m \times n$ matrix,

$$\begin{bmatrix} X_{11} & \cdots & X_{1j} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{i1} & \cdots & X_{ij} & \cdots & X_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mj} & \cdots & X_{mn} \end{bmatrix}$$

where m is the possible levels of row factors, and n is the possible levels of column factors.

We will assume that the data X_{ij} are independent normal random variables with common variance σ^2 , and that the mean value of the data depends *in an additive manner* on both its row and column. This gives,

$$\mu_{ij} = E[X_{ij}] = a_i + b_j = \mu + \alpha_i + \beta_j,$$

where the grand mean is

$$\mu = \mu_{\bullet\bullet} = a_{\bullet} + b_{\bullet} = \underbrace{\frac{1}{m} \sum_{i=1}^m a_i}_{=a_{\bullet}} + \underbrace{\frac{1}{n} \sum_{j=1}^n b_j}_{=b_{\bullet}} = \frac{1}{nm} \sum_{i=1}^m \sum_{j=1}^n \mu_{ij},$$

and where the deviation from the grand mean due to the row i is

$$\alpha_i = \mu_{i\bullet} - \mu = a_i - a_{\bullet} = \frac{1}{n} \underbrace{\sum_{j=1}^n \mu_{ij} - \mu}_{=\mu_{i\bullet}},$$

and where the deviation from the grand mean due to the column j is

$$\beta_j = \mu_{\bullet j} - \mu = b_j - b_{\bullet} = \frac{1}{m} \underbrace{\sum_{i=1}^m \mu_{ij} - \mu}_{=\mu_{\bullet j}}.$$

α_i and β_j are the deviations from the grand mean, and the sum of the deviations therefore satisfy the property

$$\sum_{i=1}^m \alpha_i = \sum_{j=1}^n \beta_j = 0. \quad (3)$$

The meaning of α_i and β_j is, say that we have to estimate an outcome. Then, if we do not have any information about to which row or which column, the

outcome belongs, then our best guess is μ . If we know that the outcome belongs to row i , then this will increase our estimate by the amount α_i . Moreover, if we also know to which column the outcome belongs, this will again increase the estimate by the amount β_j .

Estimates of μ , α_i , and β_j The estimates of μ , α_i , and β_j are $\hat{\mu}$, $\hat{\alpha}_i$ and $\hat{\beta}_j$. The estimates are

$$\begin{aligned}\hat{\mu} &= X_{\bullet\bullet} \\ \hat{\alpha}_i &= X_{i\bullet} - X_{\bullet\bullet} \\ \hat{\beta}_j &= X_{\bullet j} - X_{\bullet\bullet}\end{aligned}$$

3.3 Two-Factor Analysis of Variance: Testing Hypotheses

(See mm4 Ex. 10.23) Two hypotheses are setup. The first hypothesis tests the row effect, that is whether data is changing with the row,

$$H_0 : \text{all } \alpha_i = 0 \quad \text{versus} \quad H_1 : \exists \alpha_i \neq 0,$$

and the second hypothesis tests the column effect, that is whether data is changing with the column,

$$H_0 : \text{all } \beta_j = 0 \quad \text{versus} \quad H_1 : \exists \beta_j \neq 0.$$

An example is

	Diet 1	Diet 2
Women	7.6	19.5
	8.8	17.6
Men	22.2	30.1
	23.4	24.2

here, to test the hypothesis that the diet has the same effect on men and women, we test the row effect.

The approach is similar to one-way analysis of variance, and we find two estimators for σ^2 . One which hold whether H_0 is true or false, and a second which is only true if H_0 is true.

Remark: If there is interaction between the samples, then this should be tested, see Section 3.4. Interaction can happen if there are multiple values for each X_{ij} .

3.3.1 Test on Row Effect

The First Variance Estimator We have

$$\sum_{i=1}^m \sum_{j=1}^n \frac{(X_{ij} - \mu_{ij})^2}{\sigma^2} = \sum_{i=1}^m \sum_{j=1}^n \frac{(X_{ij} - (\mu + \alpha_i + \beta_j))^2}{\sigma^2} \sim \chi_{nm}^2.$$

To find the degrees of freedom left, when we substitute the real values with the estimates, we see that:

- To find all $\hat{\alpha}_i$ we need $m - 1$ estimates, as the last estimate comes from the property (3).
- To find all $\hat{\beta}_j$ we need $n - 1$ estimates (same reason).
- To find $\hat{\mu}$ we need 1 estimate.

In total, we need $m - 1 + n - 1 + 1 = n + m - 1$ estimates, and the degrees of freedom left is therefore $nm - (n + m - 1) = (n - 1)(m - 1)$. This gives

$$\sum_{i=1}^m \sum_{j=1}^n \frac{(X_{ij} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j))^2}{\sigma^2} = \sum_{i=1}^m \sum_{j=1}^n \frac{(X_{ij} - X_{i\bullet} - X_{\bullet j} + X_{\bullet\bullet})^2}{\sigma^2} \sim \chi_{(n-1)(m-1)}^2.$$

The first variance estimator can then be found

$$SS_e = \sum_{i=1}^m \sum_{j=1}^n (X_{ij} - X_{i\bullet} - X_{\bullet j} + X_{\bullet\bullet})^2, \quad \frac{SS_e}{\sigma^2} \sim \chi_{(n-1)(m-1)}^2, \quad E \left[\frac{SS_e}{(n-1)(m-1)} \right] = \sigma^2.$$

The Second Variance Estimator For the second estimator we consider the row averages which satisfy

$$X_{i\bullet} \sim \mathcal{N} \left(\mu, \frac{\sigma^2}{n} \right), \quad i = 1, \dots, m,$$

because β_j does not exist in the row averages and because $\alpha_i = 0$ due to the hypothesis. Using (2), then we have

$$n \sum_{i=1}^m \frac{(X_{i\bullet} - X_{\bullet\bullet})^2}{\sigma^2} \sim \chi_{m-1}^2,$$

and therefore we have

$$SS_r = n \sum_{i=1}^m (X_{i\bullet} - X_{\bullet\bullet})^2, \quad \frac{SS_r}{\sigma^2} \sim \chi_{m-1}^2, \quad E \left[\frac{SS_r}{m-1} \right] = \sigma^2.$$

Comparison of the Estimators

$$TS = \frac{\frac{SS_r}{m-1}}{\frac{SS_e}{(n-1)(m-1)}}.$$

$$p\text{-value} = \Pr[F_{m-1, (n-1)(m-1)} > TS]$$

$$= 1 - \Pr[F_{m-1, (n-1)(m-1)} < TS]$$

$$= 1 - \text{fcdf}(TS, m-1, (n-1)(m-1)).$$

3.3.2 Test on Column Effect

Instead of SS_r we use SS_c , this gives

$$SS_c = m \sum_{j=1}^n (X_{\bullet j} - X_{\bullet\bullet})^2, \quad \frac{SS_c}{\sigma^2} \sim \chi_{n-1}^2, \quad E\left[\frac{SS_c}{n-1}\right] = \sigma^2.$$

We have

$$TS = \frac{\frac{SS_c}{n-1}}{\frac{SS_e}{(n-1)(m-1)}},$$

$$p\text{-value} = \Pr[F_{n-1, (n-1)(m-1)} > TS]$$

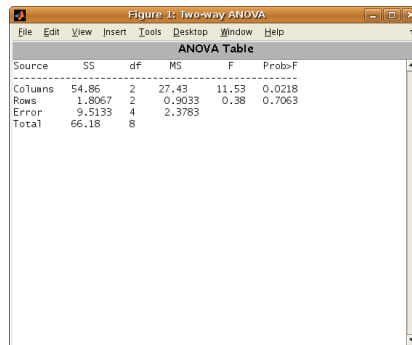
$$= 1 - \Pr[F_{n-1, (n-1)(m-1)} < TS]$$

$$= 1 - \text{fcdf}(TS, n-1, (n-1)(m-1)).$$

3.3.3 Matlab

`[pvalue, dummy, st] = anova2(X)`
`multcompare(st, 'estimate', 'row')` `multcompare(st, 'estimate', 'column')`

gives



where in row

1. SS is SS_c , $F = TS$, and $Prob > F = p$ -value.
2. SS is SS_r , $F = TS$, and $Prob > F = p$ -value.
3. SS is SS_w .

3.4 Two-Way Analysis of Variance With Interaction

(See mm4 Ex. 10.27 and 10.28) There can only be row and/or column interaction if there are more observations for each pair of factors. Say that there are l observations for each factor, then we have X_{ijk} , $k = 1, \dots, l$. We say that

$$\mu_{ij} = E[X_{ij}] = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

where the interaction of row i and column j is

$$\gamma_{ij} = \mu_{ij} - (\mu + \alpha_i + \beta_j) = \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu_{\bullet\bullet}$$

The First Variance Estimator We have

$$\sum_{k=1}^l \sum_{i=1}^m \sum_{j=1}^n \frac{(X_{ijk} - (\mu + \alpha_i + \beta_j + \gamma_{ij}))^2}{\sigma^2} \sim \chi_{nml}^2.$$

To find the degrees of freedom left, when we substitute the real values with the estimates, we see that:

- To find all $\hat{\alpha}_i$ we need $m - 1$ estimates, as the last estimate comes from the property (3).
- To find all $\hat{\beta}_j$ we need $n - 1$ estimates (same reason).
- To find $\hat{\mu}$ we need 1 estimate.
- To find $\hat{\gamma}_{ij}$ we need $(n - 1)(m - 1)$ estimates because of an extension of the property (3) which states that $\sum_{i=1}^m \gamma_{ij} = \sum_{j=1}^n \gamma_{ij} = 0$.

In total, we need $m - 1 + n - 1 + 1 + (n - 1)(m - 1) = nm$ estimates, and the degrees of freedom left is therefore $nm(l - 1)$. This gives

$$\sum_{k=1}^l \sum_{i=1}^m \sum_{j=1}^n \frac{(X_{ijk} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij}))^2}{\sigma^2} = \sum_{k=1}^l \sum_{i=1}^m \sum_{j=1}^n \frac{(X_{ijk} - X_{ij\bullet})^2}{\sigma^2} \sim \chi_{nm(l-1)}^2,$$

where

$$X_{ij\bullet} = \frac{1}{l} \sum_{k=1}^l X_{ijk}.$$

The first variance estimator can then be found

$$SS_e = \sum_{k=1}^l \sum_{i=1}^m \sum_{j=1}^n (X_{ijk} - X_{ij\bullet})^2, \quad \frac{SS_e}{\sigma^2} \sim \chi_{nm(l-1)}^2, \quad E \left[\frac{SS_e}{nm(l-1)} \right] = \sigma^2.$$

The Second Variance Estimator We should satisfy the hypothesis

$$H_0 : \text{all } \gamma_{ij} = 0 \quad \text{versus} \quad H_1 : \exists \gamma_{ij} \neq 0,$$

For the second estimator we consider the averages which satisfy

$$X_{ij\bullet} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{l}\right), \quad i = 1, \dots, m, \quad j = 1, \dots, n.$$

We have

$$l \sum_{j=1}^n \sum_{i=1}^m \frac{(X_{ij\bullet} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j))^2}{\sigma^2} = l \sum_{j=1}^n \sum_{i=1}^m \frac{(X_{ij\bullet} - X_{i\bullet\bullet} - X_{\bullet j\bullet} + X_{\bullet\bullet\bullet})^2}{\sigma^2} \sim \chi_{(n-1)(m-1)}^2,$$

and therefore we have

$$SS_{int} = l \sum_{j=1}^n \sum_{i=1}^m (X_{ij\bullet} - X_{i\bullet\bullet} - X_{\bullet j\bullet} + X_{\bullet\bullet\bullet})^2, \quad \frac{SS_{int}}{\sigma^2} \sim \chi_{(n-1)(m-1)}^2, \quad E\left[\frac{SS_{int}}{(n-1)(m-1)}\right] = \sigma^2.$$

Remark: According to Matlab *anova2*, then SS_{int} should not be multiplied with l .

Comparison of the Estimators

$$TS = \frac{\frac{SS_{int}}{(n-1)(m-1)}}{\frac{SS_e}{nm(l-1)}},$$

$$p\text{-value} = \Pr[F_{(n-1)(m-1), nm(l-1)} > TS]$$

$$= 1 - \Pr[F_{(n-1)(m-1), nm(l-1)} < TS]$$

$$= 1 - \text{cdf}(TS, (n-1)(m-1), nm(l-1)).$$

Test of Row and Column Effect with Interactions

To test row effect

$$SS_r = nl \sum_{i=1}^m (X_{i\bullet\bullet} - X_{\bullet\bullet\bullet})^2,$$

and

$$TS = \frac{\frac{SS_r}{m-1}}{\frac{SS_e}{nm(l-1)}}, \quad p\text{-value} = \Pr[F_{m-1, nm(l-1)} > TS].$$

To test column effect

$$SS_c = ml \sum_{j=1}^n (X_{\bullet j\bullet} - X_{\bullet\bullet\bullet})^2,$$

and

$$TS = \frac{\frac{SS_c}{n-1}}{\frac{SS_e}{nm(l-1)}}, \quad p\text{-value} = \Pr[F_{n-1, nm(l-1)} > TS].$$

Matlab Let

$$\mathbf{X} = \begin{bmatrix} X_{111} & \cdots & X_{1n1} \\ \vdots & \ddots & \vdots \\ X_{11l} & \cdots & X_{1nl} \\ X_{211} & \cdots & X_{2n1} \\ \vdots & \ddots & \vdots \\ X_{21l} & \cdots & X_{2nl} \\ \vdots & \ddots & \vdots \\ X_{m11} & \cdots & X_{mnl} \\ \vdots & \ddots & \vdots \\ X_{m1l} & \cdots & X_{mnl} \end{bmatrix}$$

then

$$\text{anova2}(X, l)$$

4 Nonparametric Hypothesis Tests

Nonparametric tests are used when no assumptions can be made about the underlying distribution. We consider two types of tests

- **The Sign Test:** Tests whether the median of a sample is equal to m_0 .
- **The Runs Test For Randomness:** Tests whether a sequence of 0s and 1s are a random sequence.

4.1 The Sign Test

(See mm5 Ex. 12.2) We want to test whether the mean m from the sample X_1, \dots, X_n from the distribution F is some m_0 . This gives

$$H_0 : m = m_0 \quad \text{versus} \quad H_1 : m \neq m_0$$

We know the cdf F , so $F(m) = 0.5$, that is, half of the values in X_i will be less than m and the other half greater than m . To test H_0 , we define the (counting) independent Bernoulli random variable

$$I_i = \begin{cases} 1, & \text{if } X_i < m_0, \\ 0, & \text{if } X_i \geq m_0, \end{cases}$$

and the counting of these i.i.d variables gives the test statistic

$$TS = \sum_{i=1}^n I_i \sim \text{Bin}(n, p).$$

The hypothesis can now be reformulated to

$$H_0 : p = \frac{1}{2} \quad \text{versus} \quad H_1 : p \neq \frac{1}{2}$$

That is, we should accept H_0 if $n/2 \approx TS$. With a $100(1-\alpha)$ percent confidence, we reject H_0 if

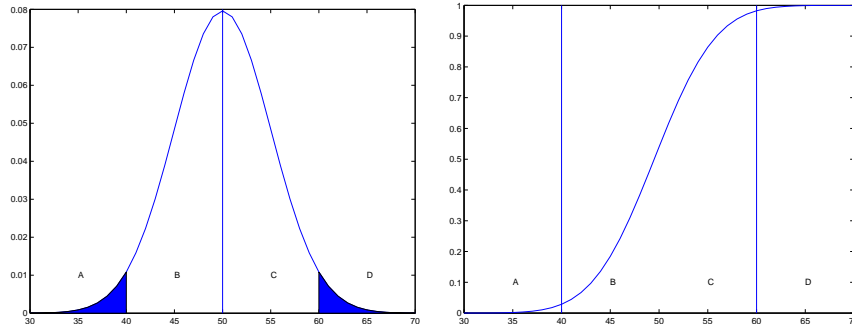
$$\underbrace{\Pr \left[\text{Bin} \left(n, \frac{1}{2} \right) \geq TS \right]}_{TS \text{ is too high}} = \Pr \left[\text{Bin} \left(n, 1 - \frac{1}{2} \right) \leq n - TS \right] \leq \frac{\alpha}{2} \quad \text{or} \quad \underbrace{\Pr \left[\text{Bin} \left(n, \frac{1}{2} \right) \leq TS \right]}_{TS \text{ is too low}} \leq \frac{\alpha}{2},$$

and the p -value is

$$\begin{aligned} p\text{-value} &= 2 \min \left\{ \Pr \left[\text{Bin} \left(n, \frac{1}{2} \right) \leq n - TS \right], \Pr \left[\text{Bin} \left(n, \frac{1}{2} \right) \leq TS \right] \right\} \\ &= \begin{cases} 2 \Pr \left[\text{Bin} \left(n, \frac{1}{2} \right) \leq TS \right], & \text{if } TS \leq \frac{n}{2} \\ 2 \Pr \left[\text{Bin} \left(n, \frac{1}{2} \right) \leq n - TS \right], & \text{if } TS \geq \frac{n}{2} \end{cases} \\ &= \begin{cases} 2 \text{binocdf}(TS, n, 0.5), & \text{if } TS \leq \frac{n}{2} \\ 2 \text{binocdf}(TS, n - TS, 0.5), & \text{if } TS \geq \frac{n}{2} \end{cases} \end{aligned}$$

Remark: If two sets of data should be compared, then by taking their difference and setting $m_0 = 0$, it can be found whether there is a difference. See mm5 Ex. 12.2.

Explanation Consider the example with $n = 100$ and $\alpha = 0.05$, then the pdf and cdf are as depicted.



$$\begin{cases} A & \Pr[\text{Bin} \left(n, \frac{1}{2} \right) \leq TS] < \frac{\alpha}{2} & \text{if } TS < \frac{n}{2} & \text{reject} \\ B & \Pr[\text{Bin} \left(n, \frac{1}{2} \right) \leq TS] \geq \frac{\alpha}{2} & \text{if } TS < \frac{n}{2} & \text{accept} \\ C & \Pr[\text{Bin} \left(n, \frac{1}{2} \right) \geq TS] \geq \frac{\alpha}{2} & \text{if } TS \geq \frac{n}{2} & \text{accept} \\ D & \Pr[\text{Bin} \left(n, \frac{1}{2} \right) \geq TS] < \frac{\alpha}{2} & \text{if } TS \geq \frac{n}{2} & \text{reject} \end{cases}$$

One-Sided Hypothesis Test for the Median (See mm5 Ex. 12.5) We set the hypothesis

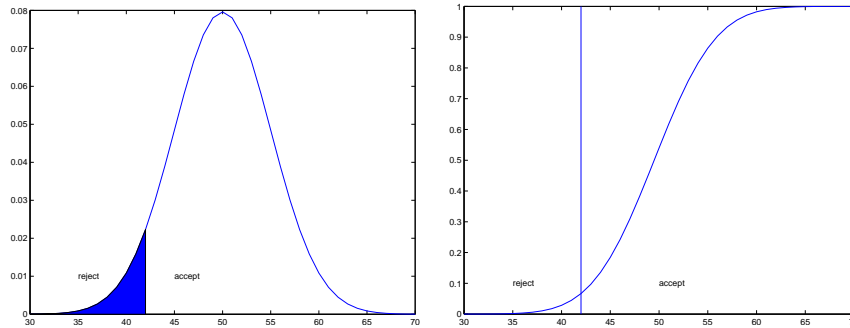
$$H_0 : m \leq m_0 \quad \text{versus} \quad H_1 : m > m_0$$

and equivalently (as before)

$$H_0 : F(m_0) \geq \frac{1}{2} \quad \text{versus} \quad H_1 : F(m_0) < \frac{1}{2}$$

(in the cdf before, $m_0 \geq 50$, and therefore $F(m_0) \geq \frac{1}{2}$.)

$$p\text{-value} = \Pr \left[\text{Bin} \left(n, \frac{1}{2} \right) \leq TS \right]$$



4.2 The Runs Test for Randomness

A sequence X_1, \dots, X_N of n 1s and m 0s ($N = n + m$) are observed.

H_0 : the sequence is random versus H_1 : the sequence is not random

The number of runs r in the sequence is the number of groups of 0s and 1s. An example is

X_1	X_2	X_3	r
0	0	1	2
0	1	0	3
1	0	0	2

Moreover, we have

$$\min r = 2$$

$$\max r = \begin{cases} 2 \min(m, n) + 1 & \text{if } m \neq n \\ 2m & \text{if } m = n \end{cases}$$

The pdf is defined as

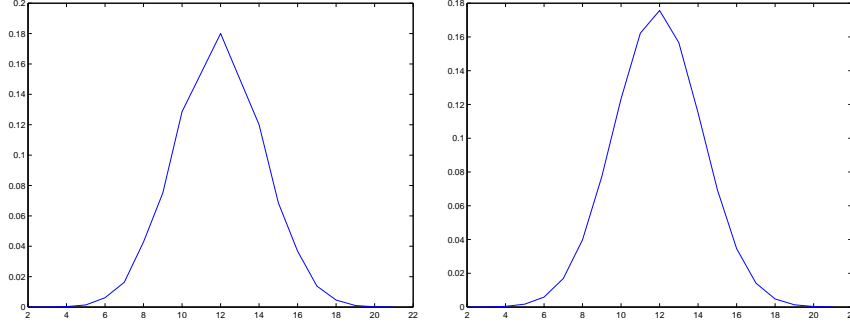
$$\Pr[R = 2k] = 2 \frac{\binom{m-1}{k-1} \binom{n-1}{k-1}}{\binom{m+n}{n}}$$

$$\Pr[R = 2k + 1] = \frac{\binom{m-1}{k-1} \binom{n-1}{k} + \binom{m-1}{k} \binom{n-1}{k-1}}{\binom{m+n}{n}}$$

It can be approximated for large n and m as a normal random variable with parameters

$$\mu = \frac{2nm}{n+m} + 1, \quad \sigma = \sqrt{\frac{2nm(2nm - n - m)}{(n+m)^2(n+m-1)}}$$

The true, and the approximation are depicted in the figures



The p -value is

$$p\text{-value} = 2 \min\{\Pr[R \geq r], \Pr[R \leq r]\}$$

$$\approx 2 \min\left\{Q\left(\frac{r - \mu}{\sigma}\right), 1 - Q\left(\frac{r - \mu}{\sigma}\right)\right\}$$

Runs Test for Non-Binary Data A random sequence of real values X_1, \dots, X_N can be transformed into a binary sequence using many approaches, where two of them are given here.

Method 1: Find the sample median $s - med$, and define the binary sequence as

$$I_i = \begin{cases} 1, & X_i \leq s - med, \\ 0, & \text{otherwise.} \end{cases}$$

Method 2: Let $I_1 = 0$ and define the binary sequence as

$$I_i = \begin{cases} 1, & X_i - X_{i-1} \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$