Review

Sequences

Statistic is about extracting useful information from a larger set of numbers (a data sequence). To do so, we need to summarize the data in ways that capture important patterns/features of the data.

Most of the computations we do in statistics are based on sums. Thus, understanding the summation formula is important

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_n$$

Examples of when the summation formula is used:

Description:	Parameter:	Estimator:
Population mean	μ	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Population variance	σ^2	$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
Population std. dev.	σ	$S_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$
Population covariance	σ_{XY}	$S_{YX} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$
Population correlation	ρ_{XY}	$R_{YX} = \frac{S_{XY}}{S_X * S_Y}$

Performing statistical analysis, these formulas (estimators) are crucial to us. All the types of statistical procedures we have used in this course relies on these expressions.

Population vs. sample

One important purpose of statistical analysis is to draw conclusions about a population using limited information from a smaller sample.

Random sampling

Our statistical procedures relies on *random sampling*. That is, objects (or individuals) must be sampled (selected) independently. In practice, there are quite many difficulties associated with sampling. For instance, when performing a survey, *volunteer response* may cause the sample to be biased towards certain types of individuals, which in turn is likely to bias the final analysis.

Types of data

It is not meaningful to apply parametric procedures (procedures based on sums) to all types of data. We distinguish between four different levels of information content

1. Nominal scale (lowest level)
2. Ordinal scale

3. Interval scale
4. Ratio scale (highest level)

Only interval and ratio scale data satisfy the requirements for using parametric procedures. Data on a nominal and ordinal scale can be handled using dummy variables. Thus, nominal and ordinal data are useful when studying segments (or groups) within the data.

Probability distributions

The important probability distributions in this course are

- the Normal distribution
- the t-distribution
- the F-distribution
- the χ^2-distribution

These distributions all belong to the same family of distributions, and are important in inferential statistics (hypothesis testing). Probability distribution are among the key components when

- Computing confidence intervals.
- Performing a hypothesis testing. Knowing that a particular test statistic approximates a certain probability distribution allows the user to evaluate the null-hypothesis.

Inference

Inferential statistics is the tool that allows us to get answers to research questions using sample information.

Confidence intervals

In general, a confidence interval takes the general form

$$PE \pm ME = [PE - ME, PE + ME]$$

where PE is a *point estimate* and ME is the so-called *margin error*. A confidence interval is stated in terms of a *confidence level*, expressed $100 \cdot (1 - \alpha)\%$.

If X is the variable of interest, a confidence interval for the mean of X is given by

$$\bar{x} \pm t_{\alpha/2} \cdot \frac{s_X}{\sqrt{n}}$$

where n is the sample size, $\bar{x}$ is the sample mean, s_X the sample standard deviation and $t_{\alpha/2}$ is the critical value. The interval is derived from a t-distribution with degrees of freedom $df = n - 1$.

A confidence interval for the difference in two means is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \cdot \sqrt{\frac{s_{X_1}^2}{n_1} + \frac{s_{X_2}^2}{n_2}}$$

where n_1 and n_2 are the samples sizes associated with the two samples, and $s_{X_1}^2$ and $s_{X_2}^2$ are the sample variances. In this case, the degrees of freedom is found as the *floor* (the number one gets when rounding down to the nearest integer value) of

$$df^* = \frac{\left(\frac{s_{X_1}^2}{n_1} + \frac{s_{X_2}^2}{n_2}\right)^2}{\frac{\left(s_{X_1}^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_{X_2}^2/n_2\right)^2}{n_2 - 1}}$$

Hypothesis testing

Before any testing can take place, a relevant hypothesis must be formulated and a significance level, denoted α, must be chosen. The significance level controls the relative frequency of how often a true null-hypothesis is rejected (type I error).

One-sample *t*-test for the mean

Depending on the research question, the test is either one-tailed or two-tailed. Formally, we state the one-tailed hypotheses in the following way

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_A &: \mu > \mu_0 \end{aligned}, \qquad \begin{aligned} H_0 &: \mu = \mu_0 \\ H_A &: \mu < \mu_0 \end{aligned}$$

while the two-tailed hypothesis takes the form

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_A &: \mu \neq \mu_0 \end{aligned}$$

The test statistic is

$$t = \frac{\bar{x} - \mu_0}{\frac{s_X}{\sqrt{n}}}$$

where μ_0 is the true mean under the null hypothesis. The test is derived from a t-distribution with degrees of freedom $df = n - 1$. When the test is performed using critical values, the rules for when to reject the null-hypothesis are:

H_0	H_A	Reject H_0 when
$\mu = \mu_0$	$\mu > \mu_0$	$t > t_\alpha$
$\mu = \mu_0$	$\mu < \mu_0$	$t < -t_\alpha$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert t\rvert > t_{\alpha/2}$

where t denotes the value of the test statistic that is computed from the realized sample.

Instead of using critical values to determine when to reject, we may use *p*-values. The rejection rule is then

Reject the null hypothesis when the *p*-value $< \alpha$

For the three formulations of the alternative hypothesis, the *p*-value is found by:

H_0	H_A	*p*-value =
$\mu = \mu_0$	$\mu > \mu_0$	$P(T \geq t \mid H_0)$
$\mu = \mu_0$	$\mu < \mu_0$	$P(T \leq t \mid H_0)$
$\mu = \mu_0$	$\mu \neq \mu_0$	$2 \cdot P(T \geq \lvert t \rvert \mid H_0)$

Two-sample *t*-test (independent samples)

Again, there are three different formulations of the hypothesis. The one-tailed formulations are:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_A &: \mu_1 > \mu_2 \end{aligned}, \qquad \begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_A &: \mu_1 < \mu_2 \end{aligned}$$

The two-tailed formulation is:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_A &: \mu_1 \neq \mu_2 \end{aligned}$$

The test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_{X_1}^2}{n_1} + \dfrac{s_{X_2}^2}{n_2}}}$$

The degrees of freedom in this case is found as the floor of

$$df^* = \frac{\left(\dfrac{s_{X_1}^2}{n_1} + \dfrac{s_{X_2}^2}{n_2}\right)^2}{\dfrac{\left(s_{X_1}^2/n_1\right)^2}{n_1 - 1} + \dfrac{\left(s_{X_2}^2/n_2\right)^2}{n_2 - 1}}$$

When the test is performed using critical values, the rules for when to reject the null-hypothesis are:

H_0	H_A	Reject H_0 when
$\mu_1 = \mu_2$	$\mu_1 > \mu_2$	$t > t_\alpha$
$\mu_1 = \mu_2$	$\mu_1 < \mu_2$	$t < -t_\alpha$
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$\lvert t \rvert > t_{\alpha/2}$

Instead of using critical values, we may use *p*-values. The rejection rule is then

Reject the null hypothesis when the *p*-value $< \alpha$

For the three formulations of the alternative hypothesis, the *p*-value is found by:

H_0	H_A	*p*-value =
$\mu_1 = \mu_2$	$\mu_1 > \mu_2$	$P(T \geq t \mid H_0)$
$\mu_1 = \mu_2$	$\mu_1 < \mu_2$	$P(T \leq t \mid H_0)$
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$2 \cdot P(T \geq \lvert t \rvert \mid H_0)$

Regression analysis

The simple regression model

$$Y = B_0 + B_1 X + \varepsilon$$

In this equation, the *intercept* B_0 (also called the *constant term*) represents the mean value of Y when $X = 0$. The value of B_1 represents the change in the mean of Y, resulting from a one-unit increase in X.

Estimating B_0 and B_1, using the OLS estimator, involves the formulas

$$b_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{XY}}{s_X^2}$$

and

$$b_0 = \bar{y} - b_1 \bar{x}$$

(these formulas can only be used for the simple regression model). The fitted regression line is

$$\hat{Y} = b_0 + b_1 X_1$$

where b_0 represents the estimated value of Y when $X = 0$. The value of b_1 is the estimated change in Y, resulting from a one-unit increase in X.

The estimated errors (residuals) can be obtained using

$$\begin{aligned}\hat{\varepsilon} &= Y - (b_0 + b_1 X_1) \\ &= Y - \hat{Y}\end{aligned}$$

Model fit

The model fit (in terms of explained variance) is

$$R^2 = r_{XY}^2$$

(this formula can only be used for the simple regression model).

The multiple regression model

In the more general case when there are k independent variables, the model is

$$Y = B_0 + B_1 X_1 + B_2 X_2 + \cdots + B_k X_k + \varepsilon$$

In this equation, B_0 represents the mean of Y when $X_1, \ldots, X_k$ are simultaneously zero. The value of the slope parameter B_1 represents the change in the expected value of Y, resulting from a one-unit increase in X_1, *holding* $X_2, \ldots, X_k$ fixed. Similar interpretation applies in case of $B_2, \ldots, B_k$.

For the general model, statistical software is needed to estimate the parameters. The fitted regression "line" takes the form

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k$$

The estimated errors can be obtained using

$$\begin{aligned}\hat{\varepsilon} &= Y - (b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k) \\ &= Y - \hat{Y}\end{aligned}$$

Model fit

Based on the data, we have

$$\begin{aligned}y_i &= b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + \hat{\varepsilon}_i \\ &= \hat{y}_i + \hat{\varepsilon}_i\end{aligned}$$

where

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}$$

From these equations (and the assumption that ε and the Xs are unrelated), we can write

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

or simply

$$TSS = ESS + RSS$$

Model fit follows from the expression

$$R^2 = 1 - \frac{RSS}{TSS}$$

Testing the individual parameters

There are three ways to formulate the alternative hypothesis. The appropriate formulation depends on the research question. A test of significance may be one-tailed

$$\begin{array}{ll} H_0 : B = B^* & H_0 : B = B^* \\ H_A : B > B^* , & H_A : B < B^* \end{array}$$

or two-tailed

$$\begin{array}{l} H_0 : B = B^* \\ H_A : B \neq B^* \end{array}$$

The test statistic is given by

$$t = \frac{b - B^*}{SE(b)}$$

The test is derived from a t-distribution with degrees of freedom $df = n - k - 1$. When the test is performed using critical values, the rules for when to reject the null-hypothesis are:

H_0	H_A	Reject H_0 when
$B = B^*$	$B > B^*$	$t > t_\alpha$
$B = B^*$	$B < B^*$	$t < -t_\alpha$
$B = B^*$	$B \neq B^*$	$\lvert t \rvert > t_{\alpha/2}$

where t to denotes the realized value of the test statistic.

Instead of using critical values to determine when to reject the null-hypothesis, we may use *p*-values. The rejection rule is

Reject the null hypothesis when the *p*-value $< \alpha$

For the three formulations of the alternative hypothesis, the *p*-value is found by:

H_0	H_A	*p*-value =
$B = B^*$	$B > B^*$	$P(T \geq t \mid H_0)$
$B = B^*$	$B < B^*$	$P(T \leq t \mid H_0)$
$B = B^*$	$B \neq B^*$	$2 \cdot P(T \geq \lvert t \rvert \mid H_0)$

Confidence intervals for the parameters

A $100 \cdot (1-\alpha)$% confidence interval for the true parameter B is given by

$$b \pm t_{\alpha/2} \cdot SE(b)$$

Testing multiple restrictions

The F-test is used to evaluate hypotheses involving more than one parameter. The testing procedure involves comparing two models, one that is *restricted* and one that is *unrestricted.* The restricted model is found by applying the restrictions from the null-hypothesis.

The F-statistic is based on the *Residual Sum of Squares* of both the restricted model (RSS_r) and the unrestricted model (RSS_{ur}), obtained from estimating the two models. The test statistic is

$$F = \frac{\frac{RSS_r - RSS_{ur}}{m}}{\frac{RSS_{ur}}{n-k-1}}$$

where m is the number of restrictions. The decision rule is to reject the null-hypothesis when $F > F_{\alpha,(m,n-k-1)}$, where $F_{\alpha,(m,n-k-1)}$ is the critical value at the α-level. Alternatively, we may use the *p*-value approach. The null-hypothesis is then rejected when the *p*-value is less than α (same rule as previously).

Non-linear forms

Quadratic transformation

Consider the model

$$Y = B_0 + B_1X_1 + B_2X_1^2 + other\ predictors + \varepsilon$$

Interpreting the effect of X on Y is not as straightforward as in the linear case.

Logarithmic transformation

Logarithmic transformation is typically applied when the considered variable is a *price* variable such as *firm sales*, *firm value*, *consumption*, *savings*, *income* (for instance, *wages*, *GDP*) etc. It is also quite common to transform so-called *count* variables. One problem is that such variables may take the value zero. One way to handle this problem is by replacing all zeros with the value one.

An important property of log-transformation is that it allows for an easy interpretation of the slope parameters in the model. First, consider the (semi-log) model

$$\log(Y) = B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k + \varepsilon$$

In this model, $100 \cdot B_1\%$ represents the (approximate) percentage change in Y for a one unit increase in X_1, holding $X_2, \ldots, X_k$ fixed (or constant). Similar interpretation applies to $B_2, \ldots, B_k$.

Second, consider the (double-log) model

$$\log(Y) = B_0 + B_1 \log(X_1) + B_2X_2 + \cdots + B_kX_k + \varepsilon$$

The parameter $B_1\%$ represents the (approximate) percentage change in Y for a one percent increase in X_1, holding $X_2, \ldots, X_k$ fixed (or constant). Note that $X_2, \ldots, X_k$ are not subject to any transformation. Thus, the interpretation of $B_2, \ldots, B_k$ remain the same as for the semi-log model.

Regression with categorical predictors

General specification of a dummy variable is

$$D = \begin{cases} 1 & \text{if some condition is satisfied} \\ 0 & \text{Otherwise} \end{cases}$$

Intercept dummy

An intercept dummy is a dummy variable that allows the intercept to take different values depending on the value of the dummy, which is either 0 or 1. Consider the model

$$Y = B_0 + B_1D + B_2X + \varepsilon$$

where D is dummy taking the value 0 or the value 1. The estimated regression line when $D = 0$ is

$$\hat{Y} = b_0 + b_2X$$

When $D = 1$, the estimated regression line becomes

$$\hat{Y} = (b_0 + b_1) + b_2X$$

Testing if there is a difference in the intercept across the two segments (or groups) in the data involves evaluating the hypothesis

$$H_0 : B_1 = 0$$
$$H_A : B_1 \neq 0$$

Slope dummy

Consider the model

$$Y = B_0 + B_1X + B_2(X \cdot D) + \varepsilon$$

The estimated regression line when $D = 0$ is

$$\hat{Y} = b_0 + b_1 X$$

When $D = 1$, the estimated regression line becomes

$$\hat{Y} = b_0 + (b_1 + b_2) \cdot X$$

Testing if there is a difference in the slope across the two segments (or groups) involves evaluating the hypothesis

$$H_0 : B_2 = 0$$
$$H_A : B_2 \neq 0$$

Note that it is also possible to combine the use intercept and slope dummies.

ANOVA-models

The ANOVA-framework allows the user to test for differences among multiple means. Thus, ANOVA represents an extension of the independent two-sample t-test.

Suppose that we have g groups. The hypothesis to be evaluated is then

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$$
$$H_A : \text{At least one } \mu_i \neq \mu_j \text{ for } i \neq j$$

The test can be performed using regression analysis. Analyzing g groups involves specifying $g - 1$ dummy variables. The model becomes (no other predictors that the dummies)

$$Y = B_0 + B_1 D_1 + B_2 D_2 + \cdots + B_{g-1} D_{g-1} + \varepsilon$$

The hypothesis is now be formulated as

$$H_0 : B_1 = 0, B_2 = 0, \ldots, B_{g-1} = 0$$
$$H_A : \text{At least one } B_j \neq 0 \text{ for } j = 1, \ldots, g - 1$$

The test statistic is

$$F = \frac{\frac{TSS - RSS}{g - 1}}{\frac{RSS}{n - g}}$$

where RSS is the residual sum of squares for the model with all the dummy variables in it, and $m - 1$ is the number of restrictions to be evaluated. The null-hypothesis is rejected when $F > F_{\alpha,(g-1,n-g)}$.

The modeling framework may be expanded to include Xs on the right-hand-side of the equation. The model is then referred to as an ANCOVA-model.

Moderation analysis

Continuous moderator

$$Y = B_0 + B_1X_1 + B_2Z + B_3(X_1 \cdot Z) + other\ predictors + \varepsilon$$

The aim of this model is to study the effect of X_1 on Y for some realization of Z (here denoted z). We can easily show that, for some fixed value z, the effect is given by the expression

$$B_1 + B_3z$$

It follows that the hypothesis to be evaluated is

$$H_0 : B_1 + B_3z = 0$$
$$H_A : B_1 + B_3z \neq 0$$

We can simplify matters by estimating the model

$$Y = B_0 + \tilde{B}_1X_1 + B_2Z + B_3(X_1 \cdot \tilde{Z}) + other\ predictors + \varepsilon$$

where $\tilde{B}_1 = B_1 + B_3z$ and $\tilde{Z} = Z - z$. By estimating this model, we can directly evaluate the effect of X_1 on Y, for some fixed value z, by the hypothesis

$$H_0 : \tilde{B}_1 = 0$$
$$H_A : \tilde{B}_1 \neq 0$$

Logistic regression analysis

The logistic regression model

Suppose that Y, the dependent variable, is a binary variable, indicating either *success* ($Y = 1$) or *failure* ($Y = 0$) of some event. The statistical framework to handle binary dependent variables is logistic regression modeling. The general logistic regression model is given by

$$\begin{aligned} P(Y = 1|X_1, \dots, X_k) &= \frac{1}{1 + e^{-(B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k)}} \\ &= \frac{e^{B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k}}{1 + e^{B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k}} \end{aligned}$$

This model may alternatively be written in terms of the log-odds (or the logit)

$$\text{logit}(P) = \log\left(\frac{P}{1 - P}\right) = B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k$$

In this expression, the effect on the odds of a one-unit increase in X_1, holding $X_2, \dots, X_k$ fixed (or constant), is given by e^{B_1}. More precisely, the associated percentage change in the odds, resulting from a one-unit increase in X_1, is $100 \cdot (e^{B_1} - 1)\%$.

Model fit

The McFadden (pseudo) R^2 is given by

$$R^2 = 1 - \frac{LL_F}{LL_0} \quad \left(= 1 - \frac{-2LL_F}{-2LL_0}\right)$$

where LL_F is the maximized log-likelihood obtained from estimating the full model (the model of interest) and LL_0 is the maximized log-likelihood obtained from estimating the so-called *null-model*. Note that the McFadden (pseudo) R^2 is not the only pseudo R^2 in logistic regression analysis.

Testing single restrictions

Evaluating the significance of the individual predictors involves the following two-tailed hypothesis

$$H_0 : B = B^*$$
$$H_A : B \neq B^*$$

The realized test statistic is

$$z = \frac{b - B^*}{SE(b)}$$

Under the null-hypothesis, the test statistic is asymptotically (when the sample size tend to infinity) standard normally distributed. The decision rule is to reject the null-hypothesis when the absolute value of z exceeds the critical value $z_{\alpha/2}$. The critical value $z_{\alpha/2}$ can be obtained from the following table:

α	$z_{\alpha/2}$ (Two-tailed test)
1 %	$z_{0.005} = 2.576$
5 %	$z_{0.025} = 1.960$
10 %	$z_{0.05} = 1.645$

Alternatively, the test can be performed using the *p*-value approach. The rejection rule is then

Reject the null hypothesis when the *p*-value $< \alpha$

Testing multiple restrictions

Testing multiple restrictions involves formulating hypotheses the same way as for the linear regression model. The test statistic is different, and is given by the expression

$$LR = -2 \cdot LL_r - (-2 \cdot LL_{ur})$$

where LL_r is the maximized log-likelihood obtained from estimating the restricted model, and LL_{ur} is the maximized log-likelihood obtained from estimating the unrestricted model. The decision rule is to reject the null-hypothesis when the LR statistic exceeds the critical value $\chi^2_{\alpha,df}$, where df is the number of restrictions being tested. Alternatively, we may use the p-value approach. The null-hypothesis is then rejected when the p-value is less than α.

Factor analysis

The general factor model

The general factor analysis model with p observed variables and q factors takes the form

$$\begin{aligned}
item_1 &= \alpha_{11}F_1 + \alpha_{12}F_2 + \cdots + \alpha_{1q}F_q + \delta_1 \\
item_2 &= \alpha_{21}F_1 + \alpha_{22}F_2 + \cdots + \alpha_{2q}F_q + \delta_2 \\
item_3 &= \alpha_{31}F_1 + \alpha_{32}F_2 + \cdots + \alpha_{3q}F_q + \delta_3 \\
\vdots \quad & \quad \vdots \qquad \vdots \qquad\qquad \vdots \qquad \vdots \\
item_p &= \alpha_{p1}F_1 + \alpha_{p2}F_2 + \cdots + \alpha_{pq}F_q + \delta_p
\end{aligned}$$

where $F_1, \ldots, F_q$ are *unobserved* (or *latent*) random variables, also referred to as *common factors*, $\delta_1, \ldots, \delta_p$ are the *unique* factors and $\alpha_{11}, \ldots, \alpha_{pq}$ are *factor loadings*. Estimation of $\alpha_{11}, \ldots, \alpha_{pq}$ is typically performed using principal component analysis (spectral decomposition).

Choosing the number of factors

Tools for choosing the appropriate number of factors are

- Choose the number of factors according to the number of eigenvalues larger than 1.
- Examining a scree plot.

Rotation

The factor solution is not unique. The various possible solutions are evaluated based on the size of the loadings and how the observed variables group. For instance, if a factor has only small loadings, say 0.3 or less (in absolute value), we may try a solution with one factor less.

Rotation greatly facilitate interpretation of the factors. Rotation works by transforming the initial solution such that the observed variables load strongly on some factors and weakly on some other.

Two types of rotation:

1. *Orthogonal rotation*: the factors remain uncorrelated (just as in the unrotated solution).
2. *Oblique rotation*: when applying this form of rotation, one allows the factors to be correlated.

The effect of rotation (either orthogonal or oblique) is summarized in the following way:

- Rotation alters the interpretation of the factors (which is in fact what we want).
- The contribution (in terms of explained variance) by each factor (SSL) will change.
- The contribution (in terms of explained variance) by the factors taken together does not change.
- The communalities does not change.

Explained variance

We have that

- the contribution of F_i (for $i = 1, \dots, q$) is found by

  $$SSL_i = a_{1i}^2 + a_{2i}^2 + \cdots + a_{pi}^2$$

  where SSL_i is the sum of the squared loadings associated with F_i. The computation holds for both the un-rotated and the rotated solution.

- the contribution of the factors taken together is the sum of contribution of the individual factors. The computation holds for both the un-rotated and the rotated solution.

- the contribution of the factors for the individual items are given by the communalities. The expression for the communality associated with $item_j$ (for $j = 1, \dots, p$) is

  $$Comm_j = a_{j1}^2 + a_{j2}^2 + \cdots + a_{jq}^2 = 1 - Var(\delta_j)$$

  where $0 \leq Comm_j \leq 1$. The computation holds for both the un-rotated and the rotated solution, given orthogonal rotation.