

1. Multivariate normal

Let k , μ , σ and ρ be constants and let $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ be a bivariate normal random vector with mean $\begin{bmatrix} \mu \\ k\mu \end{bmatrix}$ and covariance matrix $\begin{bmatrix} \sigma^2 & \rho \\ \rho & \sigma^2 \end{bmatrix}$.

(a) Find the mean and covariance of $\begin{bmatrix} X_1 + X_2 \\ X_1 - X_2 \end{bmatrix}$. Are $X_1 + X_2$ and $X_1 - X_2$ independent?

(b) For $k = 2$, $\rho = 2$ and $\mu = 1$ and $\sigma^2 = 2^2$, find:

$$P(X_1 \leq 2)$$

$$P(X_1 \leq 2 \mid X_2 = 2)$$

$$P(2X_1 - 3X_2 > 0)$$

2. Sums of squares

Let $Y_1, \dots, Y_n$ be a random sample of independent and identically distributed univariate normal variables with mean μ and variance σ^2 .

(a) What is the distribution of $\sum_{i=1}^n Y_i$?

Let $\mathbf{J}$ be an $n \times n$ matrix consisting only of ones and let $\mathbf{I}$ be the $n \times n$ identity matrix. In the following, you may use without proof that $\mathbf{I} - \frac{1}{n}\mathbf{J}$ and $\frac{1}{n}\mathbf{J}$ are symmetric and idempotent. Show that the sum $\sum_{i=1}^n (Y_i - \bar{Y})^2$ can be expressed as

$$\mathbf{Y}^T \left(\mathbf{I} - \frac{1}{n} \mathbf{J} \right) \mathbf{Y}$$

where $\mathbf{Y} = [Y_1 \cdots Y_n]^T$.

Use the above expression to argue that $\sum_{i=1}^n (Y_i - \bar{Y})^2 / \sigma^2$ is χ^2 distributed. What are the degrees of freedom?

(b) Show that, for $\mu = 0$, $\sum_{i=1}^n (Y_i - \bar{Y})^2$ and $n\bar{Y}^2$ are independent.

3. One-factor ANOVA

At university A, students arrive from from 4 different geographical regions; n_1 students from region 1, n_2 students from region 2, etc. At the start of their studies, they all complete 3 tests on prior knowledge (high school curriculum) and receive 3 test scores.

First, we consider only test scores from one of the tests (i.e. one score per student). For students $k = 1, \dots, n_l$ from geographical region l , the test score is recorded as y_{kl} . Investigators assume the following one-factor ANOVA model for test scores;

$$Y_{kl} = \mu + \alpha_l + \varepsilon_{kl}$$

for $k = 1, \dots, n_l$, $l = 1, \dots, 4$, with $\sum_l \alpha_l = 0$. All the $n_1 + n_2 + n_3 + n_4$ ε -s are assumed independent and $N(0, \sigma^2)$ -distributed.

- (a) Give short interpretations of the parameters μ , α_l and σ^2 . Your answers should be written in terms of the described investigation and not be generic definitions.

Using the model, what is the expected test score for student k from geographical region l ?

Using the model, what is the covariance in test scores for students k and k' from geographical region l ?

Since the students each completed 3 tests, the investigators want to include all of these test results in the same model. However, they realize that the assumption of independence between observations will be violated when each student contributes 3 observations to the data set. The investigators therefore define a new model which can incorporate repeated measurements for each student. Let the index $m = 1, 2, 3$ denote the test number, so that for student k from geographical region l taking test number m , the recorded test score is denoted y_{klm} . No difference in mean scores between the three tests is assumed. The following model is constructed:

$$Y_{klm} = \mu + \alpha_l + u_{kl} + \varepsilon_{klm}$$

for $k = 1, \dots, n_l$, $l = 1, \dots, 4$, $m = 1, 2, 3$ with $\sum_l \alpha_l = 0$, all ε_{klm} independent and $N(0, \sigma^2)$ -distributed, all u_{kl} independent and $N(0, \tau^2)$ -distributed. Also, the $3(n_1 + n_2 + n_3 + n_4)$ ε -s and the $n_1 + n_2 + n_3 + n_4$ u -s are assumed independent.

Use your knowledge about the normal distribution, expectation and variance to answer the following questions regarding this new type of model.

- (b) Using the model, what is the expected test for score for student k from geographical region l taking test m ?

Using the model, what is the covariance in test scores between tests m and m' for student k from geographical region l ?

Using the model, what is the covariance in test scores between students k and k' , from geographical region l , taking test m ?

4. Linear regression

Consider a linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{Y}$ is an n -dimensional random vector of responses, $\mathbf{X}$ an $n \times p$ design matrix, $\boldsymbol{\beta}$ a p -dimensional parameter (coefficient) vector and $\boldsymbol{\varepsilon}$ an n -dimensional multivariate normal vector with mean $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$, where $\mathbf{I}$ is the $n \times n$ identity matrix. Assume further that the columns of $\mathbf{X}$ are orthogonal.

- (a) Find an expression for the least squares estimator of β_j , the j th entry of $\boldsymbol{\beta}$. Show that the expression depends only on the j th column of $\mathbf{X}$, $\mathbf{x}_j$, and $\mathbf{Y}$.

Consider a two-level 2^2 factorial experiment in which the levels are coded -1 and 1 . The experiment is first conducted without replication. The design and measurements of the response are presented in the table below.

Factor 1	Factor 2	y
-1	-1	29.5
1	-1	32.7
-1	1	26.7
1	1	32.3

- (b) Express the relationship between the response and the factors (including interactions) as a multiple linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Write out the matrix $\mathbf{X}$ and discuss the interpretation of components of $\boldsymbol{\beta}$.

Show that the columns of the design matrix $\mathbf{X}$ are orthogonal.

Estimate all the components of $\boldsymbol{\beta}$.

The experiment was repeated and all measurements are presented in the table:

Factor 1	Factor 2	y_1	y_2
-1	-1	29.5	30.2
1	-1	32.7	33.5
-1	1	26.7	25.7
1	1	32.3	29.8

Two R-outputs are presented below. The first concerns fitting a *simple* linear regression to the 8 measurements and with only the levels of factor 1 as a covariate (x_1). The second concerns fitting a *multiple* linear regression to the 8 measurements but with both factors (x_1 and x_2) and their interaction (x_1x_2) as covariates.

- (c) Using the first R output (only factor 1 as a covariate), what is the hypothesis tested in the row starting with **x1**? Show how to arrive at the value of the t-test statistic (3.011) and the corresponding p -value (0.0237). Use Table 1 (below) of cumulative probabilities in the t -distribution to calculate an approximation of the p -value.

For a multiple linear model with an intercept and k other coefficients, we can test for significance of the regression using an F test with test statistic $F = \frac{SSR/k}{SSE/(n-(k+1))}$. Show that $F = T^2$ for a simple linear regression model ($k = 1$) with T being a t-test statistic as considered above.

- (d) Using the second R output (both factors and interaction), which effects are significantly different from zero if we use a significance level of $\alpha = 0.1$ and a Bonferroni correction? Do not perform a test of the intercept.

Show with calculations how to arrive at the Std. Error of 0.3619 for all coefficients in the model.

Table 1: Cumulative probabilities in the t -distribution with ν degrees of freedom

ν	$P(T \leq -4)$	$P(T \leq -3)$	$P(T \leq -2)$	$P(T \leq -1)$
4	0.008	0.020	0.058	0.187
5	0.005	0.015	0.051	0.182
6	0.004	0.012	0.046	0.178
7	0.003	0.010	0.043	0.175
8	0.002	0.009	0.040	0.173
9	0.002	0.007	0.038	0.172
10	0.001	0.007	0.037	0.170

Call:

```
lm.default(formula = y ~ x1)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.325	-1.562	0.425	1.438	2.175

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	30.0500	0.6726	44.677	8.42e-09 ***
x1	2.0250	0.6726	3.011	0.0237 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.902 on 6 degrees of freedom

Multiple R-squared: 0.6017, Adjusted R-squared: 0.5353

F-statistic: 9.064 on 1 and 6 DF, p-value: 0.02368

Call:

```
lm.default(formula = y ~ x1 * x2)
```

Residuals:

1	2	3	4	5	6	7	8
-0.35	1.25	0.50	-0.40	0.35	-1.25	-0.50	0.40

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	30.0500	0.3619	83.045	1.26e-07 ***
x1	2.0250	0.3619	5.596	0.005 **
x2	-1.4250	0.3619	-3.938	0.017 *
x1:x2	0.4000	0.3619	1.105	0.331

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.0235 on 4 degrees of freedom

Multiple R-squared: 0.9231, Adjusted R-squared: 0.8655

F-statistic: 16.02 on 3 and 4 DF, p-value: 0.01079