Statistics problems

Iakovleva Alena, DSBA 202

2022

Abstract

In this paper I decided to collect a couple of problems from our statistics course in order to use this note in the future, as it is more convenient to read the tex document, rather then hand-written text.

What do I plan to include here? I will rewise a small amount of theory and then I will solve a couple of statistics problems.

Contents

Chapter 1

Theory

1.1 ANOVA

Analysis of variance (ANOVA) is a popular tool that has an applicability and power that we can only start to appreciate in this course. The idea of analysis of variance is to investigate how variation in structured data can be split into pieces associated with components of that structure. We look only at one-way and two-way classifications, providing tests and confidence intervals which are widely used in practice.

One-way ANOVA One-way analysis of variance (one-way ANOVA) involves a continuous dependent variable and one categorical independent variable (sometimes called a factor, or treatment), where the k different levels of the categorical variable are the k different groups. We now introduce statistics associated with one-way ANOVA.

Two-way ANOVA Two-way analysis of variance (two-way ANOVA) involves a continuous dependent variable and two categorical independent variables (factors).

1.1.1 Statistics associated with one-way ANOVA

The ANOVA decomposition is

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\overline{X})^2=\sum_{j=1}^{k}(\overline{X}_{.j}-\overline{X})^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_{.j})^2$$

We have already discussed the jth sample mean and overall sample mean. The total variation is a measure of the overall (total) variability in the data from all k groups about the overall sample mean. The ANOVA decomposition decomposes this into two components.

1.1.2 Some useful formulas

1. Total variation

$$\text{Total SS} = \text{B} + \text{W} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij}^2 - nX^2$$

2. Residual (Error) SS

$$\text{Residual (Error) SS} = \text{W} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij}^2 - nX^2 - \sum_{j=1}^{k} n_j \overline{X}_{.j}^2 = \sum_{j=1}^{k} (n_j - 1){S_j}^2$$

1.1.3 One-way ANOVA table

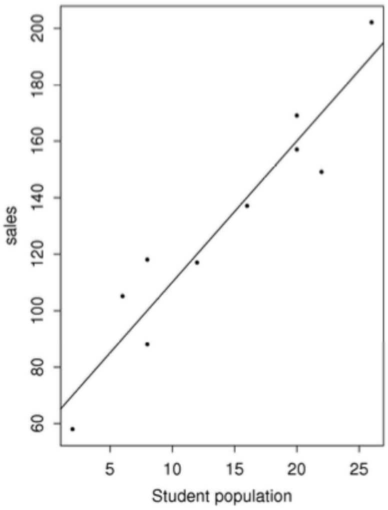
Typically, one-way ANOVA results are presented in a table as follows:

Source	DF	SS	MS	F	p-value
Factor	k - 1	B	B/(k - 1)	$\frac{B/(k-1)}{W/(n-k)}$	p
Error	n - k	W	W/(n - k)		
Total	n - 1	B + W			

1.2 Linear regression

Regression analysis is one of the most frequently-used statistical techniques. It aims to model an explicit relationship between one dependent variable, often denoted as y, and one or several regressors (also called covariates, or independent variables), often denoted as x$_1, ..., x_p$.

The goal of regression analysis is to understand how y depends on x$_1, ..., x_p$and to predict or control the unobserved y based on the observed x$_1, ..., x_p$.We start with some simple examples with p = 1.



1.2.1 Regression ANOVA

We can[**tfidf1972**] decompose the total variation of y in the simple linear regression model. It can be shown that the regression ANOVA decomposition is:

$$\sum_{i=1}^{n}(y_i-\overline{y})^2 =\sum_{i=1}^{n}\overline{\beta_i^2}(x_i-\overline{x})+\sum_{i=1}^{n}(y_i-\overline{\beta_0}-\overline{\beta_1}x_i)^2$$

where, denoting sum of squares by ‘SS’ we have:

- Total SS is

  $$\sum_{i=1}^{n}(y_i-\overline{y})^2 =\sum_{i=1}^{n}y_i^2-n\overline{y}^2$$

- Regression (explained) SS is

$$\sum_{i=1}^{n}\overline{\beta_i^2}(x_i-\overline{x})^2=\overline{\beta_i^2}(\sum_{i=1}^{n}x_i^2-n\overline{x}^2)$$

- Residual (error) SS is

$$\sum_{i=1}^{n}(y_i-\overline{\beta}_0-\overline{\beta}_1x_i)^2 =\text{Total SS - Regression SS}$$

Chapter 2

Practice

2.1 Problem 1

A dataset contains the annual cigarette consumption, x, and the corresponding mortality rate, y, due to coronary heart disease (CHD) of 21 countries. Some useful summary statistics calculated from the data are:

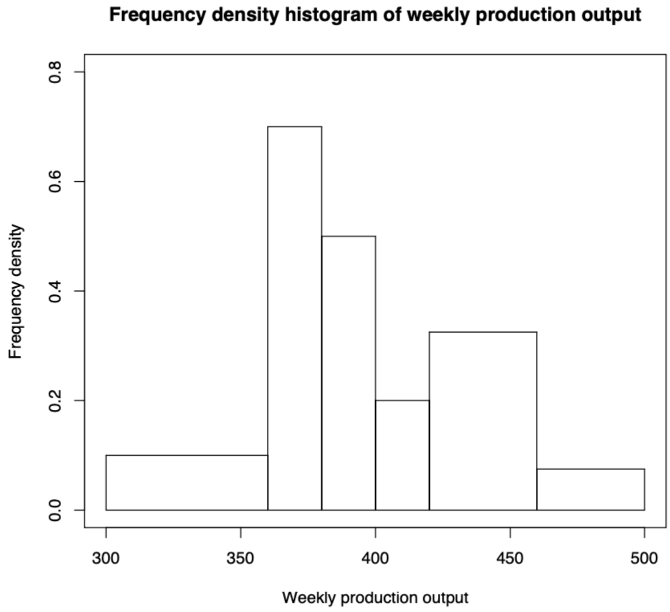
$$\sum_{i=1}^{21} x_i = 45110,\ \sum_{i=1}^{21} y_i = 042.2,\ \sum_{i=1}^{21} x_i^2 = 109957100,$$
$$\sum_{i=1}^{21} y_i^2 = 529321.58,\ \sum_{i=1}^{21} x_i y_i = 7319602.$$

Do these data support the suspicion that smoking contributes to CHD mortality? (Note the assertion ‘smoking is harmful for health’ is largely based on statistical, rather than laboratory, evidence.)

$$\overline{\beta}_1 = \frac{\sum\limits_i (x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_i (x_i-\overline{x})^2} = \frac{7319602-45110*3042.2/21}{109957100-(45110)^2/21} = 0.06$$

$\overline{\beta}_0 = 15.77$
Since t $= 0.06/0.01475 = 4.068 > 2.54 = \mathrm{t}_{0.01}$,19, we reject the hypothesis 1 = 0 at the 1 % significance level and we conclude that there is strong evidence that smoking contributes to CHD mortality.

2.2 Problem 2, Histogram



Class interval	Interval width	Frequency	Frequency density	Cumulative frequency
[300, 360)	60	6	0.100	6
[360, 380)	20	14	0.700	20
[380, 400)	20	10	0.500	30
[400, 420)	20	4	0.200	34
[420, 460)	40	13	0.325	47
[460, 500)	40	3	0.075	50

The table above includes two additional columns: (i.) ‘Frequency density’ – obtained by calculating ‘frequency divided by interval width’ (for example, 6/60 = 0.100), and (ii.) 'Cumulative frequency' – obtained by simply determining the running total of the class frequencies (for example, 6 + 14 = 20). Note the final column is not required for a histogram per se, although the computation of cumulative frequencies may be useful when determining medians and quartiles (to be discussed later in this chapter). [**Sulsky1994**]

To construct the histogram, adjacent bars are drawn over the respective class intervals such that the area of each bar is proportional to the interval frequency. This explains why equal bin widths are desirable since this reduces the problem to making the heights proportional to the interval frequency. However, you may be told to use a particular number of bins or bin widths, such that the bins will not all be of equal width. In such cases, you will need to compute the frequency density as outlined above. [1]

[1] Wow, this is my comment

Chapter 3

Thanks

My quote: ”Thanks for the opportunity to show how I can use Latex”

So, that is it! I am done with my document! I hope that you liked it!

Bibliography

[1] Sulsky D., Chen Z., Schreyer H. L. A particle method for history-dependent materials // Computer Methods in Applied Mechanics and Engineering. — 1994, V. 118. — P. 179–196.

[2] Григорьев Ю. В., Вшивков В. А., Федорук, М. П. Численное моделирование методами частиц в ячейках. — Новосибирск : Издательство СО РАН. — 2004. — 360 с.

[3] Liu G. R., Liu M. B. Smoothed particle hydrodynamics: a meshfree particle method. — Singapore : World Scientific Publishing. — 2003. — 449 p.