Psychological Test and Assessment Modeling, Volume 59, 2017 (3), 359-371

Optimal design of surveys and experiments

Dieter Rasch[1] *& Jürgen Pilz*[2]

Abstract

After a general discussion about designing experiments and surveys it is shown how the program package OPDOE can be used to determine minimal sample sizes for confidence estimation and hypotheses testing for means in the one- and two-sample problem. OPDOE is demonstrated by some examples.

Key Words: Testing hypotheses, confidence intervals, minimal sample size, experimental design

[1] *Correspondence concerning this article should be addressed to:* Dieter Rasch, PhD, University of Natural Resources and Life Sciences, Vienna, Institute of Applied Statistics and Computing, Peter-Jordan-Str. 82, 1190 Vienna, Austria; email: psychiatrie@klf-net.de

[2] Faculty of Technical Sciences, Alpen Adria University Klagenfurt

1 Introduction

The main tools of experimental research in sociology and psychology is the theory of surveys and experiments as parts of Mathematical Statistics. Mathematical Statistics developed on the fundament of Probability Theory from the end of 19th century on. At the beginning of the 20th century, Karl Pearson and Sir Ronald Aylmer Fisher were notable pioneers of this new discipline. Fisher's book (1925) was a milestone providing experimenters such basic concepts as his well-known maximum likelihood method and analysis of variance as well as notions of sufficiency and efficiency.

When we, in the sequel, speak about experiments, we understand this in the broader sense including also surveys – but see for the fundamental differences of experiments and surveys from the theory of science' point of view for instance Rasch, Kubinger, and Yanagida (2011). In concrete applications, the experiment first has to be planned, and after the experiment is finished, the analysis has to be carried out. We deal in this paper with the pre-experimental phase, i.e. the optimal planning of an experiment.

Experimental designs originated in the early years of the 20-th century mainly in agricultural field experimentation. A centre was Rothamsted Experimental Station near London, where Sir Ronald Aylmer Fisher was head of the statistical department (since 1919). There he wrote one of the first books about statistical design of experiments (Fisher, 1935); a book which was fundamental, and promoted statistical technique and application.

Everything presented in the following is, however, also very important and applicable in psychological research. The mathematical justification of the methods is not stressed, here, and proofs will be often barely sketched, rather omitted. Readers interested in this are referred to Rasch and Schott (2018).

Fisher (1935) also outlined the problem of "Lady tasting tea", now a famous design of a statistical randomized experiment which uses Fisher's exact test and is the original exposition of Fisher's notion of a null hypothesis.

We refer in the following first to Fisher's problem, that deals with soil fertility. Because soil fertility in fields varies enormously, a field is partitioned into so-called blocks (or strata in surveys) and each block subdivided into plots. It is expected that the soil within the blocks is relatively homogeneous so that the differences in the yield of the varieties planted at the plots of one block are suggested to be only due to the varieties but not due to soil differences. To ensure homogeneity of soil within blocks, the blocks must not be too large. On the other hand, the plots must be large enough so that harvesting (mainly with machines) is possible. Consequently, only a limited number of plots within the blocks is possible and only a limited number of varieties within the blocks can be tested. If all varieties can be tested in each block, we speak of a complete block design. The number of varieties is often larger than the number of plots in a block. Therefore incomplete block designs were developed, chiefly among them completely balanced incomplete block designs, ensuring that all yield differences of varieties can be estimated with equal variance using models of the analysis of variance. How all this is applicable in psychological research is shown in Rasch, Kubinger, and Yanagida (2011).

The Experimental Designs originally developed in agriculture soon were used in medicine, in psychology and in engineering or more general in all empirical sciences. Varieties were generalized to treatments, and plots to experimental units. But even today the number v of treatments or the letter y (from yield) in the models of the analysis of variance recall us to the agricultural origin.

Experimental designs are an important part in the planning (designing) of experiments. The main principles are (the three R-s):

1. Replication,
2. Randomisation
3. Reduction of the influence of noisy factors (blocking, stratification).

Statements in the empirical sciences can almost never be derived based on an experiment with only one measurement. As we often use the variance as a measure of variability of the observed character and then we need at least two observations (replications) to estimate it (in statistics the term *replication* mainly means one measurement, thus two measurements are two replications and not one measurement and one replication). Therefore, two replications are the lower bound for the number of replications. The sample size (the number of replications) has to be chosen.

Initially we consider the situation where the nuisance factors are not known or not graspable. In this case, we try to solve the problem by randomisation: this is here thought of as the unrestricted random assignment of the experimental units to the treatments (not vice versa!). Randomisation is also understood as the random selection of experimental units from a universe. Randomisation is used to keep the probability of some bias by some unknown nuisance factors as small as possible. That is, randomisation shall ensure that statistical models (as base for planning and analysing) are justified. We distinguish between pure and restricted forms of randomisation in experimental designs.

At first, we assume that the experimental material is unstructured which means there is no blocking. This is the simplest case of an experimental design. If in an experimental, design exactly n_i experimental units are randomly allocated to the i-th of v treatments ($\Sigma n_i = n$) we call this a complete or unrestricted randomisation and we call the experimental design a simple or a completely randomised experimental design. An experiment is always meant as the combination of an experimental design and some rule of randomisation.

Designing an experiment often needs some computational effort. Therefore we recommend to use program packages.

In Rasch, Herrendörfer, Bock, Victor, and Guiard (2008) on more than 2000 pages procedures and methods for nearly all practical statistical problems are given, containing numerical examples with calculations done by SAS-programs (see for instance Dembe, Partridge, and Geist (2011).

In Rasch, Kubinger, and Yanagida (2011) it was demonstrated how the package IBM SPSS Statistics can be used for the analysis of experiments and survey, respectively, and the package OPDOE of R was used for designing them.

Rasch, Pilz, Verdooren, and Gebhardt (2011) described the theory of experimental design and the practical application using examples calculated with the package OPDOE implemented in R. We describe this package in the next paragraph together with a short introduction to R.

2 The computer package OPDOE

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS systems.

The homepage of the "R Project for Statistical Computing" can be found at

https://www.r-project.org

The latest R-version (as of June 30, 2017) is R 3.4.1. A useful source for finding latest developments and packages as well as an archive of journal issues from the past describing basic packages and extensions is "The R Journal" at

https://journal.r-project.org

The Comprehensive R Archive Network (CRAN), which can be accessed from the above R homepage, is the host of many packages; one of them is OPDOE (Optimal Design of Experiments) which we describe in the sequel. To download R, we first have to choose our preferred CRAN mirror. An overview of CRAN mirrors can be obtained by accessing

https://cran.r-project.org

We recommend to use the URL

https://cran.wu.ac.at

located at the University of Business Administration Vienna. Vienna is still considered to be the "Capital of the Kingdom R", because several active members and R developers from Vienna belong to the R core group. The base system and contributed packages of R can be downloaded from the above URL, Windows and Mac users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows.

For Linux, users should check with their Linux package management system in addition to the link to CRAN indicated above.

After having R installed, you can load the OPDOE package directly from the task bar under the menu "Packages (Pakete in the German version)" by just clicking on "Install Packages (Installieren Pakete)". A full overview of available (contributed) packages in R can be obtained by clicking on "Packages" and then on "Table of available packages, sorted by name".

When installing OPDOE, the following depending packages are downloaded as well (automatically): “AlgDesign”, “gtools”, “gmp”, “mvtnorm”, “orthopolynom”, “crossdes”, “polynom”. Now, to start working with the R package OPDOE, simply type in

library (OPDOE)

after the prompt sign (“>”) at the command line interface (alternatively, you can also use a graphical user interface as provided by “R commander” or “R Studio”, respectively).

Our point of view is that the size of an experiment (or of a survey) should be determined following precision requirements fixed in advance by the researcher.

Let us assume the experimenter takes a random sample. An exact unrestricted random sample of size *n* (without replacement) is defined as a sample obtained by a unrestricted random sampling procedure, which is defined as a procedure where each element of the *N* elements of the universe has the same probability to become an element of the sample with the additional property that each of the $\binom{N}{n}$ possible subsets has the same probability to become the sample.

The base R function `sample` allows us to draw a random sample. For this, we assume that the elements of the universe are numbered from 1 to *N* and that we will randomly draw a sample of size *n* (here, we set *N=49, n=6*):

```
> set.seed(123)
> sample(49,6)
[1] 15 38 20 41 43 3
```

How OPDOE is used for sample size determination is shown in the following paragraphs. Although we restrict its application field to the case of course OPDOE can determine experimental sizes and optimal allocation as well for Analysis of Variance, Regression Analysis, and Correlation, Multiple Decisions, and also for Sequential Experimentation.

If we like to determine the sample size before an experiment begins, we have to formulate the precision needed for the analysis. We call this the precision requirement of an experiment and it contains, apart from the risks of incorrect conclusions, always the effect size δ as the difference of minimum which is of interest to be detected.

3 Estimating and testing means of normal distributions – one sample problem

In the present paper, we only discuss sample size determination via OPDOE but of course R also contains programs for a broad range of statistical analyses.

3.1 Confidence estimation

In confidence estimation, we define a random region, which covers the expectation (mean) μ with probability 1-α.

We only discuss the case that σ^2 is unknown. Then the interval is – by writing $t(f; P)$ for the P-quantile of the central t-distribution with f degrees of freedom – given by

$$[\,\overline{y} - t(n\text{-}1; 1-\frac{\alpha}{2})\frac{\mathbf{s}}{\sqrt{n}} \quad ; \quad \overline{y} + t(n-1; 1-\frac{\alpha}{2})\frac{\mathbf{s}}{\sqrt{n}}\,] \qquad (1)^3$$

Here $\overline{y}$ is the sample mean, n is the sample size and $\boldsymbol{s}$ the sample standard deviation.

The half-expected length of (1) is:

$$E(\mathbf{H}) = t(n-1; 1-\frac{\alpha}{2})\frac{E(\mathbf{s})}{\sqrt{n}} = \frac{t(n-1; 1-\frac{\alpha}{2})}{\sqrt{n}} \frac{\Gamma\left(\frac{n}{2}\right)\cdot\sqrt{2}}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{n-1}}\sigma \qquad (2),$$

$\Gamma(x)$ is the Gamma function

Our precision requirement $E(\mathbf{H}) \le \delta$ leads to the approximate equation for n:

$$n = \left\lceil t^2(n-1; 1-\frac{\alpha}{2})\frac{\sigma^2}{\delta^2} \right\rceil \qquad (3)$$

We calculate the sample size needed to construct a two-sided confidence interval for α = 0.05 and δ = 0.25σ. Because n occurs on both sides of (3) a calculation by hand must be done iteratively. In the iteration we use $n_1, n_2, \ldots$ for n and stop when n_i and n_{i+1} and this is just our n

We find for $n_0 = \infty$: $t\,(\infty; 0.975) = 1.96$, and from (2) we obtain

$$n_1 = \left\lceil \frac{1.96^2}{0.25^2} \right\rceil = \lceil 61.47 \rceil = 62\,.$$

Next we look up $t(61; 0.975) = 1.9996$, and this gives

$$n_2 = \left\lceil \frac{1.9996^2}{0.25^2} \right\rceil = \lceil 63.97 \rceil = 64$$

[3] Random variables are bold print throughout this paper.

Finally with $t(63; 0.975) = 1.9983$ n_3 becomes

$$n_3 = \left\lceil \frac{1.9983^2}{0.25^2} \right\rceil = \lceil 63.89 \rceil = 64$$

and thus $n = 64$ is the solution. The ceiling operator $\lceil\ \rceil$ accelerates the convergence of the iteration algorithm. OPDOE gives this result more easily:

```
> size.t.test(power=0.5, alpha=0.01, delta=0.4, sigma=1,
+type="two-sample")
[1] 64
```

In confidence estimation we have no type II risk. But if we, in the R-program for testing, put for power 0.05 we can formally use the R-program for testing also for the confidence estimation.

This can easily be seen if we compare (3) with the formula for the test (6). Because the *t*-distribution is symmetric at 0, we have $t(n-1;1-\beta)=0$ if $1-\beta = 0.5$ and for this equations (3) and (6) for n are identical.

3.2 Hypothesis testing

A random sample $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n$ of size n will be drawn from a normally distributed population with mean μ and variance σ^2, with the purpose of testing the null hypothesis:

$H_0 : \mu = \mu_0$ (μ_0 is a given constant)

against one of the following alternative hypotheses:

a) $H_A : \mu > \mu_0$ (one-sided alternative)

b) $H_A : \mu < \mu_0$ (one-sided alternative)

c) $H_A : \mu \neq \mu_0$ (two-sided alternative

The test statistic is

$$t = \frac{\bar{y} - \mu_0}{s}\sqrt{n} \tag{4}$$

which is non-central *t*-distributed with $n-1$ degrees of freedom (df) and non-centrality parameter

$$\lambda = \frac{\mu - \mu_0}{\sigma}\sqrt{n}\ .$$

Under the null hypothesis, the distribution is central *t*.

Given a type I risk α, H_0 is rejected if:

in case a), $t > t(n\text{-}1;1\text{-}\alpha)$,

in case b), $t < -t(n\text{-}1;1\text{-}\alpha)$,

in case c),$|t|> t(n\text{-}1;1\text{-}\alpha/2)$.

Our precision requirement is given by α and the type II risk β given $\mu\text{-}\mu_0 = \delta$.

From this, we have the requirement

$$t(n\text{-}1;1\text{-}\alpha/2) = t(n\text{-}1;\lambda;\beta) \qquad (5)$$

where $t(n\text{-}1;\lambda;\beta)$ is the β-quantile of the non-central t- distribution with $df = n - 1$ and the non-centrality parameter is

$$\lambda = \frac{\delta}{\sigma}\sqrt{n} \ .$$

Using the approximation $t(n\text{-}1;\lambda;\beta) = t(n\text{-}1, \beta) + \lambda$ leads to the approximate formula

$$n \approx \left\lceil \left[\left\{ t\left(n-1;1-\frac{\alpha}{2} \right) + t\left(n-1;1-\beta \right) \right\} \frac{\sigma}{\delta} \right]^2 \right\rceil \qquad (6)$$

From the requirement (4), the minimum sample size is calculated iteratively from the solution of

$$t(n-1;11-\frac{\alpha}{2}) = t(n-1;\frac{\delta}{\sigma}\sqrt{n};\beta) \qquad (7).$$

Our R program always gives the exact solution based on (6).

Let us calculate the minimal sample size for testing the null hypothesis:

$H_0 : \mu = \mu_0$ against $H_A : \mu \neq \mu_0$ (two-sided alternative).

Assume for example the precision requirement is $\delta = 0.8\sigma$, $\alpha = 0.05$, and $\beta = 0.01$. In the OPDOE program we have to use `power` that is 1-β and `sig.level` that is α.

Our R-program gives the solution $n = 53$.

```
> size.t.test(power=0.99,delta =0.8, sd=1, sig.level=0.01,
+alternative="two-sided”)
[1] 53
```

4 Estimating and testing means of normal distributions – two sample problem

4.1 Confidence estimation

If in the two-sample case the two variances are equal, then usually a pooled estimator $\mathbf{s}^2 = \frac{(n_x - 1)\mathbf{s}_x^2 + (n_y - 1)\mathbf{s}_y^2}{n_x + n_y - 2}$ of the common variance σ^2 is calculated from the two samples $(x_1, \mathbf{x}_2, ..., x_{nx})$ and $(y_1, \mathbf{y}_2, ..., y_{n_y})$ with sample means $\bar{x}; \bar{y}$ and sample variances $\mathbf{s}_x^2; s_y^2$.

The two-sided confidence interval is:

$$\left[\bar{\mathbf{x}} - \bar{\mathbf{y}} - t\left(n_x + n_y - 2; 1 - \frac{\alpha}{2}\right)\mathbf{s}\sqrt{\frac{n_x + n_y}{n_x n_y}}; \bar{\mathbf{x}} - \bar{\mathbf{y}} + t\left(n_x + n_y - 2; 1 - \frac{\alpha}{2}\right)\mathbf{s}\sqrt{\frac{n_x + n_y}{n_x n_y}}\right] \quad (8)$$

In this case it can be shown that optimal plans require the two sample sizes n_x and n_y to be equal. Thus $n_x = n_y = n$, and in the case where the expected half-width must be less than δ, we find n iteratively from

$$n = \left\lceil 2\sigma^2 \frac{t^2(2n-2; 1-\frac{\alpha}{2})}{\delta^2(2n-2)} \frac{2\Gamma^2\left(\frac{2n-1}{2}\right)}{\Gamma^2(n-1)} \right\rceil \quad (9)$$

We want to find the minimum size of an experiment to construct a two-sided 99% confidence interval for the difference of the expectations of two normal distributions. We assume equal variances and take independent samples from each population and define the precision by $\delta = 0.4\sigma$. Again we use the program for tests which means that the power is 0.5.

The R-program is as follows:

```
> size.t.test(power=0.5, alpha=0.01, delta=0.4, sigma=1,
+type="two-sample")
[1] 55
```

If one is not absolutely sure that the two variances are equal, one should use the confidence interval for unequal variances described below, as it was recommended by Rasch, Kubinger, and Yanagida (2011).

For the analysis never use the confidence interval (8) but (10) which is based on the estimators s_x^2 and s_y^2 of σ_x^2 and σ_y^2 respectively. The confidence interval (2.21) is only approximately a (1-α)-confidence interval (see Welch, 1947). It is given by

$$\left[\mathbf{\bar{x}}-\mathbf{\bar{y}}-t\left(f^*;1-\frac{\alpha}{2}\right)\sqrt{\frac{\mathbf{s}_x^2}{n_x}+\frac{\mathbf{s}_y^2}{n_y}};\mathbf{\bar{x}}-\mathbf{\bar{y}}+t\left(f^*;1-\frac{\alpha}{2}\right)\sqrt{\frac{\mathbf{s}_x^2}{n_x}+\frac{\mathbf{s}_y^2}{n_y}}\right] \tag{10}$$

with

$$f^*=\left\lceil\frac{\left(\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}\right)^2}{\frac{s_x^4}{(n_x-1)n_x^2}+\frac{s_y^4}{(n_y-1)n_y^2}}\right\rceil \tag{11}$$

To determine the necessary sample sizes n_x and n_y, apart from an upper bound for the half expected width δ, we need information about the two variances. Suppose that estimates s_x^2 and s_y^2 are available. For a two-sided confidence interval, we can calculate n_x and n_y approximately and iteratively:

$$n_x=\left\lceil\frac{\sigma_x(\sigma_x+\sigma_y)}{\delta^2}t^2\left(f^*;1-\frac{\alpha}{2}\right)\right\rceil \tag{12}$$

and

$$n_y=\left\lceil\frac{\sigma_y}{\sigma_x}n_x\right\rceil \tag{13}.$$

Given the minimum size of an experiment, we like to find a two-sided 99% confidence interval for the difference of the expectations of two normal distributions with unequal variances using independent samples from each population and define the precision by $\delta=0.4\sigma_x$ using (12) for n_x. If we know that $\frac{\sigma_x^2}{\sigma_y^2}=4$, we receive $n_y=\left\lceil\frac{1}{2}n_x\right\rceil$.

The R output shows that we need 33 observations – 22 from the first distribution and 11 from the second distribution:

```
> size.t.test(power=0.5, alpha=0.01, delta=0.4, sigma=1,
+ sigmas.ratio=2, type="two.sample")
[1] 22 11
```

4.2 Hypothesis testing

Let us now turn to hypothesis testing.

We have two normally distributed populations with means μ_x, μ_y and variances σ_x^2, σ_y^2, respectively. Our purpose is to take two independent random samples ($\boldsymbol{x}_{11},\ldots,\boldsymbol{x}_{1nx}$) and ($\boldsymbol{y}_{21}, \ldots, \boldsymbol{y}_{2ny}$) of sizes n_x and n_y from the two populations in order to test the null hypothesis

$H_0 : \mu_x = \mu_y$

against one of the following one- or two-sided alternative hypotheses

a) $H_A : \mu_x > \mu_y$

b) $H_A : \mu_x < \mu_y$

c) $H_A : \mu_x \neq \mu_y$

The sample sizes n_x and n_y should be determined in such a way that for a given type I risk α type II risk β does not exceed a predetermined upper bound delta as long as the alternative hypothesis holds; for the case c) we have

$\left|\mu_x - \mu_y\right| \geq \delta$.

If in the two-sample case the two variances are equal, then usually a pooled estimator $\mathbf{s}^2 = \frac{(n_x-1)\mathbf{s}_x^2 + (n_y-1)\mathbf{s}_y^2}{n_x+n_y-2}$ of the common variance σ^2 is calculated from the two samples $(x_1, \mathbf{x}_2, \ldots, x_{nx})$ and $(y_1, \mathbf{y}_2, \ldots, y_{n_y})$ with sample means $\overline{x};\overline{y}$ and sample variances $\mathbf{s}_x^2; s_y^2$.

But we never know before experimentation whether the variances are equal or not. In the analysis we therefore have always to use the Welch test as described in Rasch, Kubinger, and Yanagida (2011).

For the minimum sample size we compute an integer solution iteratively from

$$t(2n-2;1P) = t(2n-2;\lambda;\beta_0) \tag{14}$$

with P = 1- α in the one-sided cases and P = 1- $\alpha/2$ in the two-sided case and $\lambda = \frac{\mu_x - \mu_y}{\sigma}\sqrt{\frac{n}{2}}$.

Using the approximation $t(n\text{-}1;\lambda;\beta) = t(n\text{-}1,\beta) + \lambda$ leads to the approximate formula $n \approx \left\lceil 2\left[\left\{t(n-1;P) + t(n-1;1-\beta)\right\}\frac{\sigma}{\delta}\right]^2 \right\rceil$

We use here for σ the conjectured larger one of the two standard deviations, just in order to be on the safe side.

We like to know the minimal sample sizes $n_x = n_y = n$ to be drawn independently from two normal distributions for testing the null hypothesis given above. The two-sided alternative hypothesis is of interest and the precision requirements are $\delta = |\mu_x - \mu_y| = 0.9\sigma$; $\alpha = 0.05$; $\beta = 0.1$. Our R-program shows the result $n = 27$:

```
> size.t.test(power=0.9,delta=0.9,type="two.sample")
[1] 27
```

(Note that alpha=0.05 is the default value and need not be indicated extra).

For the analysis the t –test always has to be replaced by the Welch test, the reason is given in Rasch, Kubinger, and Yanagida (2011).

5 Discussion

The paper shows by some simple cases how useful the program package OPDOE is when an experiment or a survey has to be planned. But there are many other fields of experimentation as regression analysis or analysis of variance and covariance and others where OPDOE is applicable. Those interested in further application of OPDOE may look at Rasch, Pilz, Verdooren, and Gebhardt (2011) as well as at Rasch, Kubinger, and Yanagida (2011). In multivariate analysis we determine the sample size for that character for which we expect the largest variance.

Be aware that processing in another manner than planning an experiment or survey could mean that either much more effort is used than needed for a reasonable precision or the power of a test comes close to 0.5 so that it is much cheaper to abdicate any study but throw a die and accept the null hypothesis for even and reject it for odd numbers – or vice-versa.

6 References

Dembe, A. E., Partridge, J. S., Geist, L. C. (2011). Statistical software applications used in health services research: Analysis of published studies in the U.S" *BMC Health Services Research*,*11*, 252.

Fisher, R.A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.

Fisher, R.A. (1935). *The design of experiments*. Edinburgh: Oliver & Boyd..

Rasch, D., Herrendörfer, G., Bock, J., Victor, N., & Guiard, V. (2000). *Verfahrensbibliothek Versuchsplanung und -auswertung, Vol. I.* Beijing: China Statistics Press.

Rasch, D., Herrendörfer, G., Bock, J., Victor, N., & Guiard, V. (Eds.) (2008). *Verfahrensbibliothek Versuchsplanung und -auswertung* [Bibliotheca of methods of planning analyzing an experiment]. (2nd and improved edition, CD included). München: R. Oldenbourg Verlag.

Rasch, D., Kubinger K.D., & Yanagida, T. (2011), Statistics in Psychology using R and SPSS. Wiley: New York.

Rasch, D., Pilz, J., Verdooren, R.L, & Gebhardt, A. (2011). *Optimal Experimental Design with R*. Boca Raton: Chapman and Hall.

Rasch, D., & Schott, D. (2018). *Mathematical Statistics.* Oxford: Wiley.

Welch, B.L. (1947). The generalization of Students problem when several different population variances are involved. *Biometrika, 34*, 28-35.