**上海第二工业大学**

**本科毕业设计（ 论文 ）**

**学生手册**

题 目：基于Excel的毕业设计信息管理系统开发

学生姓名：　　 　 苗卫伟

学 号：　　 20134832535

学院名称：　　　　工学部

专业班级：　　　 13计科A2

指导教师：　　　 石林祥

开始日期：　2016年 11月 15日

完成日期： 2017年 05月 12日

目 录

[一、 毕业设计（论文）任务书 1](#_Toc482170015)

[二、 毕业设计（论文）开题报告 3](#_Toc482170016)

[三、 毕业设计（论文）中期检查表 7](#_Toc482170017)

[四、 毕业设计（论文）审阅表 8](#_Toc482170018)

[五、 毕业设计（论文）评阅表 9](#_Toc482170019)

[六、 毕业（论文）指导记录表 10](#_Toc482170020)

[七、 毕业设计（论文）答辩记录表 13](#_Toc482170021)

[八、 外文翻译稿和原稿 15](#_Toc482170022)

# 毕业设计（论文）任务书

|  |  |  |  |
| --- | --- | --- | --- |
| 题　目 | 基于Excel的毕业设计信息管理系统开发 | | |
| 姓　名 | 苗卫伟 | 学　　号 | 20134832535 |
| 题目类型 | 模拟课题（ ）实践课题（√）实验课题（ ）科研课题（ ）其他（ ） | | |
| 1. 课题主要研究（设计）内容： 2. 毕业设计开题和毕业设计中期答辩管理 3. 毕业设计最终答辩及结题管理及毕业设计成绩分析 4. 系统架构的设计 5. Excel交互界面和展示界面的设计 | | | |
| 二、工作进度要求（分阶段提出具体时间要求）：  第1阶段（2016年11月10日~2016年11月30日）：  向学生介绍课题，布置毕业设计（论文）任务；  第2阶段（2016年12月1日~2016年12月15日）：  确认毕业设计任务，以及相应的系统功能模块；  第3阶段（2016年12月1日~2017年1月10日）：  写开题报告，进行开题答辩，拟定解决毕业设计的关键问题  第4阶段（2017年1月11日~2017年2月21日）：  进行需求分析，基本确定技术路线与目标，熟悉开发工具与平台，外文翻译资料；；提交开题报告。  第5阶段（2017年2月22日~2017年4月13日）：  进行各模块的开发，调式；并接受中期检查。  第6阶段（2017年4月14日~2017年4月27日）：  进行论文的撰写；  第7阶段（2017年4月28日～2017年5月11日）：  交论文、英译中等文档，提供指导教师审阅，并按要求修改论文；  第8阶段（2017年5月12日～2017年5月22日）：  毕业设计（论文）评阅、审查。  第9阶段（2017年5月23日）：  毕业答辩。 | | | |
| 三、应查阅的主要参考文献：  [1] Christian Nagel，Jay Glynn，Morgan Skinner著，李铭，黄静 译，C#高级编程，清华大学出版社，2014年10月 第1版  [2] 刘增杰 著，MySQL 5.7从入门到精通，清华大学出版社，2016年9月第1版  [3] 陈明 著，数据库系统及应用--SQL Server 2000，清华大学出版社  [4] 王斌会 著，Excel应用与数据统计分析，暨南大学出版社，2011年3月 第1版  [5] 麦克格拉斯 著，李永伦 译，VSTO开发者指南，机械工业出版社，2009年1月 第1版  [6] 郑丽敏 编，Excel数据处理与分析，人民邮电出版社，2012 ，第四版  [7] 郑阿奇 主编 , 袁永福, 张小勇编著, 精通C#编程, 电子工业出版社, 2011年，第4版  [8] (美) 沃森(Watson, K.) 等著，齐立波，黄俊伟 译，清华大学出版社，2014年1月，第1版  [9] Excel Home 著，Excel 2013应用大全，人民邮电出版社，2015年8月，第1版  [10] ERIC CARTER，ERIC LIPPERT著，Visual Studio Tools for Office: Using C# with Excel, Word, Outlook, and InfoPath，艾迪生-韦斯利出版公司，2005年9月第1版  [11] Robert Sheldon, Geoff Moes著，Beginning MySQL,，Wrox Press出版，2005年5月21日，第1版 | | | |
| 四、毕业设计（论文）预期成果或结论性观点  本设计（论文）完成后将提供一套可以正常运行的系统，包含可以正常运行的可执行文件，全部源代码和使用说明。 | | | |
| 五、毕业设计（论文）完成提交方式（设计、作品照片、实物、模型、技术文档或论文、含有技术文档或论文的光盘等）  1. 书面提交毕业设计说明书（论文）；  2. 书面提交不少3000字的译文资料及原文；  3. 包含软件源代码、可执行程序、相关的文档、任务书及设计说明书（论文）的光盘； | | | |
| 指导教师： 石林祥 审核人：  2017 年 1 月 8 日 年 月 日 | | | |

# 毕业设计（论文）开题报告

|  |  |  |  |
| --- | --- | --- | --- |
| 题　目 | 基于Excel的毕业设计信息管理系统开发 | | |
| 姓　名 | 苗卫伟 | 学　号 | 20134832535 |
| 1. 本课题的研究目的和意义   目的：  以减少教师不必要的劳动和重复性工作提高教师工作效率，方便学院对本科生毕业设计的管理，使学校和导师对学生毕业设计的管理与指导具有一定的实用价值  意义：   1. 方便老师和毕业生对毕业设计题目的拟定和选择 2. 基于Excel的设计让系统操作更加便利 3. 基于Excel的展示更方便老师对毕业设计的审核和成绩分析 | | | |
| 本课题的主要研究内容（提纲）   1. 毕业设计开题管理，向学生介绍课题及课题的确定 2. 毕业设计中期答辩管理 3. 毕业设计最终答辩及结题管理 4. 系统结构的设计 5. Excel交互界面的设计 6. Excel展示界面的设计 | | | |
| 文献综述（国内外研究情况及其发展）：  毕业设计的选题，实现，审批无论对老师还是学生都是非常重要的工作， 毕业生毕业设计的管理是大学的主要管理工作之一，涉及到校、系、师、生的诸多方面，随着教学体制的不断改革，尤其是学分制、选课制的展开和深入，毕业生工作地点的差异，毕业生论文管理工作日趋繁重、复杂。迫切需要研制开发一个基于Internet 网咯的综合网上毕业生论文管理软件。  国外高校一般具有较大规模的稳定的技术队伍来提供服务与技术支持， 纵观这些系统主要有以下几个特点：先进性：实现网络化，甚至一些著名院校都实现跨校访问历年毕业论文。通用性：各专业都能使用。方便性：无论身在何地，只要有网络，便可实现论文提交及查询。灵活性：题库随时更新。技术上趋于网页：无论国外还是国内大多数毕业设计管理系统都是基于web的，很少会使用客户端，web有着很多的有点比如不需要在本地安装任何软件或插件就可以正常的运行，但是相对于开发客户端来说web成本高而且还需要有专业人员去维护整个系统。相比之下客户端的开发成本就比较低而且不需要专业的人员不维护系统，而本系统采用Excel插件的形式，这将会使成本比较低廉，后期不需要专业人员维护，方便实用。  国内高校信息化建设相对起步较晚，尤其是毕业生论文方面。在数字校园理论逐步应用的过程中，各高校一方面不断投资购建各种硬件、系统软件和网络，另一方面也不断开发实施了各类教学、科研、办公管理等应用系统，形成了一定规模的信息化建设体系。但是，由于整体信息化程度相对落后，经费短缺，理论体系不健全等原因，国内高校教务管理系统在机构设置、服务范围、服务质量及人员要求上与国外高校相比都有一定的差距。纵观目前国内研究现状，在安全性和信息更新化方面存在有一定的不足，各种现有系统资料单独建立，共享性差；在以管理者为主体的方式中，信息取舍依赖管理者对于信息的认知与喜好，较不容易掌握用户真正的需求，也因此无法完全满足用户的需求。  目前国家的教育体制也正处在不断改革、创新的阶段，我国教育部门充分吸取国外优秀的教学模式，结合国内多年的办学经验，逐步探索出适合中国特色的教学形式，国家教育部面向各级各类学校开展了全面学分制改革。因此，毕业设计管理软件应充分依托校园网，实现教务信息的集中管理、分散操作、信息共享，使传统的教务管理朝数字化、无纸化、智能化、综合化的方向发展，并为进一步实现完善的计算机教务管理系统和全校信息系统打下良好的基础。 | | | |
| 1. 拟解决的关键问题 2. Excel功能区交互界面的设计 3. 如何在Excel上进行展示 4. 对学生毕业设计成绩进行怎么样的分析 5. 怎么实现自动生成毕业设计学生手册 6. 怎么实现自动化的配置软件的运行环境 | | | |
| 1. 研究思路和方法 2. Excel功能区通过VSTO或者是Excel DNA建立 3. 展示区基础信息直接通过Excel表格展示，分析数据采用数据表等形式展示 4. 把客户端和服务端分开不直连数据库这样是数据更安全的传输 5. 毕业设计学生手册采用模板的形式运行用户自定义 6. 采用WIX技术来实现软件的自动化安装和运行环境的配置 | | | |
| 1. 本课题的进度安排   第1阶段（2016年11月10日~2016年11月30日）：  向学生介绍课题，布置毕业设计（论文）任务；  第2阶段（2016年12月1日~2016年12月15日）：  确认毕业设计任务，以及相应的系统功能模块；  第3阶段（2016年12月1日~2017年1月10日）：  写开题报告，进行开题答辩，拟定解决毕业设计的关键问题  第4阶段（2017年1月11日~2017年2月21日）：  进行需求分析，基本确定技术路线与目标，熟悉开发工具与平台，外文翻译资料；；提交开题报告。  第5阶段（2017年2月22日~2017年4月13日）：  进行各模块的开发，调式；并接受中期检查。  第6阶段（2017年4月14日~2017年4月27日）：  进行论文的撰写；  第7阶段（2017年4月28日～2017年5月11日）：  交论文、英译中等文档，供指导教师审阅，并按要求修改论文；  第8阶段（2017年5月12日～2017年5月20日）：  毕业设计（论文）评阅、审查。  第9阶段（2017年5月20日）：  毕业答辩。 | | | |
| 1. 参考文献   [1] Christian Nagel，Jay Glynn，Morgan Skinner著，李铭，黄静 译，C#高级编程，清华大学出版社，2014年10月 第1版  [2] 刘增杰 著，MySQL 5.7从入门到精通，清华大学出版社，2016年9月第1版  [3] 陈明 著，数据库系统及应用--SQL Server 2000，清华大学出版社  [4] 王斌会 著，Excel应用与数据统计分析，暨南大学出版社，2011年3月 第1版  [5] 麦克格拉斯 著，李永伦 译，VSTO开发者指南，机械工业出版社，2009年1月 第1版  [6] 郑丽敏 编，Excel数据处理与分析，人民邮电出版社，2012 ，第四版  [7] 郑阿奇 主编 , 袁永福, 张小勇编著, 精通C#编程, 电子工业出版社, 2011年，第4版  [8] (美) 沃森(Watson, K.) 等著，齐立波，黄俊伟 译，清华大学出版社，2014年1月，第1版  [9] Excel Home 著，Excel 2013应用大全，人民邮电出版社，2015年8月，第1版  [10] ERIC CARTER，ERIC LIPPERT著，Visual Studio Tools for Office: Using C# with Excel, Word, Outlook, and InfoPath，艾迪生-韦斯利出版公司，2005年9月第1版  [11] Robert Sheldon, Geoff Moes著，Beginning MySQL,，Wrox Press出版，2005年5月21日，第1版 | | | |
| 指导教师意见  指导教师（签名）：  年 月 日 | | | |
| 所在系（所）意见  负 责 人（签章）：  年 月 日 | | | |

# 毕业设计（论文）中期检查表

|  |  |  |  |
| --- | --- | --- | --- |
| 题　目 | 基于Excel的毕业设计信息管理系统开发 | | |
| 姓　名 | 苗卫伟 | 学　号 | 20134832535 |
| 原毕业设计（论文）的进度计划：  第1阶段（2016年11月10日~2016年12月1日）：  向学生介绍课题，布置毕业设计（论文）任务；  第2阶段（2016年12月1日~2016年12月31日）：  确认毕业设计任务，以及相应的系统功能模块；  第3阶段（2016年1月1日~2016年2月16日）：  进行需求分析，基本确定技术路线与目标，熟悉开发工具与平台，外文翻译资料；  第4阶段（2017年3月1日~2017年4月21日）：  进行各模块的开发，调式； | | | |
| 中期报告（已完成的研究内容，所取得阶段性成果，下一步工作计划和研究内容等）完成情况：   1. 创建数据库 2. 数据库基础资料的建立 3. 完成系统结构的设计 4. 完成Excel交互界面的设计 5. 完成Excel展示界面的设计 6. 完成后台逻辑 | | | |
| 指导教师意见：  **指导教师签字：**  年 月 日 | | | |

# 毕业设计（论文）审阅表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 题　目 | 基于Excel的毕业设计信息管理系统开发 | | | |
| 姓　名 | 苗卫伟 | | 姓　名 | 苗卫伟 |
| 审 阅 成 绩 | |  | | |
| 毕业设计（论文）评语 | | | | |
| 备注： | | | | |
| 审阅人（签名 ）：　　　　　　　　　　　　　　　　　　　　　　　　　　　 年 月 日 | | | | |

# 毕业设计（论文）评阅表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 题　目 | 基于Excel的毕业设计信息管理系统开发 | | | |
| 姓　名 | 苗卫伟 | | 姓　名 | 苗卫伟 |
| 评 阅 成 绩 | |  | | |
| 毕业设计（论文）评语 | | | | |
| 备注： | | | | |
| 评阅人（签名 ）： 评阅人职称：　　　　　　　　　　　　　　　　　　　　　　　　　　　　 年 月 日 | | | | |

# 毕业（论文）指导记录表

| 题　目 | 基于Excel的毕业设计信息管理系统开发 | | |
| --- | --- | --- | --- |
| 姓　名 | 苗卫伟 | 学　号 | 20134832535 |
| 指 导 内 容  学生签名： 指导老师签名： | | | |
| 指 导 内 容  学生签名： 指导老师签名： | | | |
| 指 导 内 容  学生签名： 指导老师签名： | | | |
| 指 导 内 容  学生签名： 指导老师签名： | | | |
| 指 导 内 容  学生签名： 指导老师签名： | | | |
| 指 导 内 容  学生签名： 指导老师签名： | | | |
| 指 导 内 容  学生签名： 指导老师签名： | | | |
| 指 导 内 容  学生签名： 指导老师签名： | | | |
| 指 导 内 容  学生签名： 指导老师签名： | | | |
| 指 导 内 容  学生签名： 指导老师签名： | | | |
| 指 导 内 容  学生签名： 指导老师签名： | | | |
| 指 导 内 容  学生签名： 指导老师签名： | | | |

# 毕业设计（论文）答辩记录表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 题　　目 | | 基于Excel的毕业设计信息管理系统开发 | | |
| 姓　　名 | | 苗卫伟 | 学　号 | 20134832535 |
| 答辩日期 | |  | 地　　点 |  |
| 答辩记录 | 用简明语言记录答辩过程，答辩委员会提出问题和学生回答情况(可附页)：  **记录员签字： 　　　　 答辩组长签字：**  年 月 日 年 月 日 | | | |

（接下页）（接上页）

|  |  |
| --- | --- |
| 评  语 | 答辩小组毕业设计（论文）的评语（选题意义，所用资料和数据可靠性；创新性成果及学术水平；写作的规范化和逻辑性；存在的主要不足之处，答辩情况。） |
| 答辩成绩 |  |
| **答辩组成员签字**：  **答辩组长签字：**    年 月 日 | |

# 外文翻译稿和原稿

## 外文资料翻译—原文部分

**Using Excel for Statistical Data Analysis - Caveats**

**ABSTRACT**

We used Excel to do some basic data analysis tasks to see whether it is a reasonable alternative to using a statistical package for the same tasks.  We concluded that Excel is a poor choice for statistical analysis beyond textbook examples, the simplest descriptive statistics, or for more than a very few columns.  The problems we encountered that led to this conclusion are in four general areas:

* Missing values are handled inconsistently, and sometimes incorrectly.
* Data organization differs per analysis, forcing you to reorganize your data in many ways if you want to do many different analyses.
* Many analyses can only be done on one column at a time, making it inconvenient to do the same analysis on many columns.
* Output is poorly organized, sometimes inadequately labeled, and there is no record of how an analysis was accomplished.

Excel is convenient for data entry, and for quickly manipulating rows and columns prior to statistical analysis. However, when you are ready to do the statistical analysis, we recommend the use of a statistical package such as SAS, SPSS, Stata, Siesta or Minitab.

**Introduction**

Excel is probably the most commonly used spreadsheet for PCs. Newly purchased computers often arrive with Excel already loaded. It is easily used to do a variety of calculations, includes a collection of statistical functions, and a Data Analysis Tool Pak. Thus, if you suddenly find you need to do some statistical analysis, you may turn to it as the obvious choice. We decided to do some testing to see how good Excel would serve as a Data Analysis application.

To present the results, we will use a small example. The data for this example is fictitious. It was chosen to have two categorical and two continuous variables, so that we could test a variety of basic statistical techniques. Since almost all real data sets have at least a few missing data points, and since the ability to deal with missing data correctly is one of the features that we take for granted in a statistical analysis package, we introduced two empty cells in the data:

Each row of the spreadsheet represents a subject. The first subject received Treatment 1, and had Outcome 1. X and Y are the values of two measurements on each subject. We were unable to get a measurement for Y on the second subject, or on X for the last subject, so these cells are blank. The subjects are entered in the order that the data became available, so the data is not ordered in any way.

We used this data to do some simple analyses and compared the results with a standard statistical package. The comparison considered the accuracy of the results as well as the ease with which the interface could be used for bigger data sets - i.e. more columns. We used SPSS as the standard, though any of the statistical packages OIT supports would do equally well for this purpose. In this article when we say "a statistical package," we mean SPSS, SAS, STATA, SYSTAT, or Minitab.

Most of Excels statistical procedures are part of the Data Analysis tool pack, which is in the Tools menu. It includes a variety of choices including simple descriptive statistics, t-tests, correlations, 1 or 2-way analysis of variance, regression, etc. If you do not have a Data Analysis item on the Tools menu, you need to install the Data Analysis Tool Pak.  Search in Help for "Data Analysis Tools" for instructions on loading the Tool Pak.

Two other Excel features are useful for certain analyses, but the Data Analysis tool pack is the only one that provides reasonably complete tests of statistical significance. Pivot Table in the Data menu can be used to generate summary tables of means, standard deviations, counts, etc. Also, you could use functions to generate some statistical measures, such as a correlation coefficient. Functions generate a single number, so using functions you will likely have to combine bits and pieces to get what you want. Even so, you may not be able to generate all the parts you need for a complete analysis.

Unless otherwise stated, all statistical tests using Excel were done with the Data Analysis Tool Pak. To check a variety of statistical tests, we chose the following tasks:

* Get means and standard deviations of X and Y for the entire group, and for each treatment group.
* Get the correlation between X and Y.
* Do a two-sample t-test to test whether the two treatment groups differ on X and Y.
* Do a paired t-test to test whether X and Y are statistically different from each other?
* Compare the number of subjects with each outcome by treatment group, using a chi-squared test.

These tasks are routine for a data set of this nature, and all of them could be easily done using any of the above listed statistical packages.

**General Issues**

**Enable the Analysis Tool Pak**

The Data Analysis Tool Pak is not installed with the standard Excel setup.  Look in the Tools menu.  If you do not have a Data Analysis item, you will need to install the Data Analysis tools.  Search Help for "Data Analysis Tools" for instructions.

**Missing Values**

A blank cell is the only way for Excel to deal with missing data.  If you have any other missing value codes, you will need to change them to blanks.

**Data Arrangement**

Different analyses require the data to be arranged in various ways.  If you plan on a variety of different tests, there may not be a single arrangement that will work.  You will probably need to rearrange the data several ways to get everything you need.

**Dialog Boxes**

Choose Tools/Data Analysis, and select the kind of analysis you want to do.  The typical dialog box will have the following items:  
     Range:  Type the upper left and lower right corner cells. e.g. A1:B100.  You can only choose adjacent rows and columns.  Unless there is a checkbox for grouping data by rows or columns (and there usually is not), all the data is considered as one glop.

Labels - There is sometimes a box you can check off to indicate that the first row of your sheet contains labels.  If you have labels in the first row, check this box, and your output MAY be labeled with your label.  Then again, it may not.

Output location - New Sheet is the default.  Or, type in the cell address of the upper left corner of where you want to place the output in the current sheet.  New Worksheet is another option, which I have not tried.  Ramifications of this choice are discussed below.

Other items, depending on the analysis.

**Output location**

The output from each analysis can go to a new sheet within your current Excel file (this is the default), or you can place it within the current sheet by specifying the upper left corner cell where you want it placed.  Either way is a bit of a nuisance.  If each output is in a new sheet, you end up with lots of sheets, each with a small bit of output.  If you place them in the current sheet, you need to place them appropriately; leave room for adding comments and labels; changes you need to make to format one output properly may affect another output adversely.  Example:  Output from Descriptive has a column of labels such as Standard Deviation, Standard Error, etc.  You will want to make this column wide to be able to read the labels.  But if a simple Frequency output is right underneath, then the column displaying the values being counted, which may just contain small integers, will also be wide.

**Results of Analyses**

**Descriptive Statistics**

The quickest way to get means and standard deviations for an entire group is using Descriptive in the Data Analysis tools. You can choose several adjacent columns for the Input Range (in this case the X and Y columns), and each column is analyzed separately. The labels in the first row are used to label the output, and the empty cells are ignored. If you have more, non-adjacent columns you need to analyze, you must repeat the process for each group of contiguous columns. The procedure is straightforward, can manage many columns reasonably efficiently, and empty cells are treated properly.

To get the means and standard deviations of X and Y for each treatment group requires the use of Pivot Tables (unless you want to rearrange the data sheet to separate the two groups).  After selecting the (contiguous) data range, in the Pivot Table Wizard's Layout option, drag Treatment to the Row variable area, and X to the Data area.  Double click on “Count of X” in the Data area, and change it to Average.  Drag X into the Data box again, and this time change Count to Steve.  Finally, drag X in one more time, leaving it as Count of X.  This will give us the Average, standard deviation and number of observations in each treatment group for X.  Do the same for Y, so we will get the average, standard deviation and number of observations for Y also.  This will put a total of six items in the Data box (three for X and three for Y). As you can see, if you want to get a variety of descriptive statistics for several variables, the process will get tedious.

A statistical package lets you choose as many variables as you wish for descriptive statistics, whether they are contiguous. You can get the descriptive statistics for all the subjects together, or broken down by a categorical variable such as treatment. You can select the statistics you want to see once, and it will apply to all variables chosen.

**Correlations**

Using the Data Analysis tools, the dialog for correlations is much like the one for descriptive - you can choose several contiguous columns, and get an output matrix of all pairs of correlations. Empty cells are ignored appropriately. The output does NOT include the number of pairs of data points used to compute each correlation (which can vary, depending on where you have missing data), and does not indicate whether any of the correlations are statistically significant. If you want correlations on non-contiguous columns, you would either must include the intervening columns, or copy the desired columns to a contiguous location.

A statistical package would permit you to choose non-contiguous columns for your correlations. The output would tell you how many pairs of data points were used to compute each correlation, and which correlations are statistically significant.

**Two-Sample T-test**

This test can be used to check whether the two treatment groups differ on the values of either X or Y. To do the test, you need to enter a cell range for each group. Since the data were not entered by treatment group, we first need to sort the rows by treatment. Be sure to take all the other columns along with treatment, so that the data for each subject remains intact. After the data is sorted, you can enter the range of cells containing the X measurements for each treatment. Do not include the row with the labels, because the second group does not have a label row. Therefore, your output will not be labeled to indicate that this output is for X. If you want the output labeled, you must copy the cells corresponding to the second group to a separate column, and enter a row with a label for the second group. If you also want to do the t-test for the Y measurements, you’ll need to repeat the process. The empty cells are ignored, and other than the problems with labeling the output, the results are correct.

A statistical package would do this task without any need to sort the data or copy it to another column, and the output would always be properly labeled to the extent that you provide labels for your variables and treatment groups. It would also allow you to choose more than one variable at a time for the t-test (e.g. X and Y).

**Paired t-test**

The paired t-test is a method for testing whether the difference between two measurements on the same subject is significantly different from 0. In this example, we wish to test the difference between X and Y measured on the same subject. The important feature of this test is that it compares the measurements within each subject. If you scan the X and Y columns separately, they do not look obviously different. But if you look at each X-Y pair, you will notice that in every case, x is greater than Y. The paired t-test should be sensitive to this difference. In the two cases where either X or Y is missing, it is not possible to compare the two measures on a subject. Hence, only 8 rows are usable for the paired t-test.

When you run the paired t-test on this data, you get a t-statistic of 0.09, with a 2-tail probability of 0.93. The test does not find any significant difference between X and Y. Looking at the output more carefully, we notice that it says there are 9 observations. As noted above, there should only be 8. It appears that Excel has failed to exclude the observations that did not have both X and Y measurements. To get the correct results copy X and Y to two new columns and remove the data in the cells that have no value for the other measure. Now re-run the paired t-test. This time the t-statistic is 6.14817 with a 2-tail probability of 0.000468. The conclusion is completely different!

Of course, this is an extreme example. But the point is that Excel does not calculate the paired t-test correctly when some observations have one of the measurements but not the other. Although it is possible to get the correct result, you would have no reason to suspect the results you get unless you are sufficiently alert to notice that the number of observations is wrong. There is nothing in online help that would warn you about this issue.

Interestingly, there is also a TTEST function, which gives the correct results for this example. Apparently, the functions and the Data Analysis tools are not consistent in how they deal with missing cells. Nevertheless, I cannot recommend the use of functions in preference to the Data Analysis tools, because the result of using a function is a single number - in this case, the 2-tail probability of the t-statistic. The function does not give you the t-statistic itself, the degrees of freedom, or any number of other items that you would want to see if you were doing a statistical test.

A statistical package will correctly exclude the cases with one of the measurements missing, and will provide all the supporting statistics you need to interpret the output.

**Crosstabulation and Chi-Squared Test of Independence**

Our final task is to count the two outcomes in each treatment group, and use a chi-square test of independence to test for a relationship between treatment and outcome.  To count the outcomes by treatment group, you need to use Pivot Tables. In the Pivot Table Wizard's Layout option, drag Treatment to Row, Outcome to Column and to Data.  The Data area should say "Count of Outcome" – if not, double-click on it and select "Count". If you want percent’s, double-click "Count of Outcome", and click Options; in the “Show Data As” box which appears, select "% of row".  If you want both counts and percent’s, you can drag the same variable into the Data area twice, and use it once for counts and once for percent’s.

Getting the chi-square test is not so simple, however. It is only available as a function, and the input needed for the function is the observed counts in each combination of treatment and outcome (which you have in your pivot table), and the expected counts in each combination. Expected counts? What are they? How do you get them? If you have sufficient statistical background to know how to calculate the expected counts, and can do Excel calculations using relative and absolute cell addresses, you should be able to navigate through this. If not, you’re out of luck.

If you surmounted the problem of expected counts, you can use the Chicest function to get the probability of observing a chi-square value bigger than the one for this table. Again, since we are using functions, you do not get many other necessary pieces of the calculation, notably the value of the chi-square statistic or its degrees of freedom.

No statistical package would require you to provide the expected values before computing a chi-square test of independence. Further, the results would always include the chi-square statistic and its degrees of freedom, as well as its probability. Often you will get some additional statistics as well.

**Additional Analyses**

The remaining analyses were not done on this data set, but some comments about them are included for completeness.

**Simple Frequencies**

You can use Pivot Tables to get simple frequencies.  (see Crosstabulations for more about how to get Pivot Tables.)  Using Pivot Tables, each column is considered a separate variable, and labels in row 1 will appear on the output.  You can only do one variable at a time.  
   
Another possibility is to use the Frequencies function.  The main advantage of this method is that once you have defined the frequencies function for one column, you can use Copy/Paste to get it for other columns.  First, you will need to enter a column with the values you want counted (bins).  If you intend to do the frequencies for many columns, be sure to enter values for the column with the most categories.  e.g., if 3 columns have values of 1 or 2, and the fourth has values of 1,2,3,4, you will need to enter the bin values as 1,2,3,4. Now select enough empty cells in one column to store the results - 4 in this example, even if the current column only has 2 values.  Next choose Insert/Function/Statistical/Frequencies on the menu.  Fill in the input range for the first column you want to count using relative addresses (e.g. A1:A100).  Fill in the Bin Range using the absolute addresses of the locations where you entered the values to be counted (e.g. $M$1: $M$4). Click Finish. Note the box above the column headings of the sheet, where the formula is displayed.  It starts with "= FREQUENCIES (".  Place the cursor to the left of the = sign in the formula, and press Ctrl-Shift-Enter.  The frequency counts now appear in the cells you selected.

To get the frequency counts of other columns, select the cells with the frequencies in them, and choose Edit/Copy on the menu. If the next column you want to count is one column to the right of the previous one, select the cell to the right of the first frequency cell, and choose Edit/Paste (Ctrl-V).  Continue moving to the right and pasting for each column you want to count.  Each time you move one column to the right of the original frequency cells, the column to be counted is shifted right from the first column you counted.

If you want percent’s as well, you’ll have to use the Sum function to compute the sum of the frequencies, and define the formula to get the percent for one cell.  Select the cell to store the first percent, and type the formula into the formula box at the top of the sheet - e.g. = N1\*100/N$5 - where N1 is the cell with the frequency for the first category, and N5 is the cell with the sum of the frequencies.  Use Copy/Paste to get the formula for the remaining cells of the first column.  Once you have the percent for one column, you can Copy/Paste them to the other columns.  You’ll need to be careful about the use of relative and absolute addresses! In the example, above, we used N$5 for the denominator, so when we copy the formula down to the next frequency on the same column, it will still look for the sum in row 5; but when we copy the formula right to another column, it will shift to the frequencies in the next column.

Finally, you can use Histogram on the Data Analysis menu. You can only do one variable at a time.  As with the Frequencies function, you must enter a column with "bin" boundaries. To count the number of occurrences of 1 and 2, you need to enter 0,1,2 in three adjacent cells, and give the range of these three cells as the Bins on the dialog box.    The output is not labeled with any labels you may have in row 1, nor even with the column letter.   If you do frequencies on lots of variables, you will have difficulty knowing which frequency belongs to which column of data.

**Linear Regression**

Since regression is one of the more frequently used statistical analyses, we tried it out even though we did not do a regression analysis for this example. The Regression procedure in the Data Analysis tools lets you choose one column as the dependent variable, and a set of contiguous columns for the independents. However, it does not tolerate any empty cells anywhere in the input ranges, and you are limited to 16 independent variables. Therefore, if you have any empty cells, you will need to copy all the columns involved in the regression to new columns, and delete any rows that contain any empty cells. Large models, with more than 16 predictors, cannot be done at all.

**Analysis of Variance**

In general, the Excel's ANOVA features are limited to a few special cases rarely found outside textbooks, and require lots of data re-arrangements.

**One-way ANOVA**

Data must be arranged in separate and adjacent columns (or rows) for each group.  Clearly, this is not conducive to doing 1-ways on more than one grouping.  If you have labels in row 1, the output will use the labels.

**Two-Factor ANOVA Without Replication**

This only does the case with one observation per cell (i.e. no Within Cell error term).  The input range is a rectangular arrangement of cells, with rows representing levels of one factor, columns the levels of the other factor, and the cell contents the one value in that cell.   
**Requesting Many Analyses**

If you had a variety of different statistical procedures that you wanted to perform on your data, you would almost certainly find yourself doing a lot of sorting, rearranging, copying and pasting of your data. This is because each procedure requires that the data be arranged in a way, often different from the way another procedure wants the data arranged. In our small test, we had to sort the rows to do the t-test, and copy some cells to get labels for the output. We had to clear the contents of some cells to get the correct paired t-test, but did not want those cells cleared for some other test. And we were only doing five tasks. It does not get better when you try to do more. There is no single arrangement of the data that would allow you to do many different analyses without making many different copies of the data. The need to manipulate the data in many ways greatly increases the chance of introducing errors.

Using a statistical program, the data would normally be arranged with the rows representing the subjects, and the columns representing variables (as they are in our sample data). With this arrangement, you can do any of the analyses discussed here, and many others as well, without having to sort or rearrange your data in any way. Only much more complex analyses, beyond the capabilities of Excel and the scope of this article would require data rearrangement.

**Working with Many Columns**

What if your data had not 4, but 40 columns, with a mix of categorical and continuous measures? How easily do the above procedures scale to a larger problem?

At best, some of the statistical procedures can accept multiple contiguous columns for input, and interpret each column as a different measure. The descriptive and correlations procedures are of this type, so you can request descriptive statistics or correlations for many continuous variables, if they are entered in adjacent columns. If they are not adjacent, you need to rearrange columns or use copy and paste to make them adjacent.

Many procedures, however, can only be applied to one column at a time. T-tests (either independent or paired), simple frequency counts, the chi-square test of independence, and many other procedures are in this class. This would become a serious drawback if you had more than a handful of columns, even if you use cut and paste or macros to reduce the work. In addition to having to repeat the request many times, you must decide where to store the results of each, and make sure it is properly labeled so you can easily locate and identify each output.

Finally, Excel does not give you a log or other record to track what you have done. This can be a serious drawback if you want to be able to repeat the same (or similar) analysis in the future, or even if you’ve simply forgotten what you’ve already done.

Using a statistical package, you can request a test for as many variables as you need at once. Each one will be properly labeled and arranged in the output, so there is no confusion as to what’s what. You can also expect to get a log, and often a set of commands as well, which can be used to document your work or to repeat an analysis without having to go through all the steps again.

**Summary**

Although Excel is a fine spreadsheet, it is not a statistical data analysis package. In all fairness, it was never intended to be one. Keep in mind that the Data Analysis Tool Pak is an "add-in" - an extra feature that enables you to do a few quick calculations. So, it should not be surprising that that is just what it is good for - a few quick calculations. If you attempt to use it for more extensive analyses, you will encounter difficulties due to any or all the following limitations:

* Potential problems with analyses involving missing data. These can be insidious, in that the unwary user is unlikely to realize that anything is wrong.
* Lack of flexibility in analyses that can be done due to its expectations regarding the arrangement of data. This results in the need to cut/paste/sort/ and otherwise rearrange the data sheet in various ways, increasing the likelihood of errors.
* Output scattered in many different worksheets, or all over one worksheet, which you must take responsibility for arranging in a sensible way.
* Output may be incomplete or may not be properly labeled, increasing possibility of misidentifying output.
* Need to repeat requests for some analyses multiple times to run it for multiple variables, or to request multiple options.
* Need to do some things by defining your own functions/formulae, with its attendant risk of errors.
* No record of what you did to generate your results, making it difficult to document your analysis, or to repeat it later, should that be necessary.

If you have more than about 10 or 12 columns, and/or want to do anything beyond descriptive statistics and perhaps correlations, you should be using a statistical package. There are several suitable ones available by site license through OIT, or you can use them in any of the OIT PC labs. If you have Excel on your own PC, and don’t want to pay for a statistical program, use Excel to enter the data (with rows representing the subjects, and columns for the variables). All the mentioned statistical packages can read Excel files, so you can do the (time-consuming) data entry at home, and go to the labs to do the analysis

## 外文资料翻译—原文部分

**使用Excel进行统计数据分析——注意事项**

**摘要**

我们使用Excel进行一些基本的数据分析任务，以查看是否是相同任务使用统计软件包的合理替代方法。 我们得出结论，Excel是除了教科书的例子，最简单的描述性统计数据之外的统计分析，或者是少于几列的统计分析的不良选择。 我们遇到的问题导致了以下四个总结：

* 对处理缺失值的处理方法不一致,有时不正确。
* 数据组织根据分析不同，如果您想进行许多不同的分析，迫使您以多种方式重新组织数据。
* 许多分析只能一次在一列中进行，因此对多列进行相同的分析是不方便的。
* 输出组织得很糟糕,有时标记不足,没有记录的分析是如何完成的。

Excel对于数据输入非常方便，并且在统计分析之前快速地处理行和列。然而,当你准备做统计分析,我们建议使用SAS等统计软件包SPSS占据,午睡或一款统计软件。

**介绍**

Excel可能是pc最常用的电子表格。新买的电脑通常都是用Excel下载的。它很容易被用于进行各种计算，包括统计函数的集合，以及一个数据分析工具Pak。因此，如果你突然发现你需要做一些统计分析，你可能会把它当作一个显而易见的选择。我们决定做一些测试，看看Excel是如何作为一个数据分析应用程序的。

为了展示结果，我们将使用一个小示例。本例中的数据是虚构的。它被选择有两个分类和两个连续变量，这样我们就可以测试各种基本的统计技术。因为几乎所有的真实数据集至少有一些缺失的数据点,既然能够正确处理缺失数据的特点之一,就是我们认为理所当然的在统计分析中包中,我们介绍了两个空细胞数据:

电子表格的每一行都代表一个主题。第一个主题接受治疗1，结果1。X和Y是两个测量值的值。我们无法在第二个主题上得到Y的测量值，也无法得到最后一个题目的X，所以这些单元格是空的。这些数据被输入到数据可用的顺序中，因此数据不是以任何方式排序的。

我们使用这些数据做一些简单的分析，并将结果与标准的统计软件包进行比较。这个比较考虑了结果的准确性，以及接口可以用于更大的数据集——也就是更多的列。我们使用SPSS作为标准，尽管OIT支持的任何统计软件包都能在这个目的上做得很好。在这篇文章中,当我们说“我们的意思是SPSS统计软件包,”,SAS占据,SYSTAT或者一款统计软件。

大多数优秀的统计程序都是数据分析工具包的一部分，它在工具菜单中。它包括各种选择,包括简单的描述性统计,t,相关性,1或双向方差分析、回归,等等。如果你没有一个数据分析工具菜单上的项目,您需要安装Pak数据分析工具。搜索“数据分析工具”，以获得关于装载工具的说明。

另外两个Excel特性对于某些分析是很有用的，但是数据分析工具包是唯一能够提供相当完整的统计意义测试的工具。数据透视表数据菜单中可以用来生成汇总表的意思,标准差,计数,等等。同时,你可以用函数来生成一些统计的措施,如相关系数。函数生成一个单一的数字，因此使用函数可能需要将位和片段组合起来才能得到您想要的结果。即便如此，您可能无法生成完整的分析所需的所有部分。

另外两个Excel特性对于某些分析是有用的，但是除非另有说明，所有使用Excel的统计测试都是用数据分析工具Pak完成的。为了检查各种统计测试，我们选择了以下任务:

* 得到X和Y的均值和标准差为整个集团,和每个治疗组。
* 得到X和Y之间的相关性。
* 做一个两个示例学习测试是否两个处理组不同X和Y的结果是否相同。
* 做一个配对t检验测试是否X和Y是统计上的不同吗?
* 比较对象的数量和每个结果处理组,使用卡方测试。

这些任务对于这种性质的数据集来说是例行程序，所有这些任务都可以轻松地使用上面列出的统计软件包。

**一般问题**

**启用分析工具Pak**

数据分析工具Pak没有安装在标准的Excel设置中。查看工具菜单。如果您没有一个数据分析项目，您将需要安装数据分析工具。搜索帮助“数据分析工具”用于指导。

**缺省值**

一个空白单元格是Excel处理丢失数据的唯一方法。如果您有任何其他丢失的值代码，您将需要将它们更改为空。

**数据排列**

不同的分析要求数据以不同的方式进行排列。如果您计划进行各种不同的测试，可能不会有一个单独的安排可以工作。您可能需要重新排列数据以获得所需的所有信息。

**对话框**

选择工具/数据分析，并选择您想要做的分析。典型的对话框将会有以下项目:

输入范围:输入左上角和右下角的单元格。例如:A1:B100。您只能选择相邻的行和列。除非是分组数据行或列的复选框(通常是没有),所有的数据被认为是一个难吃的东西。

标签——有时你可以检查一个盒子，以表明你的表的第一行包含标签。如果您在第一行有标签，请检查这个框，并且您的输出可能被标记为您的标签。然而，它可能不会。

输出位置——新表是默认值。或者，输入你想要将输出放在当前表格的左上角的单元地址。新的工作表是另一种选择，我没有尝试过。下面将讨论这一选择的影响。

其他项目，取决于分析。

**输出的位置**

每个分析的输出可以去一个新表在你当前的Excel文件(这是默认的),或者你可以把它在当前表左上角单元格指定你想要放置的地方。这两种方式都有点令人讨厌。如果每个输出都在一个新表中，那么就会得到大量的表，每个输出都有少量的输出。如果你把他们在当前的表,你需要把他们适当,离开房间添加评论和标签,更改你需要做出正确格式一个输出可能会影响另一个输出不利。示例:来自描述性的输出有一列标签，例如标准偏差、标准错误等。您将希望使这个列宽，以便能够读取标签。但是，如果一个简单的频率输出就在下面，那么显示被计数的值的列，它可能只包含小整数，也将是宽的。

**分析的结果**

**描述性统计**

对于整个团队来说，获得方法和标准偏差的最快方法是在数据分析工具中使用描述性的方法。您可以为输入范围选择几个相邻的列(在本例中为X和Y列)，并且分别对每个列进行分析。第一行中的标签用于对输出进行标记，而空的单元则被忽略。如果您需要分析更多的非相邻列，则必须为每组连续的列重复这个过程。这个过程很简单，可以有效地管理许多列，并且对空单元进行适当的处理。

为了得到每个治疗组的X和Y的均值和标准偏差，需要使用主数据表(除非您想重新安排数据表来分隔这两个组)。在选择(连续)数据范围后，在Pivot表向导的布局选项中，将处理拖到行变量区域，并将X拖到数据区域。在数据区域双击“X的计数”，并将其更改为平均值。再次将X拖到数据框中，这次更改为Steve。最后,拖动X在一个更多的时间,把它作为计数的X这将给我们的平均水平,标准偏差和观测的数量在每个治疗组为Y X做同样的事情,所以我们会得到平均值,标准差和Y的观测数量。这将总共6个条目的数据框(3 X和Y)为3分。正如你所看到的,如果你想获得各种描述性统计的几个变量,这个过程会乏味。

一个统计软件包可以让您选择尽可能多的变量，无论它们是连续的。你可以把所有科目的描述性统计数据，或者分类变量，比如治疗，分类。您可以选择您想要查看的统计信息，它将应用于所选择的所有变量。

**相关性**

使用数据分析工具，关联的对话框与描述性的对话框非常相似——你可以选择几个连续的列，并得到所有的相关系数的输出矩阵。空单元被适当地忽略。输出并不包括对数据点的数量用于计算每个相关(可以不同,这取决于你有缺失的数据),并且不显示任何相关性是否显著。如果您想要在非连续的列上进行关联，那么您要么必须包含中间的列，要么将所需的列复制到一个连续的位置。

一个统计包将允许您为相关性选择非连续的列。输出会告诉你用了多少个数据点来计算每一种相关性，而这些相关性在统计学上是显著的。

**两个示例学习任务**

这个测试可以用来检查两个治疗组在X或y值上是否有不同，为了进行测试，您需要为每个组输入一个单元格范围。由于数据没有被治疗组输入，我们首先需要通过治疗来对这些行进行排序。确保将所有其他的列和处理一起处理，这样每个主题的数据都保持不变。在对数据进行排序后，您可以输入包含每个处理的X度量的单元范围。不要将行与标签包含在一起，因为第二组没有标签行。因此,输出将不会标记,以表明该输出是x。如果你想要输出标签,你必须对应于第二组的细胞复制到一个单独的列中,并输入一行与第二组的标签。如果您还想对Y度量进行t测试，那么您需要重复这个过程。空单元被忽略，除了标记输出的问题之外，结果是正确的。

统计软件包将这个任务没有任何需要或将其复制到另一个数据列进行排序,并输出总是正确标记,你提供标签变量和治疗组。它还将允许您选择一个以上的变量在时间t检验(例如X和Y)。

**配对t检验**

成对的t-test是一种测试方法，用于测试同一主题上的两个度量之间的差异是否显著不同于0。在本例中，我们希望测试在同一主题上测量的X和Y之间的差异。这个测试的一个重要特性是它比较了每个主题的度量。如果分别扫描X和Y列，它们看起来不会明显不同。但如果你看看每个y,你会发现在每一情况下,X大于y的配对t检验对这种差异应该敏感。在这两种情况中，如果X或Y缺失，就不可能比较这两种方法。因此,只有8行可用配对t检验。

当您运行的配对t检验数据,得到t统计量为0.09,2-tail概率为0.93。这个测试并没有发现X和y之间有什么显著的差异，更仔细地观察输出结果，我们注意到有9个观察结果。如上所述，应该只有8个。似乎Excel没有排除那些没有X和Y测量值的观测结果。为了得到正确的结果，将X和Y复制到两个新的列中，并删除那些对其他度量没有任何价值的单元格中的数据。现在重新运行配对t检验。这一次的t统计量是6.14817 2-tail概率为0.000468。结论是完全不同的！

当然，这是一个极端的例子。但重点是，当一些观测结果有一个测量值而不是另一个测量值时，Excel不能正确计算配对t测试。虽然有可能得到正确的结果，但是如果你没有足够的警觉，注意到观察的数量是错误的，你就没有理由怀疑你得到的结果。在网上的帮助中，没有什么可以警告你这个问题。

有趣的是，这里也有一个最简单的函数，它给出了这个示例的正确结果。显然，函数和数据分析工具在处理丢失的单元时是不一致的。不过,我不建议使用的函数优先于数据分析工具,因为使用一个函数的结果是一个数字——在这种情况下,该2-tail t统计量的概率。这个函数没有给出t统计量，自由度，或者其他任何你想要看的东西如果你在做一个统计测试的话。

一个统计包将正确地排除缺少的一个度量的情况，并将提供您需要的所有支持统计数据来解释输出。

**Crosstabulation独立和卡方测试**

我们的最后一项任务是计算每个治疗组的两个结果，并使用卡方测试独立性来测试治疗和结果之间的关系。要计算治疗组的结果，您需要使用Pivot表。在Pivot表向导的布局选项中，将处理拖到行、结果到列和数据。数据区域应该说“计数结果”——如果不是,双击它,选择“数”。如果你想要的百分比,双击“计数的结果”,并单击选项;“显示数据”框中,选择“行%”。如果您想要两个计数和%的值，您可以将同一个变量拖到数据区域两次，然后使用一次，一次用于计数，一次用于百分数。

然而，获得卡方测试并不是那么简单。只是作为一个函数,输入所需的函数是观察到的数量在每个组合治疗和结果(数据透视表中),和预期的数量在每个组合。预期的数量?它们是什么?你是怎么得到它们的?如果您有足够的统计背景，知道如何计算预期计数，并且可以使用相对和绝对的单元地址进行Excel计算，那么您应该能够导航到这一点。如果没有，那你就倒霉了。

如果你超越预期数量的问题,你可以使用最时髦的功能观察卡方值的概率大于这个表。同样，由于我们使用的是函数，所以没有得到很多其他必要的计算量，特别是卡方统计量的值或者它的自由度。

在计算一个卡方的独立性测试之前，没有任何统计软件包要求您提供预期的值。此外，结果还将包括卡方统计数据及其自由度，以及它的概率。通常你也会得到一些额外的统计数据。

**额外的分析**

其余的分析没有在这个数据集上完成，但是关于它们的一些注释包含在完整性中。

**简单的频率数据分析**

你可以使用透视表来获得简单的频率。(见Crosstabulations更多关于如何获得数据透视表。)使用主表，每个列都被认为是一个单独的变量，而行1中的标签将出现在输出中。你一次只能做一个变量。

另一种可能是使用频率函数。这种方法的主要优点是，一旦定义了一个列的频率函数，就可以使用复制/粘贴来获取其他列。首先，您需要输入您想要计算的值(垃圾箱)的列。如果您打算为许多列执行频率，请确保以大多数类别为列输入值。如。,如果3列有值的1或2,第四个值1,2,3,4,您将需要输入本值为1,2,3,4。现在，在一个列中选择足够的空单元来存储结果——在本例中是4，即使当前列只有2个值。下一个选择插入菜单上/功能/统计/频率。填写你想要的第一列的输入范围计算使用相对地址(例如A1:A100)。填写本范围使用绝对地址的位置输入值计算(例如$ M $ 1:M 4美元)。单击Finish。请注意该表格的列标题上方的方框，其中显示了公式。它始于“=频率(”。把光标放在等号的左边的公式,并按Ctrl-Shift-Enter。频率计数现在出现在你选择的细胞中。

要获得其他列的频率计数，选择带有频率的单元格，并在菜单上选择编辑/复制。在下一篇专栏文章如果你想计数是一列右边的前一个,选择细胞的右边第一个频率细胞,并选择编辑/粘贴(ctrl - v)。继续向右移动，并粘贴到您想要计算的每一列。每次将一个列移动到原始频率单元格的右边时，要计算的列将从您计算的第一个列中移动。

如果你想要%的值，你必须使用Sum函数来计算频率的和，并定义一个公式来得到一个单元格的百分比。选择细胞存储第一个百分比,并在公式框中输入公式的表-例如= N1 \* 100 / N 5美元——N1与频率为第一类细胞,它们是细胞频率之和。使用复制/粘贴来获得第一列中剩余单元格的公式。一旦你有了一个列的百分比，你就可以复制/粘贴到其他的列。您需要小心使用相对和绝对地址！在上面的示例中,我们使用N 5美元为分母,所以当我们复制到下一个频率的公式在同一列,它仍然会寻找在第五排总和;但当我们将正确的公式复制到另一个列,它将在下一篇专栏文章中频率。

最后，你可以在数据分析菜单上使用Histogram。你一次只能做一个变量。与频率函数一样，您必须输入一个带有“bin”边界的列。数出现的次数1和2中,您需要输入0,1,2在三个相邻的细胞,并给出这三种细胞的范围为对话框上的垃圾箱。输出没有被标记为第1行中的任何标签，甚至不包括列字母。如果你在很多变量上做频率，你就很难知道哪个频率属于

**线性回归**

由于回归是比较常用的统计分析之一，所以我们尝试了它，尽管我们没有对这个例子进行回归分析。数据分析工具中的回归过程允许您选择一个列作为独立变量，并为独立变量选择一组连续的列。但是，它不能容忍输入范围内的任何空单元，并且您只能使用16个独立变量。因此，如果您有任何空单元，您将需要复制回归到新列中的所有列，并删除包含任何空单元的任何行。有超过16个预测因子的大型模型根本就不可能完成。

**概要**

尽管Excel是一个很好的电子表格，但它不是一个统计数据分析包。公平地说，它从来没有打算成为一个。记住,数据分析工具Pak是一个“插件”一个额外的功能,使您可以做一些快速的计算。所以，这并不奇怪，这就是它的好处一些快速计算。如果您尝试使用它进行更广泛的分析，您将遇到困难，因为任何或所有以下的限制:

* 分析涉及缺失数据的潜在问题。这些可能是阴险的，因为粗心大意的用户不太可能意识到任何事情都是错误的；
* 在分析缺乏灵活性,可以由于其预期完成有关数据的安排。这导致需要剪切/粘贴/排序/否则以不同的方式重新排列数据表,增加了出错的可能性。
* 输出分散在许多不同的工作表,或在一个工作表,你必须负责安排在一个合理的方式。
* 输出可能不完整或可能不是适当的标签,增加识别错输出的可能性。
* 需要重复多次请求一些分析运行多个变量,或请求多个选项。
* 需要做一些事情通过定义自己的函数/公式,伴随着它的错误的风险。
* 没有记录你生成你的结果,使得它难以记录分析,或重复以后,应该是必要的。

如果您有超过10或12列的列，并且/或想要做超出描述统计和相关性的任何事情，那么您应该使用一个统计软件包。有几个合适的网站通过OIT提供网站许可证，或者你可以在任何OIT PC实验室使用它们。如果你有Excel在自己的个人电脑,不想支付统计程序,使用Excel来输入数据(代表研究对象行,列变量)。所有提到的统计包可以读取Excel文件,所以你可以在家做(时间)的数据条目,并去实验室做分析。

**参考文献**

[1] 作者Eva Goldwater 题目Using Excel for Statistical Data Analysis - Caveats