临床研究论文图表制作入门

EmpowerStats 临床科研设计与数据分析
系列教程 1

共 1 小时 15 分钟
配套免费视频课
快速开启临床研究新世界的大门

扫码参加

陈星霖/ **EmpowerStats** 技术团队

http://www.empowerstats.net/empowerU/

2017 年 11 月第一版；2021 年 7 月 24 日更新版

【适用人群】

零基础起步的学友，接受新知识最快

不限专业与学历

【学习目标】

完成 SCI 论文常用统计分析图表制作方法
了解预测模型和机器学习等图表制作方法

【课程介绍】

- ✓ **极简教程**：多年实践经验的总结。常规 3 个月到半年的学习目标**在 1 小时内完成。**
- ✓ **跨入一流**：如果你是数据分析的新手，对很多统计方法不了解，不会编程，对分析思路掌握也不多，**让你登上巨人的肩膀，很快达到一流水平。**
- ✓ **快速飞越**：如果你有多年数据分析的经验，有很强的统计功底与编程技巧，熟练使用 SAS、R、SPSS 等统计软件，**本课程让你如虎添翼**，你的数据分析将更快、更深入、更全面。
- ✓ **接轨国际**：现代临床流行病学科研设计与数据分析思路，是发表高质量 SCI 论文的基础。
- ✓ **实例与模板化教学：零基础学友也可做出全部图表。**

【课前准备】

1. **教材准备：本课程是网易云视频课（扫首页二维码可免费参加），课时 1 处下载文字版教程（不定期更新）。方法：电脑登录后在视频左下角“参考资料”处下载（数据、文献教程等）。**

2. 软件准备：采用易佰统计软件演示。该软件可以免费下载，基本模块终身免费，研究生毕业或发表 3-5 分 SCI 论文，**用免费版足矣。本课程 90%内容均用免费版易佰软件完成操作。**高级模块需要付费购买（200 元/月，优惠价格请参考软件官网）。新用户注册后有完整版权限，完整版到期后：点击分析项目，创建新项目，就可以终身免费使用基本模块。易佰软件基于 R 语言，获得国际认可，已发表 SCI 论文三百余篇。

临床研究论文图表制作入门

1 让统计分析变得简单：基于 R 语言的全自动分析工具.....3
1.1 第一步：免费下载安装注册易侕软件.....3
1.2 第二步：用软件自带的练习数据创建分析项目.....3
1.3 第三步：开始统计分析，一键实现“各种组间比较”.....5
1.3.1 操作小技巧“shift”或“ctrl”键全选或单选.....5
1.3.2 巧用变量注解文件，自动完成图表注解.....6
2 站在巨人肩膀上：关联分析四个表的制作.....8
3 一图抵万言：平滑曲线拟合图的制作.....10
4 随访数据的生存分析（K-M 曲线、Cox 模型分析）.....11
4.1 第一步：下载练习数据创建分析项目.....11
4.2 第二步：明确分析思路.....11
4.3 第三步：操作 “全自动流行病学分析”模块.....11
5 预测模型（列线图、ROC 曲线、校准曲线）制作（了解）.....13
5.1 第一步：下载练习数据创建分析项目.....13
5.2 第二步：明确分析思路.....13
5.3 第三步：操作“预测模型与 ROC 分析模块”.....13
6 机器学习系列图表制作（了解）.....16
7 软件常用基本操作.....20
7.1 连续变量分组（等分、按指定切点分）.....20
7.2 小操作技能：分组切点的具体数值如何查看呢？.....20
7.3 分类变量取值重编码.....21
7.4 设置小数点位数.....21
7.5 赋值错误更正.....22
7.6 选择分析人群与输出数据文件.....23
8 格式说明《数据》和《变量注解》.....24
8.1 《数据》文件格式要求.....24
8.2 《变量注解》文件要求.....25

1　让统计分析变得简单：基于 R 语言的全自动分析工具

课程适合首次接触易侕统计软件的学友。带你走进一个新世界，体会让数据分析变得如此简单。

1.1　第一步：免费下载安装注册易侕软件

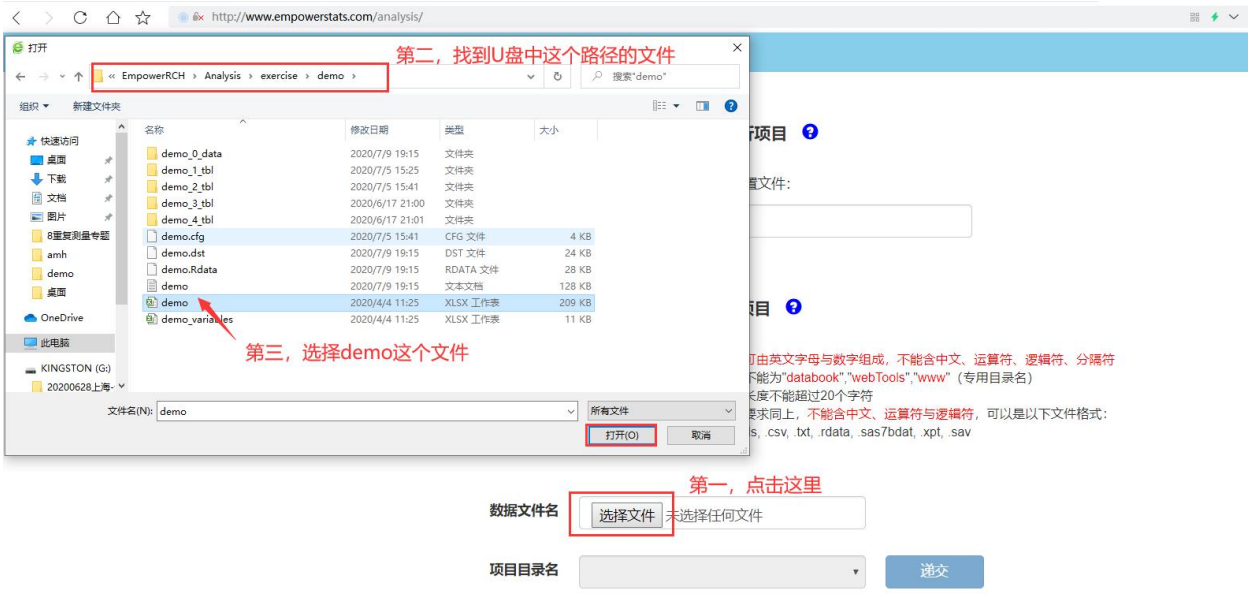
下载网址 http://www.empowerstats.net/cn/

安装 3.0 版。按官网流程安装，如需帮助请见附件《易侕软件安装说明》。后台是 R 语言程序，获得国际认可，已有多篇 SCI 论文发表。

1.2　第二步：用软件自带的练习数据创建分析项目

点击新建分析项目下的“选择文件”：

找到 EmpowerRCH-Analysis-exercise-demo 文件夹的 demo.XLS 文件，打开。

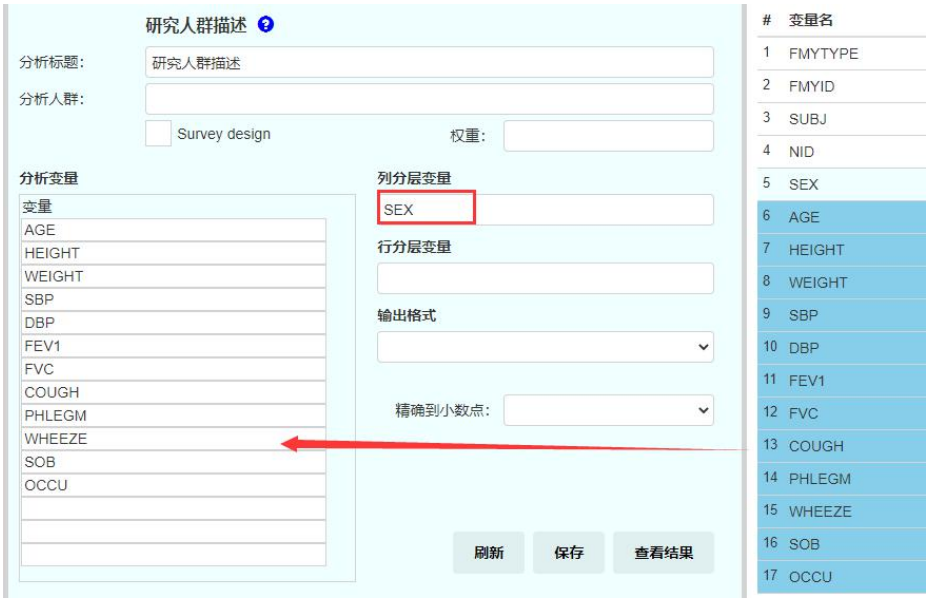


会自动填充“项目目录名”，点击递交：

数据就读取成功了：

1.3 第三步：开始统计分析，一键实现“各种组间比较”

选“数据分析”-“论文图表”的“研究人群描述”模块。要分析的变量拖入“分析变量”。SEX 拖入“列分层变量”。点击“查看结果”。



1.3.1 操作小技巧“shift”或“ctrl”键全选或单选

选择多个变量：点击变量名，按住“shift”键选择另外一个变量名，这样两个变量名之间的变量都被自动选择。如要选择的变量不是相连在一起的，按住“ctrl”键选择下一个变量名，继续按住“ctrl”键可以继续选择下一个变量。

表一就这样做成了！易侃软件自动识别变量类型，连续变量得出均数、标准差；分类变量得出例数百分比。自动做组间比较，得出 p 值。下表全选复制到 word 中简单编辑，就可以用于发表了。如果对结果解读有疑问，可以看系列教程 3 的图表结果解读部分。

SEX	1	2	Standardize diff.	P-value	P-value*
N	417	415			
AGE	38.4 ± 14.7	37.3 ± 13.1	0.1 (-0.1, 0.2)	0.242	0.481
HEIGHT	1.6 ± 0.1	1.5 ± 0.1	2.0 (1.8, 2.2)	<0.001	<0.001
WEIGHT	56.4 ± 7.1	50.4 ± 6.8	0.9 (0.7, 1.0)	<0.001	<0.001
SBP	132.1 ± 22.0	128.5 ± 23.0	0.2 (0.0, 0.3)	0.026	<0.001
DBP	69.6 ± 11.7	69.3 ± 11.0	0.0 (-0.1, 0.2)	0.725	0.937
FEV1	3.8 ± 1.3	3.1 ± 0.8	0.8 (0.6, 0.9)	<0.001	<0.001
FVC	5.1 ± 1.1	3.9 ± 0.8	1.3 (1.1, 1.5)	<0.001	<0.001
COUGH			0.3 (0.2, 0.4)	<0.001	-
0	338 (81.2%)	376 (91.5%)			
1	78 (18.8%)	35 (8.5%)			
PHLEGM			0.2 (0.1, 0.4)	<0.001	-
0	322 (77.4%)	355 (86.4%)			
1	94 (22.6%)	56 (13.6%)			
WHEEZE			0.1 (-0.1, 0.2)	0.321	-
0	360 (86.5%)	365 (88.8%)			
1	56 (13.5%)	46 (11.2%)			
SOB			0.2 (0.0, 0.3)	0.016	-
0	311 (74.8%)	276 (67.2%)			
1	105 (25.2%)	135 (32.8%)			
OCCU			0.0 (-0.1, 0.2)	0.512	-
0	197 (47.4%)	204 (49.6%)			
1	219 (52.6%)	207 (50.4%)			

表中结果: Mean+SD / N(%)
P值*: 如是连续变量，用Kruskal Wallis秩和检验得出, 如计数变量有理论数<10，用Fisher精确概率检验得出.

1.3.2 巧用变量注解文件，自动完成图表注解

上表还有问题：变量名 AGE、BMI 分别代表什么？SMOKE： 0、1 分别代表什么？不是很清楚。可以用变量注解文件，如果投英文期刊，注解成英文的；如果投中文，就注解成中文的。

	A	B	C
1	Variable Name	Code	Label
2	FMYTYPE		Family type
3		0	0
4		1	1
5	FMYID		Family ID
6	SUBJ		Subject ID
7	NID		NID
8		1	1
9		2	2
10		11	11
11		12	12
12		13	13
13		14	14
14		15	15
15		16	16
16		17	17
17	SEX		Gender
18		1	Male
19		2	Female
20	AGE		AGE
21	HEIGHT		HEIGHT
22	WEIGHT		WEIGHT
23	SBP		Systolic blood pressure
24	DBP		Diastolic blood pressure
25	FEV1		FEV1
26	FVC		FVC
27	COUGH		COUGH
28		0	No
29		1	Yes
30	PHLEGM		PHLEGM
31		0	No
32		1	Yes
33	WHEEZE		WHEEZE
34		0	No
35		1	Yes

变量注解文件包括三列：变量名（Variable Name）、取值编码（Code）和注解（注解）。例如性别 1 是男、2 是女；SBP 是连续变量，就直接写名称 Systolic blood pressure。

操作：点击**数据操作-读入/导入变量注解文件**，重新运行研究人群模块就注解好了。

研究人群描述

Gender	Male	Female	Standardize diff.	P-value	P-value*
N	417	415			
AGE	38.44 ± 14.67	37.31 ± 13.13	0.08 (-0.05, 0.22)	0.242	0.481
HEIGHT	1.64 ± 0.06	1.53 ± 0.05	2.01 (1.84, 2.18)	<0.001	<0.001
WEIGHT	56.44 ± 7.08	50.39 ± 6.77	0.87 (0.73, 1.02)	<0.001	<0.001
Systolic blood pressure	132.09 ± 22.01	128.52 ± 22.96	0.16 (0.02, 0.30)	0.026	<0.001
Diastolic blood pressure	69.62 ± 11.67	69.34 ± 10.95	0.02 (-0.11, 0.16)	0.725	0.937
FEV1	3.84 ± 1.25	3.05 ± 0.77	0.76 (0.61, 0.91)	<0.001	<0.001
FVC	5.14 ± 1.15	3.89 ± 0.76	1.29 (1.13, 1.45)	<0.001	<0.001
COUGH			0.30 (0.16, 0.44)	<0.001	-
No	338 (81.25%)	376 (91.48%)			
Yes	78 (18.75%)	35 (8.52%)			

2　站在巨人肩膀上：关联分析四个表的制作

“全自动流行病学分析”模块，轻松实现论文中常用的图表。

首先明确分析思路：

- 结局变量 Y: SBP（收缩压） 、DBP（舒张压）
- 危险因素 X: BMI （体重指数）
- 混杂因素 Z: SEX（性别）、AGE（年龄）、PSMK（被动吸烟）、 ALH（饮酒）

操作：还是用上一节的数据，选择“数据分析”-“论文图表”-“全自动快速探查分析”的**“全自动流行病学分析”模块**。 把相应变量拖入输入框内，点击“查看结果”。

研究人群特征

BMI Tertile	Low	Middle	High	P-value
AGE	40.1 + 15.7	36.5 + 13.4	37.1 + 12.9	0.006
BMI	18.8 + 1.1	21.0 + 0.6	23.7 + 1.5	<0.001
SBP	132.0 + 24.9	128.1 + 20.5	130.8 + 22.0	0.123
DBP	70.4 + 12.4	68.6 + 10.8	69.5 + 10.7	0.174
SEX				0.001
1	56.7	52.1	40.4	
2	43.3	47.9	59.6	
PSMK				0.849
0	42.6	41.1	43.5	
1	57.4	58.9	56.5	

单因素分析

Exposure	Statistics	SBP	DBP
BMI	21.2+2.3	0.1 (-0.6, 0.8) 0.8117	0.0 (-0.3, 0.4) 0.8681
BMI Tertile			
Low	263 (33.2%)	0	0
Middle	265 (33.4%)	-3.9 (-7.8, -0.1) 0.0463	-1.8 (-3.8, 0.1) 0.0617
High	265 (33.4%)	-1.2 (-5.0, 2.7) 0.5512	-1.0 (-2.9, 1.0) 0.3248
SEX			
1	417 (50.1%)	0	0
2	415 (49.9%)	-3.5 (-6.7, -0.4) 0.0273	-0.3 (-1.8, 1.3) 0.7503
AGE	37.9+13.9	0.8 (0.7, 0.9) <0.0001	0.3 (0.2, 0.3) <0.0001
AGE Tertile			
Low	273 (32.9%)	0	0

分层分析

Sub-group	N	SBP	DBP
X= BMI			
SEX			
1	394	-0.6 (-1.7, 0.5) 0.2778	-0.4 (-1.0, 0.2) 0.1731
2	399	0.7 (-0.2, 1.6) 0.1387	0.3 (-0.1, 0.8) 0.1448
AGE Tertile			
Low	268	0.2 (-0.5, 1.0) 0.5364	0.0 (-0.5, 0.4) 0.8196
Middle	262	0.5 (-0.4, 1.3) 0.3037	0.3 (-0.2, 0.9) 0.2115
High	263	0.6 (-0.8, 2.0) 0.3902	0.2 (-0.5, 0.9) 0.6053
PSMK			
0	334	-0.6 (-1.7, 0.6) 0.3382	-0.1 (-0.7, 0.4) 0.6133
1	454	0.5 (-0.3, 1.4) 0.2307	0.2 (-0.3, 0.6) 0.4583

多元回归分析

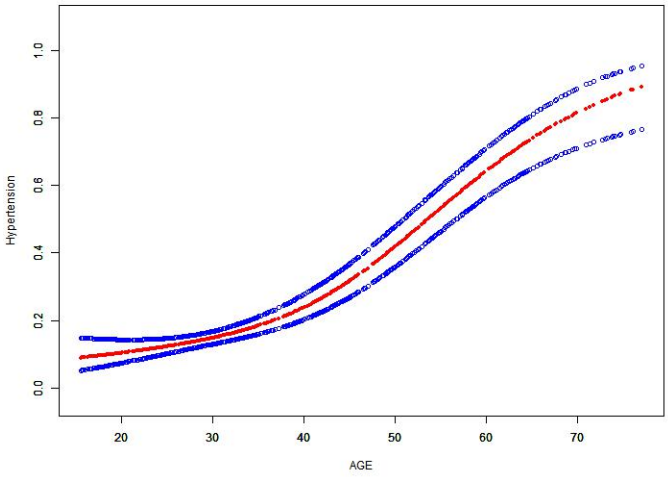
	Non-adjusted	Adjust I	Adjust II
Y= SBP			
BMI	0.1 (-0.6, 0.8) 0.8117	0.8 (0.2, 1.4) 0.0127	1.0 (0.4, 1.7) 0.0007
BMI Tertile			
Low	0	0	0
Middle	-3.9 (-7.8, -0.1) 0.0463	-0.8 (-4.2, 2.6) 0.6367	-0.2 (-3.6, 3.2) 0.9072
High	-1.2 (-5.0, 2.7) 0.5512	2.2 (-1.3, 5.7) 0.2150	3.3 (-0.2, 6.7) 0.0615
Y= DBP			
BMI	0.0 (-0.3, 0.4) 0.8681	0.3 (-0.1, 0.6) 0.1199	0.3 (0.0, 0.7) 0.0441
BMI Tertile			
Low	0	0	0
Middle	-1.8 (-3.8, 0.1) 0.0617	-0.8 (-2.6, 1.1) 0.4249	-0.5 (-2.4, 1.3) 0.5747
High	-1.0 (-2.9, 1.0) 0.3248	0.2 (-1.7, 2.0) 0.8565	0.5 (-1.4, 2.4) 0.6055

直接输出论文常用的四个表。软件自动识别变量类型，自动选择统计方法。自己不选就不会选错。 如果运行不出，可以少放几个调整变量。分类变量某一组样本量少，分层后运行不出来。

3　一图抵万言：平滑曲线拟合图的制作

用易佰软件可以轻松做出平滑曲线拟合图，是文章的加分项。

例如分析年龄（age）和高血压（HBP）的关系。使用“数据分析”-“论文图表”菜单下的“平滑曲线拟合”模块。是否高血压（HBP）放在应变量处，暴露变量为 age，点击查看结果。得出的图有 PDF（高清矢量图）和 PNG（图片）格式，均存放在工作路径的文件夹里。



4　随访数据的生存分析（K-M 曲线、Cox 模型分析）

使用易佾“全自动流行病学分析”模块，轻松实现随访数据生存分析的常用图表。

4.1　第一步：下载练习数据创建分析项目

下载 DataSurv.xls 数据和 DataNameSurv.xls 变量说明后创建分析项目。操作同上一节。

4.2　第二步：明确分析思路

- 结局变量 Y:　Death（死亡）
- 随访时间 T:　Time（随访时间）
- 危险因素 X:　CRP （C 反应蛋白的水平）
- 混杂因素 Z:　SEX（性别）、AGE（年龄）、PDduration（某疾病的病程）、MMSE（认知评分）、 ALBUMIN（白蛋白）、NSAIDS（用药情况）

4.3　第三步：操作 “全自动流行病学分析”模块

把相应变量拖入输入框内，勾选“Cox 模型生存分析”，给出随访时间变量 TIME。点击“查看结果”。 注意：这里结局指标的编码是 0 和 1,1 表示发生事件（例如死亡）。

得出论文中常用的四个表：研究人群描述、单因素分析、分层分析和多元回归分析。

研究人群特征

CRP Tertile	Low	Middle	High	P-value
Age	67.6 + 9.3	69.3 + 9.0	70.2 + 9.2	0.139
PDduration	7.6 + 5.4	7.6 + 5.6	8.4 + 5.6	0.512
MMSE	25.1 + 4.4	25.1 + 4.7	23.1 + 5.6	0.007
Albumin, mg/dl	4.2 + 0.3	4.1 + 0.4	3.9 + 0.4	<0.001
CRP	0.2 + 0.1	0.4 + 0.1	3.4 + 5.1	<0.001
Sex				0.025
Male	33.3	39.8	51.7	
Female	66.7	60.2	48.3	
NSAIDs				0.854

单因素分析

Exposure	Statistics	Death
CRP	1.5+3.4	1.1 (1.0, 1.1) 0.003
CRP Tertile		
Low	87 (27.8%)	1.0
Middle	108 (34.5%)	0.9 (0.4, 1.9) 0.712
High	118 (37.7%)	2.3 (1.1, 4.4) 0.018
Age	69.1+9.2	1.1 (1.0, 1.1) 0.001
Age Tertile		

分层分析

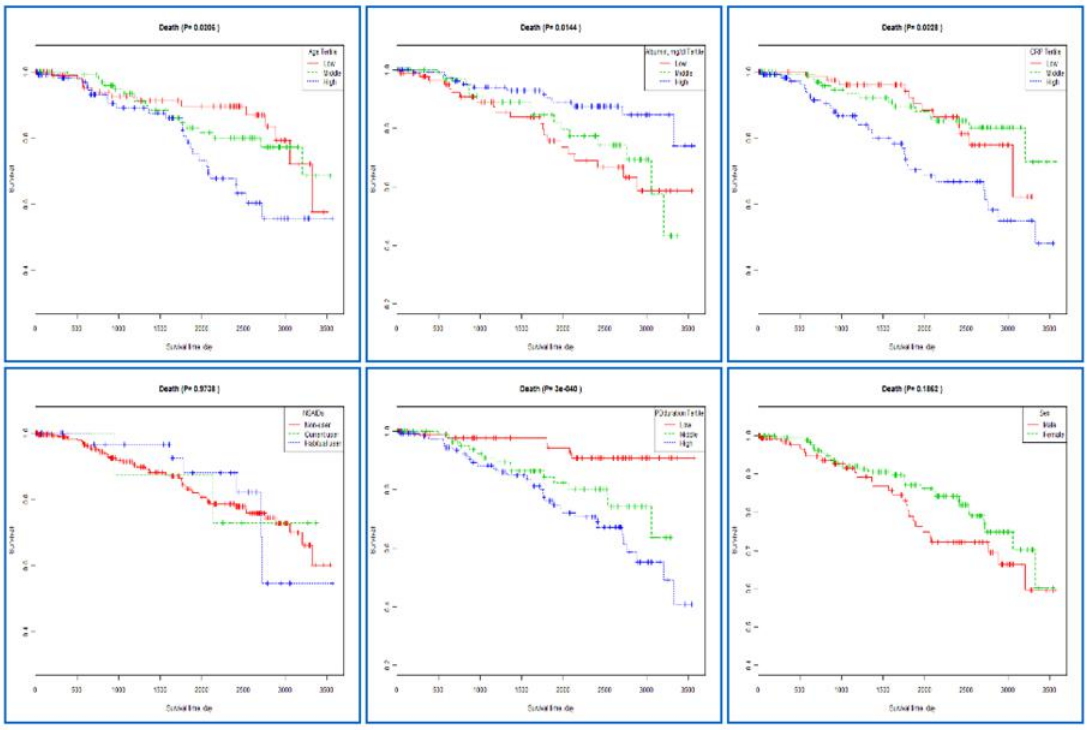
Sub-group	N	Death
X= CRP		
Age Tertile		
Low	94	1.1 (1.0, 1.2) 0.238
Middle	111	1.3 (1.1, 1.5) 0.001
High	108	1.0 (1.0, 1.1) 0.205
Sex		
Male	133	1.1 (1.0, 1.2) 0.002
Female	180	1.0 (1.0, 1.1) 0.142

多元回归分析

	Non-adjusted	Adjust I	Adjust II
CRP	1.1 (1.0, 1.1) 0.003	1.1 (1.0, 1.1) 0.027	1.1 (1.0, 1.1) 0.031
CRP Tertile			
Low	1.0	1.0	1.0
Middle	0.9 (0.4, 1.9) 0.712	0.8 (0.3, 2.0) 0.630	0.9 (0.3, 2.4) 0.850
High	2.3 (1.1, 4.4) 0.018	1.9 (0.8, 4.2) 0.135	2.2 (0.9, 5.1) 0.071

表中数据：β (95%CI) Pvalue / OR (95%CI) Pvalue
结果变量: Death
暴露变量: CRP; CRP Tertile
Adjust I adjust for: Age; Sex (Male, Female); PDduration; MMSE; Albumin, mg/d
Adjust II adjust for: Age (smooth); Sex (Male, Female); PDduration (smooth);
user, Habitual user)

同时得出 KM 生存曲线图，如果是连续变量，自动按三等分组分析。也可以自己设定切点分组，具体操作请见后文。



5　预测模型（列线图、ROC 曲线、校准曲线）制作（了解）

易侕软件可以基于 R 或 SAS 强大的程序包，一键实现预测模型的变量筛选、模型比较、建模与验证、ROC 曲线、列线图 Nomogram、计算机模拟 Bootstrap 等。可以说是一网打尽各种统计花活。**是文章的加分图表。属于高级功能需要购买。**（约 3 千元/年，具体优惠请见官网）

5.1　第一步：下载练习数据创建分析项目

研究目的：建立乳腺癌生存状态的预测模型
测量指标：患者基线的临床资料和影像学指标等、随访患者的预后（死亡）

A	B	C	D	E	F	G	H	I	J
id	age	pathsize	histgrad	er	pr	pathscat	ln_yesno	time	status
80	64	0.5	2	0	0	1	0	12.4	0
1169	37	2	3	1	1	1	1	12.4	1
466	71	1.4	2	1	1	1	0	12.6	0
576	62	1.5	4	1	0	1	0	12.6	0

红框内是结局 status（0=存活或删失，1=死亡）和随访时间（单位：月）。每一行是一个患者，例如第 1 行编号 80 的患者，年龄 64 岁随访到 12.4 个月时没有死亡。第 2 行编号 1169 的患者随访到 12.4 个月时发生了死亡。下载数据操作同前。

5.2　第二步：明确分析思路

用基线资料和影像学指标等（多个 X），对死亡（Y）进行预测。

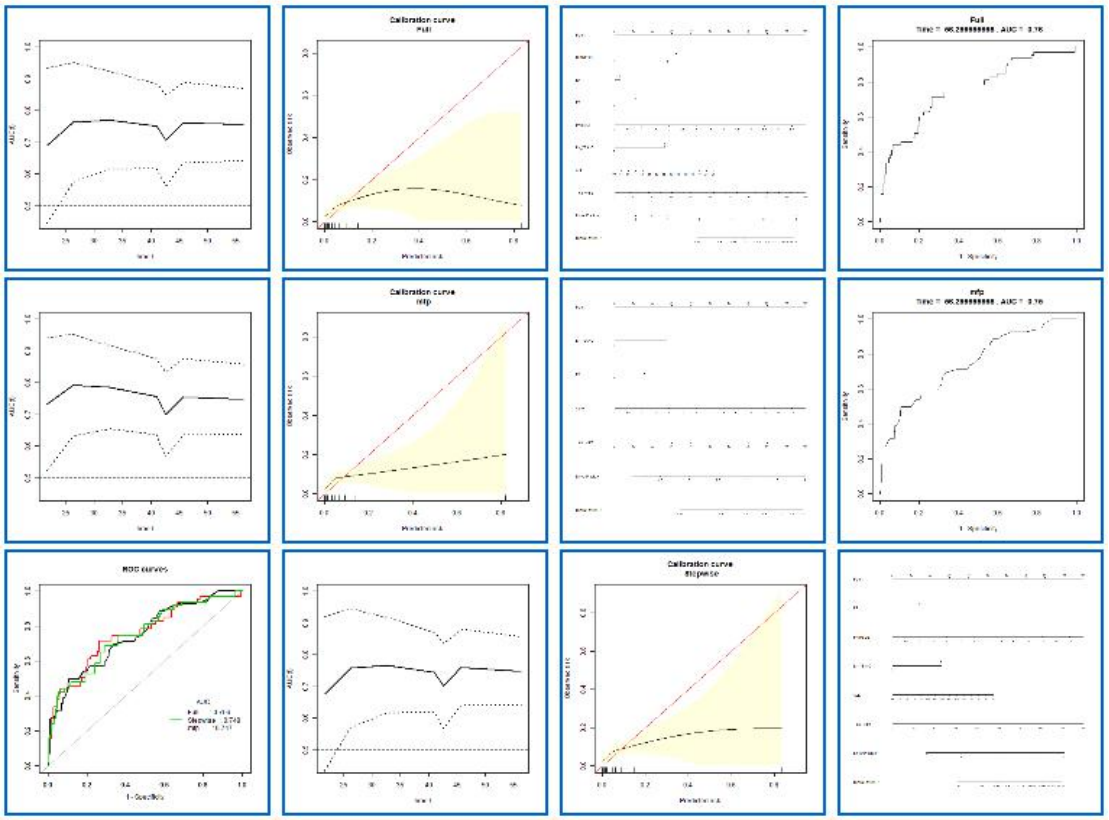
5.3　第三步：操作“预测模型与 ROC 分析模块”

选择数据分析-诊断试验与预测方程-预测模型与 ROC 分析模块。操作设置如下，点击查看结果。

- 结果变量：status（结局）
- 回归模型选：Cox 回归模型
- 时间变量：time（随访时间）
- 模型自变量：histgrad，ER,PR，pathscat，ln_yesno，age

结果出来了，得出 Cox 比例风险模型分析建立的预测模型和系列图表，包括：

患者生存概率预测列线图（Nomorgram）和 ROC 曲线。列线图是在 Cox 比例风险模型基础上绘制。自动得出了发表 SCI 论文所需的大部分图表：包括 ROC 曲线（单个模型）、ROC 曲线（多个模型比较）、time-dependent AUC 曲线拟合图、校准曲线 Calibration curve。



运用常规和高级统计分析方法，建立所有可能的模型，并从中自动选择最佳模型。

- ✓ Model 0：多变量分数多项式（Multivariable Fractional Polynomials，MFP）模型，考虑到非线性关系建模
- ✓ Model 1：全自变量的预测模型（full model）
- ✓ Model 2：逐步回归基于 AIC 原则筛选自变量建立简化的模型(stepwise model)。易侕软件自动调用 R 软件包中的 STEP()函数以计算 AIC 信息统计量为准则（Akaile Information Criterion，赤池信息准则），选取 AIC 信息统计量最小的模型为最终模型，本案例分析中最终模型选入 4 个变量：PR，pathscat，ln_yesno 和 age，部分结果截图如下。

Model 2: Stepwise (stepAIC) selected model from observed data
-0.45126*(PR=1) +0.44078*PATHSIZE +0.79125*(LN_YESNO=1) -0.02322*AGE

	coeff.	se.	HR	Low 95%CI	High 95%CI	P value
factor(PR)1	-0.4513	0.3014	0.6368	0.3528	1.1496	0.1343
PATHSIZE	0.4408	0.1153	1.5539	1.2396	1.9479	0.0001
factor(LN_YESNO)1	0.7912	0.2912	2.2061	1.2468	3.9037	0.0066
AGE	-0.0232	0.0122	0.9770	0.9539	1.0008	0.0577

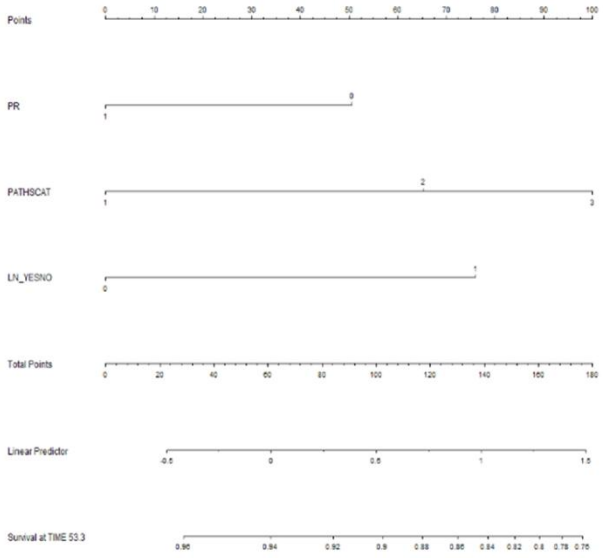
rsq = 0.0455, maxrsq = 0.5171 , Log likelihood -277.6811

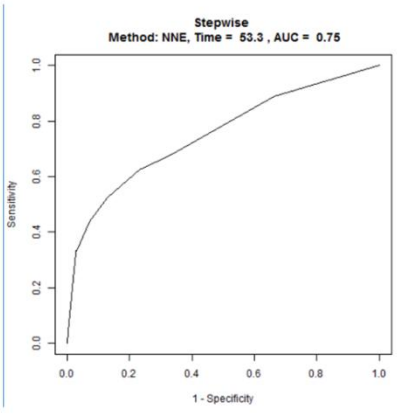
concordant	discordant	tied.risk	tied.time	std(c-d)	C	se(C)
15436	5770	10	2	1888.0373	0.7278	0.0445

Obs	Events	Model L.R.	d.f.	P.value	Score	Score P	R2	g	gr
815	50	37.95	4	<0.0001	45.5925	0	0.088	0.8962	2.4503

- ✓ 如果定义了 bootstrap 次数，采用重抽样的方法，产生 n 个新样本，分别构建回归模型。即 BS full model 与 BS stepwise model

图中给出了各变量不同取值对应的评分（Pionts），据此计算个体总评分（Total Points），这里默认的是时间 75%位数对应的生存时间。如果要得到个体其他时间点（例如 3 年或 5 年）的生存概率操作也很简单，在诊断试验与 ROC 分析模块输入“预测时间”设置就行，就不演示了。





6 机器学习系列图表制作（了解）

易侕软件可以基于 R 或 SAS 强大的程序包，一键实现机器学习建模。**属于高级功能需要购买。**

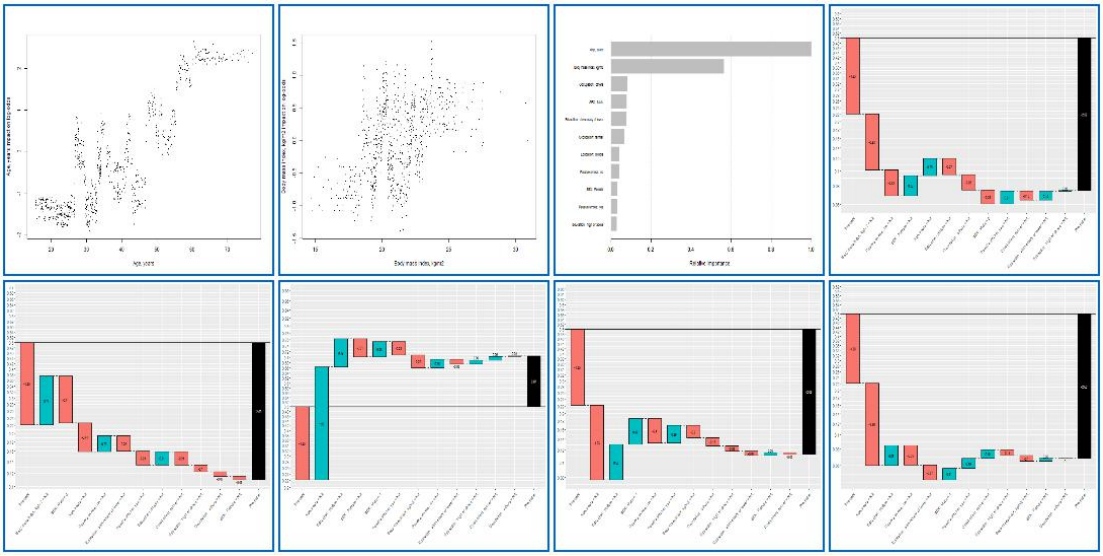
第一步：运用易侕软件自带的练习数据（demo）创建分析项目。操作同前。

第二步：明确分析思路：

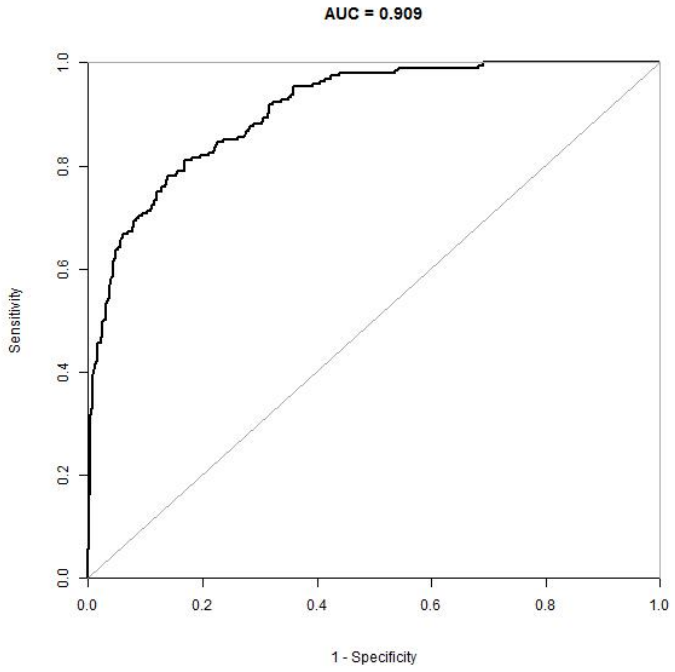
- 结局变量 Y：HBP（是否高血压）
- 自因素 X：SEX（性别）、BMI 等

第三步：操作：选择数据分析-诊断试验与预测方程-机器学习与预测分析模块。“结果变量”放入是否发生高血压，“自变量”框内放入可以用于预测的变量。其他设置用默认的，查看结果。

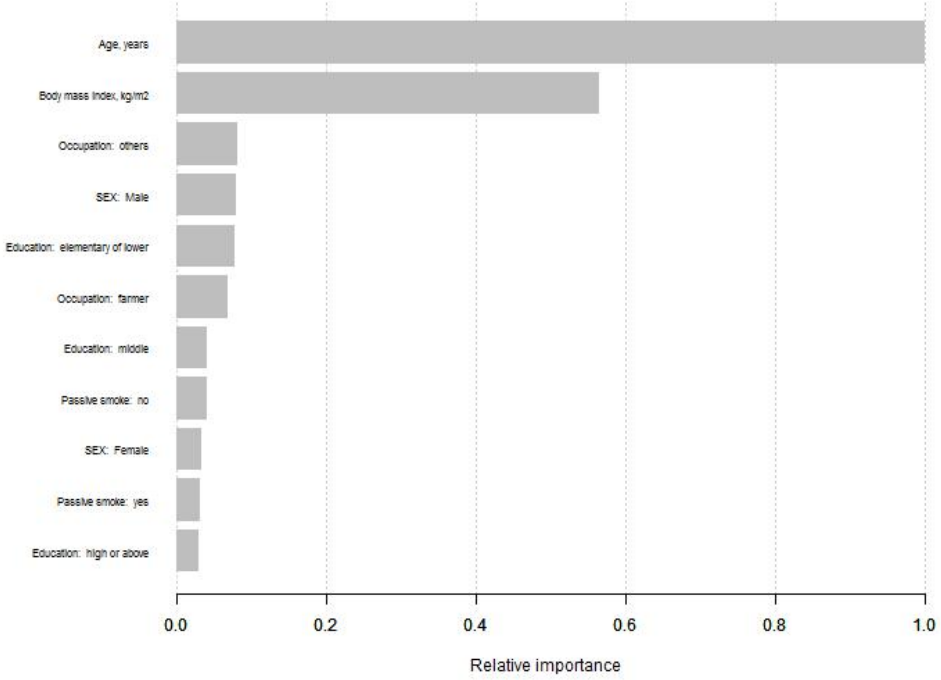
直接得出一系列的图表结果：



下图是 ROC 曲线：



下图是各指标对预测的重要程度排序图。下图可以看到排在第一位的是 age，如果把 BMI 对结局 Y 的重要性当做 1（100%），排在第二位的 BMI 约为 0.55。



除了直观的图片结果，还有表的结果（部分截图）：

机器学习与预测

Predict for: High BP
Predictors: SEX; Age, years; Body mass index, kg/m2; Occupation; Education; Passive smoke
SET params as: booster=gbtree, objective=binary:logistic, max_depth=10, min_child_weight=9.47417389415205, subsample=0.746097089722753, colsample_bytree=0.628952796338126

Variable importance

Feature	Gain	Cover	Frequency
Age, years	0.48785293	0.35510042	0.34319527
Body mass index, kg/m2	0.27571811	0.35294774	0.35502959
Occupation: others	0.03990816	0.02659491	0.02564103
SEX: Male	0.03859949	0.04662882	0.04733728
Education: elementary of lower	0.03778082	0.03362801	0.03550296
Occupation: farmer	0.03355167	0.06231527	0.06706114
Education: middle	0.02033221	0.02540244	0.02366864
Passive smoke: no	0.02027207	0.02684833	0.03155819
SEX: Female	0.01608313	0.01175535	0.01380671
Passive smoke: yes	0.01504638	0.02890401	0.03155819
Education: high or above	0.01485503	0.02987472	0.02564103

Observations of testing data: 832
Testing Set: Prediction results

Testing Set: Confusion Matrix

	Observed	
Prediction	0	1
0	571	74
1	27	121

C:\Users\T530\AppData\Roaming\EmpowerRCH\EmpowerRCH_examples\demo_48_tbl\demo_48_tbl.htm

易侑软件后台是 R 语言，具体的 R 软件包可以查看帮助文件如下表。例如，当结局指标是时间依赖的生存状态，需要在文章中写是用 CoxBoost 的 R package 做的。可以在结果输出界面查看并调用 R 编改程序，爱好学习 R 的朋友可以研究。

一、模块功能、所用 R 软件包与主要参数设置：

学习对象与目标	所用的 R package	学习器参数设置与调试
1. 两分类或多分类	XGBoost XGBoostExplainer MLR	可自动调试参数与手置参数设置 xgboost 所用参数
2. 有序多分类	XGBoost	可手置参数设置 xgboost 所用参数， 自动调试参数无作用
3. 连续性变量	XGBoost	可自动调试参数与手置参数设置 xgboost 所用参数
4. 时间依赖的生存状态	CoxBoost	自动寻找最佳参数 stepno 与 penalty， 自动调试参数与手置参数均无作用
5. 多重标签(多个两分类属性)	MLR MLDR	自动比较选择用 Binary relevance 或 Classifier chains 分类方法， 自动调试参数与手置参数均无作用
6. 集聚分析	cluster factoextra	自动寻找类数， 自动调试参数与手置参数均无作用

二、页面设置：结果变量、Objective、时间变量：

学习对象与目标	结果变量与时间变量
1. 两分类或多分类	需要两分类或多分类结果变量（数字型）,时间变量置空
2. 有序多分类	需要多分类结果变量（数字型）,时间变量置空
3. 连续性变量	需要连续性结果变量，时间变量置空
4. 生存状态	需要两分类（0=censor 1=event）状态变量与时间变量
5. 多重标签	结果变量置空，系统自动将数据中变量名为 LABEL 打头的两分类变量做为结果变量。 因此，须先将结果变量命名为 LABEL1-LABLELn,分别表示 n 个两分类的属性（0=false 1=ture）， 时间变量无作用
6. 集聚分析	结果变量、时间变量均置空

本章为操作基础课，体会让数据分析变得如此简单，如果你动手做出了本课程的图表，统计分析水平已经很高了。关于结果临床意义的解读，详见后续课程。

7　软件常用基本操作

打开易侕软件自带的练习数据，学习常用操作：**分析项目 - 打开现有项目 - 练习用的项目，选第二行（后缀是 demo）的练习数据。**

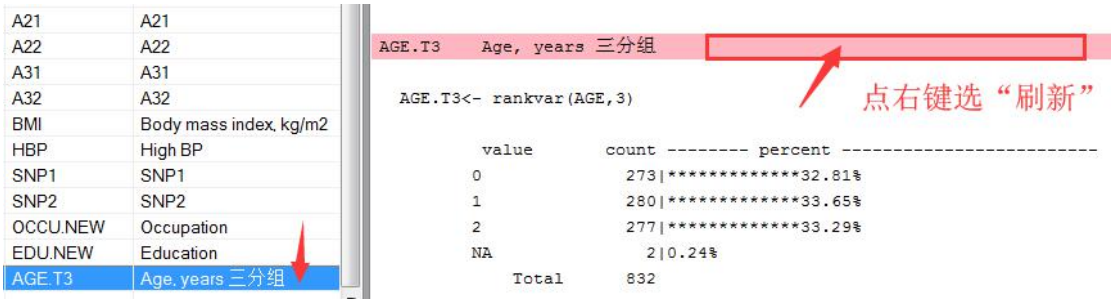
7.1　连续变量分组（等分、按指定切点分）

可以等分组（常用三等分、四等分）、按百分位数两分组、按切点分组（临床切点）等。

在变量名 Age 上点右键，选“连续性变量分组”

勾选“三分组”的选框，点击保存：

在变量的最下面自动创建好了新变量：AGE.T3。如果要查看变量分布，可在右边粉色区域点右键选“刷新”。如果没有出现变量分布，表示指标创建失败，需要删除后重新创建。



操作就完成了。其它操作（按百分位数两分组、**按切点**分组）大家自行练习。**注意，手动输入切点是，需要用英文输入法。**

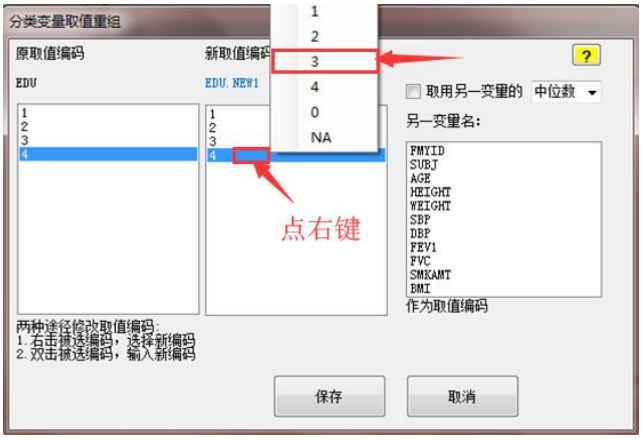
7.2　小操作技能：分组切点的具体数值如何查看呢？

可以用研究人群描述模块，分析变量是原始连续变量，选择列分层变量为分组变量，**输出内容与格式选第二项 2: Mean(SD) Median (Min-MAx) / N(%)。**这样每组的最小值 - 最大值(Min-MAx)就是对应的分组切点值。

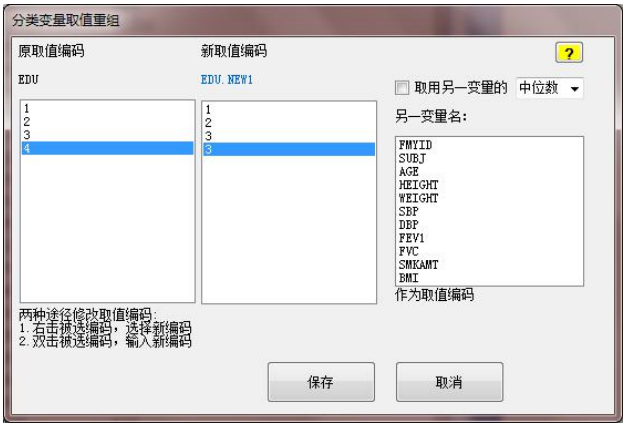
7.3 分类变量取值重编码

用途：将多个分组合并；回归分析时，分类变量的参照组编码=0。

1. 在变量 OCCU（职业）上点右键，选“分类变量取值重编码”
2. 在右边框的 4 上面点右键，选 3。



3. 这样“新取值编码”最后一行就变成了 3，表示原来的 4 和 3 合并为一组，编码为 3。点击保存。自动创建好了新变量：OCCU,NEW1。刷新后检查变量分布。



7.4 设置小数点位数

输出结果的数据是四舍五入后的，可以在操作设置界面调整小数点位数：

单因素分析

	Statistics	Systolic BP, mmhg	
Body mass index, kg/m2	21.2 ± 2.3	0.1 (-0.6, 0.8)	0.812

⇨

单因素分析

	Statistics	Systolic BP, mmhg	
Body mass index, kg/m2	21.202 ± 2.300	0.083 (-0.601, 0.767)	0.81166

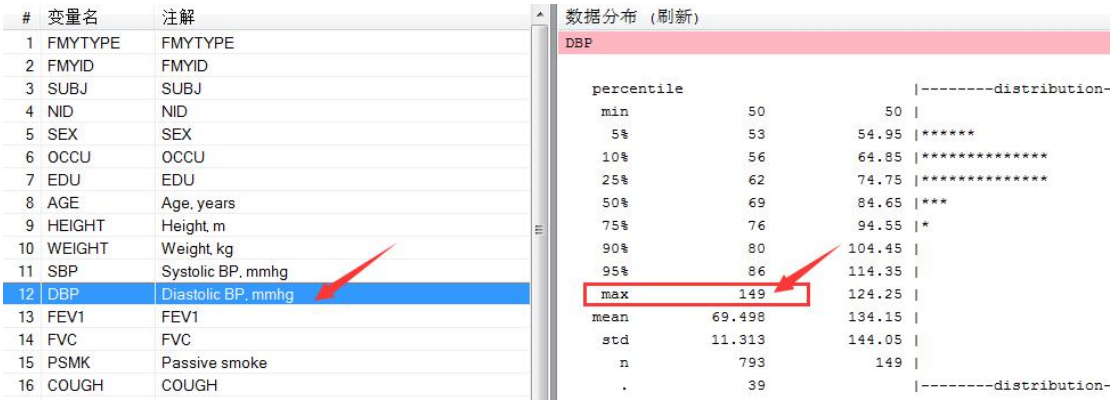
7.5 赋值错误更正

做数据分析前，必须要看每个变量的分布，**通过查看最大值和最小值，发现数据中的异常值，**进行数据清理。数据清理非常重要！

数据导入易侕后，会自动得出变量分布图（见下图右侧）：连续性变量显示三列。

- 第一列是百分位数、最小值（min）、最大值（max）、均数（mean）、标准差（std）、例数（n）、缺失数据（用 . 表示）
- 第二列是第一列对应的取值
- 第三列是右侧频数分布图每组的中位数。

通过看变量分布进行数据清理。例如：点击舒张压（DBP），看到最大值是 149。



#	变量名	注解
1	FMYTYPE	FMYTYPE
2	FMYID	FMYID
3	SUBJ	SUBJ
4	NID	NID
5	SEX	SEX
6	OCCU	OCCU
7	EDU	EDU
8	AGE	Age, years
9	HEIGHT	Height, m
10	WEIGHT	Weight, kg
11	SBP	Systolic BP, mmhg
12	DBP	Diastolic BP, mmhg
13	FEV1	FEV1
14	FVC	FVC
15	PSMK	Passive smoke
16	COUGH	COUGH

数据分布 (刷新)

DBP

percentile			
min	50	50	|
5%	53	54.95	|******
10%	56	64.85	|**********
25%	62	74.75	|**********
50%	69	84.65	|***
75%	76	94.55	|*
90%	80	104.45	|
95%	86	114.35	|
max	149	124.25	|
mean	69.498	134.15	|
std	11.313	144.05	|
n	793	149	|
.	39		|

如果想把舒张压大于 140 的均按缺失处理，可以在原始 Excel 中修改，或在赋值错误更正模块修改。操作：在 DBP 上点右键，**选“赋值错误更正”。更正条件输入＞140，正确赋值输入 NA，表示缺失。注意：用英文输入法。点保存后**自动创建了新变量。注意刷新后检查变量分布。

7.6 选择分析人群与输出数据文件

可以限制分析对象在某一类人群中，例如在 sex=1 的人群中做曲线拟合。

操作方法 1：在分析人群对应的空白处点右键，出现下图对话框。可以设置条件：

(1) 拖入变量名

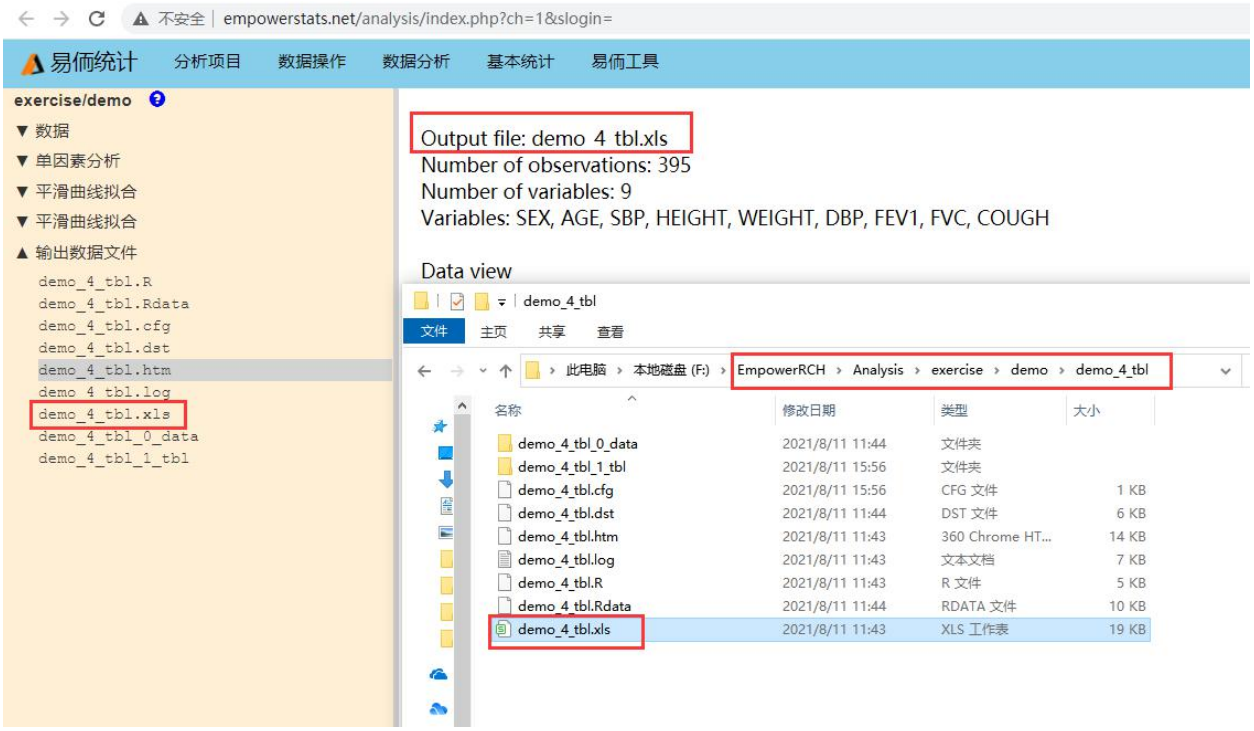
(2) 点击下拉菜单，选择条件为=，＞，＜，≥或≤

(3) 输出数值或变量名，注意需要用英文输入法

(4) 如果有多个条件，可以点击“添加一行”的设置

说明：有的电脑需要在分析人群对应的空白处点**两次右键。**

操作方法 2：在数据操作-输出与查看-输出数据文件模块操作，可以输出数据文件到一个新的数据中，再重新创建分析项目。



8 格式说明《数据》和《变量注解》

数据分析前需要准备好《数据》和《变量注解》。**前面课程已分享实例。**

8.1 《数据》文件格式要求

	A	B	C	D	E	F	G
1	FMYTYPE	FMYID	SUBJ	NID	SEX	AGE	HEIGHT
2	0	1	1	11	1	17.3	1.64
3	0	1	3	1	1	44.8	1.7
4	0	1	2	2	2	43.1	1.5
5	0	2	5	1	1	38.9	1.65
6	0	2	4	2	2	35.9	1.59
7	0	3	6	2	2	36	1.57
8	0	4	9	13	1	39.5	1.64
9	0	4	10	14	1	30.1	1.68
10	0	4	12	1	1	68.7	1.59
11	0	4	7	11	2	33.8	NA
12	0	4	8	12	2	42.5	1.47

1. 指标的变量名为第一行（不要多行）
2. 通常每个患者一行，每个指标一列
3. 连续变量（如年龄）用原始数据，不加单位
4. 分类变量（如血型、分组）用 0,1,2 表示，不用中文和字母
5. 分组变量对照组编码为“0”

6. 缺失变量用空格或者“NA”表示
7. 不参与分析的变量（如姓名）不用列出
8. 随访数据（详见生存分析部分实例）：
   - ✓ 数据中有随访时间变量
   - ✓ 结局变量（0=未发生结局事件或截尾数据；1=发生结局），例如以肺癌死亡为结局，0=随访时未死亡或失访，1=死亡
9. 若患者的指标在不同时点测量多次（了解，详见重复测量专题课程）：
   - ✓ 每个患者一个唯一的编号（ID）
   - ✓ 每次测量一行
   - ✓ 变量中必须有每次测量的时间（具体日期或者时点）

8.2 《变量注解》文件要求

	A	B	
1	变量名	取值编码	注解
2	FMYTYPE		FMYTYPE
3	FMYID		FMYID
4	SUBJ		SUBJ
5	NID		NID
6	SEX		Gender
7		1	Male
8		2	Female
9	AGE		Age (years)
10	HEIGHT		Height (m)
11	WEIGHT		Weight (kg)

1. 包括“变量名”、“取值编码”和“注解”三列
2. “变量名”尽量简短，例如：X1、X2、X3…，**不能用 NA 或 Na 作为变量名，变量名不要出现空格、运算符号等。易佰 3.0 版变量名需要大写。**
3. “取值编码”
   - ✓ 分类型变量列出编码，连续型变量此处空白
   - ✓ 编码从变量名对应的第二行开始顺序填写
   - ✓ 一般用 0,1,2……顺次编码，**对照组用 0 编码**
4. 注解：表示取值编码代表的意义，如上例中 X2 为 SEX（1= Male，2= Female）
   - ✓ 若文章输出图标需要显示为英文，则“注解”为英文；
   - ✓ 若文章输出图标需要显示为中文，则“注解”为中文；

最常见也是最耗时的是整理数据，包括数据清理和手工制作变量注解文件。
推荐用易佰 DataWeb 数据采集管理系统收集数据，可以自动制作符合要求的《数据》和《变量注解》文件，节省你的时间。目前录入 2000 例研究对象 100 个变量用免费版就足够了。
http://www.empowerstats.net/dataweb/

我们一直在做一件事，而且只专注这件事

给您一个巨人的肩膀，将临床经验转化为证据，为祖国的临床科研做出贡献！

易侕学院官网：http://www.empowerstats.net/empowerU/

易侕统计官网：http://www.empowerstats.net/cn/

易侕 DataWeb 官网：http://www.empowerstats.net/dataweb/

联系人邮箱：陈博士 cindylpf@126.com