社会统计学及SPSS软件应用
STATISTICS WITH SPSS

Instructor:王荣欣

Email: rxwang@qq.com

周一3-4节、单周周四3-4节，3A106-2

2020年12月28日

CONTENTS

Threats to interval validity:

1. history effect
2. maturation
3. testing effects
4. instrumentation
5. regression to the mean
6. selection bias
7. differential attrition
8. treatment contamination
9. resentful demoralization

1. Observational (non-experimental) studies
2. Experimental studies. 就实验过程和限制性条件而言（样本是否可被随机分组），可分为
   (1) 标准实验：避免ommitted variable bias or endogeneity bias
      - 随机对照实验（Randomized Controlled Trial）
   (2) 准实验（quasi-experiment）
   (3) 自然实验（natural experiment）

SELECTION BIAS

1 由共同原因造成的相关性，称为混杂偏误（confounding bias）
   (1) The true causal effect X→Y is "mixed" with the spurious correlation between X and Y induced by the fork X←Z→Y.
   (2) Z is a confounder of the proposed causal relationship between X and Y.

2 由以共同结果为条件造成的相关性，称为样本选择偏误（sample selection bias）
   (1) 由于采用非随机样本造成的偏误。
      1. 非随机样本可能来自共同的个体的自我选择行为，也可能是由于调查者的抽样规则造成的。

1. If you want to know, whether caregivers are particularly prone to depression, you need a sample of caregivers and noncaregivers.
2. If you want to know what factors lead people to migrate, you need a sample of both those who do and those who do not migrate.
3. If you want to evaluate the efficacy of a program, you need a sample of places where the program has and has not been implemented (or data before and after implementation) (p.382, 中文版第360页)

Endogeneity bias (and its close cousin, sample-selection bias) is a form of omitted-variable bias. (p.389, 中文版第367页)

1. Instrumental variable
2. Heckman selection model
3. Endogenous switching regression model
4. Propensity score matching

PROPENSITY SCORE MATCHING

What is the effect of attending an elite university on subsequent occupational status?

1. To carry out a multiple regression of occupational status on attendance at elite versus other universities plus a set of variables controlling for family background, high school performance and so on.
2. The difficulty is that attending an elite university tends to be so highly correlated with the control variables that controlling for ***confounding factors*** fails to hold them constant, because there are **few** people with low values on the control variables who attend elite universities.
3. Propensity score, a scalar summary of the degree of similarity between cases with respect to a large number of covariates.

1 计数回归指的是因变量是计数数据的回归分析。

2 有什么样的统计分布，可以描述顾客每月光顾超市的次数？

计数变量

1 社会现象中，往往研究的是在某一个特点时间和空间分布的**某个事情的发生数目**。

2 计数变量（count variable：事件的发生数）
(1) 某地区不同时间中死亡的数量
(2) 社区一年中的发案次数
(3) 足球比赛中的进球数量
(4) 大学教师在某时间内发表的文章数量

3 这些资料的分布都服从poisson分布。

计数变量（COUNT VARIABLE）

1. 个人的迁移次数
2. 个人晋升次数
3. 个人每年的看病次数
4. 个人每天抽几根烟
5. 妇女的曾生子女数
6. 癌症病人体内肿瘤的个数

既然是数数，它就必须是非负的整数，也不能是小数。

客户关系管理中的RFM模型

1. 最近一次消费（Recency）
2. **消费频率（Frequency）**：一定时间内客户到访的次数。
3. 消费金额（Monetary）

什么样的统计分布，可以描述顾客每月光顾超市的次数？

1 目标：找到一种可以描述非负整数的概率分布。

(1) 二项分布（binomial distribution）：有上界
(2) **泊松分布**
(3) 其他分布

1 根据泊松分布，顾客每个月光顾超市X次的概率为：

$$P(X) = \frac{e^{-\lambda}\lambda^{X}}{X!}$$

2 λ是顾客平均每个月光顾超市的次数。

3 X是整数（X=0，1，2，3……），但λ可以是小数。

1 二项分布：指在n次Bernoulli实验中，概率为π的事件A，出现X次的概率分，记做：$X \sim B(n, \pi)$。

2 Bernoulli实验是一种最简单的独立重复实验，每次实验只会出现对立的两种结果中的一种，如生或死，如投中和未投中等。其公式为：

$$P(X) = \frac{n!}{X!(n-X)!}\pi^X(1-\pi)^{(n-X)}$$

1 Poisson分布是经常出现的离散型概率分布，是二项分布n很大而π很小时（$n \to +\infty, \pi \to 0$）的特殊形式。

2 它是二分类资料在n次实验中发生X次某种结果的小概率分布规律（即描述对象是罕见事件，发生率很低）。

$$\lim_{n\to+\infty} P(X) = \lim_{n\to+\infty} \frac{n!}{X!(n-X)!}\pi^X(1-\pi)^{(n-X)} = \frac{e^{-\lambda}\lambda^X}{X!}$$

1 在足够多的n次独立Bernoulli实验中，罕见事件出现X次的概率分布，Poisson分布符合函数：

$$P(X) = \frac{e^{-\lambda}\lambda^{X}}{X!}$$

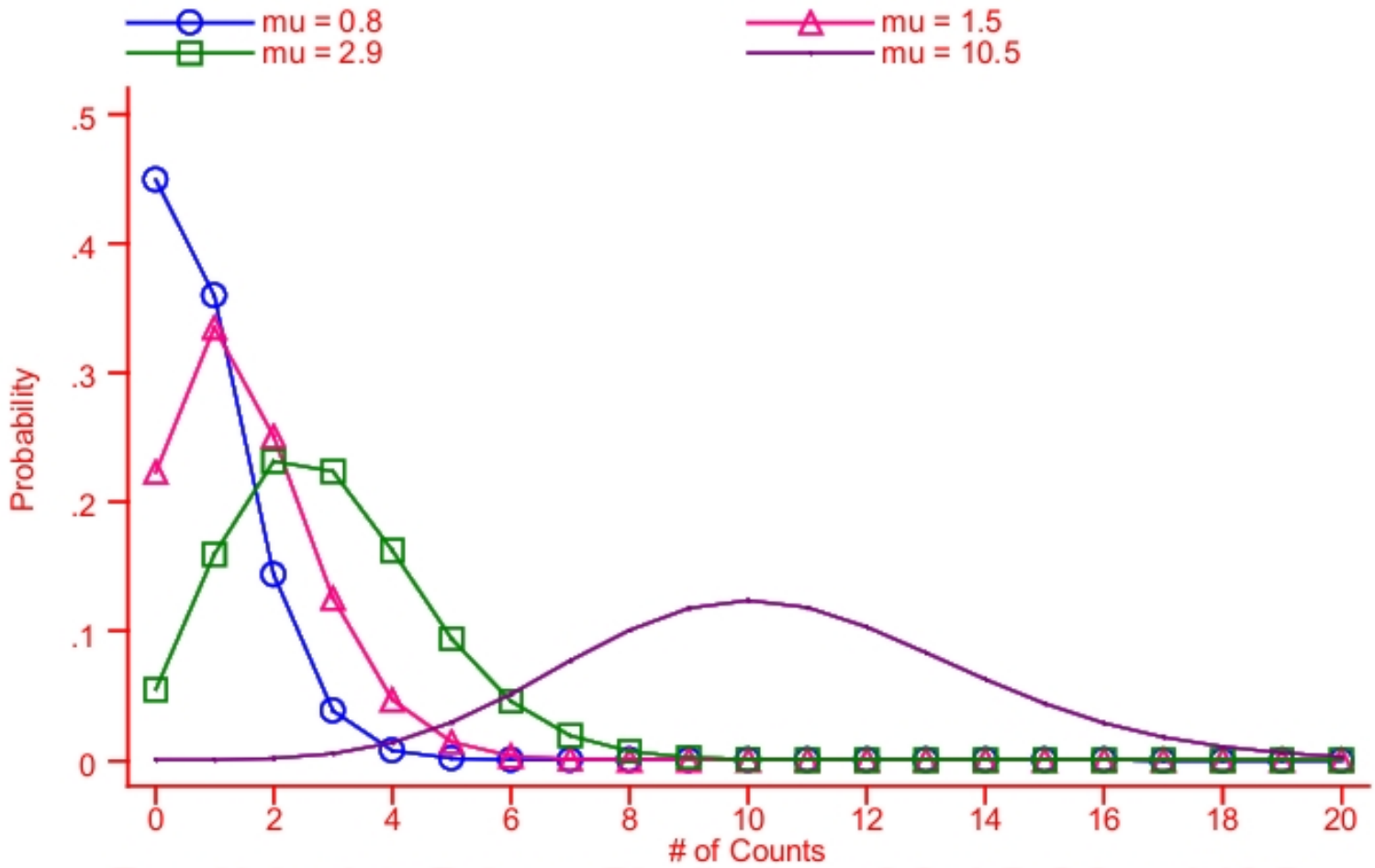
(1) λ表示总体均值，记作$X \sim P(\lambda)$，总有$\sum P(X) = 1$。
(2) X=0，1，2，3……

2 要想确定一个泊松分布，我们只需要知道λ这一个参数就可以了。

3 它的概率分布完全由一个单一的参数λ决定。

Poisson分布的一个重要特征：Poisson分布的总体均值λ等于总体方差σ，$E(y_i) = \lambda$，$Var(y_i) = \lambda$。

1. 当$\lambda \leqslant 1$时，X越大，P(X)越小；
2. 当$\lambda > 1$时，随X增大，P(X)先增大而后变小，所以λ越大则趋于正态分布，λ越小则越呈偏态分布；
3. 实际应用中，$\lambda \geqslant 20$就可将其看作是正态分布。



Four Univariate Poisson Distributions: 0.8, 1.5, 2.9 and 10.5

能否用普通线性回归？

1. 顾客光顾超市的次数是一个具有数值意义的变量，但它不是连续的。
2. OLS回归要求因变量为连续、无界、正态分布。

3. 计数变量却是离散的、有界的，而且往往非正态分布。
4. 对计数因变量做OLS回归，违反其基本假定，导致估计的有效性、一致性、无偏性都会发生问题。

如何“建模”？

1. 如果自变量X能够影响因变量Y，那么它必须通过影响λ来实现。

2. 所以，只要能够建立一种λ和Xb之间的函数关系，那么就可以获得一个关于计数数据的回归模型。

能否这样“建模”？

$$\lambda(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \epsilon$$

1. 简单的线性模型能否满足要求？
2. 提示：
   (1) 等号的左边是？
   (2) 等号的右边是？

怎样进行变换?

1. 对$\lambda(X)$进行某种变换，让变换后的$\lambda(X)$可以取任意值，那么线性模型就适用了。

2. 答案：对数变换。

1 假定Y_i符合 Poisson分布，解释变量为X_{i1}, X_{i2}, X_{i3}……

2 平均数λ_i的对数是解释变量的**线性函数**：

$$ln(\lambda_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_k X_{ik}$$

3 λ_i就是自变量线性函数的自然指数：

$\lambda_i = exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_k X_{ik})$，保证$\lambda$非负。

4 λ被称为“发生率比”（incidence rate ratio）。

(1) 表示X增加一个单位，事件的平均发生次数将是原来的多少倍。

表 2　社会资本和社会网络回归方程分析结果

自变量	社会资本（对数）B	网络规模（对数）B	网络顶端（对数）B	网络差异（对数）B	与领导层纽带关系 Exp（B）	与经理层纽带关系 Exp（B）	与知识层纽带关系 Exp（B）
阶级阶层地位							
行政领导层	0117***	0121***	0123***	0121***	2185***	1136*	2114***
经理层	0115***	0117***	0118***	0118***	2118***	1172***	1189***
专业技术层	0118***	0124***	0121***	0128***	1159***	1148***	3148***
职员层	0111***	0110	0115***	0114***	1160***	1131*	1152***
技术工人层	0106***	0107	0108***	0105	1122*	1116	1133**
雇主层	0104	0130*	-0102	0109	1105	0196	1117
自雇层	0100	0108	0101	-0106	1108	0189	0193
非技术工人层（参考项）							

Figure 3.2: 哪个因变量应该用泊松回归

表 2　社会资本和社会网络回归方程分析结果

社会资本（对数）B	网络规模（对数）B	网络顶端（对数）B	网络差异（对数）B	与领导层纽带关系 Exp（B）	与经理层纽带关系 Exp（B）	与知识层纽带关系 Exp（B）

Figure 3.3: 边燕杰论文《城市居民社会资本的来源及作用》

因变量	变量标签	变量值
社会资本	因子得分值	连续变量
网络规模	拜年人数	连续变量
网络顶端	最高声望	连续变量
网络差异	职业个数	连续变量
与领导层纽带关系	有无拜年	有=1，没有=0
与经理层纽带关系	有无拜年	有=1，没有=0
与知识层纽带关系	有无拜年	有=1，没有=0

自变量	变量标签	变量值
阶级阶层地位	领导层、 经理层、 专业技术层等	8种地位
职业活动交往	科层关联度、市场关联度	连续变量
城市	广州和厦门等	5个城市

处理计数数据（COUNT DATA）的回归模型

只要能找到一种没有上界、取值为非负整数的概率分布，就可以获得一个回归模型。

1 泊松回归

(1) Y的均值和方差都等于λ。这种情况称为等离散(equidispersion)。

2 负二项泊松分布（Negative Binomial Poisson Distribution）

(1) 方差大于均值，即过离散（overdispersion）。

3 零膨胀的泊松回归（Zero Inflated Poisson Distribution）

1. poisson y x1 x2 x3, irr
   （irr表示汇报incidence rate ratio，即汇报回归系数的指数形式）
2. predict yhat /*计算y的预测值*/
3. poisgof
   /*代表poisson goodness of fit，估计模型拟合优度*/
4. listcoef
   /*listing coefficients，列出模型估计的系数。需要先下载*/
   (1) findit spost13 /*寻找下载地址*/
   (2) 或者net install spost13 /*下载安装命令*/

用Sherry McKibben and Dudley Poston对1万多名中国妇女研究的数据(1997)来做示范。

中国妇女曾生子女数据 s-d_ceb.dta		
变量	变量标签	变量值
livebir	Number of live births	子女数（个）
menarche	first menstruation in months	月经初潮（月）
educ	education completed in years	教育年限（年）
policy	first pregnancy occurred after 1980	是=1，否=0
rural	if live in rural area	是=1，否=0

研究假设

A hypothesis is a testable statement of the proposed relationship between the independent variable, which measures the cause, and the dependent variable, which measures the effect.（Pollock, 2015）

H1：妇女初潮的年龄越大，曾生子女数越多。
H2：受教育年限越多，曾生子女数越少。
H3：如果初孕晚于1980年，生育数便较少。
H4：农村妇女比城市妇女生育数更多。

什么是好的研究假设？

1. 讨论两个或以上变量（e.g. 明天天气会变暖）
2. 要指出自变量和因变量的关系 （e.g. 降水和温度）
3. 要解释自变量和因变量是怎样的关系 （e.g. 正相关、负相关...）
4. 可以通过实证方法验证
5. 假设有推广性或一般性（generality）

```
poisson livebir menarche educ policy rural
```

```
Iteration 0:   log likelihood = -15438.903
Iteration 1:   log likelihood =   -15438.9
Iteration 2:   log likelihood =   -15438.9

Poisson regression                              Number of obs   =      10919
                                                LR chi2(4)      =    1765.71
                                                Prob > chi2     =     0.0000
Log likelihood =   -15438.9                     Pseudo R2       =     0.0541

------------------------------------------------------------------------------
     livebir |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    menarche |   .0031248   .0003397     9.20   0.000      .002459    .0037907
        educ |  -.0352641   .0019389   -18.19   0.000     -.0390642    -.031464
      policy |  -.2471372   .0156392   -15.80   0.000     -.2777894   -.2164849
       rural |   .2560273   .0216347    11.83   0.000      .2136239    .2984306
       _cons |   .1902786   .0707621     2.69   0.007      .0515875    .3289698
------------------------------------------------------------------------------
```

可以计算泊松模型的拟合优度(Poisson goodness of fit)并对 $\mathbf{H_0}$：“观测数据分布与泊松模型预测分布之间无差异”进行检验。这一统计检验与通常的检验不同，因为我们期望它不显著，即无差异假设如果不能被拒绝，则说明模型预测比较好，如果统计性显著则表明模型预测很糟糕，不能较好地拟合观测数据。所以，研究人员这时其实希望卡方值要小，显著度要比 0.05 大。

当拟合模型还在内存中时，用命令 **poisgof** (代表 **pois**son **g**oodness **o**f **f**it)要求估计模型拟合优度。

```
poisgof
```

```
      Deviance goodness-of-fit   =  5271.248
      Prob > chi2(10914)         =     1.0000
```

尽管从图形比较看出模型对观测值的拟合并不是很好，但是从输出的模型拟合优度检验却很好，显著度为 1>>0.05，说明模型拟合得非常好，不能拒绝“模型预测与观测数据无差异”的假设。也就是说，观测数据可以认为是服从泊松分布的。

```
listcoef, help
```

poisson (N=10919): Factor Change in Expected Count

Observed SD: 1.0712442

livebir	b	z	P>|z|	e^b	e^bStdX	SDofX
menarche	0.00312	9.198	0.000	1.0031	1.0706	21.8175
educ	-0.03526	-18.188	0.000	0.9654	0.8544	4.4618
policy	-0.24714	-15.802	0.000	0.7810	0.8945	0.4511
rural	0.25603	11.834	0.000	1.2918	1.1151	0.4256

1 由于泊松分布的特殊性质，用λ来帮助理解估计结果，即$\lambda = E(Y|X) = e^{x\beta}$。

2 在其他变量不变的情况下，X每增加一个单位，将导致Y的期望频数（expected counts）是原来的e^{β}倍，即增加（$e^{\beta} - 1$）倍。

表下注明，e^b 一栏就是 IRR，即 exp(b)，其意义是 X 上一个单位的增加导致期望计数上的乘数变化（即 factor change）。这与发生比率的概念相同。

比如，变量 rural 的 IRR 为 1.2918（$=e^{0.25603}$），表明：在其他自变量不变时，农村妇女与城市妇女相比，期望曾生子女数要乘上 1.29，即增加 0.29 倍。

在其他不变条件下，1980 年独生子女政策出台后才初孕的妇女的生育数只有更早进入生育的妇女的生育数的 0.78 倍，即生育量只相当于更早生育者的 78%。

在其他不变条件下，妇女初潮年龄每推后 1 个月将导致其生育数乘以 1.003 倍。

Figure 3.4: 系数的解释

```
listcoef, help
```

```
poisson (N=10919): Factor Change in Expected Count

 Observed SD: 1.0712442

------------------------------------------------------------------------------
     livebir |      b         z     P>|z|     e^b    e^bStdX      SDofX
-------------+----------------------------------------------------------------
   menarche |   0.00312     9.198   0.000    1.0031    1.0706    21.8175
      educ |  -0.03526   -18.188   0.000    0.9654    0.8544     4.4618
    policy |  -0.24714   -15.802   0.000    0.7810    0.8945     0.4511
     rural |   0.25603    11.834   0.000    1.2918    1.1151     0.4256
------------------------------------------------------------------------------
```

1. 变量rural的IRR：1.2918$=e^{0.25603}$，表明：在其他自变量不变时，农村妇女的生育数是城市妇女的1.29倍（即增加0.29倍）。
2. 1980年后才初孕的妇女的生育数为更早进入生育的妇女0.78倍（即生育量只相当于更早生育者的78%）。
3. 妇女初潮年龄推后1个月将导致其生育数乘以1.003倍。
4. 妇女教育年数增加1年将导致其生育数减少3.5%。

负二项泊松分布

1 如果随机变量Y_i过度发散，方差大于均值，即过离散（overdispersion），则属于负二项泊松分布，这时应使用负二项回归模型（Negative Binomial Regression）来代替泊松回归模型进行估计。

1. nbreg y x1 x2 x3, irr
   （irr表示汇报incidence rate ratio，即汇报回归系数的指数形式）
2. nbreg命令的结果中包含了过度分散效应的likelihood ratio test of alpha。
   (1) 零假设是等分散。
   (2) 若发现alpha显著（$P < 0.05$），则说明该数据存在过度分散的状况。
   (3) 提示我们以负二项回归拟合数据要优于泊松回归。

零膨胀的泊松回归

在社会科学的实际研究中，经常发现观察事件发生次数中含有大量的零值，即许多观察个体在观察单位时间、空间、面积内没有发生某一随机事件。

1. 一年内的住院次数
2. 一生的离婚次数
3. 一生的坐牢次数
4. 15-49岁育龄女性的生育子女数
5. 女性人工流产次数

零膨胀的泊松回归

1. 首先，决定事件数取零或取正整数。
   (1) 二分Logit模型
2. 其次，若事件数取正整数，则具体决定选择哪个正整数。
   (1) 泊松模型或负二项模型

若计数数据中0的个数太多（约大于50%），这时可以考虑零膨胀的泊松回归。

1 零膨胀泊松模型

(1) zip y x1 x2 x3, inflate(x1 x2 x3) vuong

(2) “inflate( )”列出影响0、1过程中的所有变量。

(3) vuong统计比较ZIP和泊松回归哪个更好。
若检验结果所对应的P值小于0.05，表明ZIP更好。

2 零膨胀负二项模型

(1) zinb y x1 x2 x3, inflate(x1 x2 x3) vuong

1. 存在过度分散