社会统计学及SPSS软件应用
STATISTICS WITH SPSS

Instructor:王荣欣

Email: rxwang@qq.com

周二3-4节、单周周四3-4节，3A106-2

2020年10月8日

CONTENTS

参数检验

1 单样本的假设检验
- 单样本Z检验
- 单样本T检验
- 单样本χ^2检验（适用总体方差和标准差）

2 双样本的假设检验
- Paired-samples test
- Independent sample test

3 多样本的假设检验
- One-way ANOVA
- Two-way or Multi-way ANOVA

非参数检验（eg. χ^2检验）

答疑

否定域与显著性水平（3）

1. Decision rule 1: Reject the null hypothesis H_o: $\mu \geq \mu_0$ if Z $\leq c$, where c is the α quantile of a standard normal distribution.
2. Decision rule 2: Reject the null hypothesis H_o: $\mu \leq \mu_0$ if Z $\geq c$
3. Decision rule 3: Reject the null hypothesis H_o: $\mu = \mu_0$ if Z $\leq -c$ or $Z \geq c$

0.025 quantile of a standard normal distribution is c=-1.96. If you want the probability of a Type I error to be α=0.025, the critical value is -1.96.

If α=0.005, then the critical value is -2.58.

例子：面包的均重

面包的标准重量是500g。
连续购买10天的面包重量依次是 492 496 497 490 493 490 495 492 489 494

例子：面包的均重（双尾检验）

1. H_0: 面包的均重是500g（面包师傅没有克扣面粉）。
   - 或H_0: 面包的均重既不大于500g，也不小于500g。
2. H_1: 面包的均重不是500g。

```
. ttest 面包重量=500

One-sample t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
面包重量 |      10       492.8    .8537499    2.699794    490.8687    494.7313
------------------------------------------------------------------------------
    mean = mean(面包重量)                                         t =  -8.4334
Ho: mean = 500                                 degrees of freedom =       9

   Ha: mean < 500               Ha: mean != 500                 Ha: mean > 500
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

例子：面包的均重（单尾检验）

单边检验（one-sided test），亦称单尾检验（one-tailed test）。

1. H_0: 面包的均重不小于500g。
2. H_1: 面包的均重小于500g。（备择假设是“$<$”的形式，是左侧检验）

例子：面包的均重（单尾检验）

1. H_0: 面包的均重不大于500g。
2. H_1: 面包的均重大于500g。（备择假设是“$>$”的形式，是右侧检验）

双尾检验

研究问题：总体的平均年龄μ是否为24岁？

1. H_0: 总体参数是24。
2. H_1: 总体参数不是24（可能大于24，也有可能小于24）。

双尾检验

研究问题：工资的差别与教育程度的差别是否有关（有显著关系）？

1. H_0: 工资的差别与教育程度的差别没有显著关系。
2. H_1: 工资的差别与教育程度的差别有显著关系。

单尾检验

1. H_0: 总体中两个变量之间的关系大于0（是正相关）。
2. H_1: 总体中两个变量之间的关系是负相关。
   （备择假设是“$<$”的形式，是左侧检验）

单尾检验

1. H_0: 总体中两个变量之间的关系小于0（是负相关）。
2. H_1: 总体中两个变量之间的关系是正相关。
   （备择假设是“$>$”的形式，是右侧检验）

T TEST

Dependent variable is continuous, and independent variable is dichotomous(eg. 男性/女性、城市/农村).

1. One-sample T test
2. Paired-samples T test
3. Independent sample T test

Test the significance by $*p \leq 0.1, **p \leq 0.05, ***p \leq 0.01$, and if it is significant, it means findings from a sample can be found in the population.

ONE-SAMPLE T TEST

The purpose of the t test for a single **sample mean** is to determine whether the mean for a random sample of participants differs significantly from a known value (**population mean**) or a hypothetical value （假定总体均值知道）.

PAIRED-SAMPLES T TEST

A researcher drew a random sample from a population and administered a depression scale to the sample. This administration of the scale yielded **pre-test scores**, for which the researcher computed a mean. Then, the researcher administered a new antidepressant drug to the sample. Next, the researcher administered the depression scale again, which yielded **posttest scores**. As a result, for each pretest score earned by an individual, there is an associated posttest score for the same individual. These sets of scores are paired scores.

通过对比自变量两个平均数的差异，判断这种差异在总体中是否存在，即是否具有显著性，从而判断自变量和因变量之间是否有关系。

ONE-SAMPLE T TEST

Example 2.1

假设有人提出*1988*年全国城市居民平均收入为*3800*元，而在*1988*年的*CHIP*数据中，我们发现居民的年收入均值为*3687*元，标准差为*3489*元。
那么在*0.05*的显著性水平下，这一样本结果与*3800*元的提法一致吗？

1. 首先，建立$H_0 : \mu = 3800$ 和 $H_1 : \mu \neq 3800$
2. 根据样本数据，计算T检验统计量
3. stata 命令：ttest earn =3800
4. 结论：否定H_0，即认为全国城市居民平均收入不是3800元。

```
. ttest earn=3800

One-sample t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    earn |   20421    3687.005     24.4163    3489.139    3639.147    3734.862
------------------------------------------------------------------------------
    mean = mean(earn)                                             t =  -4.6279
Ho: mean = 3800                                 degrees of freedom =    20420

   Ha: mean < 3800               Ha: mean != 3800                 Ha: mean > 3800
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Figure 2.1: One-sample T Test

Example 2.2

在1988年的CHIP数据中，有女性9700人，其年收入均值为3329元，标准差为3197元；有男性10721人，其年收入均值为4011元，标准差为3704元。

那么在0.05的显著性水平下，是否存在收入的性别差异？

1. 首先，建立$H_0 : \mu_1 = \mu_2$ 和 $H_1 : \mu_1 \neq \mu_2$
2. 检验方差是否相等（F检验）
3. 根据样本数据，计算T检验统计量
4. stata 命令：ttest earn, by(sex) unequal
5. 结论：否定H_0，即认为男性和女性的年收入均值是不相等

N_1=9700，$\bar{y_1}$=3329元，s_1=3197；N_2=10721，$\bar{y_2}$=4011元，s_2=3704元
那么，在0.05 的显著性水平下，是否存在收入的性别差异？

1. 首先，建立$H_0 : \mu_1 = \mu_2$ 和 $H_1 : \mu_1 \neq \mu_2$
2. 若满足H_0，男性和女性所有可能样本平均收入的分布为同一分布，这一分布的均值应该相等。
3. 尽管从同一分布得到男性和女性样本平均收入可能不同，但二者的距离不应该很远。若二者的距离太远，我们就认为在H_0下是不可能的，即推翻H_0。那么，多远叫“远”？
4. 若t值超过了设定的临界点（即$p < 0.05$），我们就认为二者的距离已经足够远了，从而推翻H_0。

```
. ttest earn, by(sex) unequal

Two-sample t test with unequal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
male	10721	4010.848	35.77724	3704.456	3940.718	4080.978
female	9700	3329.074	32.45828	3196.77	3265.449	3392.699
combined	20421	3687.005	24.4163	3489.139	3639.147	3734.862
diff		681.7748	48.30684		587.0895	776.4601

```
    diff = mean(male) - mean(female)                              t =  14.1134
Ho: diff = 0                     Satterthwaite's degrees of freedom =  20373.8

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Figure 2.2: Two-sample T Test

Analysis of Variance

比较多个独立总体的均值是否有差异。

Dependent variable is is interval or ratio, and independent variable is non-dichotomous.

One-way ANOVA

Example 2.3

在*CGSS2003*数据中，对收入（取对数）进行方差分析，检验不同教育程度的收入水平是否存在差异。

1. 首先，建立$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ 和 $H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$
2. 根据样本数据，计算F检验统计量
3. stata 命令：anova lninc edu
4. 结论：否定H_0，即认为不同教育程度的收入水平存在差异。

```
. tab edu,sum(lninc)
```

RECODE of educ (教育程度)	Summary of lninc		
	Mean	Std. Dev.	Freq.
0	5.8131884	.77492148	9
1	6.2615059	.7317007	52
2	6.3786467	.83584724	138
3	6.594358	.6006642	167
4	7.0642606	.64383691	92
Total	6.5706113	.75885336	458

Figure 2.3: 描述性统计

```
anova lninc edu
```

```
                           Number of obs =       458     R-squared     =  0.1434
                           Root MSE      =  .705445     Adj R-squared =  0.1358

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+------------------------------------------------------
                   Model |   37.730583     4  9.43264576       18.95     0.0000
                         |
                     edu |   37.730583     4  9.43264576       18.95     0.0000
                         |
                Residual |   225.436718   453  .497652798
              -----------+------------------------------------------------------
                   Total |   263.167301   457  .575858426
```

Figure 2.4: One-way ANOVA

Two-way or Multi-way ANOVA

Example 2.4

在2003年的CGSS数据中，对收入（取对数）进行方差分析，检验不同教育程度、不同性别、不同党员身份的收入水平是否存在差异。

1. 首先，建立H_0
2. 根据样本数据，计算F检验统计量
3. stata 命令：anova lninc edu sex party
4. 结论：否定H_0

```
. anova lninc edu sex party
```

Number of obs =	452	R-squared	=	0.1762	
Root MSE	= .694816	Adj R-squared	=	0.1651	

Source	Partial SS	df	MS	F	Prob > F
Model	45.9401838	6	7.65669731	15.86	0.0000
edu	32.4662959	4	8.11657397	16.81	0.0000
sex	3.87457706	1	3.87457706	8.03	0.0048
party	2.95558709	1	2.95558709	6.12	0.0137
Residual	214.832278	445	.482769164		
Total	260.772462	451	.578209449		

Figure 2.5: Multi-way ANOVA

归纳与总结

测量层次	单个变量	分类+连续	分类+连续
针对样本	＞＝＜	两个平均数差异的比较	多个平均数差异的比较
针对总体	标准误、置信区间	双样本T检验	ANOVA：F检验

Nonparametric statistics for analyzing data that

1. are not normally distributed
2. are measured at the **nominal or ordinal** level
3. were not randomly selected
4. were selected from a very small sample

HYPOTHESIS TESTING WITH CHI-SQUARE

两个分类变量的独立性。

1. H_0：两个变量之间没有关系。
2. 如果卡方检验的结果是有显著性差异，则说明两个变量之间不是独立的，是有相关性的。

CHI-SQUARE

1. By the end of 1896, Pearson wanted to develop a goodness of fit test for asymmetrical distributions for biologists and economists, which culminated in his chi-square goodness of fit test in 1900.
2. Chi-square distribution can be used for discrete data that takes on any-shaped distribution, such as asymmetrical, binomial or Poisson distributions.
3. R.A.Fisher formulated the ”degrees of freedom” in 1922 to determine if the chi-square results were statistically significant or not.

卡方分布

若X_1，X_2，…X_n为来自标准正态分布的n个样本，则它们的平方和$X_1^2+X_2^2+X_3^2...+X_n^2$就记作$\chi^2$，且$\chi^2$服从参数为n的卡方分布，记作$\chi^2$服从$\chi^2(n)$。

1. Observed frequencies
2. Expected frequencies are those numbers that we expect to observe in each cell（单元格） if there is no association between the variables.

CHI-SQUARE TEST(OBSERVED)

		性别		
		男	女	总数
是否经理	否	290	100	390
	是	80	4	84
总数		370	104	474

CHI-SQUARE TEST(EXPECTED)

		性别		
		男	女	总数
是否经理	否	304.4	85.6	390
	是	65.6	18.4	84
总数		370	104	474

单元格的预期值，等于该单元格所在行的边缘值（marginals）乘以它所在列的边缘值，除以大总数（grand total）。

卡方检验的两个要素

1 卡方值：衡量预期与观察之间差距的指标。

$$\chi^2 = \sum \frac{(observed\ \ value - expected\ \ value)^2}{expected\ \ value}$$

2 自由度：指的是任意度。2×2列联表有四个单元格，这个表格的自由度是1，意味着只有一个单元格可以“任意取值”，只有一个单元格是“自由的”。

degree of freedom = (r-1)(c-1)

r = number of rows, c = number of columns

思考题

1 简述列联表中两个变量相互独立的条件。

思考题

Association between child abuse and substance abuse by parent of abusers (n=49)

Substance Abuse by Parent of Abusers	Child Abuse	
	Yes	No
Yes	18	19
No	2	10

$\chi^2 = 3.84(1), p < 0.05$, read as "Chi-square with one degree of freedom equals 3.84".

A Typology of Statistical Test

Statistical Test	Dependent Variable	Independent Variable
One sample t-test	interval or ratio	
Independent samples t-test	interval or ratio	nominal with only two categories
Paired samples t-tests	interval or ratio	nominal with only two categories
ANOVA	interval or ratio	nominal with more than two categories
χ^2	nominal or ordinal	nominal or ordinal

INFERENTIAL STATISTICS

1. Hypothesis testing is a scientific procedure for making rational decisions about two different claims. （先假设总体的情况，然后进行抽样，检验原有的假设是否成立）
   - 假设检验则是先对总体参数提出一个假设，然后利用样本信息来判断这一假设是否成立。
2. Estimation theory is a branch of statistics that deals with estimating the values of parameters. （先看样本情况，再问总体的情况）
   - 参数估计是利用样本信息来推断未知的总体参数。

估计（ESTIMATION）

估计是指从总体中随机抽取一个样本，利用样本统计量推算总体参数的过程。

- 点估计（point estimation）是指根据样本数据中计算出的样本统计量对未知的总体参数进行估计，得到的是一个确切的值。
- 区间估计（interval estimation）是指对总体未知参数的估计是基于数据计算出的一个取值范围。

区间估计

1. Z分布与总体均值的区间估计
2. T分布与总体均值的区间估计
3. χ^2分布与总体均值的区间估计
4. F分布与总体均值的区间估计

CONFIDENCE INTERVAL

A confidence interval is used to estimate a population parameter when we have no idea what the parameter value is. We are simply interested in using sample statistics to find and estimate the values of population parameters. What is a good estimate of a parameter of a variable?

confidence interval（$1-\alpha$）一般设为90%、95%和99%。
相对应的显著性水平（α）为0.1、0.05和0.01。

1. if confidence level is 90%, then $\mu = \bar{y} \pm 1.65 \times SE$
2. if confidence level is 95%, then $\mu = \bar{y} \pm 1.96 \times SE$
3. if confidence level is 99%, then $\mu = \bar{y} \pm 2.58 \times SE$

CONFIDENCE LIMITS FOR μ

$1-\alpha$	α	**Z**	**LCL**	**UCL**
0.90	0.10	1.645	$\bar{y}-1.645\times SE$	$\bar{y}+1.645\times SE$
0.95	0.05	1.96	$\bar{y}-1.96\times SE$	$\bar{y}+1.96\times SE$
0.99	0.01	2.58	$\bar{y}-2.58\times SE$	$\bar{y}+2.58\times SE$

$1-\alpha$ 置信水平

α 显著性水平

Example 3.1

Given: An interval/ratio variable Y

$\bar{y} = 27$
$\sigma_Y = 10$, *n=25*

Research question: What is the value of the population mean, μ*?*

1

$z = \frac{\bar{y}-\mu}{\frac{\sigma}{\sqrt{n}}}$

$\bar{y}$ - μ = $\pm 1.96 \times SE$ （取z值1.96，即有95%的信心）

$\mu = \bar{y} \pm 1.96 \times SE$

$SE = \frac{\sigma_Y}{\sqrt{n}} = 2$ ($\sigma_Y = 10$, n=25)

$\mu = 27 \pm 1.96 \times 2 = (23.08, 30.92)$

结论：The confidence interval of μ (that is the average age of population) at the confidence level of 95% (that is 1-α) is (23.08, 30.92).

思考题

在其他条件相同的情况下，95%的置信区间比90%的置信区间更宽，还是更窄？

CONFIDENCE INTERVAL

Example 3.2

数据：*chip.dta*

```
sum earn
mean earn
display 3489.139/sqrt(20421)
display 3687.005-1.96*24.416295
display 3687.005+1.96*24.416295
```

```
. sum earn
```

Variable	Obs	Mean	Std. Dev.	Min	Max
earn	20,421	3687.005	3489.139	154.8	76300

```
. mean earn

Mean estimation                 Number of obs    =     20,421
```

	Mean	Std. Err.	[95% Conf.	Interval]
earn	3687.005	24.4163	3639.147	3734.862

```
. display 3489.139/sqrt(20421)
24.416295

. display 3687.005-1.96*24.416295
3639.1491

. display 3687.005+1.96*24.416295
3734.8609
```