统计学（Stata实现）

第二次上机课

赵震宇（2120253538）
南开大学 国际经济研究所
zzynankai@outlook.com
xishanyu2.github.io
2025 年 12 月 11 日

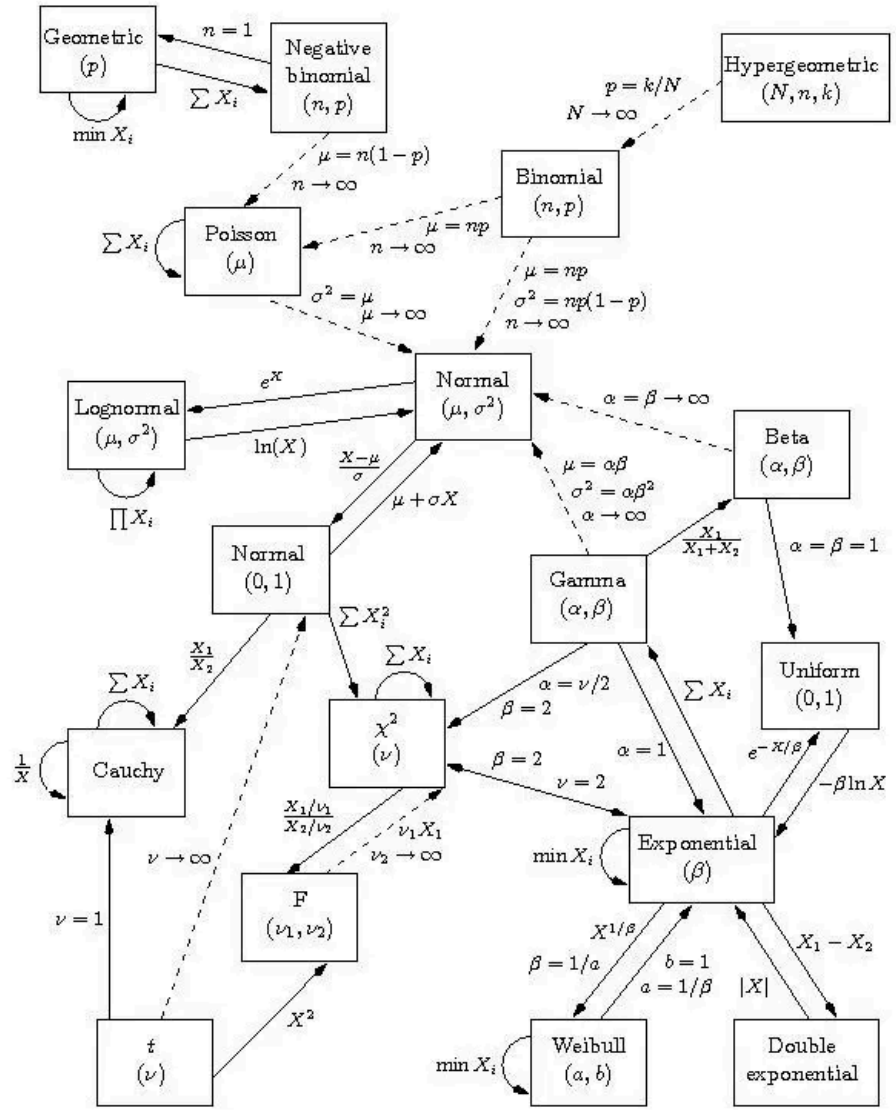
CONTENT

目录

1. 概率分布

15种离散和连续概率分布理论及软件实现

- 扎马步-常用概率分布函数详解：Stata篇
- 扎马步-常用概率分布函数详解：R篇
- 扎马步-常用概率分布函数详解：Python篇



Relationship among distributions. Solid lines represents transformations and special cases, while dashed lines represents limits. Adapted from Leemis, L. M. (1986). Relationships Among Common Univariate Distributions, *The American Statistician* **40**: 143-146.

1.1 离散分布

1.1.1伯努利分布 X ~ Bernoulli(p)

$$P(X = k) = p^k(1-p)^{1-k}$$

```
* 参数设置
local p = 0.7
* 计算 PMF (P(X=1) 和 P(X=0))
* 使用二项分布函数binomialp(n,k,p), 其中试验次数 n=1
display "P(X=1): " binomialp(1, 1, `p')
display "P(X=0): " binomialp(1, 0, `p')
```

1.1.2 二项分布 X ~ Binomial(n,p)

- 二项分布描述了**多次独立伯努利**试验中的成功次数：

$$P(X = k) = C_n^k p^k (1-p)^{n-k}$$

```
* 参数设置
local n = 10 // 试验次数
local p = 0.3 // 成功概率
* 计算 PMF (例: k=3)
display "P(X=3): " binomialp(`n', 3, `p')
```

- 当试验次数n很大且成功概率p适中时，根据中心极限定理，二项分布可以用**正态分布**来近似。
- 当n很大，p很小，使得 λ =np的值适中时，二项分布可以用**泊松分布**来近似。

1.1.3 泊松分布 X ~ Poisson(λ)

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

```
* 参数设置
local lambda = 3.5
* 计算 PMF (例: k=4)
display "P(X=4): " poissonp(`lambda', 4)
```

- 泊松分布是**二项分布**在n→∞, p→0, np= λ 条件下的极限。
- 当 λ 较大时(通常 λ >20)，泊松分布可以用**正态分布**来近似。

1.2 连续分布

1.2.1 正态分布 X ~ N(μ,σ²)

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

```
* 参数设置
local mu = 10 // 均值
local sigma = 2 // 标准差
* 计算 PDF (例: x=11)
display "PDF at x=11: " normalden(11, `mu', `sigma')
* 计算 CDF (例: x=11)
display "CDF at x=11: " normal((11 - `mu')/`sigma')
```

1.2.2 标准正态分布 X ~ N(0,1)

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

```
* Stata练习：正态分布曲线下μ±1.96σ区间的面积占总面积的百分比?
```

- 中心极限定理指出，从任何具有有限均值和有限方差的总体中抽取足够大的独立随机样本时，样本均值的分布将近似于正态分布，无论原始总体的分布形状如何。
- 通常样本大小n≥30被认为“足够大”，但具体取决于原始分布的偏度和峰度。
- 使得正态分布在假设检验和置信区间构建中扮演核心角色。

1.2.3 绘制正态分布的pdf曲线

```
twoway function y=normalden(x),range(-5 5) xline(0) ytitle("概率密度")
```

```
twoway function z=normalden(x,1,2),range(-5 10) xline(1) ytitle("概率密度")
```

1.2.4 绘制标准正态分布的CDF曲线

```
twoway function y=normal(x), range(-5 5) title("累计分布")
```

1.2.5 计算标准正态分布的概率、反函数值

```
/*
Stata练习:
已知学生的统计考试成绩服从均值为72，标准差等于8的正态分布，求:
（1）学生成绩不及格的概率;
（2）学生成绩处于65-80之间的概率。
*/
```

1.2.6 标准正态分布表、标准正态分布分位数表

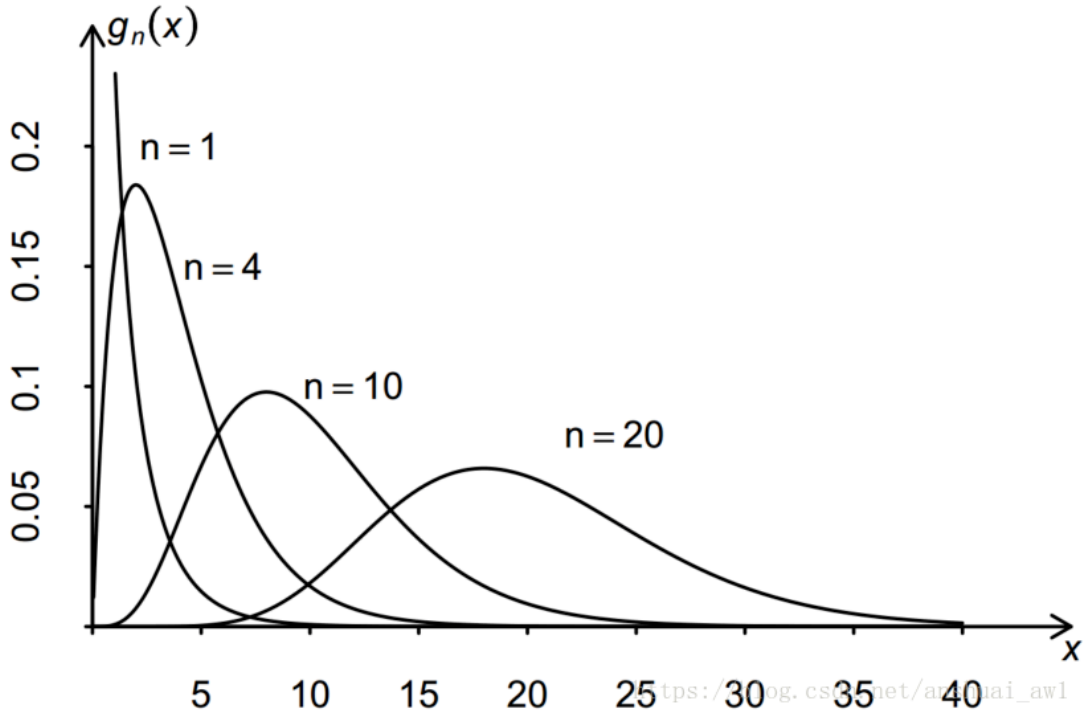
`ztable`

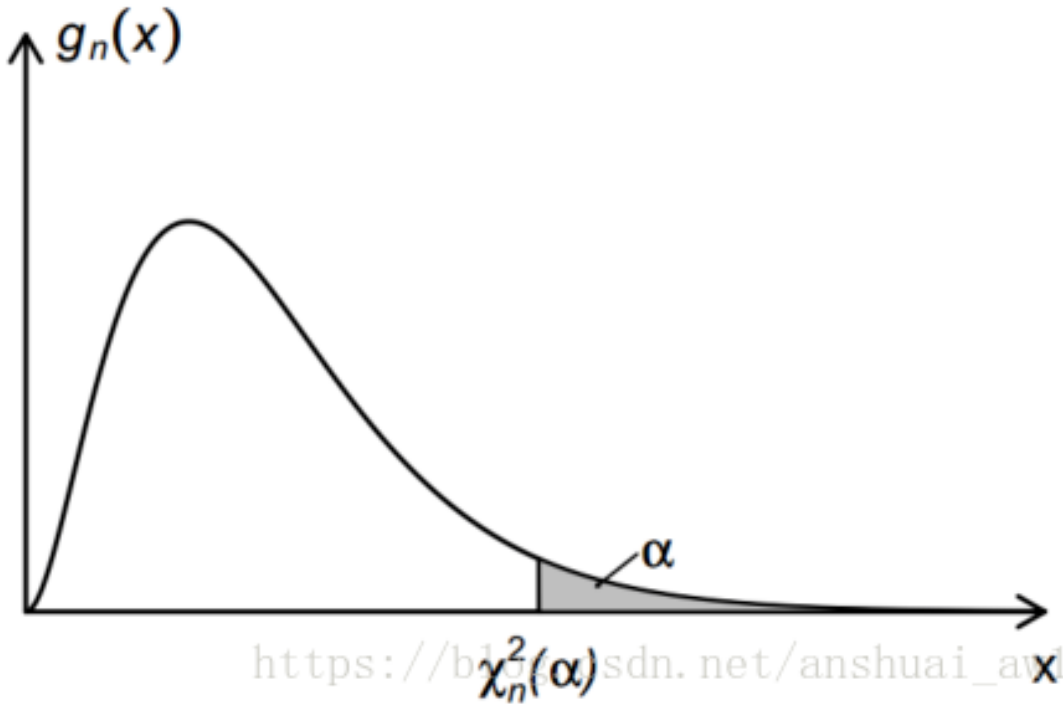
```
net install probtabl, replace from(https://stats.idre.ucla.edu/stat/stata/ado/teach)
```

2. 抽样分布

2.1 χ^2分布

设$X_1, X_2, \cdots, X_n \sim NID(0,1)$，令$X = \sum_{i=1}^{n} X_i^2$，则称$X$是自由度为$n$的$\chi^2$变量，其分布称为自由度为$n$的$\chi^2$分布，记为$X \sim \chi_n^2$.





2.1.2 绘制χ^2分布的pdf曲线

```
twoway function chi3=chi2den(3,x),range(0 20) ///
|| function chi5=chi2den(5,x),range(0 20) lp(dash) ytitle("概率密度")
```

2.1.3 计算χ^2分布的概率、反函数值

`chi2`, `chi2tail`, `invchi2`, `invchi2tail`

2.1.4 χ^2分布临界值表

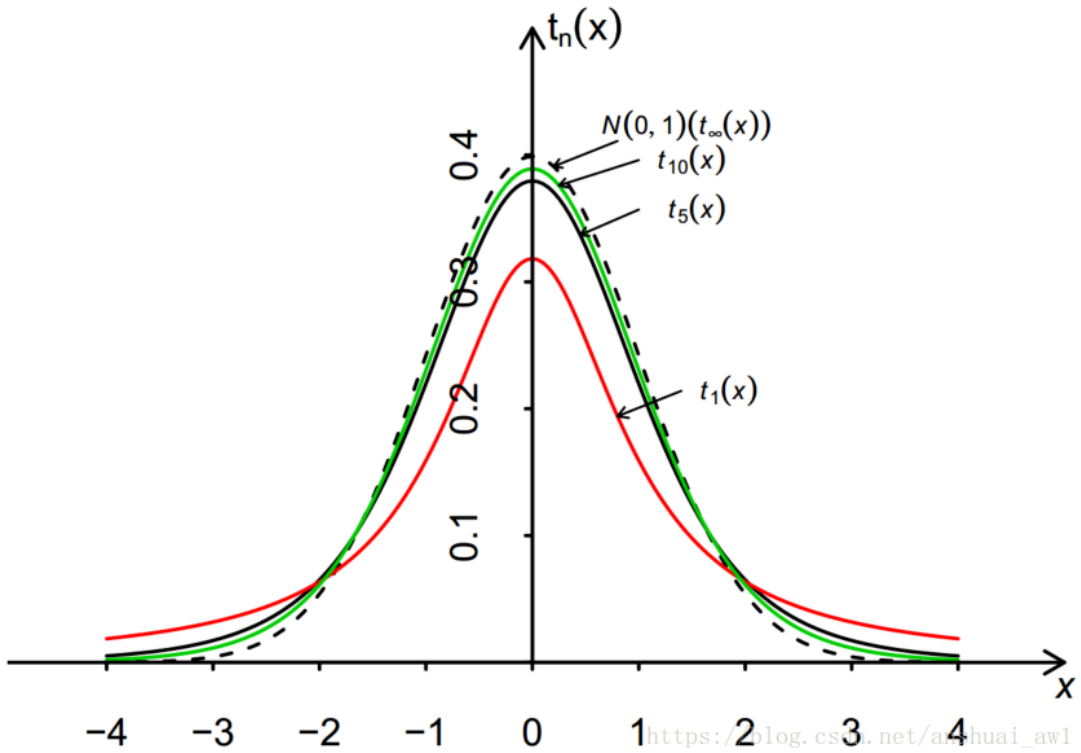
`chitable`, `chitable [df]`

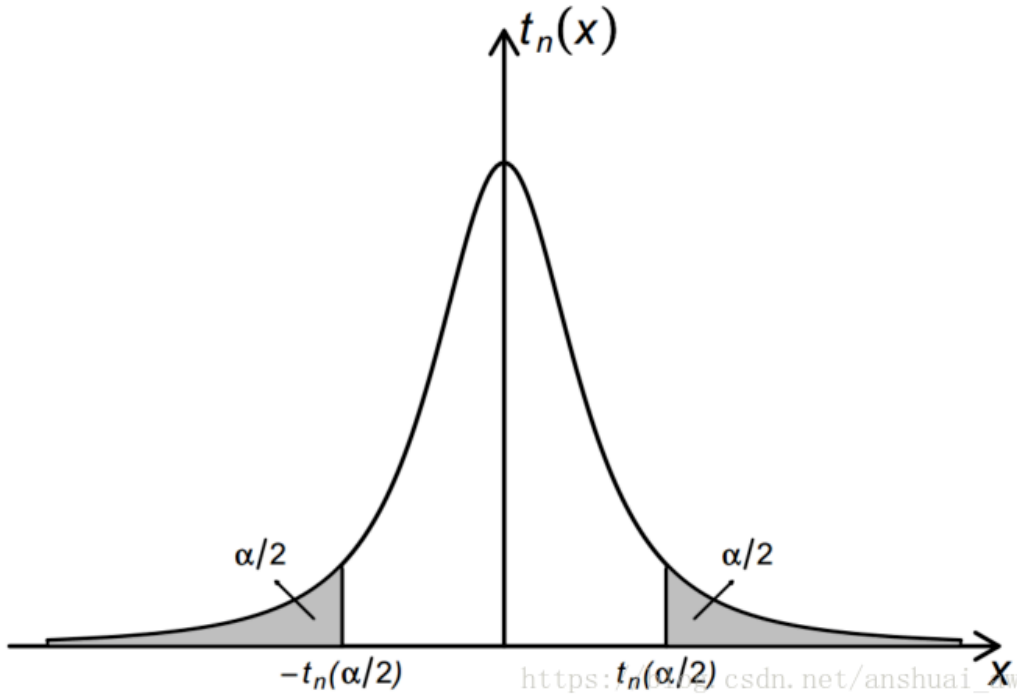
2.2 t分布

设随机变量$X \sim N(0,1), Y \sim \chi^2_n$，且$X$和$Y$独立，则称

$$T = \frac{X}{\sqrt{Y/n}}$$

为自由为n的t变量，其分布称为由为n的t分布，记为$T \sim t_n$.


$t_n(x)$
$N(0,1)(t_\infty(x))$
$t_{10}(x)$
$t_5(x)$
$t_1(x)$
0.4
0.3
0.2
0.1
x
−4 −3 −2 −1 0 1 2 3 4


$t_n(x)$
α/2
α/2
$-t_n(\alpha/2)$
$t_n(\alpha/2)$
x

2.2.2 绘制t分布的pdf曲线

```
twoway function t=tden(1,x),range(-5 5) ///
|| function y=normalden(x),range(-5 5) xline(0) ytitle("概率密度")
```

2.2.3 计算t分布的概率、反函数值

`t`, `ttail`, `invt`, `invttail`

```
* Stata练习：已知随机变量X服从自由度为10的t分布，计算P(|X|≤2)。
```

2.2.4 t分布临界值表

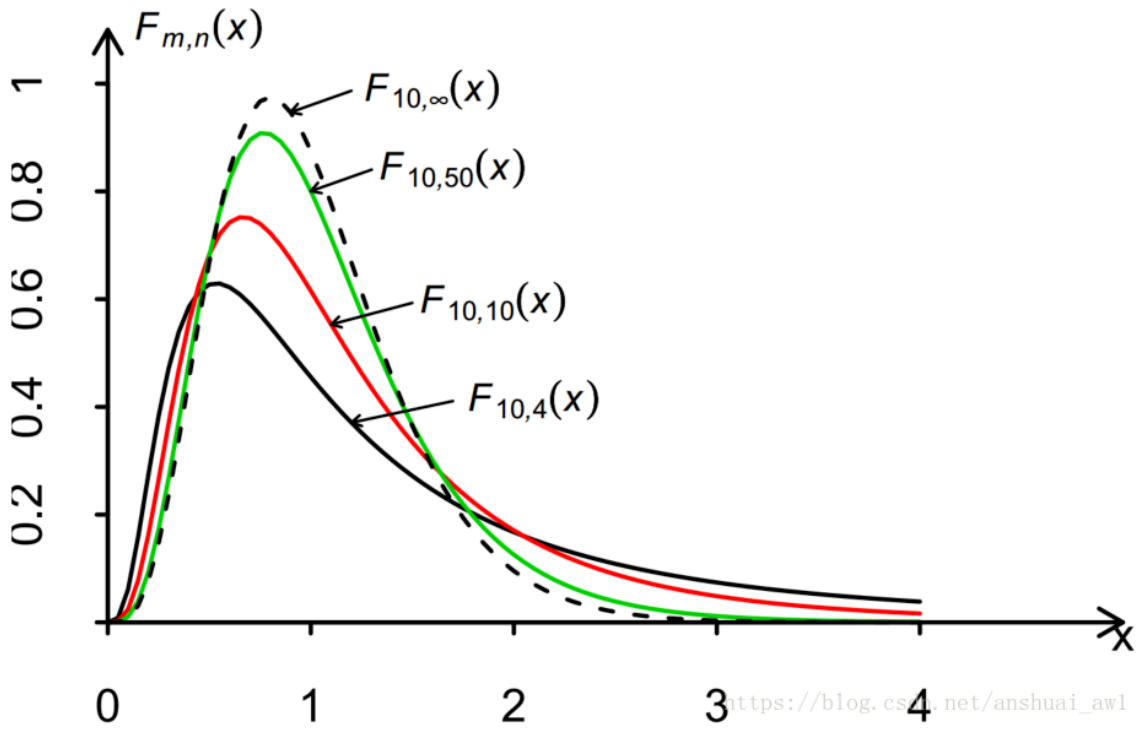
`ttable`, `ttable [df]`

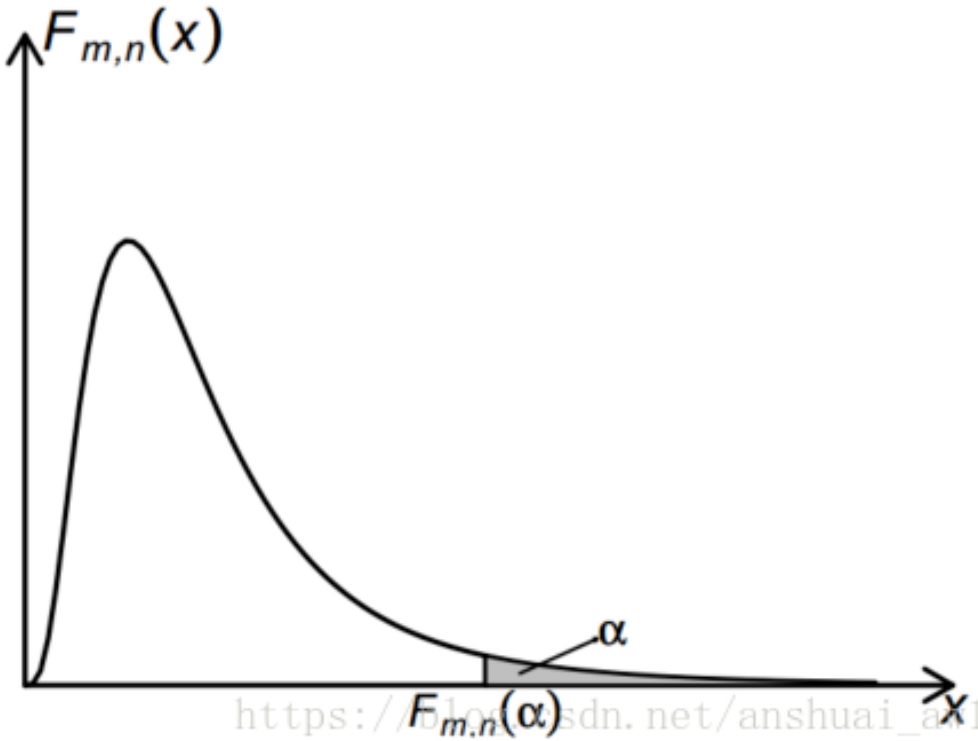
2.3 F分布

设随机变量$X \sim \chi^2_m, Y \sim \chi^2_n$, 且$X$和$Y$独立，则称

$$F = \frac{X/m}{Y/n}$$

为自由度分别是m和n的F变量，其分布称为自由度分别是m和n的F分布 ， 记为$F \sim F_{m,n}$





2.3.2 绘制F分布的pdf曲线

```
twoway function F20=Fden(10,20,x),range(0 5) ///
|| function F5=Fden(10,5,x),range(0 5) lp(dash) ytitle("概率密度")
```

2.3.3 计算F分布的概率、反函数值

`F`, `Ftail`, `invF`, `invFtail`

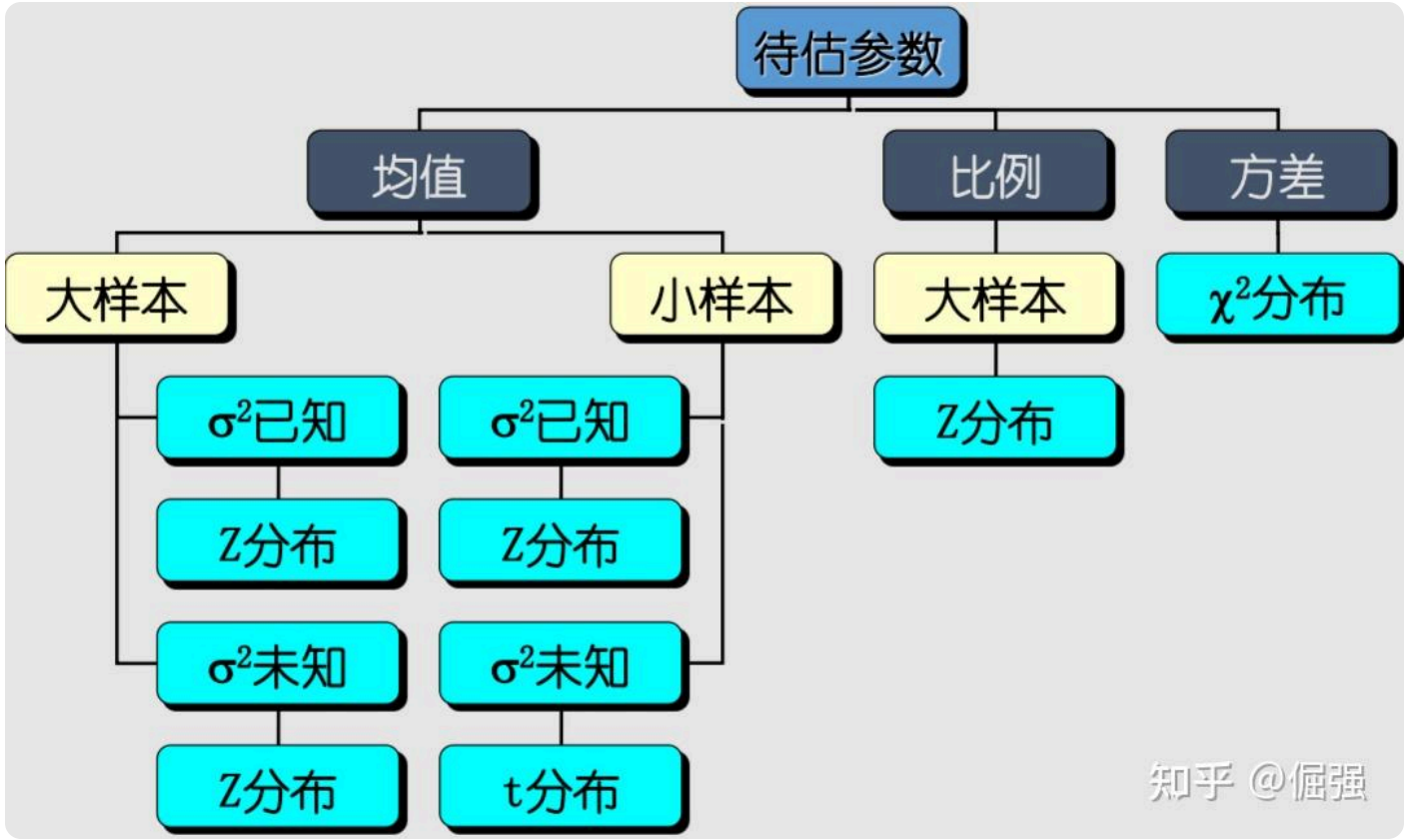
2.3.4 F分布临界值表

`ftable`, `ftable [ df1 df2 ] [, alpha() ]`

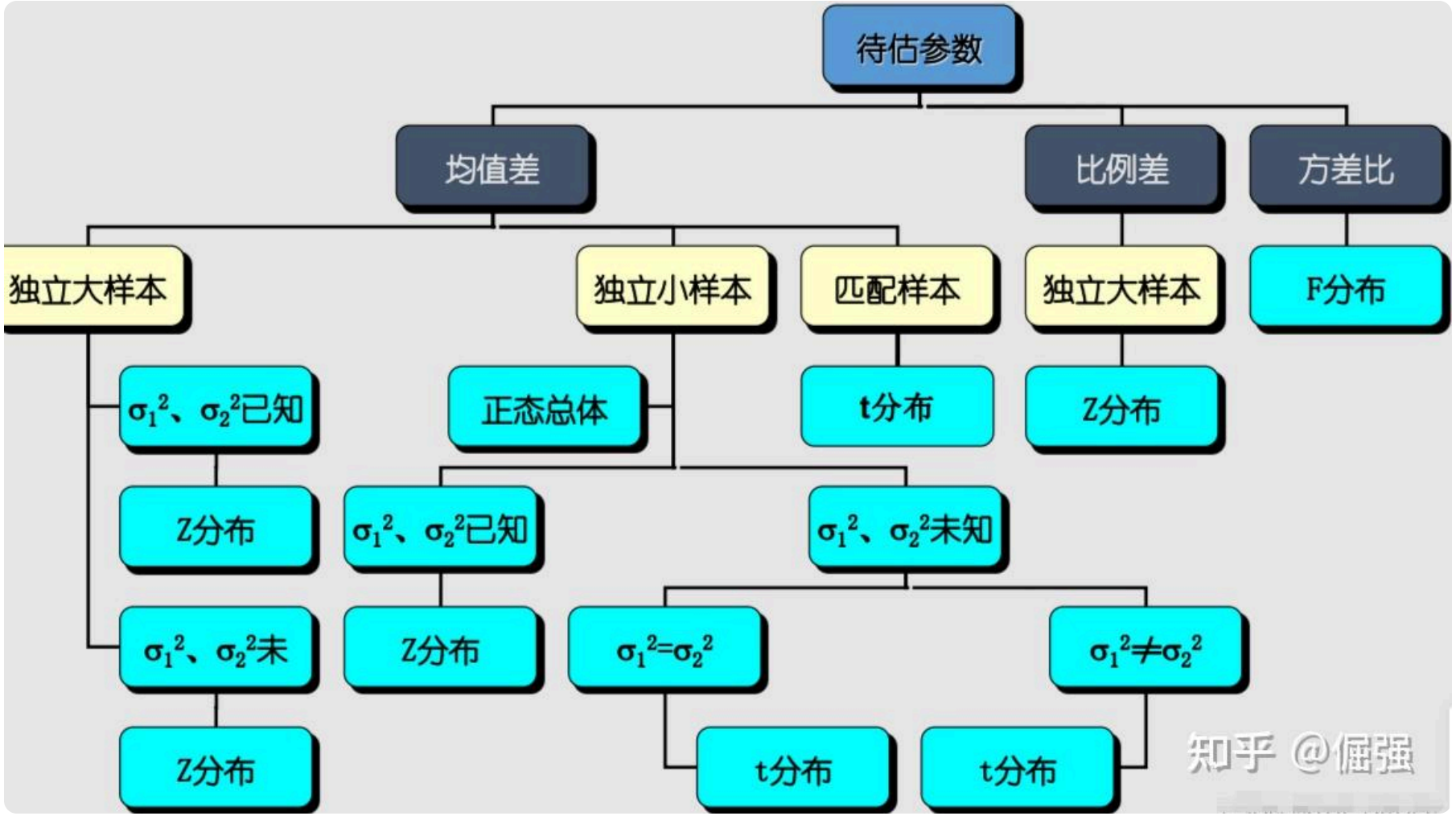
2.4 中心极限定理：蒙特卡罗模拟（略）

3. 参数估计

一个总体参数的区间估计



两个总体参数的区间估计



3.1 总体均值

3.1.1 原始数据已知

```
ci mean x //计算变量x均值的95%的置信区间
ci mean x, level(90) //计算变量x均值的90%的置信区间
```

3.1.2 原始数据未知

```
cii mean n x s //n样本容量，x样本均值，s样本标准差
```

3.2 样本比例

```
cii prop n np
```

```
* ci -- Confidence intervals for means, proportions, and variances

    Confidence intervals for means, normal distribution

        ci means [varlist] [if] [in] [weight] [, options]

        cii means #obs #mean #sd [, level(#)]

    Confidence intervals for proportions

        ci proportions [varlist] [if] [in] [weight] [, prop_options options]

        cii proportions #obs #succ [, prop_options level(#)]

    Confidence intervals for variances

        ci variances [varlist] [if] [in] [weight] [, bonett options]

        cii variances #obs #variance [, level(#)]
```

3.3 均值之差

3.3.1 样本数据已知

```
ttest x1 = x2
```

3.3.2 样本数据未知

```
ttesti n1 x1 s1 n2 x2 s2
```

```
* ztest -- z tests (mean-comparison tests, known variance)

    One-sample z test

        ztest varname == # [if] [in] [, onesampleopts]

    Two-sample z test using groups

        ztest varname [if] [in] , by(groupvar) [twosamplegropts]

    Two-sample z test using variables

        ztest varname1 == varname2 [if] [in], unpaired [twosamplevaropts]
```

```
Paired z test

    ztest varname1 == varname2 [if] [in] , sddiff(#) [level(#)]

    ztest varname1 == varname2 [if] [in] , corr(#) [pairedopts]


Immediate form of one-sample z test

    ztesti #obs #mean #sd #val [, level(#)]


Immediate form of two-sample unpaired z test

    ztesti #obs1 #mean1 #sd1 #obs2 #mean2 #sd2 [, level(#)]
```

```
* ttest -- t tests (mean-comparison tests)

    One-sample t test

        ttest varname == # [if] [in] [, level(#)]

    Two-sample t test using groups

        ttest varname [if] [in] , by(groupvar) [options1]

    Two-sample t test using variables

        ttest varname1 == varname2 [if] [in], unpaired [unequal welch level(#)]

         // unequal: unpaired data have unequal variances
```

```
Paired t test

    ttest varname1 == varname2 [if] [in] [, level(#)]

Immediate form of one-sample t test

    ttesti #obs #mean #sd #val [, level(#)]

Immediate form of two-sample t test

    ttesti #obs1 #mean1 #sd1 #obs2 #mean2 #sd2 [, options2]
```

```
* prtest -- Tests of proportions

    One-sample test of proportion

        prtest varname == #p [if] [in] [, onesampleopts]

    Two-sample test of proportions using groups

        prtest varname [if] [in] , by(groupvar) [twosamplegropts]

    Two-sample test of proportions using variables

        prtest varname1 == varname2 [if] [in] [, level(#)]
```

```
Immediate form of one-sample test of proportion

    prtesti #obs1 #p1 #p2 [, level(#) count]

Immediate form of two-sample test of proportions

    prtesti #obs1 #p1 #obs2 #p2 [, level(#) count]
```

```
* sdtest -- Variance-comparison tests

    One-sample variance-comparison test

        sdtest  varname == # [if] [in] [, level(#)]

    Two-sample variance-comparison test using groups

        sdtest  varname [if] [in] , by(groupvar) [level(#)]

    Two-sample variance-comparison test using variables

        sdtest  varname1 == varname2 [if] [in] [, level(#)]

        // Test that var1 and var2 have same standard deviation
```

```
Immediate form of one-sample variance-comparison test

    sdtesti #obs {#mean | . } #sd #val [, level(#)]


Immediate form of two-sample variance-comparison test

    sdtesti #obs1 {#mean1 | . } #sd1 #obs2 {#mean2 | . } #sd2 [, level(#)] // Test sd1=sd2
```

第七章教材课后练习题

7.1 利用下面的信息，构建总体均值的置信区间。

(1) 总体服从正态分布，已知 σ=500，n=15，$\bar{x}$=8900，置信水平为95%。

(2) 总体不服从正态分布，已知 σ=500，n=35，$\bar{x}$=8900，置信水平为95%。

(3) 总体不服从正态分布，σ 未知，n=35，$\bar{x}$=8900，s=500，置信水平为90%。

(4) 总体不服从正态分布，σ 未知，n=35，$\bar{x}$=8900，s=500，置信水平为99%。

7.2 某大学为了解学生每天上网的时间，在全校7500名学生中采取重复抽样方法随机抽取36人，调查他们每天上网的时间，得到数据`习题7.2.xlsx`（单位:小时），求该校大学生平均上网时间的置信区间，置信水平分别为90%，95%和99%。

7.3 某企业生产的袋装食品采用自动打包机包装，每袋标准重量为100克。现从某天生产的一批产品中按重复抽样随机抽取50包进行检查，测得每包重量`习题7.3.xlsx`，已知食品每包的重量服从正态分布，求：

(1)确定该种食品平均重量的95%的置信区间。

(2)如果规定食品重量低于100克属于不合格，确定该批食品合格率的95%的置信区间。

7.4 假设总体服从正态分布，利用数据`习题7.4.xlsx`构建总体均值μ的99%的置信区间。

7.5 利用下面的样本数据构建总体比例π的置信区间。

(1)n=44，p=0.51，置信水平为99%。

(2)n=300，p=0.82，置信水平为95%。

(3)n=1150，p=0.48，置信水平为90%。

7.6 在一项家电市场调查中，随机抽取了200个居民户，调查他们是否拥有某一品牌的电视机，其中拥有该品牌电视机的家庭占23%。求总体比例的置信区间，置信水平分别为90%和95%。

7.7 一位银行的管理人员想估计每位顾客在该银行的月平均存款额。他假设所有顾客月存款额的标准差为1000元，要求的估计误差在200元以内，置信水平为99%。应选取多大的样本?

7.8 某居民小区共有居民500户，小区管理者准备采用一种新的供水设施，想了解居民是否赞成。采取重复抽样方法随机抽取了50户，其中有32户赞成，18户反对。

(1)求总体中赞成该项改革的户数比例的置信区间(α=0.05)。

(2)如果小区管理者预计赞成的比例能达到80%，估计误差不超过10%，应抽取多少户进行调查(α=0.05)?

7.9 根据下面的样本结果，计算总体标准差 σ 的90%的置信区间。

(1) $\bar{x}$=21，s=2，n=50。

(2) $\bar{x}$=1.3，s=0.02，n=15。

(3) $\bar{x}$=167，s=31，n=22。

7.10 顾客到银行办理业务时往往需要等待一段时间，而等待时间的长短与多种因素有关，比如，银行业务员办理业务的速度，顾客排队的方式等。为此，某银行准备采取两种排队方式进行试验，第一种排队方式是：所有顾客都进入一个等待队列；第二种排队方式是：顾客在三个业务窗口处列队三排等待。为比较哪种排队方式使顾客等待的时间更短，银行各随机抽取10名顾客，他们在办理业务时所等待的时间如`习题7.10.xlsx`（单位:分钟）求:

(1)构建第一种排队方式等待时间标准差的95%的置信区间。

(2)构建第二种排队方式等待时间标准差的95%的置信区间。

(3)根据(1)和(2)的结果，你认为哪种排队方式更好?

7.11 从两个正态总体中分别抽取两个独立的随机样本，它们的均值和标准差如下表所示：

来自总体1的样本	来自总体2的样本
$\bar{x}_1 = 25$	$\bar{x}_2 = 23$
$s_1^2 = 16$	$s_2^2 = 20$

(1) 设$n_1 = n_2 = 100$，求$\mu_1 - \mu_2$的95%的置信区间。

(2) 设$n_1 = n_2 = 10$，$\sigma_1^2 = \sigma_2^2$，求$\mu_1 - \mu_2$的95%的置信区间。

(3) 设$n_1 = n_2 = 10$，$\sigma_1^2 \neq \sigma_2^2$，求$\mu_1 - \mu_2$的95%的置信区间。

(4) 设$n_1 = 10, n_2 = 20$，$\sigma_1^2 = \sigma_2^2$，求$\mu_1 - \mu_2$的95%的置信区间。

(5) 设$n_1 = 10, n_2 = 20$，$\sigma_1^2 \neq \sigma_2^2$，求$\mu_1 - \mu_2$的95%的置信区间。

7.12 表`习题7.12.xlsx`是由四对观察值组成的随机样本。

(1)计算A与B各对观察值之差，再利用得出的差值计算$\bar{d}$和s_d。

(2)设μ_1和μ_2分别为总体A和总体B的均值，构建$\mu_d = \mu_1 - \mu_2$的95%的置信区间。

7.13 一家人才测评机构对随机抽取的10名小企业的经理人用两种方法进行自信心测试，得到的自信心测试分数如`习题7.13.xlsx`，构建两种方法平均自信心得分之差$\mu_d = \mu_1 - \mu_2$的95%的置信区间。

7.14 从两个总体中各抽取一个$n_1 = n_2 = 250$的独立随机样本，来自总体1的样本比例为$p_1 = 40\%$，来自总体2的样本比例为$p_2 = 30\%$。

求：

(1) 构建$\pi_1 - \pi_2$的90%的置信区间。

(2) 构建$\pi_1 - \pi_2$的95%的置信区间。

7.15 生产工序的方差是工序质量的一个重要度量。当方差较大时，需要对工序进行改进以减小方差。习题7.15.xlsx是两部机器生产的袋装茶重量的数据（单位:克），构建两个总体方差比σ_1^2/σ_2^2的95%的置信区间。

7.16 根据以往的生产数据，某种产品的废品率为2%。如果置信区间为95%，估计误差不超过4%，应抽取多少样本?

A tutorial on how to add the 95% CI to a two-way line plot

GitHub

Objective

To add 95% CI to the Stata connected line plot, we will use the Medical Expenditure Panel Survey (MEPS) from the Agency for Healthcare Research and Quality (AHRQ). The data has already been downloaded and cleaned.

We are interested in plotting the average total healthcare expenditures along with the 95% CI of responders from 2008 to 2019.

Step 1: Load the data into Stata

First, we will need to load the data into Stata.

```
clear all
use adjusted_combined_data_post.dta
```

Step 2: Visualize the data

Let's look at the number of subjects we have across the years. There are twelve years of data from 2008 - 2019.

```
tab year
```

Step 3: Two-way line plot

We can create a two-way line plot using the `connected` command. We'll plot he average total healthcare expenditure again the years (2008-2019).

We estimate the mean total healthcare expenditure using the `egen` command by year. Then we use the `graph twoway connected` code to generate the two-way line plot. Before you plot the figure, make sure that you sort by year. Otherwise, you will be a spaghetti-like plot where Stata will try to draw the line across the years if they are not sorted or in order.

```
egen mean_expenditure = mean(totexp), by(year)
sort year
graph twoway connected mean_expenditure year, ///
      title("Total healthcare expenditures (2008-2019)") ///
      ytitle("Average total healthcare expenditure ($)") ///
      xtitle("Time (year)") ///
      xlab(2008(1)2019) ///
      ylab(, labsize(small) nogrid) ///
      graphregion(color(white)) ///
      bgcolor(white)
graph export "connected1.png", replace
```

Step 4: Estimate the 95% CI

So far, our plot only illustrates the average total healthcare expenditure across time. But we want to plot the average total healthcare expenditures with 95% confidence intervals (CI). The 95% CI is important because it will give us some visual indicator of the variance surrounding each value.

We will need to estimate the 95% CI for each value. We can get the 95% CI for each value by using the `ci means` command.

```
ci means mean_expenditure
```

This gives us the 95% CI for the pooled total healthcare expenditure. If we want to get the 95% CI for each year, we need to specify the year. We use the `forvalues` loop to cycle through the years and estimate the mean, standard error, and 95% CI for the total healthcare expenditure at each year (2008-2019).

```
forvalues i = 2008/2019 {
        ci mean totexp if year == `i'
  }
```

Step 5: Generate the high and low values of the 95% CI

We start by adding two new variables: `high` and `low`, which are the high and low ends of the 95% CI. Then we estimate the 95% CI for the total healthcare expenditure at each year. The data are stored in Stata and can be retrieved using the `r()` syntax. For example, if I want the lower bound of the 95% CI in 2009, then I write `replace low = r(lb) if year == 2009`.

To make this process more efficient, we use the `forvalues` loop for each year (2008-2019).

```
gen high = .
gen low = .

forvalues i = 2008/2019 {
        ci mean totexp if year == `i'
        replace high = r(ub) if year == `i'
        replace low = r(lb) if year == `i'
}
```

Step 6: Plot the 95% CI

Now, that we have the 95% CI for total healthcare expenditure at each year, we can plot these using the `rcap` command. This allows us to create error bars using the variables we generated for the upper `high` and lower `low` bounds of the 95% CI.

```
graph twoway (rcap low high year) ///
             (connected mean_expenditure year, color(navy) msize(small) ///
                        title("Total healthcare expenditures (2008-2019)") ///
                        ytitle("Average total healthcare expenditure ($)") ///
                        xtitle("Time (year)") ///
                        xlab(2008(1)2019, labsize(vsmall)) ///
                        ylab(, labsize(vsmall) nogrid) ///
                        graphregion(color(white)) ///
                        bgcolor(white) ///
                          color(navy))
graph export "connected_errorbars.png", replace
```

Conclusions

Adding the 95% CI allows the reader to visualize the variance surrounding the mean total healthcare expenditure per year. This added dimension to the two-way line plot is useful for publication-quality images. However, these series of commands can be used for any continuous data type.

延申阅读

- Stata绘图：高级柱状图-均值和置信区间-cibar-coefpl

4. 假设检验

女士品茶：统计学如何变革科学和生活

1920年的剑桥大学，某个风和日丽的下午，一群科学家正悠闲地享受下午茶时光。正如往常一样准备冲泡奶茶的时候，有位女士突然说：“冲泡的顺序对于奶茶的风味影响很大。先把茶加进牛奶里，与先把牛奶加进茶里，这两种冲泡方式所泡出的奶茶口味截然不同。**我可以轻松地辨别出来**。”在场的绝大多数人对这位女士的“胡言乱语”嗤之以鼻。然而，其中一位身材矮小、戴着厚眼镜的先生却不这么看，他对这个问题很有兴趣。这个人就是费歇尔(R.A.Fisher)。

Fisher的思路是：他首先假该设女士**没有这个能力**（这个假设被称为原假设）。随后，Fisher将8杯已经调制好的奶茶**随机地**放到那位女士的面前，看看这位女士能否正确地品尝出不同的茶。

用p表示该女士每次答对的概率，用随机变量X表示女士答对的次数；在n次实验中，女士答对k次的概率可以用**二项分布**来描述。

在原假设下，女士并没有鉴别的能力，能否答对完全靠蒙——此时，p=0.5（类似于抛硬币）。

可以计算出n=8、p=0.5时女士答对k次的概率，如下表所示：

答对的次数 (k)	概率 $P(X=k)$	累计概率 $P(X\leq k)$
0	0.0039	0.0039
1	0.0313	0.0352
2	0.1094	0.1445
3	0.2188	0.3633
4	0.2734	0.6367
5	0.2188	0.8555
6	0.1094	0.9648
7	0.0313	0.9961
8	0.0039	1.0000

现在问题来了：如果**实际观测到女士连续答对了8次**（即k=8），那么，她到底有没有鉴别能力？或者，从概率的角度来看：原假设p=0.5到底对不对？

从表中可以看出：如果原假设p=0.5成立（即女士没有鉴别能力），那么99.61%的情况下，女士蒙对的次数应该**小于或者等于7次**。

而现在实际观测到的结果是**女士连续答对了8次**——这说明了什么？当p=0.5时，“连续答对8次”的概率比较低，仅为0.39%；而只有当p大于0.5时（比如，接近于1），发生“连续答对8次”这种事情的概率才比较高。也就是说，女士**极有可能**具备鉴别能力。在一次观测中，当小概率事件发生时，我们有足够的理由怀疑**原假设**（p=0.5）的正确性。

有人可能会问：如果该女士确实没有鉴别能力，而仅仅是那天运气比较好、连续蒙对了8次——有没有这种可能？

有这种可能。虽然本例中发生这种情况的概率比较小，仅为0.39%。

这就意味着：我们的判断有可能是错误的。如果女士确实没有鉴别能力（原假设为真），而我们根据观测到的样本做出了“拒绝原假设”的判断——那我们就犯了**第Ⅰ类错误**。

Source: 假设检验——这一篇文章就够了

Step1：**提出原假设和备择假设**

Step2：**构造检验统计量**

Step3：**选择显著性水平，给出拒绝域**

Step4：**做出判断**

4.1 总体均值

4.1.1 单个样本均值

```
ttest x = c
```

4.1.2 两个匹配样本均值

```
ttest x1 = x2
```

4.1.3 两个独立样本均值

```
ttest x1 = x2, unpaired
ttest x1 = x2, unpaired unequal //两个总体的方差不同
```

4.2 总体比例

4.2.1 单样本比例

```
prtest p = c
```

4.2.2 两个样本比例

```
prtesti n1 p1 n2 p2
```

4.3 总体方差

```
sdtest x = c
sdtest x1 == x2
```

第八章教材课后练习题

8.1 已知某炼铁厂的铁水含碳量服从正态分布$N(4.55, 0.108^2)$，现在测定了9炉铁水，其平均含碳量为4.484。如果估计方差没有变化，能否认为现在生产的铁水平均含碳量为4.55$(\alpha = 0.05)$?

8.2 有一种元件，要求其使用寿命不得低于700小时。现从一批这种元件中随机抽取36件，测得其平均寿命为680小时。已知该元件寿命服从正态分布， σ=60小时，试在显著性水平0.05下确定这批元件是否合格。

8.3 某地区小麦的一般生产水平为亩产250千克，其标准差为30千克。现用一种化肥进行试验，从25个地块抽样，平均产量为270千克。这种化肥是否使小麦明显增产($\alpha = 0.05$)?

8.4 糖厂用自动打包机打包，每包的标准重量是100千克。每天开工后需要检验一次打包机工作是否正常。某日开工后测得9包重量`习题8.4.xlsx`(单位:千克)，已知每包的重量服从正态分布，试检验该日打包机工作是否正常($\alpha = 0.05$)。

8.5 某种大量生产的袋装食品按规定不得少于250克。今从一批该食品中任意抽取50袋，发现有6袋低于250克。若规定不符合标准的比例超过5%就不得出厂，问该批食品能否出厂($\alpha = 0.05$)?

8.6 某厂家在广告中声称，该厂生产的汽车轮胎在正常条件下行驶距离超过目前的平均水平25000公里。对一个由15个轮胎组成的随机样本做了试验，得到样本均值和标准差分别为27000公里和5000公里。假定轮胎寿命服从正态分布，问该厂家的广告所声称的内容是否真实($\alpha = 0.05$)?

8.7 某种电子元件的寿命x服从正态分布。现测得16只元件的寿命`习题8.7.xlsx`(单位:小时)，是否有理由认为元件的平均寿命显著地大于225小时($\alpha = 0.05$)?

8.8 随机抽取9个单位，测得结果为`习题8.7.xlsx`，以$\alpha = 0.05$的显著性水平对下述假设进行检验：$H_0 : \sigma^2 \leq 100, H_1 : \sigma^2 > 100$。

8.9 A，B两厂生产同样的材料。已知其抗压强度服从正态分布，且$\sigma_A^2 = 63^2$，$\sigma_B^2 = 57^2$。从A厂生产的材料中随机抽取81个样品，测得$\bar{x}_A = 1070kg/cm^2$；从B厂生产的材料中随机抽取64个样品，测得$\bar{x}_B = 1020kg/cm^2$。根据以上调查结果，能否认为A，B两厂生产的材料的平均抗压强度相同($\alpha = 0.05$)？

8.10 装配一个部件可以采用不同的方法，所关心的问题是哪一个方法的效率更高。劳动效率可以用平均装配时间来反映。现从不同的装配方法中各抽取12件产品，记录各自的装配时间`习题8.10.xlsx`(单位：分钟)，两总体为正态总体，且方差相同，问这两种方法的装配时间有无显著差别($\alpha = 0.05$)？

8.11 调查了339名50岁以上的人，在205名吸烟者中有43个患慢性气管炎，在134名不吸烟者中有13人患慢性气管炎。调查数据能否支持“吸烟者容易患慢性气管炎这种观点($\alpha = 0.05$)?

8.12 为了控制贷款规模，某商业银行有个内部要求，平均每项贷款的数额不能超过60万元。随着经济的发展，贷款规模有增大的趋势。银行经理想了解在同样的项目条件下贷款的平均规模是否明显地超过60万元，故一个n=144的随机样本被抽出，测得至$\bar{x}$=68.1万元，s=45。在$\alpha = 0.01$的显著性水平下采用P值进行检验。

8.13 有一种理论认为服用阿司匹林有助于减少心脏病的发生，为了进行验证，研究人员把自愿参与实验的22000人随机平均分成两组，一组人员每星期服用三次阿司匹林(样本1)，另一组人员在相同的时间服用安慰剂(样本2)。持续3年之后进行检测，样本1中有104人患心脏病，样本2中有189人患心脏病。以$\alpha = 0.05$的显著性水平检验服用阿司匹林是否可以降低心脏病发生率。

8.14 某工厂制造螺栓，规定螺栓口径为7.0cm，方差为0.03cm。今从一批螺栓中抽取80个测量其口径，得平均值为6.97cm，方差为0.0375cm。假定螺栓口径服从正态分布，问这批螺栓是否达到规定的要求($\alpha = 0.05$)?

8.15 有人说在大学中男生的学习成绩比女生的学习成绩好。现从一所学校中随机抽取25名男生和16名女生，对他们进行相同题目的测试。测试结果表明，男生的平均成绩为82分，方差为56分，女生的平均成绩为78分，方差为49分。假设显著性水平$\alpha = 0.02$，从上述数据中能得到什么结论?

欢迎交流~

zzynankai@outlook.com

Bilibili: 西山yu

xishanyu2.github.io