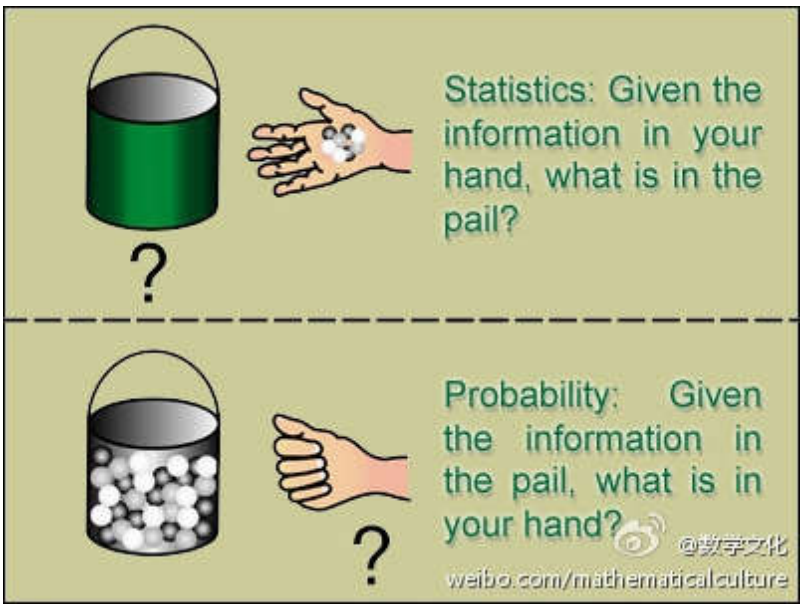
数理统计

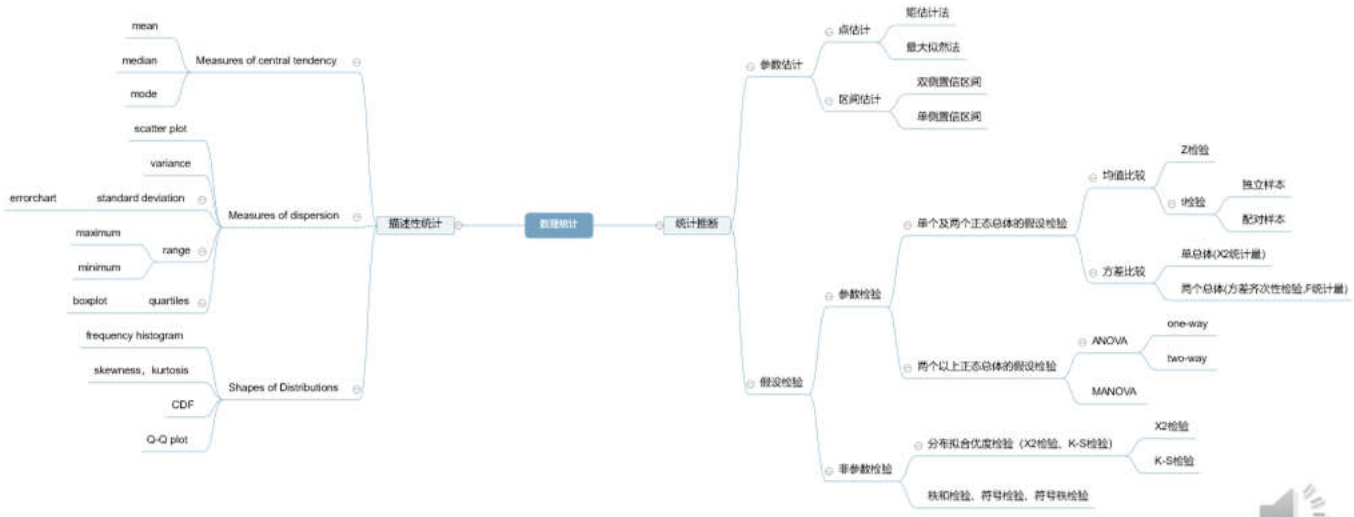
样本、统计量、抽样分布

统计：给出你手中的信息（样本），桶（总体）里有什么?

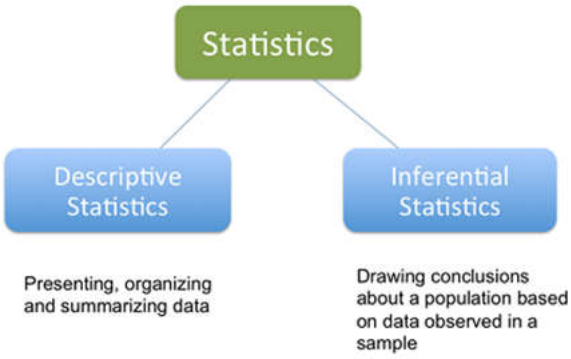
概率：给出桶中的信息，你手里有什么?



- 数理统计 要点



descriptive statistics & statistical inference （描述性统计和统计推断）



The former focus on summarizing and illustrating the features of a collection of observed data, which is referred to as a sample. The sample is drawn from a population, denotes the total set of similar individuals, items, or events of our experiment interests. Contrary to descriptive statistics, statistical inference further deduces the characteristics of a population from the given samples.

前者侧重于总结和说明被观测数据集合（样本）的特征。样本是从总体中抽取的，表示与我们实验相关的相似个体或事件的总集合。与描述性统计相反，统计推断从给定的样本中进一步推导出总体的特征。

呈现、组织和汇总数据

根据样本中观测到的数据得出关于总体的结论

SPSS：分析 > 描述统计 > 描述

描述性统计是用于数据分析的术语，它有助于以有意义的方式描述、显示或总结数据，比如数据可能出现的模式。但是，描述性统计并不允许我们在所分析的数据之外得出结论，或者就我们可能做出的任何假设得出结论。它只是描述数据的一种方式。

Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that, for example, patterns might emerge from the data. Descriptive statistics do not, however, allow us to make conclusions beyond the data we have analysed or reach conclusions regarding any hypotheses we might have made. They are simply a way to describe our data.

Descriptives

[DataSet1] C:\Users\eleve\Desktop\ST_UG_2021\data\ch3\Employee data.sav

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Current Salary	474	$15,750	$135,000	$34,419.57	$17,075.661	2.125	.112	5.378	.224
Valid N (listwise)	474								

SPSS：分析 > 描述统计 > 探索

综合性分析

一、总体和样本

1. 总体

一个统计问题总有它明确的研究对象.

研究对象的全体称为总体，

总体中每个成员称为个体，

总体中所包含的个体的个数称为总体的容量.

研究某批灯泡的质量

总体 { 有限总体 / 无限总体

定义：

设X是具有分布函数F的随机变量，若$X_1, X_2, \cdots, X_n$是具有同一分布函数F的、相互独立的随机变量，则称$X_1, X_2, \cdots, X_n$为从分布函数F（或总体F、或总体X）得到的容量n为的简单随机样本，简称样本，它们的观察值$x_1, x_2, \cdots, x_n$称为样本值，又称为X的n个独立的观察值.

第二节　样本及抽样分布

- 统计量与经验分布函数
- 统计三大抽样分布
- 几个重要的抽样分布定理

一、统计量

1. 统计量 (Statistic)

由样本值去推断总体情况，需要对样本值进行“加工”，这就要构造一些样本的函数，它把样本中所含的（某一方面）的信息集中起来.

两个统计量

几个常见统计量

样本平均值 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

它反映了总体均值的信息

样本方差 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

$= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$

它反映了总体方差的信息

样本标准差 $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$

样本k阶原点矩

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

k=1,2,…

它反映了总体k阶矩的信息

样本k阶中心矩

$$B_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^k$$

它反映了总体k阶中心矩的信息

统计量的观察值

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i ; s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2} ; \alpha_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \; k = 1,2,\cdots$$

$$b_k = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^k \; k = 1,2,\cdots$$

请注意：

若总体X的k阶矩$E(X^k)=\mu^k$存在，则当$n\to\infty$时，

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i{}^k\xrightarrow{p}\mu^k \qquad k=1,2,\cdots.$$

事实上由$X_1,X_2,\cdots,X_n$独立且与X同分布，
有$X_1{}^k,X_2{}^k,\cdots,X_n{}^k$独立且与$X^k$同分布，$E(X_i^k)=\mu^k$
$k=1,2,\cdots,n$再由辛钦大数定律可得上述结论.

再由依概率收敛性质知，可将上述性质推广为

$$g(A_1,A_2,\cdots,A_k)\xrightarrow{p}g(\mu_1,\mu_2,\cdots,\mu_k)$$

其中g为连续函数.

这就是矩估计法的理论根据.

statistics vs parameters

统计量（样本） vs 参数（总体）

Properties of samples, such as the mean or standard deviation, are not called parameters, but statistics. Inferential statistics are techniques that allow us to use these samples to make generalizations about the populations from which the samples were drawn.

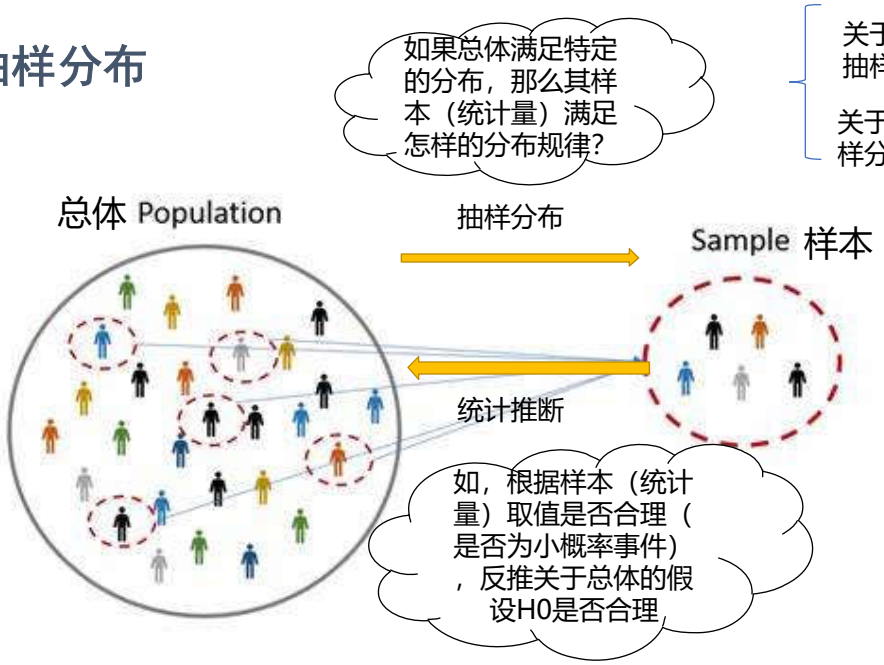
样本的属性如平均值或者标准差，不称为参数而是被称作统计量。推断统计是使我们能够利用样本对样本所来自的总体进行概括/推断的技术。

Statistic 统计量		Parameter 参数	
样本均值	$\bar{x}$ (sample mean)	μ (population mean)	总体均值
样本标准差	s (sample standard deviation)	σ (population standard deviation)	总体标准差

使用统计量评估参数

We use statistics to estimate parameters.

二、抽样分布



统计三大抽样分布

1、 χ^2 分布

χ^2分布是由正态分布派生出来的一种分布.

定义: 设 $X_1, X_2, \cdots, X_n$ 相互独立, 都服从标准正态分布$N(0,1)$, 则称随机变量:

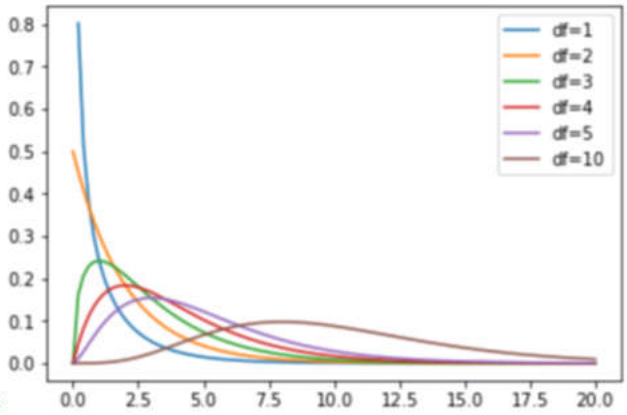
$$\chi^2 = X_1^{\,2} + X_2^{\,2} + \cdots + X_n^{\,2}$$

所服从的分布为自由度为 n 的 χ^2 分布.

自由度（degree of freedom）

记为 $\chi^2 \sim \chi^2(n)$

卡方分布 -
不同自由度下
的PDF曲线




```
import numpy
import scipy.stats
import matplotlib.pyplot  as plt
x=numpy.linspace(0,200,100)
plt.plot(x,scipy.stats.chi2.pdf(x,df=1))
plt.plot(x,scipy.stats.chi2.pdf(x,df=2))
plt.plot(x,scipy.stats.chi2.pdf(x,df=3))
plt.plot(x,scipy.stats.chi2.pdf(x,df=4))
plt.plot(x,scipy.stats.chi2.pdf(x,df=5))
plt.plot(x,scipy.stats.chi2.pdf(x,df=10))
plt.legend(['df=1','df=2','df=3','df=4','df=5','df=10'])
plt.show()
```

χ^2分布的性质

1. 设 $X_1, X_2, \cdots, X_n$ 相互独立, 都服从正态分布 $N(\mu, \sigma^2)$, 则 $\chi^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi^2(n)$

2. 设 $X_1 \sim \chi^2(n_1), X_2 \sim \chi^2(n_2)$, 且$X_1$,$X_2$相互独立, 则$X_1 + X_2 \sim \chi^2(n_1 + n_2)$.

这个性质叫χ^2分布的可加性.

3. 若 $\chi^2 \sim \chi^2(n)$, 则当n充分大时, $\frac{X - n}{\sqrt{2n}}$ 的分布近似正态分布**N(0,1).**

(应用中心极限定理可得)

4.若$\chi^2 \sim \chi^2(n)$,χ^2分布的数学期望与方差,

$$E(X)=n,\ D(X)=2n.$$

事实上，由$X_i \sim N(0,1)$,故$E(X_i{}^2) = D(X_i) = 1$

$$D(X_i^2) = E(X_i^4) - [E(X_i^2)]^2 = 3 - 1 = 2$$

$$E(\chi^2) = \sum_{i=1}^{n} E(X_i{}^2) = n, D(\chi^2) = \sum_{i=1}^{n} D(X_i{}^2) = 2n.$$

证明定理3：当n充分大时，$\frac{X-n}{\sqrt{2n}} \sim \mathrm{N}(0,1)$

中心极限定理：

$\bar{X}$ 近似地 $N(\mu, \sigma^2/n)$ 或 $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ 近似地 $N(0,1)$

将χ^2/n看作 n 个$\chi^2(1)$随机变量的均值

选学内容： χ^2分布的PDF（概率密度函数）解析式

Usually a positive whole number

The pdf of the χ^2 distribution with k degrees of freedom:

$$f(x) = \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)} \quad \text{for } x \geq 0$$

$\mu = k$

$\sigma^2 = 2k$

An Introduction to the χ^2 Distribution

2、t 分布 （学生分布）

定义：设$X\sim N(0,1)$，$Y\sim \chi^2(n)$，且X与Y相互独立，则称变量

$$t=\frac{X}{\sqrt{Y/n}}$$ 所服从的分布为自由度为 n 的 t 分布.

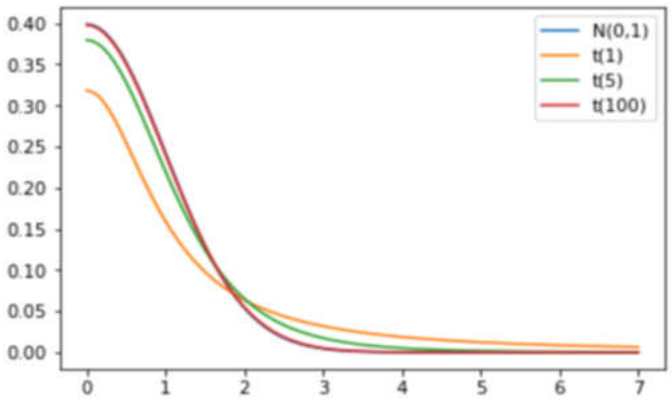
记为 $t\sim t(n)$.

t分布的性质：

1. 具有自由度为n的t分布$t\sim t(n)$，其数学期望与方差为：$E(t)=0, D(t)=n/(n-2)\ (n>2)$

2. 即当n足够大时，$t \overset{近似}{\sim} N(0,1)$.

t 分布




```
import numpy
import scipy.stats
import matplotlib.pyplot as plt
x=numpy.linspace(0,7,100)

plt.plot(x,scipy.stats.norm.pdf(x))
plt.plot(x,scipy.stats.t.pdf(x,df=1))

plt.plot(x,scipy.stats.t.pdf(x,df=5))
plt.plot(x,scipy.stats.t.pdf(x,df=100))
plt.legend(['N(0,1)','t(1)','t(5)','t(100)'])
plt.show()
```

3、*F*分布

定义： 设 $U\sim\chi^2(n_1), V\sim\chi^2(n_2),$

U 与V相互独立，则称随机变量

$$F=\frac{U/n_1}{V/n_2}$$

服从自由度为n_1及n_2 的F分布，n_1称为第一自由度，n_2称为第二自由度，记作

$$F\sim F(n_1,n_2)$$

由定义可见，

$$\frac{1}{F}=\frac{V/n_2}{U/n_1}\sim F(n_2,n_1)$$

F 分布的性质

1. *F*分布的数学期望为:

$$E(F)=\frac{n_2}{n_2-2} \qquad \text{若} n_2>2$$

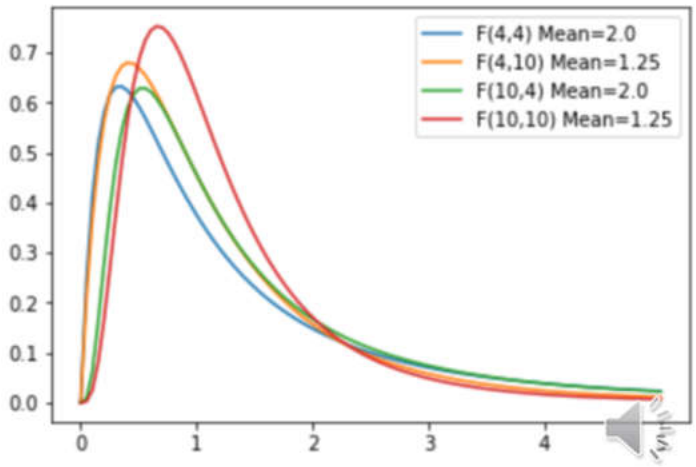
即它的数学期望并不依赖于第一自由度n_1.

```
# F distribution

import numpy
import scipy.stats
import matplotlib.pyplot  as plt
x=numpy.linspace(0,5,100)

plt.plot(x,scipy.stats.f.pdf(x,dfn=4,dfd=4))
plt.plot(x,scipy.stats.f.pdf(x,dfn=4,dfd=10))
plt.plot(x,scipy.stats.f.pdf(x,dfn=10,dfd=4))
plt.plot(x,scipy.stats.f.pdf(x,dfn=10,dfd=10))
plt.legend(['F(4,4)
Mean='+str(scipy.stats.f.mean(4,4)),\
        'F(4,10)
Mean='+str(scipy.stats.f.mean(4,10)),\
        'F(10,4)
Mean='+str(scipy.stats.f.mean(10,4)),\
        'F(10,10)
Mean='+str(scipy.stats.f.mean(10,10))])
plt.show()
```

F分布的PDF



三、重要的抽样分布定理

设总体X的均值为μ，方差为σ^2，$X_1, X_2, \cdots, X_n$是来自总体的一个样本，则样本均值$\bar{X}$和样本方差S^2有

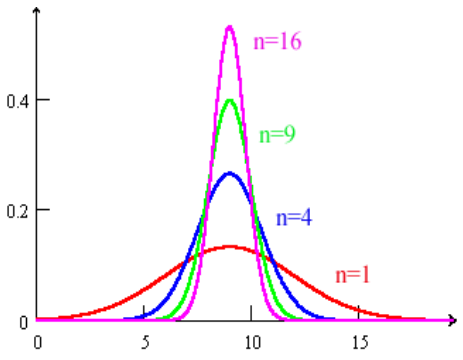
$$E(\bar{X}) = \mu,$$
$$D(\bar{X}) = \frac{\sigma^2}{n},$$
$$E(S^2) = \sigma^2$$

定理 1 (样本均值的分布)

设 $X_1, X_2, \ldots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本，$\bar{X}$是样本均值，则有

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

即 $$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$



定理 2 (样本方差的分布)

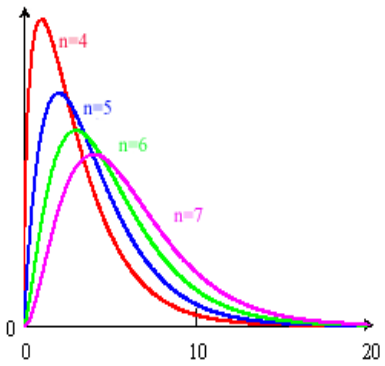
设$X_1, X_2, \ldots, X_n$是来自正态总体 $N(\mu, \sigma^2)$ 的样本，$\bar{X}$和S^2分别为样本均值和样本方差，则有

(1) $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

(2) $\bar{X}$与S^2独立.

n取不同值时 $\dfrac{(n-1)S^2}{\sigma^2}$

的分布



定理 3 (样本均值的分布)

设$X_1, X_2, \ldots, X_n$是取自正态总体 $N(\mu, \sigma^2)$ 的样本，$\bar{X}$和S^2 分别为样本均值和样本方差，则有

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

注意： σ^2 与 S^2 的区别

定理 3 证明（t-distribution的推导）（选学）

Suppose we are about to draw a random sample of n observations from a normally distributed population.

假设我们要从正态分布的总体中抽取n个样本。

若总体方差σ^2未知，n个样本的统计量 $t=\frac{\bar{X}-\mu}{\sqrt{S^2/n}}$ 服从的分布为自由度(df)为 $n-1$的 t 分布.

The population standard deviation is almost always unknown

$$Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

has the standard normal distribution

标准正态分布

The sample standard deviation

$$\frac{\bar{X}-\mu}{S/\sqrt{n}}$$

$$t=\frac{\bar{X}-\mu}{\sqrt{S^2/n}}=\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n\times\frac{S^2}{\sigma^2}}}\sim\frac{N(0,1)}{\sqrt{\chi^2(n-1)}}$$

$$because\quad \frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$$

自由度为n-1的t分布

$$t=\frac{\bar{X}-\mu}{S/\sqrt{n}}$$

has the t distribution with $n-1$ degrees of freedom

S^2 The denominator of the sample variance is n - 1

定理 4 (两总体样本均值差、样本方差比的分布)

设$X\sim N(\mu_1,\sigma^2)$，$Y\sim N(\mu_2,\sigma^2)$，且X与Y独立，$X_1, X_2, \ldots, X_{n_1}$是来自$X$的样本，$Y_1, Y_2, \ldots, Y_{n_2}$是取自$Y$的样本，

$\bar{X}$和$\bar{Y}$分别是这两个样本的样本均值，S_1^2和S_2^2分别是这两个样本的样本方差,则有

1、$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}\sim F(n_1-1,n_2-1)$，$because\quad \frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$

2、$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$$

小结

学习了统计量的概念，几个重要的统计量及其分布，即抽样分布.

抽样分布

设$X_1,\cdots,X_n$相互独立，且均服从正态分布$N(0,1)$，

χ^2分布

则称随机变量$\chi^2=\sum_{i=1}^{n}X_i{}^2$服从自由度为$n$的$\chi^2$分布，

记为$\chi^2\sim\chi^2(n)$.

t分布

设$X\sim N(0,1)$，$Y\sim\chi^2(n)$，且X与Y相互独立，则称

随机变量$t=\dfrac{X}{\sqrt{Y/n}}$服从自由度为n的t分布，记为$t\sim t(n)$.

F分布

设$U\sim\chi^2(n_1)$，$V\sim\chi^2(n_2)$，U与V相互独立，则称

随机变量$F=\dfrac{U/n_1}{V/n_2}$服从自由度为（n_1,n_2）的分布，

记为$F\sim F(n_1,n_2)$.

抽样分布定理

已知总体方差前提下，样本均值的分布

设$X\sim N(\mu,\sigma^2)$，$X_1,\cdots,X_n$是来自总体X的样本，则样本均值$\bar{X}$有$\bar{X}\sim N(\mu,\sigma^2/n)$.

样本方差 | 未知总体方差前提下，均值的分布

设$X_1,\cdots,x_n$是来自总体$N(\mu,\sigma^2)$的样本，$\bar{X},S^2$分别是样本均值和样本方差，则有

(1) $\dfrac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$

(2) $\bar{X}$与S^2独立.

(3) $\dfrac{\bar{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$

两总体均值比较、方差齐次性检验

设$X_1,\cdots,X_{n_1}$与$Y_1,\cdots,Y_{n_2}$分别来自总体$N(\mu_1,\sigma_1^2)$和$N(\mu_2,\sigma_2^2)$的样本，且这两个样本相互独立.$\bar{X},\bar{Y}$分别是这两个样本的样本均值；S_1^2,S_2^2分别是这两个样本的样本方差，则有

1、$\dfrac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t(n_1+n_2-2)$

2、$\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}\sim F(n_1-1,n_2-1)$