统计学基础

本章主要内容

1. 统计学基本概念：统计学介绍、统计数据和变量的类型、常用基本概念
2. 描述性统计分析：集中趋势的度量、离散趋势的度量、偏态与峰态
3. 由正态分布导出的几个重要分布：卡方分布、t分布、F分布
4. 常用统计量的分布：均值、比例、方差

本章主要内容

1. 统计学基本概念：统计学介绍、统计数据的类型、常用基本概念
2. 描述性统计分析：集中趋势的度量、离散趋势的度量、偏态与峰态
3. 由正态分布导出的几个重要分布：卡方分布、t分布、F分布
4. 常用统计量的分布：均值、比例、方差

统计学介绍

统计学，statistics，一门关于数据的学科，选择适当的统计学方法分析数据，并从数据中得出有用的信息，最后给出结论。

统计学方法：参数估计、假设检验、方差分析、回归分析、逻辑回归、聚类分析、主成分分析、因子分析、时间序列分析等。

统计学应用：

- 估计市民的平均工资（参数估计）、判断产品质量是否规定标准（假设检验）
- 分析不同行业之间的服务质量是否有差异（方差分析）
- 商业银行预测不良贷款（回归分析）
- 预测企业员工是否离职（逻辑回归）
- 用户分群管理（聚类分析）

统计数据的类型

统计数据主要三种类型：数值数据、分类数据和顺序数据。

- 数值数据：用数字来描述事物，现实处理的大多数数据都是数值型数据，例如产品销售额，数值型，而且是连续的数据；例如产品数量，数值型，离散型数据；年龄，数值型，离散型数据。
- 分类数据：主要指类别型数据，例如性别数据，里面都是男或者女；行业数据，里面可能有互联网、汽车、金融、银行、餐饮等。这类数据只是对事物进行分类，并没有顺序。
- 顺序数据：主要是指有顺序的类别变量，例如产品等级，可以分为次品、二等品、一等品等；例如对事物的态度，有不感兴趣、感兴趣、喜欢、非常喜欢等，这类数据除了对事物进行分类之外，还有一个程度的递进。

变量的类型

变量，说明现象某种特征的概念，例如销量、年龄、性别、行业、产品等级、受教育程度等。

变量分为三种类型：数值型变量、分类变量和顺序变量。

- 数值型变量：例如销量、年龄等，取值为数值型数据。
- 分类变量：例如，性别、行业等，取值为分类数据。
- 顺序变量：例如产品等级、对事物的态度等，取值为顺序数据。

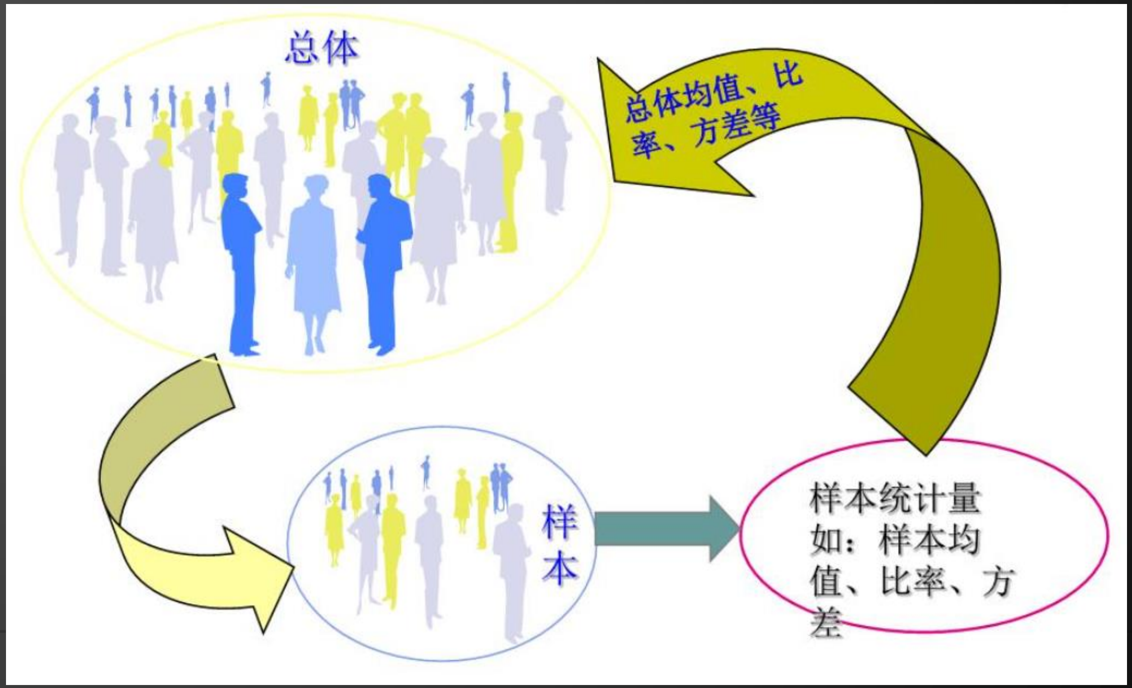
说明：R语言中，分类变量和顺序变量被称为因子。

几个基本概念

四个概念：总体、样本、参数、统计量（也叫估计量）

例如，要估计北京市民的平均月收入，统计全体市民？不好操作，那就抽取1000人进行统计。

- 总体：全体北京市民
- 样本：抽取的1000人
- 参数：全体北京市民的平均月收入
- 统计量：抽取的1000人样本的平均月收入



本章主要内容

1. 统计学基本概念：统计学介绍、统计数据的类型、常用基本概念
2. 描述性统计分析：集中趋势的度量、离散趋势的度量、偏态与峰态
3. 由正态分布导出的几个重要分布：卡方分布、t分布、F分布
4. 常用统计量的分布：均值、比例、方差

描述性统计分析

集中趋势的度量

- 众数
- 平均数
- 中位数
- 四分位数

离散趋势的度量

- 四分位差
- 极差
- 方差
- 标准差

偏态与峰态的度量

- 偏态系数
- 峰态系数

描述性统计分析

集中趋势的度量

- 众数
- 平均数
- 中位数
- 四分位数

离散趋势的度量

- 四分位差
- 极差
- 方差
- 标准差

偏态与峰态的度量

- 偏态系数
- 峰态系数

众数

- 众数：表示总体中出现次数最多的数值。

例如，在某城市中随机抽取9个家庭，调查得到每个家庭的人均月收入数据如下（单位：元）。

1080 750 1080 1080 850 960 2000 1250 1630

- 在以上数据中，1080出现的次数最多，出现了3次，所以众数是1080。

- 多个众数的情况：

1080 750 1080 850 850 960 2000 1250 1630

1080和850出现的次数相同，都是两次，所以这组数据有两个众数，分别是1080和850。

- 在Excel中，可以通过函数MODE.SNGL来求众数。如果有多个众数，用MODE.MULT

平均数：均值

- 平均数，其实就是均值，一组数据相加，再除以数据的个数得到的结果就是均值。
- 例如，有数据：1080 750 1080 1080 850 960 2000 1250 1630

$$\bar{x} = \frac{1080+750+1080+1080+850+960+2000+1250+1630}{9} = 1186.67$$

- 在Excel中，可以通过函数average来求均值。
- 除了均值之外，还有加权平均数、几何平均数。

中位数

- 将总体中的各个数据按照升序排列，居于中间位置的数值，便是中位数。
- 例如，对于数据：1080　750　1080　1080　850　960　2000　1250　1630
- 按照升序排列后：750　850　960　1080　1080　1080　1250　1630　2000

中间位置上的数据为：1080，所以中位数为1080。

如果有偶数个数据，则中位数是中间位置两个数字的平均数。

例如，有以下数据，750　850　960　1080　1080　1080

中间位置上有两个数字，960和1080，中位数为960和1080的平均，即1020。

中位数

总结:

- 将总体中的各个数据按照升序排列，居于中间位置的数值，便是中位数。
- 如果数据为奇数项，中位数是中间位置的数值
- 如果数据为偶数项，中位数是中间位置两个数值的平均数

Excel中，可以通过函数median来求中位数。

四分位数

- 把所有数值由小到大排列，分成四等份，处于三个分割点位置的数值就是四分位数。

例如，有数据：

1080 750 1080 1080 850 960 2000 1250 1630

- 将原始数据按照升序排列后，

750 850 960 1080 1080 1080 1250 1630 2000

第一个四分位数

$$Q_1 = \frac{n+3}{4}$$

第二个四分位数

$Q_3 =$ 中位数

第三个四分位数

$$Q_3 = \frac{3n+1}{4}$$

四分位数

- 750 850 960 1080 1080 1080 1250 1630 2000

第一个四分位数

$$Q_1 = \frac{n+3}{4}$$

第二个四分位数

$$Q_3 = \text{中位数}$$

第三个四分位数

$$Q_3 = \frac{3n+1}{4}$$

- 第一个四分位数：也叫下四分位数，位置为(9+3)/4=3，所以第一个四分位数为960；
- 第二个四分位数：中位数，为1080；
- 第三个四分位数：也叫上四分位数，位置为(3*9+1)/4=7，所以第三个四分位数为1250。

用Excel公式QUARTILE.INC可以很容易计算四分位数。

描述性统计分析

集中趋势的度量

- 众数
- 平均数
- 中位数
- 四分位数

离散趋势的度量

- 四分位差
- 极差
- 方差
- 标准差

偏态与峰态的度量

- 偏态系数
- 峰态系数

四分位差

- 四分位差：也叫四分位距，是上四分位数和下四分位数之差。
- 750　850　960　1080　1080　1080　1250　1630　2000
- 第一个四分位数：下四分位数，为960
- 第三个四分位数：上四分位数，为1250
- 四分位差：$Q_d = 1250 - 960 = 290$

- 四分位差反映了数据中间50%的离散程度，其数值越小，表示数据越集中，数值越大，表示数据越分散。

极差

- 极差：表示一组数据中最大值与最小值之差。

例如，有数据：

1080 750 1080 1080 850 960 2000 1250 1630

其中，最大值为2000，最小值为750，所以极差=2000-750=1250

- 在Excel中，我们可以先用max和min分别计算出最大值和最小值，然后作差即可。

方差

- 方差(variance)反映的数据波动性，用数学语言表示就是，各变量值与其均值离差平方的均值。数学公式为：

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

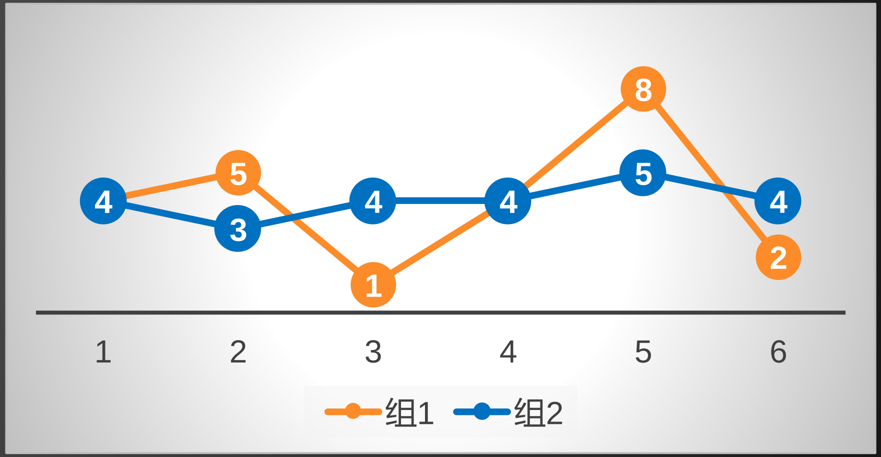
- 说明：这里的$n-1$表示自由度，自由度表示一组数据中可以自由取值的个数，如果样本数据的个数为n，这n个数据可以自由取值，则自由度为n，但是当样本数据的均值$\bar{x}$确定后，只有$n-1$个数据可以自由取值，所有自由度就是$n-1$。
- 对于n个样本数据，样本数据的均值$\bar{x}$确定相当于1个约束条件，自由度为$n-1$，按照这个逻辑，对n个样本数据附加k个约束条件，则自由度为$n-k$。
- 在Excel中，通过公式VAR.P可以计算出方差。

方差

- 假设有以下两组数据，试比较它们的离散程度。

组1	4	5	1	4	8	2	4
组2	4	3	4	4	5	4	4

- 通过Excel公式VAR.P可以很容易得到这两组数据的方差。
- 组1：$\mathrm{D}(X_1) = 4.29$
- 组2：$\mathrm{D}(X_2) = 0.29$
- 结论：数据组1的离散程度大于数据组2。



标准差

- 标准差(Standard Deviation)就是方差开方得到。

$$std = \sqrt{var} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- 由于方差是在原来数据的基础上进行了平方，所以单位发生了变化，标准差的单位则和原来的数据一致，所以在实际分析时，标准差使用得更多。
- 在Excel中，通过公式STDEV.P可以得到标准差。

例如，针对前面给出的两组数据，可以计算出它们的标准差分别为：

- 组1：$\sigma_1 = 2.07$
- 组2：$\sigma_2 = 0.53$

描述性统计分析

集中趋势的度量

- 众数
- 平均数
- 中位数
- 四分位数

离散趋势的度量

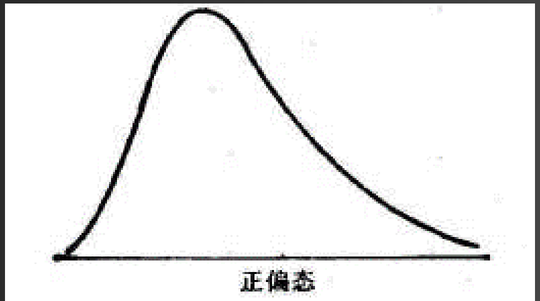
- 四分位差
- 极差
- 方差
- 标准差

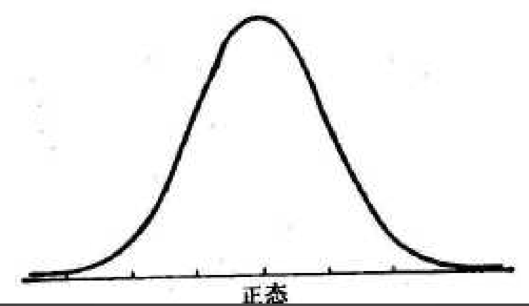
偏态与峰态的度量

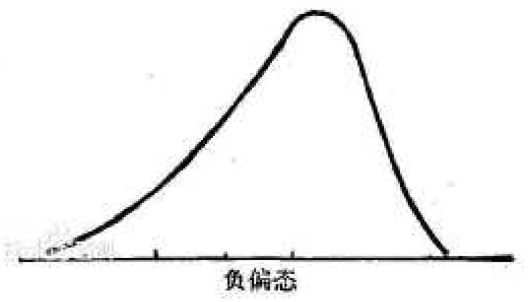
- 偏态系数
- 峰态系数

偏态

- 偏态(skewness)是对数据分布对称性的测度，如下图所示。







举个例子，学员的考试成绩，

- 正态：即正态分布，大多数学员的考试成绩中等，成绩特别高的很少，特别低的也很少。
- 正偏态：大多数学员的考试成绩偏低，成绩中等很少，成绩特别高的更少。
- 负偏态：大多数学员的考试成绩偏高，成绩中等的很少，成绩特别低的更少。

偏态

我们可以通过偏态系数来衡量偏态，计算公式如下。

$$sk = \frac{n\sum(x_i - \bar{x})^2}{(n-1)(n-2)s^3}$$

其中，s为样本的标准差。

- 当$sk > 0$时，分布是正偏态的。
- 当$sk = 0$时，分布是对称的。
- 当$sk < 0$时，分布是负偏态的。

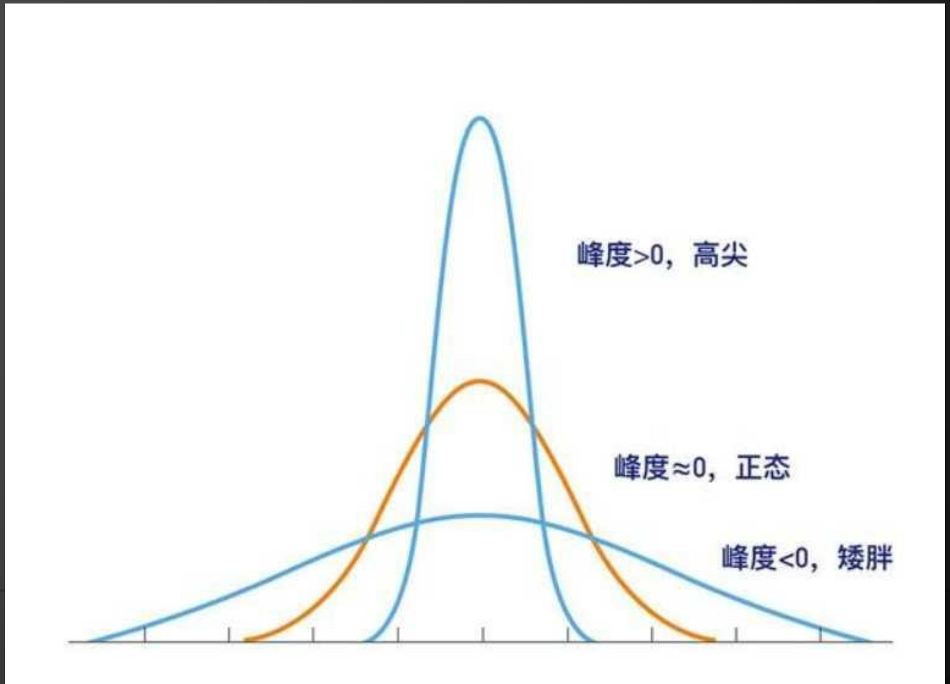
在Excel中，通过skew公式可以很容易计算出偏态系数。

峰态

- 峰态表示数据分布的扁平程度的度量。例如，不同峰态的分布如下图所示。
- 打个比方，学员的考试成绩，正态意味着大多数学员的考试成绩中等，成绩特别高的很少，特别低的也很少。“高尖”的分布形态以为几乎所有学员的考试成绩中等，成绩特别高和特别低的几乎没有。“矮胖”的分布形态意味着有一部分学员的考试成绩中等，成绩特别高和特别低的也有不少。
- 用峰态系数可以衡量峰态，峰态系数用K来表示。

1. 当$K < 0$时，分布比较高尖，为尖峰分布。
2. 当$K > 0$时，分布比较矮胖，为平峰分布。

- 在Excel中，可以通过公式KURT来计算峰态系数。

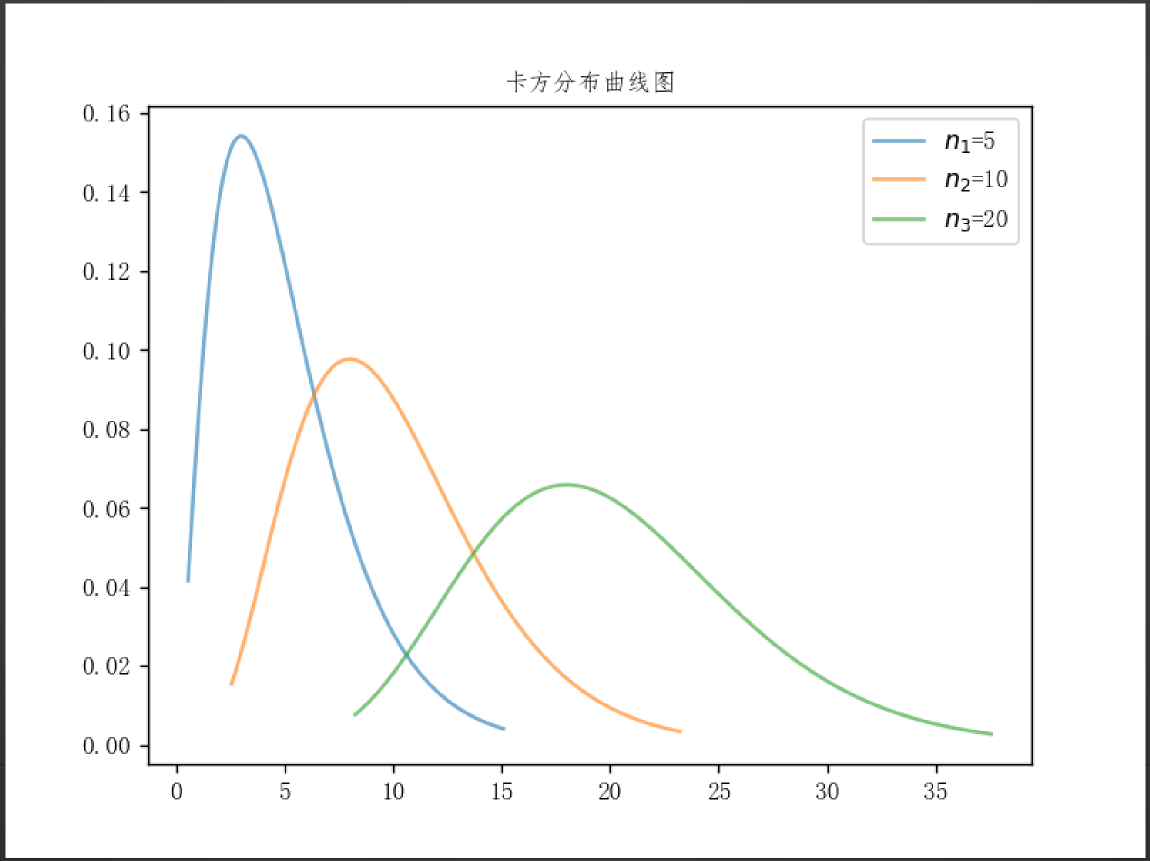


本章主要内容

1. 统计学基本概念：统计学介绍、统计数据的类型、常用基本概念
2. 描述性统计分析：集中趋势的度量、离散趋势的度量、偏态与峰态
3. 由正态分布导出的几个重要分布：卡方分布、t分布、F分布
4. 常用统计量的分布：均值、比例、方差

卡方分布

定义：随机变量$X_1, X_2, \cdots, X_n$相互独立，且$X_i(i = 1,2,\cdots,n)$服从标准正态分布$N(0,1)$，则它们的平方和$\sum_{i=1}^{n} X_i^2$服从自由度为n的χ^2分布。



卡方分布

χ^2分布两边不对称，左边的称为左尾分布，右边的称为右尾分布，如下图所示。

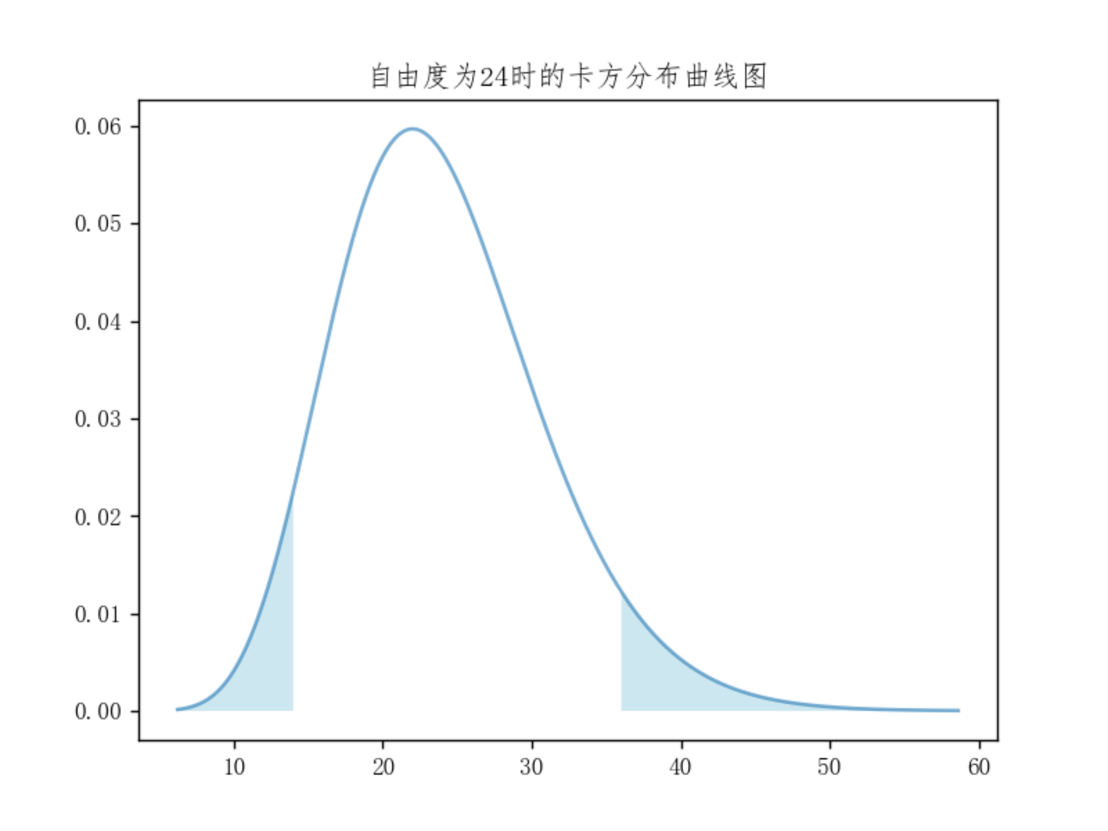
χ^2分布的分位点（分位数）：

- 右边的分位点：也上分位点，假设左右两边阴影部分的概率相等，总和$\alpha = 0.05$，则其对应的分位点为$\chi^2_{\alpha/2}(24)$。
- 左边的分位点：对应的分位点为 $\chi^2_{1-\alpha/2}(24)$，

在Excel中，计算χ^2分布的分位点的公式为：

- 右边的分位点：CHISQ.INV.RT
- 左边的分位点：CHISQ.INV

说明：后续内容会用到！



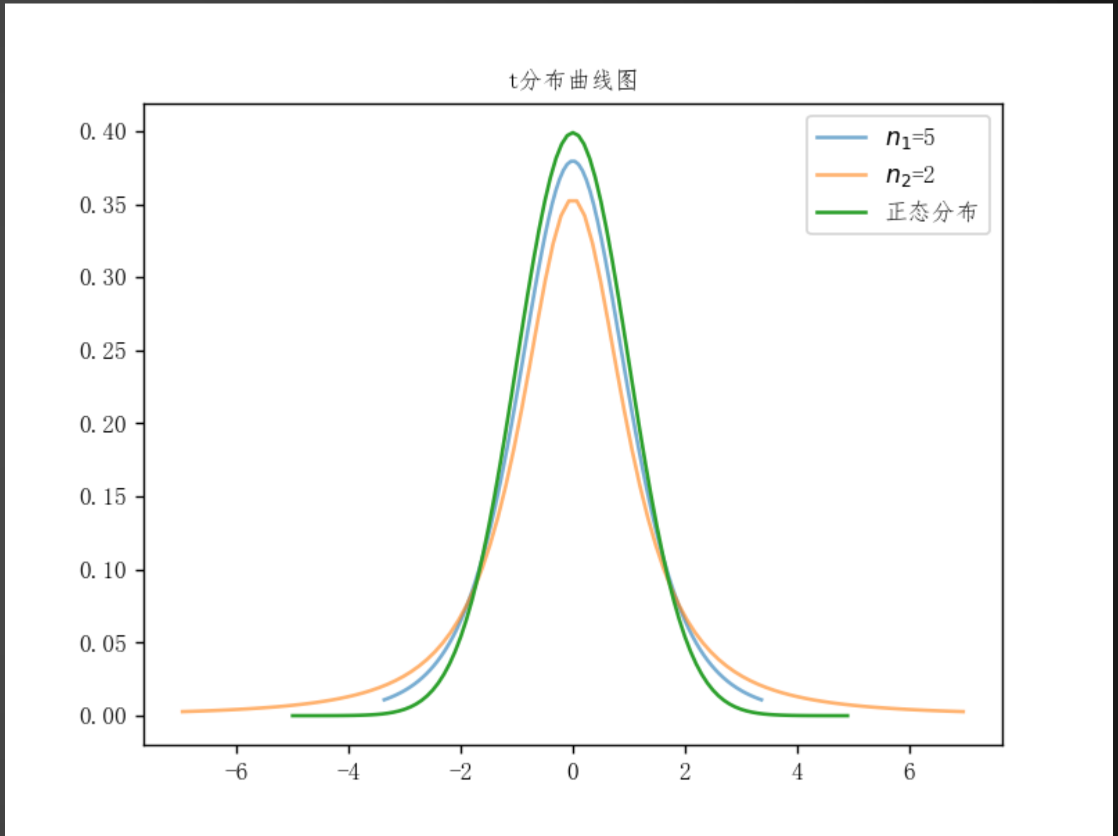
t分布

t分布也称学生氏分布，是戈赛特于1908年在一篇以 “Student”为笔名的论文中首次提出的。

定义：设随机变量$X\sim N(0,1), Y\sim\chi^2(n)$，且$X$与$Y$独立，则

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

其分布称为t分布，记为$t(n)$，其中，n为自由度。

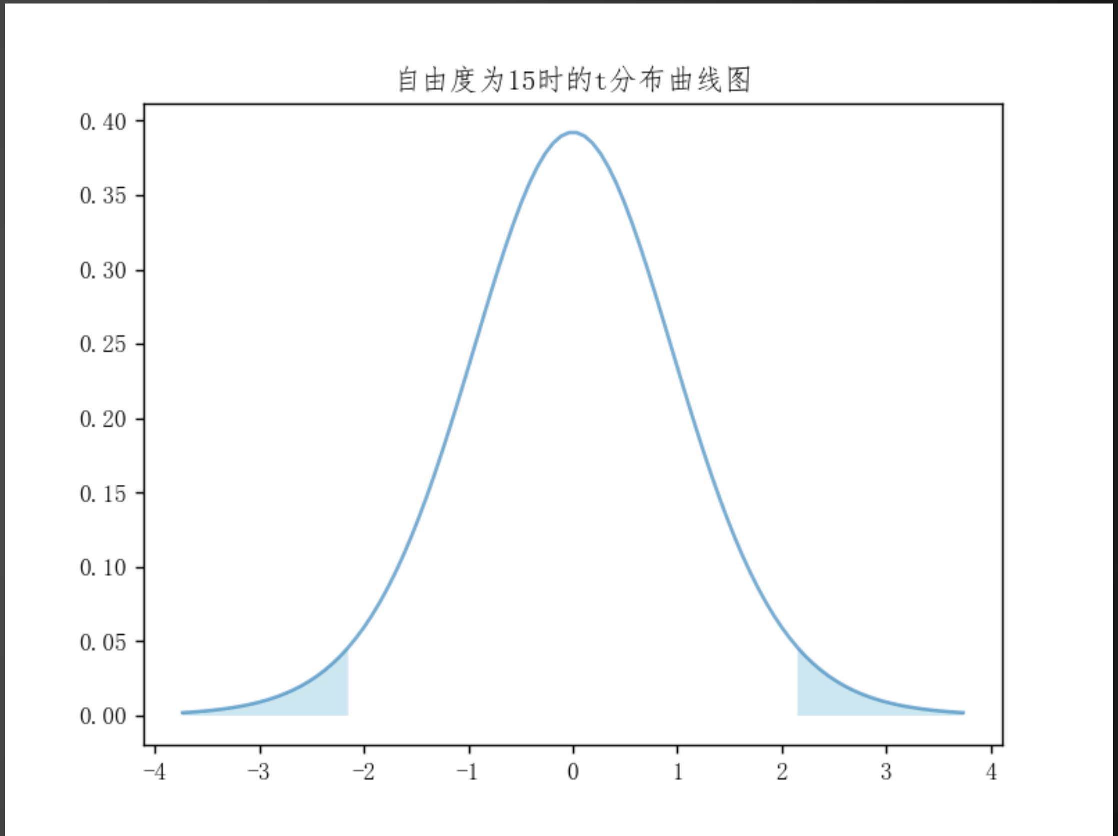


从分布图可以看出，由于t分布是关于$x=0$对称的。

t分布

t分布的分位点：

- 由于t分布左右两边对称，所以计算分位点只需要计算一边即可。
- 右图中两边阴影部分代表的概率为$\alpha = 0.05$，则对应的分位点为$t_{\alpha/2}(15)$。
- 用Excel公式T.INV可以获得t分布的分位点



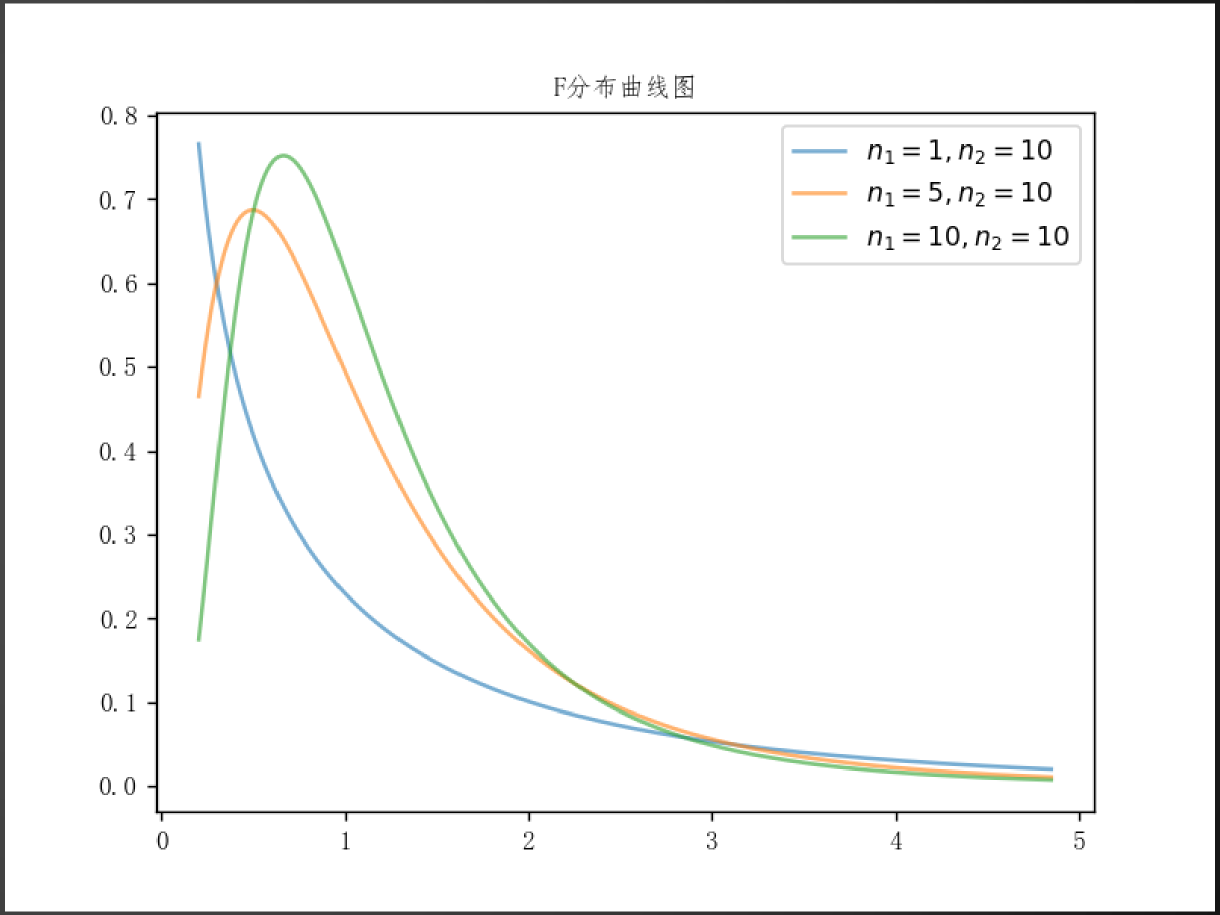
F分布

定义：设随机变量Y和Z相互独立，且Y和Z分别服从自由度为n_1和n_2的χ^2分布，则随机变量X：

$$X=\frac{Y/n_1}{Z/n_2}=\frac{n_2Y}{n_1Z}$$

则称X服从第一自由度为n_1，第二自由度为n_2的F分布，记为$X \sim F(n_1,n_2)$

F分布在后续课程中，例如参数估计、方差分析中有着重要的应用。



F分布

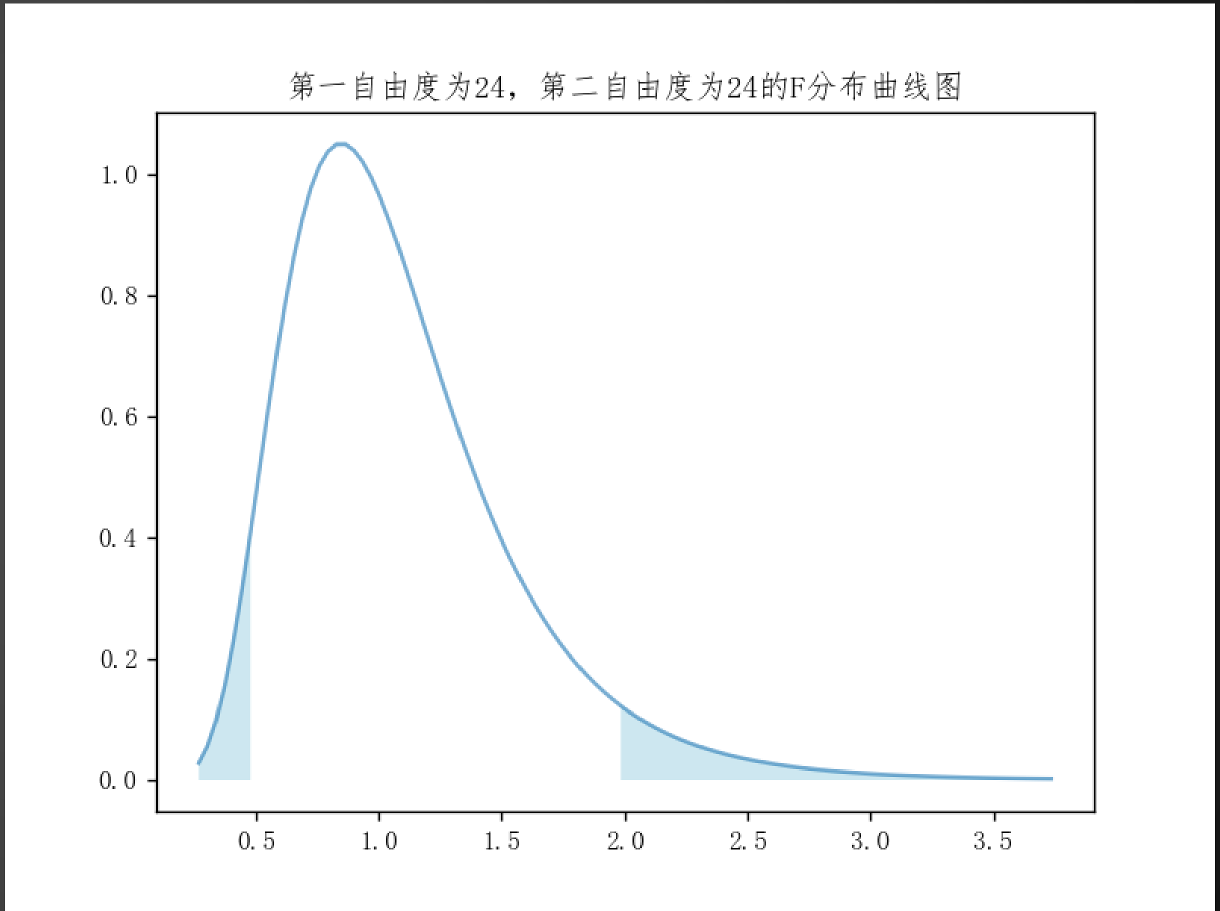
F分布的分位点：假设左右两边阴影部分的概率相等，总和$\alpha = 0.05$

- 右边的分位点记为$F_{\alpha/2}(n_1, n_2)$
- 左边的分位点记为$F_{1-\alpha/2}(n_1, n_2)$
- 且有如下有关系

$$F_{1-\alpha/2}(n_1, n_2) = \frac{1}{F_{\alpha/2}(n_1, n_2)}$$

Excel中，计算F分布的分位点公式为：

- 右边： F.INV.RT
- 左边： F.INV



本章主要内容

1. 统计学基本概念：统计学介绍、统计数据的类型、常用基本概念
2. 描述性统计分析：集中趋势的度量、离散趋势的度量、偏态与峰态
3. 由正态分布导出的几个重要分布：卡方分布、t分布、F分布
4. 常用统计量的分布：均值、比例、方差

常用统计量的分布

样本常用统计量包括样本均值、样本比例、样本方差。

- 样本均值：一个样本均值的分布、两个样本均值之差的分布
- 样本比例：一个样本比例的分布、两个样本比例之差的分布
- 样本方差：一个样本方差的分布、两个样本方差比的分布

常用统计量的分布

样本常用统计量包括样本均值、样本比例、样本方差。

- 样本均值：一个样本均值的分布、两个样本均值之差的分布
- 样本比例：一个样本比例的分布、两个样本比例之差的分布
- 样本方差：一个样本方差的分布、两个样本方差比的分布

样本均值的分布

- 当总体分布为正态分布$N(\mu,\sigma^2)$时，有以下结论：

样本均值$\bar{X}$的抽样分布仍然服从正态分布，期望为μ，方差为σ^2/n，即

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

- 如果总体不是正态分布，当n充分大时（$n\geq 30$），根据中心极限定理，样本均值$\bar{X}$的抽样分布近似服从正态分布。

中心极限定理：从均值为μ，方差为σ^2的任意一个总体中抽取样本量为n的样本，当n充分大时，样本均值$\bar{X}$的抽样分布近似服从期望为μ，方差为σ^2/n的正态分布。

两个总体均值之差的分布

实际业务中，会遇到比较两个均值之差的问题，例如比较两个学校的平均分数、比较两组工人组装产品的平均时间等。

设$\bar{X}_1$是独立地抽自总体$X_1 \sim N(\mu_1, \sigma_1^2)$的容量为$n_1$的样本的均值

设$\bar{X}_2$是独立地抽自总体$X_2 \sim N(\mu_2, \sigma_2^2)$的容量为$n_2$的样本的均值，则有

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

$$D(\bar{X}_1 - \bar{X}_2) = D(\bar{X}_1) + D(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

常用统计量的分布

样本常用统计量包括样本均值、样本比例、样本方差。

- 样本均值：一个样本均值的分布、两个样本均值之差的分布
- 样本比例：一个样本比例的分布、两个样本比例之差的分布
- 样本方差：一个样本方差的分布、两个样本方差比的分布

样本比例的分布

已知总体比例为π，从该总体中抽取n个样本，在这n个样本中，具有某一特征的样本个数为X，则样本比例用p表示，则$p=\frac{X}{n}$

当n充分大时（$n\geq 30$），p近似服从正态分布，均值为π，方差为$\frac{\pi(1-\pi)}{n}$，即

$$p\sim N(\pi,\frac{\pi(1-\pi)}{n})$$

说明：推导需要用到二项分布、渐进分布等知识，这里略去。

两个样本比例之差的分布

样本1：从具有参数π_1的二项总体中抽取的包含n_1个观测值的样本

样本2：从具有参数π_2的二项总体中抽取的包含n_2个观测值的样本

则两个样本比例差为：

$$p_1 - p_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$$

当n_1和n_2很大时，该比例之差近似服从正态分布，均值和方差分别为：

$$E(p_1 - p_2) = \pi_1 - \pi_2$$

$$D(p_1 - p_2) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

常用统计量的分布

样本常用统计量包括样本均值、样本比例、样本方差。

- 样本均值：一个样本均值的分布、两个样本均值之差的分布
- 样本比例：一个样本比例的分布、两个样本比例之差的分布
- 样本方差：一个样本方差的分布、两个样本方差比的分布

样本方差的分布

样本方差的分布比较复杂，这里只讨论当总体分布为正态分布时，样本方差的分布。

当$X_1, X_2, \cdots, X_n$为来自正态总体$N(\mu, \sigma^2)$的样本，则样本方差S^2的分布为：

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

即服从自由度为$n-1$的卡方分布。

两个样本方差比的分布

跟之前一样，只讨论两个总体为正态分布时的情况。

设$X_1, X_2, \cdots X_{n_1}$是来自正态总体$N(\mu_1, \sigma_1^2)$的一个样本，设$Y_1, Y_2, \cdots Y_{n_2}$是来自正态总体$N(\mu_2, \sigma_2^2)$的一个样本，且，$X_i(i = 1,2, \cdots, n)$与$Y_i(i = 1,2, \cdots, n)$相互独立，则

$$\frac{S_x^2/S_y^2}{\sigma_1^2/\sigma_2^2} = \frac{S_x^2/\sigma_1^2}{S_y^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

其中，

$\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ ， $\bar{Y} = \frac{1}{n_1}\sum_{i=1}^{n_2} Y_i$

$S_x^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2$， $S_y^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2$

$F(n_1 - 1, n_2 - 1)$是第一自由度为$n_1 - 1$，第二自由度为$n_2 - 1$的F分布。