样本及抽样分布

统计学概述

在概率论中，我们多研究的随机变量，它的分布都是假设已知的。如果你已经知道了随机变量X是的分布和参数，你去推导它的期望、方差等数字特征，去推导它其他一些性质，去推导X的平方是什么分布，或推导和另一个随机变量Y相加又是什么分布。这些工作属于**概率论**范畴。

但在数理统计中，我们研究的随机变量，它的分布是未知的，或者是某些参数不知道，人们通过对所研究的随机变量进行重复独立的观察，得到许多观察值，对这些数据进行分析，从而对所研究的随机变量的分布做出种种推断。比如，实际工作中有个随机变量Z，你不知道是什么分布，你看到了一些试验值，觉得Z可能是正态分布，于是你假设Z是正态分布，你用试验数据，推断出它的均值可能是1，方差可能是4，然后做假设检验，看看这一结论在多大程度上可靠，如果认为可靠，用这个结论来做分析，或者预测将要进行的试验结果。这叫**统计**。

概率论是统计推断的基础，在给定数据生成过程下观测、研究数据的性质，是**推理**；而统计推断则根据观测的数据，反向思考其数据生成过程。预测、分类、聚类、估计等，都是统计推断的特殊形式，强调对于数据生成过程的研究，是**归纳**。

抽样：总体与样本

总体

总体，是指由许多有某种共同性质的事物组成的集合，会在此集合中选出样本进行统计推断，选取样本的方式可能会用乱数或是其他抽样方式。

例如要针对所有乌鸦的共有特性进行研究，总体是目前存在、以前曾经存在或是未来可能存在的所有乌鸦。**但是，因为时间的限制、地域可取得性的限制、以及研究者的有限资源等，不可能观测总体中的每一个，因此研究者会从总体中产生样本，再由样本的特性去了解总体的特性。**

产生样本的目的之一就是为了要知道**总体的特性**，包括

- 总体均值： $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$
- 总体标准差： $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$

样本

研究中，从总体中抽取（观察或调查）一部分的个体称为样本。

样本容量

样本容量是指一个样本中所包含的单位数，一般用n表示，它是抽样推断中非常重要的概念。样本容量的大小与推断估计的准确性有着直接的联系，即在总体既定的情况下，样本容量越大其统计估计量的代表性误差就越小，反之,样本容量越小其估计误差也就越大。

最常用的样本统计量——样本均值

根据样本构造的不含未知参数的函数为**统计量**，样本均值是一个统计量。我们可以用样本均值描述一个样本，多个样本则会有多个样本均值。

描述统计

直方图和箱线图

直方图

略

分位

Q1：四分位

Q3：四分之三分位

IQR：四分位差

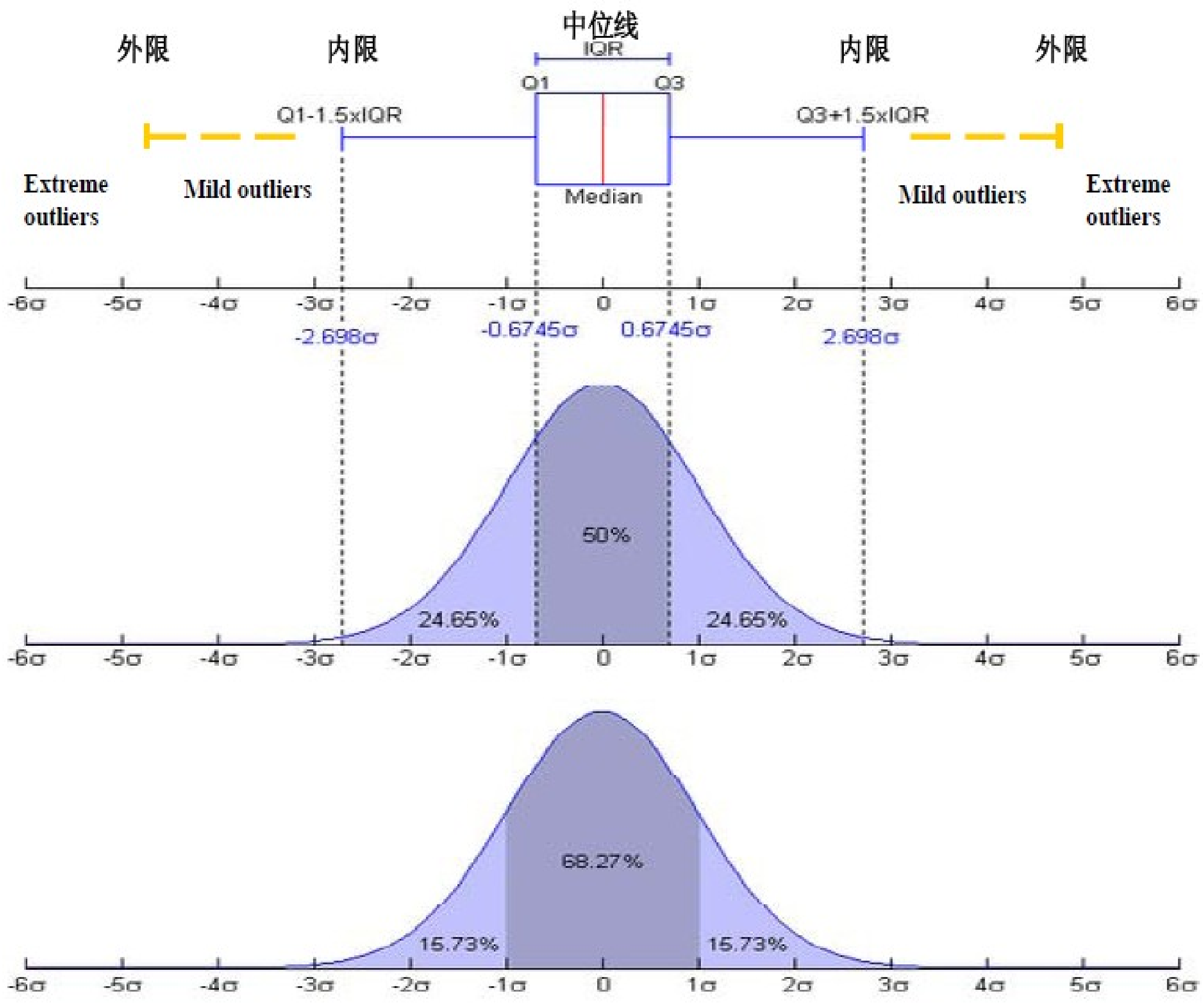
1. 几乎 50% 的数据在 IQR 间。
2. IQR 受到数据集中每一个值的影响。
3. IQR 不受异常值的影响。
4. 均值不一定在IQR中。

异常值Outlier

$Outlier < Q1 - 1.5 \times IQR$

$Outlier > Q3 + 1.5 \times IQR$

箱线图Boxplot



集中趋势

均值Mean

当数据中出现异常值时，均值无法描述分布中心；

中位数Medium

众数也很难描述分布中心；

众数Mode

中位数不会考虑到所有的数据，对异常值的鲁棒性更好。在处理高偏斜分布时，中位数通常能够最好地反映出集中趋势。

正偏斜分布与负斜分布

正斜分布靠左： $mode < medium < mean$

负斜分布靠右： $mean < medium < mode$

鲁棒性Robust

即使偏离了基准也不会受太大的影响。

离散程度

方法

找出任意两个值之间差的平均值：数值过多

找出每个值与最大值或最小值之间差的平均值：容易受异常值干扰

找出每个值与数据集均值之间差的平均值：适合

概念

离均差：$x_i - \bar{x}$

平均偏差：$\sum \frac{x_i - \bar{x}}{n} = 0$

平均绝对偏差：$\sum \frac{|x_i - \bar{x}|}{n} = 0$

（总体）方差(平均平方偏差)：$DX = \sum \frac{(x_i - \bar{x})^2}{n} = E(X - EX)^2 = EX^2 - (EX)^2$

（总体）标准差：$\sigma = \sqrt{DX}$

贝塞尔校正

比如在高斯分布（正态分布）中，我们抽取一部分的样本，用样本的方差来估计总体的方差。由于样本主要是落在 $x = \mu$ 中心值附近，那么样本方差一定小于总体的方差（因为高斯分布的边沿抽取的数据很少）。为了能弥补这方面的缺陷，那么我们把公式的 n 改为 $n-1$,以此来提高方差的数值。这种方法叫做贝塞尔校正系数。

当我们用小样本数据的标准差去估计总体的标准差的时候采用 $n-1$，但是这个小样本数据的实际标准差还是用 n 的那个公式 的，不要混淆了数据的实际标准差。

无偏性证明

对于一个随机变量 X 进行 n 次抽样，获得样本 $x_1, x_2, \ldots, x_n$ ，那么样本均值：$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

有偏的样本方差为:

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

无偏的样本方差为：

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

为了证明 s^2 的无偏性， 我们拿出样本方差种的一部分来进行单独分析，

$$\begin{aligned}
&\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2
\end{aligned}$$

同理，我们有

$$\begin{aligned}
&\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n}\left[(x_i - \mu) - (\bar{x} - \mu)\right]^2 \\
&= \sum_{i=1}^{n}\left[(x_i - \mu)^2\right]^2 - n(\bar{x} - \mu)^2
\end{aligned}$$

对上式两侧取期望，我们有

$$
\begin{aligned}
&\mathrm{E}\left(\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\right)\\
=&\mathrm{E}\left(\sum_{i=1}^{n}\left[\left(x_i-\mu\right)-\left(\overline{x}-\mu\right)\right]^2\right)\\
=&\mathrm{E}\left(\sum_{i=1}^{n}\left(x_i-\mu\right)^2-n\left(\overline{x}-\mu\right)^2\right)\\
=&\sum_{i=1}^{n}\mathrm{E}\left(\left(x_i-\mu\right)^2\right)-n\,\mathrm{E}\left(\left(\overline{x}-\mu\right)^2\right)\\
=&\sum_{i=1}^{n}\mathrm{Var}\left(x_i\right)-n\,\mathrm{Var}\left(\overline{x}\right)
\end{aligned}
$$

因为 $\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，于是我们有 $\mathrm{Var}\left(\overline{x}\right)=\frac{1}{n}\mathrm{Var}\,x$

因此

$$
\begin{aligned}
&\mathrm{E}\left(\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\right)\\
&=\sum_{i=1}^{n}\mathrm{Var}\left(x_i\right)-n\,\mathrm{Var}\left(\overline{x}\right)\\
&=\left(n-1\right)\mathrm{Var}\left(x\right)
\end{aligned}
$$

最后，我们有

$$
\begin{aligned}
\mathrm{E}\left(s^2\right)&=\mathrm{E}\left(\sum_{i=1}^{n}\frac{\left(x_i-\overline{x}\right)^2}{n-1}\right)\\
&=\frac{1}{n-1}\mathrm{E}\left(\sum_{i=1}^{n}\left[\left(x_i-\mu\right)-\left(\overline{x}-\mu\right)\right]^2\right)\\
&=\mathrm{Var}\left(x\right)
\end{aligned}
$$

可见 S^2 是对 $Var\left(X\right)$ 的无偏估计。

3σ 原则

数值分布在 $(\mu-\sigma,\mu+\sigma)$ 中的概率为 0.6827。

数值分布在 $(\mu-2\sigma,\mu+2\sigma)$ 中的概率为 0.9545。

数值分布在 $(\mu-3\sigma,\mu+3\sigma)$ 中的概率为 0.9973。

链接。

归一化：标准正态分布

样本均值的频数直方图的数字不能直接看出比例排名，所以引入**频率直方图**，但直方图固有弊端在于会缺少部分信息，所以需要缩小组距以增加信息，但过小又没有了直方图意义，所以引入**概率分布图——标准正态分布**。

Z 值：标准差数量

公式

$$z = \frac{x - \mu}{\delta}$$

含义

无论值是多少，我们都可以将其转换为与均值的标准差。通过将正态分布中的值转换为 z，就可以知道小于或大于该值得百分比。

例如某个值与平均值相差 1 个标准偏差 σ，则无论是哪种正态分布，我们都知道大约 80% 的值 < 该值。

标准正态分布

我们可以将任何正态分布转化为标准正态分布，通过 Z 值进行分析，再按照任何方式扩展。

Z 值表

链接。

抽样分布

样本统计量

根据样本构造的不含未知参数的函数为统计量。

样本均值

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

样本方差

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

样本标准差

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

样本K阶(原点)矩

$$A_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k, \quad k = 1, 2, \ldots$$

样本K阶中心矩

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

抽样分布

在使用统计量进行统计推断时，需要知道统计量的分布，比如样本均值的分布。

统计量的分布，叫做**抽样分布**。总体分布函数已知时，样本分布是确定的，但是：

1. 通常，我们是不知道总体分布的；
2. 要求出统计量的精确分布是困难的。

虽然总体不知道时，我们很难确定，解决这种问题需要学习非参数统计。然而，有两种情况是比较好研究的：

1. 对于正太总体分布，其常用的统计量的分布是可以推断出来的。
2. 对于一般总体分布，我们可以由大数定律和中心极限定理得到其样本均值统计量的期望、分布和方差等。

正态总体的常用统计量的分布

假设总体 $X \sim N\left(\mu, \sigma^2\right)$。

我们可能会用到各种各样的**统计量**，但归根结底是这些统计量满足**四种典型的分布**，即 z（即正态分布）、χ^2、t 和 F 分布，每一个分布对应一种检验方法，即 z 检验、检 χ^2 验、t 检验和 F 检验。

这些统计量大多都是一个样本或多个样本的样本均值 $\bar{X}$、样本方差 S^2、总体均值 μ 和总体方差 σ^2 这些元素组成的，比如 $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$，$\dfrac{\bar{X}-\mu}{S/\sqrt{n}}$ 等，但我们在选择时，一定要只有被检验一个参数不知道，所以，如果我们想用第一个统计量，那么除了 $\bar{X}$，n这俩一定知道的参数之外，如果我们要检验 μ，那么就必须知道总体标准差 σ。换句话说，如果我们知道总体标准差 σ，那么我们就可以选择 $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ 这个统计量，并根据其满足的标准正态分布规律对总体均值 μ 进行假设检验（或求置信区间）。

但是实际情况中，总体标准差 σ 我们大多不知道，这个时候就不能使用 $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ 这个统计量了，而由于我们能求出样本标准差 S，那么就可以选择 $\dfrac{\bar{X}-\mu}{S/\sqrt{n}}$ 这个统计量，这个统计量需要知道的关于总体的信息（参数）更少，但也服从 $t(n-1)$ 分布。也就是只需要知道总体满足正态分布即可，而不需要知道其总体方差 σ，就可以对总体均值 μ 进行检验。

我们了解并学习这 4 种分布，是因为这 4 种分布，其分布函数和密度函数都能很好地进行量化，正态分布就是最好的例子，其他三种只是学习之前我们不常接触而已。

以下是对这四种分布的详细的介绍。

样本均值的正态分布

样本均值是最常用的统计量之一，一般用于 z–**检验**，用以检验总体均值。

统计量：$\bar{X}=\dfrac{1}{n}\sum_{i=1}^{n}X_i$，或 $z=\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$

统计量分布：$\bar{X}\sim N\left(\mu,\dfrac{\sigma^2}{n}\right)$，或 $z\sim N(0,1)$。

卡方分布

定义

设 $X\sim N(0,1)$，则称统计量

$\chi^2=X_1^2+X_2^2+\ldots+X_n^2$

服从自由度为 n的 χ^2 分布，记为$\chi^2\sim\chi^2(n)$，自由度指上式右端包含的独立变量的个数。

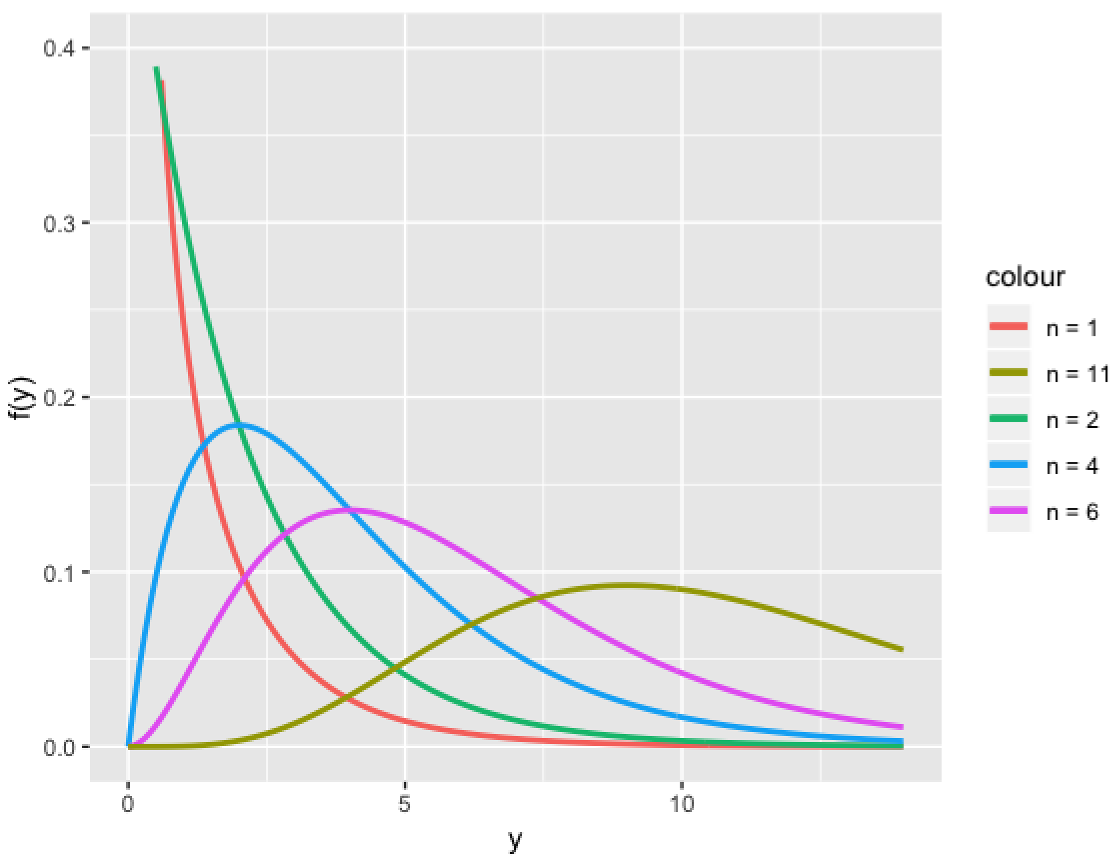
密度函数

$$f_n(y) = \frac{1}{2^{n/2}\Gamma(n/2)} y^{n/2-1} e^{-y/2}$$，其中 $x \geq 0$；

当 $x \leq 0$ 时，$f_k(x) = 0$。这里 Γ 代表Gamma函数。

> 推导见书P139

图形



卡方分布的可加性

设 $\chi_1^2 \sim \chi^2(n_1), \chi_2^2 \sim \chi^2(n_2)$，并且 χ_1^2, χ_2^2 相互独立，则有 $\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2)$

卡方分布的数学期望和方差

$\chi^2 \sim \chi^2(n)$，则 $E(\chi^2) = n, D(\chi^2) = 2n$ 。

卡方分布上的分位点

对于给定的 $\alpha, 0 < \alpha < 1$，满足条件：

$$P\{\chi^2 > \chi^2_\alpha(n)\} = \int_{\chi^2_\alpha(n)}^{\infty} f(y)dy = \alpha$$

卡方分布表

卡方分布表

费希尔曾证明，当 n 充分大时，近似地有 $\chi^2_\alpha(n) \approx \frac{1}{2}(z_\alpha + \sqrt{2n-1})^2$ 。

利用前式可以求得当 $n > 40$ 时卡方分布上 α 分位点的近似值。

t 分布

定义

设 $X \sim N(0,1)$ ，$Y \sim \chi^2(n)$，且 X，Y 相互独立，则称随机变量(统计量)

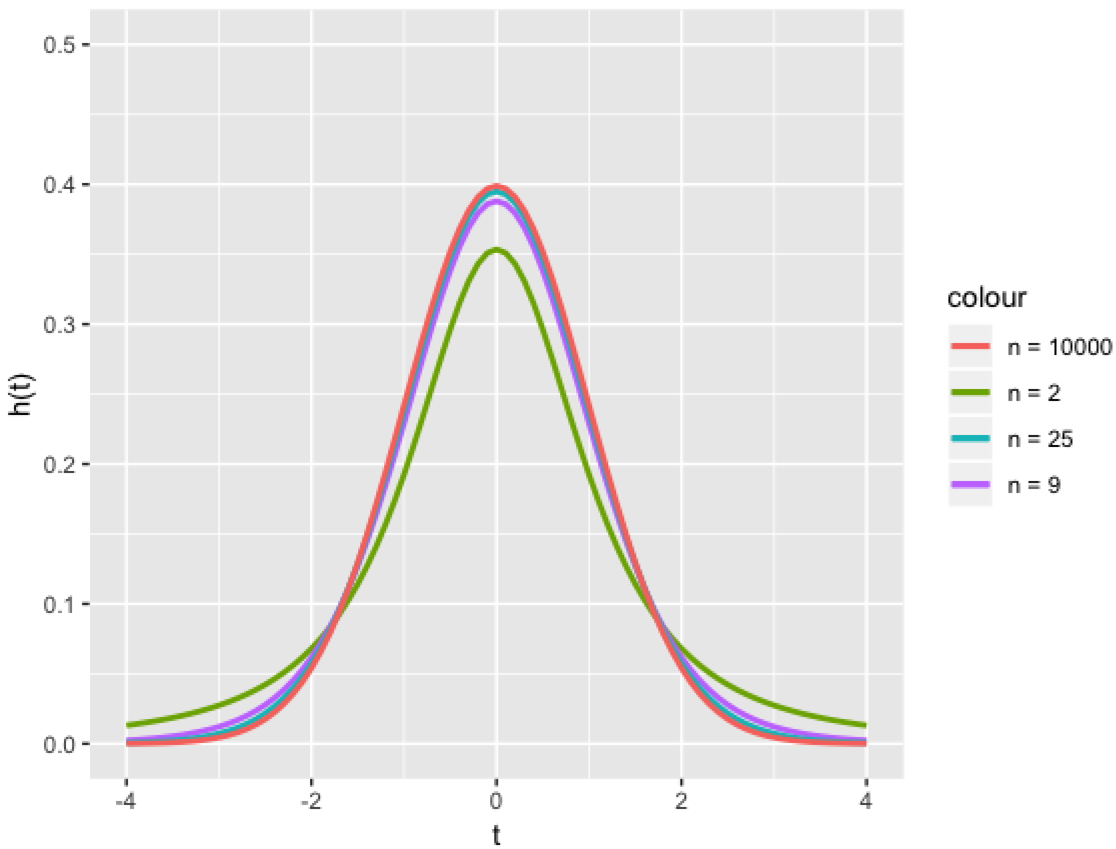
$$t = \frac{X}{\sqrt{Y/n}}$$

服从自由度为n的t分布，即为 $t \sim t(n)$，t 分布又称学生氏(student)分布。

密度函数

$$h(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{\pi n}\Gamma(n/2)}(1+\frac{t^2}{n})^{-(n+1)/2}, -\infty < t < \infty$$

图像



$h(t)$ 的图形关于 $t=0$ 对称，当 n 充分大时，其图形类似于标准正态变量概率密度的图形。但对于较小 n ，t 分布与 $N(0,1)$ 分布相差较大。

t 分布分位点

对于给定的 $\alpha, 0<\alpha<1$，满足条件：

$$P\{t>t_\alpha(n)\}=\int_{t_\alpha(n)}^{\infty} h(t)dt=\alpha$$

的点 $t_a\,(n)$ 就是 $t(n)$ 分布上的 α 分位点。

$t_{1-a}\,(n)=-t_a\,(n)$

且当 $n>45$ 时，对于常用的 α 的值，就用正态近似：$t_a(n)=z_\alpha$

F 分布

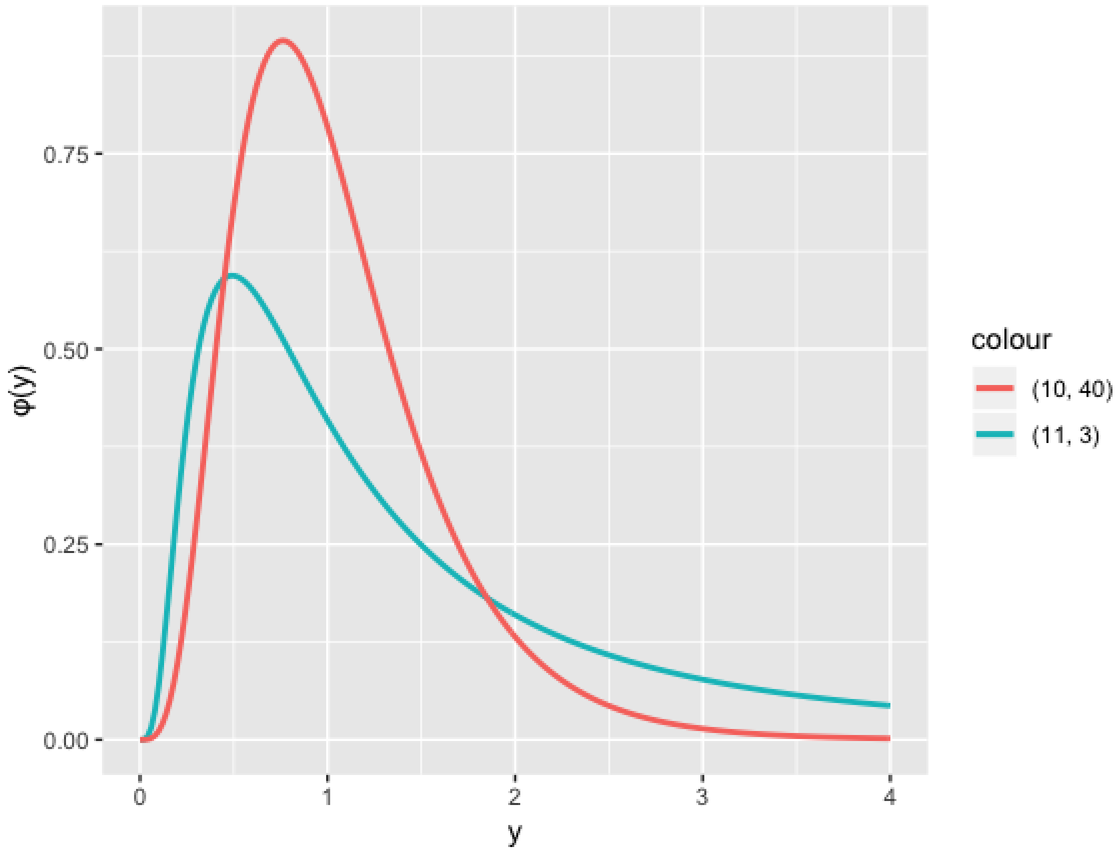
定义

$U\sim\chi^2(n_1),V\sim\chi^2(n_2)$，且 U,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 F 分布，记为 $F\sim F(n_1,n_2)$。

概率密度

$\varphi(y)=\dfrac{\Gamma[(n_1+n_2)/2](n_1/n_2)^{n_1/2}y^{(n_1/2)-1}}{\Gamma(n_1/2)\Gamma(n_2/2)[1+(n_1y/n_2)]^{(n_1+n_2)/2}},y>0$，其他为 0。

图形



由定义可知，若 $F \sim F(n_1, n_2)$，则 $\frac{1}{F} \sim F(n_2, n_1)$

还有性质：$F_{1-\alpha}(n_1, n_2) = \frac{1}{F_\alpha(n_2, n_1)}$

分位点

对于给定的 $\alpha, 0 < \alpha < 1$，满足条件：

$$P\{F > F_\alpha(n_1, n_2)\} = \int_{F_\alpha(n_1,n_2)}^{\infty} \varphi(y)dy = \alpha$$

的点 $F_\alpha(n_1, n_2)$ 就是 $F(n_1, n_2)$ 分布的 **上 α 分位点**。

类似地有卡方分布，t 分布，F 分布的**下分位点**。

常用统计量的分布

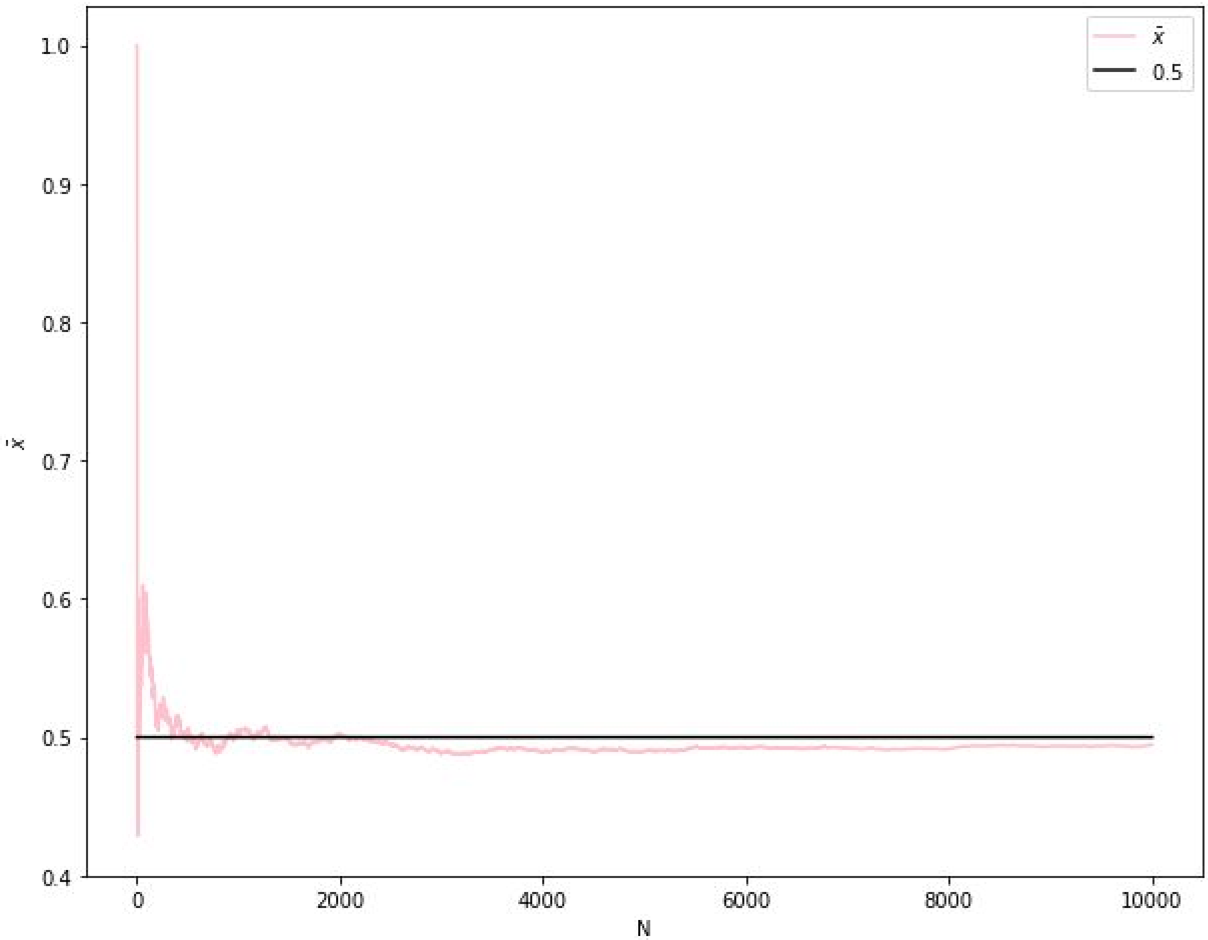
1. $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
2. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
3. $\bar{X}$ 与 S^2 相互独立
4. $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$
5. $\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$
6. 当 $\sigma_1^2=\sigma_2^2=\sigma^2$ 时，$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_\omega\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$，其中 $S_\omega^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}, S_\omega=\sqrt{S_\omega^2}$

一般总体样本均值的分布

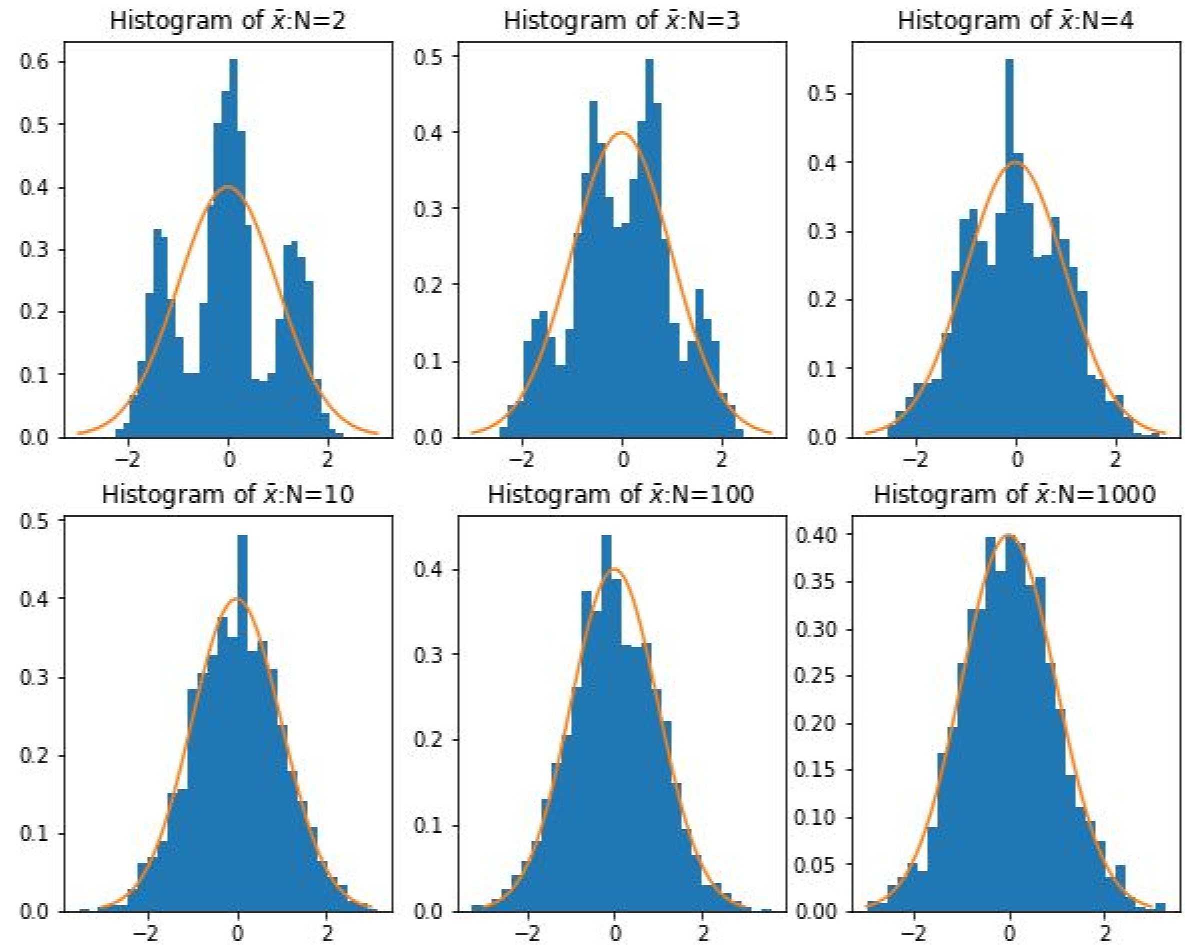
大数定律和中心极限定理回顾

在概率论中，我们已经了解了**大数定律**和**中心极限定理**（详见前面的章节）：

大数定律讲的是样本均值收敛到总体均值（就是期望），像这个图一样：



而中心极限定理告诉我们，当样本量足够大时，样本均值的分布慢慢变成正态分布，就像这个图：



示例

X 代表掷骰子点数的随机变量，$X=1,2,...,6,EX=3.5$，我们做一次试验时掷 2 次骰子，即样本容量为 2，做一次实验的话是一个样本，2 个数字的均值是一个统计量，叫样本均值。

对于这个实验，我们知道总体分布或分布律，为比如一个样本 $(1,4)$，样本均值=2.5，也就是观察值 =2.5。我们可以发现，只做一次试验，样本统计量的观察值是不等于总体 X 的均值 $EX=3.5$ 的。

但是，只要我们试验的次数足够多，比如又做了 100 次试验，得到 100 个样本：$(4,6),(3,1),(1,2)...$ 样本均值的观察值依次为：5，2，1.5，... 大数定律说的就是这些样本均值依概率收敛于总体期望，即 $\frac{1}{100}(2.5+5+2+1.5...)\approx 3.5=EX$ ，用依概率收敛的符号表示即 $\bar{X}_n \xrightarrow{p} \mu$。

中心极限定理是说，当样本量足够大时，这些样本均值的观察值是满足正态分布的。

总结

随机变量 $X, EX=\mu, DX=\sigma^2$。则独立同分布情况下，若样本量很大，由中心极限定理，样本均值 $\bar{X}$ 近似地服从参数为 $N(\mu, \frac{\sigma^2}{n})$ 的正态分布。

样本容量如果增大 n 倍，其标准差会缩小为 $\frac{1}{\sqrt{n}}$，分布也会变窄。

应用

1. 对于一个随机变量 $X, EX=\mu, DX=\sigma^2$。若设定样本容量为 n，我们可以得到样本均值 $\bar{X}$ 的满足参数为 $N(\mu, \frac{\sigma^2}{n})$ 的正态分布，换个说法，$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$。
2. 在分布确定、有了抽样分布的基础上，当我们实际得到一个样本，我们想检验这个样本是否正常。
3. 既然 μ，σ 和 n 均已知，那么 $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ 是一个统计量，即 z，由于单位正态分布天然的计算和观察优势，我们可以利用 z 得到出现此样本的概率。
4. 比如，我们得到 $z > z_0$ 的概率只有 0.01，那么我们可以认为这是不正常的。因为小概率事件在一次试验中是很难发生的，但也确实有可能发生，比如这里发生的几率就是 0.01。
5. 所以我们如果假定，一次试验当原假设为真时，我们不接受它的概率为 0.05，也就是说弃真错误 $\alpha=0.05$，我们就会抛弃这个样本，觉得它是假的，也就是说我们认为这个样本不正常。另一种说法是，我们有 0.95 的把握认为这个样本是不正常的。
6. 这就是假设检验的基本思想，具体会在之后的章节提到。