统计量与分布

续本达

清华大学 工程物理系

2023-11-13

复习

统计学基本概念

统计学

收集、分析、表述和解释数据的科学

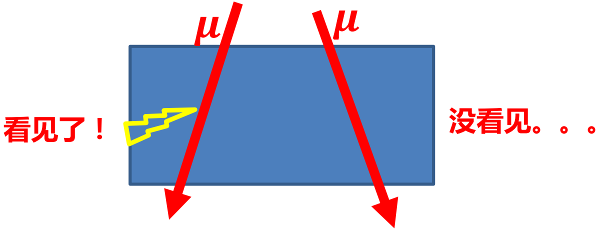
描述统计学

- 列表法：频数分布表、频率分布表
- 图示法：
  1. 直方图：有四种画法，纵轴的标度不同
  2. 箱线图
- 样本分位数：设有容量为 n 的样本观察值 $(x_1, x_2, \cdots, x_n)$，样本 p 分位数 $(0 < p < 1)$ 记为 x_p，
  1. 至少有 np 个观察值小于或等于 x_p；
  2. 至少有 $n(1-p)$ 个观察值大于或等于 x_p。

随机样本

引子

探测器效率：搭建一个宇宙线缪子（$\mu^{\pm}$）探测器。要求探测效率 $\epsilon > 95\,\%$.



探测器效率：

$$\epsilon = \frac{n_{\text{obs}}}{n_{\text{tot}}}$$

总体

定义 (总体)

研究对象全体元素组成的集合
所研究的对象的某个 (或某些) 数量指标的全体，它是一个随机变量 (或多维随机变量)，记为 X。可以有限，也可以无限。

X 的分布函数和数字特征称为总体的分布函数和数字特征.

三层含义

1. 研究对象的全体
2. 数据
3. 分布

总体

定义 (总体)

研究对象全体元素组成的集合
所研究的对象的某个 (或某些) 数量指标的全体，它是一个随机变量 (或多维随机变量)，记为 X。可以有限，也可以无限。

X 的分布函数和数字特征称为总体的分布函数和数字特征.
三层含义

1. 研究对象的全体
2. 数据
3. 分布

定义

个体

组成总体的每一个元素，即总体的每个数量指标，可看作随机变量 X 的某个取值。用 X_i 表示。

样本

从总体中抽取的部分个体.

- 用 $(X_1, X_2, \cdots, X_n)$ 表示， n 为样本容量。
- 称 $(x_1, x_2, \cdots, x_n)$ 为总体 X 的一个容量为 n 的样本观测值.

样本空间

样本所有可能取值的集合.

定义

个体

组成总体的每一个元素，即总体的每个数量指标，可看作随机变量 X 的某个取值。用 X_i 表示。

样本

从总体中抽取的部分个体.

- 用 $(X_1, X_2, \cdots, X_n)$ 表示， n 为样本容量。
- 称 $(x_1, x_2, \cdots, x_n)$ 为总体 X 的一个容量为 n 的样本观测值.

样本空间

样本所有可能取值的集合.

定义

个体

组成总体的每一个元素，即总体的每个数量指标，可看作随机变量 X 的某个取值。用 X_i 表示。

样本

从总体中抽取的部分个体.

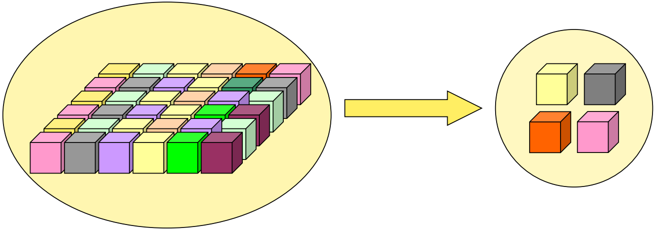
- 用 $(X_1, X_2, \cdots, X_n)$ 表示， n 为样本容量。
- 称 $(x_1, x_2, \cdots, x_n)$ 为总体 X 的一个容量为 n 的样本观测值.

样本空间

样本所有可能取值的集合.

样本的生成

从总体中抽取，并作为总体代表的一部分总体单位的集合体。



- 样本取自总体，不唯一

简单随机样本

若总体 X 的样本 $(X_1, X_2, \cdots, X_n)$ 满足:

1. $X_1, X_2, \cdots, X_n$ 与 X 有相同的分布
2. $X_1, X_2, \cdots, X_n$ 相互独立

则称 $(X_1, X_2, \cdots, X_n)$ 为 简单随机样本 .

抽样方法

- 对有限总体，放回抽样所得到的样本为简单随机样本。
- 如果放回抽样不方便, 常用不放回抽样代替，条件是 $N/n \geq 10$。N 为总体中个体总数， n 为样本容量。

简单随机样本

若总体 X 的样本 $(X_1, X_2, \cdots, X_n)$ 满足:

1. $X_1, X_2, \cdots, X_n$ 与 X 有相同的分布
2. $X_1, X_2, \cdots, X_n$ 相互独立

则称 $(X_1, X_2, \cdots, X_n)$ 为 简单随机样本 .

抽样方法

- 对有限总体，放回抽样所得到的样本为简单随机样本。
- 如果放回抽样不方便, 常用不放回抽样代替，条件是 $N/n \geq 10$。N 为总体中个体总数， n 为样本容量。

统计量

定义

样本 $(X_1, X_2, \cdots, X_n)$ 的不含有未知参数的连续函数 $g(X_1, X_2, \cdots, X_n)$ 称为 **统计量** 。

1. 利用样本的函数进行统计推断
2. 样本是随机变量，统计量也是随机变量。

例

考虑 $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$ ，如果参数 μ, σ^2 已知，则它是统计量，否则不是。

定义

样本 $(X_1, X_2, \cdots, X_n)$ 的不含有未知参数的连续函数 $g(X_1, X_2, \cdots, X_n)$ 称为 **统计量** 。

1. 利用样本的函数进行统计推断
2. 样本是随机变量，统计量也是随机变量。

例

考虑 $\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$ ，如果参数 μ, σ^2 已知，则它是统计量，否则不是。

常见统计量

设 $(X_1, X_2, \cdots, X_n)$ 是来自总体 X 的容量为 n 的样本，分别称下列统计量为

样本均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

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

样本标准差 $\sqrt{S^2}$

样本 k 阶原点矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$，例如 $A_1 = \overline{X}$

样本 k 阶中心矩 $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$

$$B_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2 \equiv S_n^2$$

常见统计量

设 $(X_1, X_2, \cdots, X_n)$ 是来自总体 X 的容量为 n 的样本, 分别称下列统计量为

样本均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

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

样本标准差 $\sqrt{S^2}$

样本 k 阶原点矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$, 例如 $A_1 = \overline{X}$

样本 k 阶中心矩 $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$

$$B_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2 \equiv S_n^2$$

常见统计量

设 $(X_1, X_2, \cdots, X_n)$ 是来自总体 X 的容量为 n 的样本，分别称下列统计量为

样本均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

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

样本标准差 $\sqrt{S^2}$

样本 k 阶原点矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$，例如 $A_1 = \overline{X}$

样本 k 阶中心矩 $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$

$$B_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2 \equiv S_n^2$$

常见统计量

设 $(X_1, X_2, \cdots, X_n)$ 是来自总体 X 的容量为 n 的样本，分别称下列统计量为

样本均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

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

样本标准差 $\sqrt{S^2}$

样本 k 阶原点矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$，例如 $A_1 = \overline{X}$

样本 k 阶中心矩 $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k$

$$B_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2 = \frac{n-1}{n}S^2 \equiv S_n^2$$

常见统计量

设 $(X_1, X_2, \cdots, X_n)$ 是来自总体 X 的容量为 n 的样本，分别称下列统计量为

样本均值 $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$

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

样本标准差 $\sqrt{S^2}$

样本 k 阶原点矩 $A_k = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^k$，例如 $A_1 = \overline{X}$

样本 k 阶中心矩 $B_k = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i-\overline{X})^k$

$$B_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2 = \frac{n-1}{n}S^2 \equiv S_n^2$$

样本均值的一阶性质

定理 (偏差之和)

若把样本中的数据与样本均值之差称为偏差，则样本所有偏差之和为零，即 $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$。

$$\sum_{i=1}^{n}(X_i - \overline{X}) = \sum_{i=1}^{n} X_i - n\overline{X} = n\overline{X} - n\overline{X} = 0$$

样本均值的一阶性质

定理 (偏差之和)

若把样本中的数据与样本均值之差称为偏差，则样本所有偏差之和为零，即 $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$。

$$\sum_{i=1}^{n}(X_i - \overline{X}) = \sum_{i=1}^{n} X_i - n\overline{X} = n\overline{X} - n\overline{X} = 0$$

样本均值的二阶性质

定理 (偏差平方和最小)

数据观察值与样本均值的偏差平方和最小，即在形如 $\sum_{i=1}^{n}(X_i-c)^2$ 的函数中，$\sum_{i=1}^{n}(X_i-\overline{X})^2$ 最小。

$$
\begin{aligned}
\sum_{i=1}^{n}(X_i-c)^2 &= \sum_{i=1}^{n}(X_i-\overline{X}+\overline{X}-c)^2 \\
&= \sum_{i=1}^{n}(X_i-\overline{X})^2 + n(\overline{X}-c)^2 + 2\sum_{i=1}^{n}(X_i-\overline{X})(\overline{X}-c) \\
&= \sum_{i=1}^{n}(X_i-\overline{X})^2 + n(\overline{X}-c)^2 \geq \sum_{i=1}^{n}(X_i-\overline{X})^2
\end{aligned}
$$

样本均值的二阶性质

定理 (偏差平方和最小)

数据观察值与样本均值的偏差平方和最小，即在形如 $\sum_{i=1}^{n}(X_i-c)^2$ 的函数中，$\sum_{i=1}^{n}(X_i-\overline{X})^2$ 最小。

$$
\begin{aligned}
\sum_{i=1}^{n}(X_i-c)^2 &= \sum_{i=1}^{n}(X_i-\overline{X}+\overline{X}-c)^2 \\
&= \sum_{i=1}^{n}(X_i-\overline{X})^2 + n(\overline{X}-c)^2 + 2\sum_{i=1}^{n}(X_i-\overline{X})(\overline{X}-c) \\
&= \sum_{i=1}^{n}(X_i-\overline{X})^2 + n(\overline{X}-c)^2 \geq \sum_{i=1}^{n}(X_i-\overline{X})^2
\end{aligned}
$$

样本均值的二阶性质

定理 (偏差平方和最小)

数据观察值与样本均值的偏差平方和最小，即在形如 $\sum_{i=1}^{n}(X_i-c)^2$ 的函数中，$\sum_{i=1}^{n}(X_i-\overline{X})^2$ 最小。

$$\sum_{i=1}^{n}(X_i-c)^2 = \sum_{i=1}^{n}(X_i-\overline{X}+\overline{X}-c)^2$$

$$= \sum_{i=1}^{n}(X_i-\overline{X})^2 + n(\overline{X}-c)^2 + 2\sum_{i=1}^{n}(X_i-\overline{X})(\overline{X}-c)$$

$$= \sum_{i=1}^{n}(X_i-\overline{X})^2 + n(\overline{X}-c)^2 \geq \sum_{i=1}^{n}(X_i-\overline{X})^2$$

样本均值的二阶性质

定理 (偏差平方和最小)

数据观察值与样本均值的偏差平方和最小，即在形如 $\sum_{i=1}^{n}(X_i - c)^2$ 的函数中，$\sum_{i=1}^{n}(X_i - \overline{X})^2$ 最小。

$$
\begin{aligned}
\sum_{i=1}^{n}(X_i - c)^2 &= \sum_{i=1}^{n}(X_i - \overline{X} + \overline{X} - c)^2 \\
&= \sum_{i=1}^{n}(X_i - \overline{X})^2 + n(\overline{X} - c)^2 + 2\sum_{i=1}^{n}(X_i - \overline{X})(\overline{X} - c) \\
&= \sum_{i=1}^{n}(X_i - \overline{X})^2 + n(\overline{X} - c)^2 \geq \sum_{i=1}^{n}(X_i - \overline{X})^2
\end{aligned}
$$

极限分布

定理 (中心极限定理)

设 $X_1, X_2, \cdots, X_n$ 是来自某个总体的样本，$\overline{X}$ 是样本均值。

1. 若总体分布为 $N(\mu, \sigma^2)$，则 $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$。
2. 若总体分布未知或不是正态分布，但 $\mathrm{E}[X] = \mu$，$\mathrm{Var}[X] = \sigma^2$ 存在，则 n 较大时，$\overline{X}$ 的渐进分布为 $N\left(\mu, \frac{\sigma^2}{n}\right)$。

一二阶矩的性质

矩

设总体 X 具有二阶矩，即 $\mathrm{E}[X] = \mu$，$\mathrm{Var}[X] = \sigma^2 < +\infty$，$X_1, X_2, \cdots, X_n$ 是从这个总体得到的样本，$\overline{X}$ 和 S^2 分别是样本均值和样本方差，则

$$\mathrm{E}[\overline{X}] = \mu$$

$$\mathrm{Var}[\overline{X}] = \frac{\sigma^2}{n}$$

$$\mathrm{E}[S^2] = \sigma^2$$

一二阶矩的性质

矩

设总体 X 具有二阶矩，即 $\mathrm{E}[X]=\mu$，$\mathrm{Var}[X]=\sigma^2<+\infty$，$X_1,X_2,\cdots,X_n$ 是从这个总体得到的样本，$\overline{X}$ 和 S^2 分别是样本均值和样本方差，则

$$\mathrm{E}[\overline{X}]=\mu$$

$$\mathrm{Var}[\overline{X}]=\frac{\sigma^2}{n}$$

$$\mathrm{E}[S^2]=\sigma^2$$

一二阶矩的性质

矩

设总体 X 具有二阶矩，即 $\mathrm{E}[X]=\mu$，$\mathrm{Var}[X]=\sigma^2<+\infty$，$X_1,X_2,\cdots,X_n$ 是从这个总体得到的样本，$\overline{X}$ 和 S^2 分别是样本均值和样本方差，则

$$\mathrm{E}[\overline{X}]=\mu$$
$$\mathrm{Var}[\overline{X}]=\frac{\sigma^2}{n}$$
$$\mathrm{E}[S^2]=\sigma^2$$

样本方差 S^2 与样本二阶中心矩 S_n^2

$$\sum_{i=1}^{n}(X_i-\overline{X})^2=\sum_{i=1}^{n}X_i^2-n\overline{X}^2$$

$$\Longrightarrow S_n^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2$$

$$\mathrm{E}(S_n^2)=\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right)-\mathrm{E}(\overline{X}^2)$$

$$=\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right)-[\operatorname{Var}(\overline{X})+\mathrm{E}^2(\overline{X})]$$

$$=\frac{n}{n}[\operatorname{Var}X+\mathrm{E}^2(X)]-\left(\frac{1}{n}\operatorname{Var}X+\mathrm{E}^2(X)\right)=\frac{n-1}{n}\operatorname{Var}X$$

$$\Longrightarrow \mathrm{E}(S^2)=\operatorname{Var}X$$

样本方差 S^2 与样本二阶中心矩 S_n^2

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$$

$$\Longrightarrow S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

$$\mathrm{E}(S_n^2) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \mathrm{E}(\overline{X}^2)$$

$$= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - [\mathrm{Var}(\overline{X}) + \mathrm{E}^2(\overline{X})]$$

$$= \frac{n}{n}[\mathrm{Var}\, X + \mathrm{E}^2(X)] - \left(\frac{1}{n}\mathrm{Var}\, X + \mathrm{E}^2(X)\right) = \frac{n-1}{n}\mathrm{Var}\, X$$

$$\Longrightarrow \mathrm{E}(S^2) = \mathrm{Var}\, X$$

样本方差 S^2 与样本二阶中心矩 S_n^2

$$\sum_{i=1}^{n}(X_i-\overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$$

$$\Longrightarrow S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

$$\begin{aligned}
\mathrm{E}(S_n^2) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \mathrm{E}(\overline{X}^2) \\
&= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - [\mathrm{Var}(\overline{X}) + \mathrm{E}^2(\overline{X})] \\
&= \frac{n}{n}[\mathrm{Var}\, X + \mathrm{E}^2(X)] - \left(\frac{1}{n}\mathrm{Var}\, X + \mathrm{E}^2(X)\right) = \frac{n-1}{n}\mathrm{Var}\, X
\end{aligned}$$

$$\Longrightarrow \mathrm{E}(S^2) = \mathrm{Var}\, X$$

样本方差 S^2 与样本二阶中心矩 S_n^2

$$\sum_{i=1}^{n}(X_i-\overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$$

$$\Longrightarrow S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

$$\begin{aligned}
\mathrm{E}(S_n^2) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \mathrm{E}(\overline{X}^2) \\
&= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - [\mathrm{Var}(\overline{X}) + \mathrm{E}^2(\overline{X})] \\
&= \frac{n}{n}[\mathrm{Var}\, X + \mathrm{E}^2(X)] - \left(\frac{1}{n}\mathrm{Var}\, X + \mathrm{E}^2(X)\right) = \frac{n-1}{n}\mathrm{Var}\, X
\end{aligned}$$

$$\Longrightarrow \mathrm{E}(S^2) = \mathrm{Var}\, X$$

统计量的分布

概念

统计量是仅依赖于样本的随机变量，所以它有概率分布。

定义

统计量 $T_n = g(X_1, X_2, \cdots, X_n)$ 的分布称为 **抽样分布** 。

例 (样本均值的分布)

设总体 $X \sim N(\mu, \sigma^2)$， μ 和 σ^2 已知，样本为 $(X_1, \cdots, X_n)$，

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

$$\text{或} \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

即正态总体的样本均值服从正态分布，与总体相比，均值相同，方差减小为 $\frac{1}{n}$.

概念

统计量是仅依赖于样本的随机变量，所以它有概率分布。

定义

统计量 $T_n = g(X_1, X_2, \cdots, X_n)$ 的分布称为 **抽样分布** 。

例 (样本均值的分布)

设总体 $X \sim N(\mu, \sigma^2)$， μ 和 σ^2 已知，样本为 $(X_1, \cdots, X_n)$，

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

$$\text{或} \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

即正态总体的样本均值服从正态分布，与总体相比，均值相同，方差减小为 $\frac{1}{n}$.

χ^2 分布

定义

定义 设 $X_1, X_2, \cdots, X_n$ 相互独立，且都服从标准正态分布 $N(0,1)$，则

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$$

n 为自由度，即求和中独立变量的个数。

Gamma 分布特例

$\chi^2(n)$ 是 $\mathrm{Ga}\left(\frac{n}{2}, \frac{1}{2}\right)$，概率密度为

$$f(y) = \frac{\frac{1}{2}^{\frac{n}{2}}}{\Gamma(n/2)} y^{n/2-1} e^{-y/2}$$

$$\mathrm{Ga}(\alpha, \lambda) : f(x; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, x \geq 0$$

定义

定义 设 $X_1, X_2, \cdots, X_n$ 相互独立，且都服从标准正态分布 $N(0,1)$，则

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$$

n 为自由度，即求和中独立变量的个数。

Gamma 分布特例

$\chi^2(n)$ 是 $\mathrm{Ga}\left(\frac{n}{2}, \frac{1}{2}\right)$，概率密度为

$$f(y) = \frac{\frac{1}{2}^{\frac{n}{2}}}{\Gamma(n/2)} y^{n/2-1} e^{-y/2}$$

$$\mathrm{Ga}(\alpha, \lambda) : f(x; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, x \geq 0$$

证明要项

1. 我们证明过若 $X \sim N(0,1)$，则 $X^2 \sim \chi^2(1) = \mathrm{Ga}\left(\frac{1}{2}, \frac{1}{2}\right)$。
2. 伽玛分布的可加性：若 $X_1 \sim \mathrm{Ga}(\alpha_1, \lambda), X_2 \sim \mathrm{Ga}(\alpha_2, \lambda)$，且 X_1, X_2 相互独立，则 $Z = X_1 + X_2 \sim \mathrm{Ga}(\alpha_1 + \alpha_2, \lambda)$.
3. n 个服从 $\chi^2(1)$ 分布的独立随机变量之和服从 $\chi^2(n)$。

χ^2 分布的性质

1. 对于 $X \sim \chi^2(n)$，$\mathrm{E}(X) = n$, $\mathrm{Var}(X) = 2n$
2. 若 $X_1 \sim \chi^2(n_1)$, $X_2 \sim \chi^2(n_2)$, X_1, X_2 相互独立，则 $X_1 + X_2 \sim \chi^2(n_1 + n_2)$

证明要项

1. 我们证明过若 $X \sim N(0,1)$，则 $X^2 \sim \chi^2(1) = \mathrm{Ga}\left(\frac{1}{2}, \frac{1}{2}\right)$。
2. 伽玛分布的可加性：若 $X_1 \sim \mathrm{Ga}(\alpha_1, \lambda), X_2 \sim \mathrm{Ga}(\alpha_2, \lambda)$，且 X_1, X_2 相互独立，则 $Z = X_1 + X_2 \sim \mathrm{Ga}(\alpha_1 + \alpha_2, \lambda)$.
3. n 个服从 $\chi^2(1)$ 分布的独立随机变量之和服从 $\chi^2(n)$。

χ^2 分布的性质

1. 对于 $X \sim \chi^2(n)$，$\mathrm{E}(X) = n, \mathrm{Var}(X) = 2n$
2. 若 $X_1 \sim \chi^2(n_1), X_2 \sim \chi^2(n_2)$, X_1, X_2 相互独立，则 $X_1 + X_2 \sim \chi^2(n_1 + n_2)$

证明要项

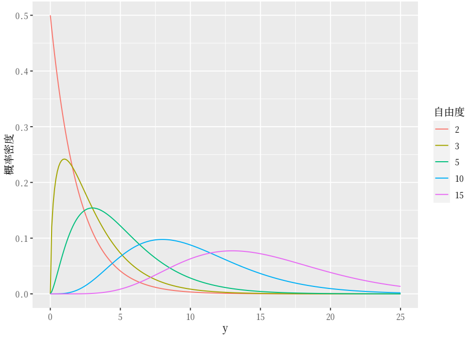
1. 我们证明过若 $X \sim N(0,1)$，则 $X^2 \sim \chi^2(1) = \mathrm{Ga}\left(\frac{1}{2}, \frac{1}{2}\right)$。
2. 伽玛分布的可加性：若 $X_1 \sim \mathrm{Ga}(\alpha_1, \lambda), X_2 \sim \mathrm{Ga}(\alpha_2, \lambda)$，且 X_1, X_2 相互独立，则 $Z = X_1 + X_2 \sim \mathrm{Ga}(\alpha_1 + \alpha_2, \lambda)$.
3. n 个服从 $\chi^2(1)$ 分布的独立随机变量之和服从 $\chi^2(n)$。

χ^2 分布的性质

1. 对于 $X \sim \chi^2(n)$，$\mathrm{E}(X) = n, \mathrm{Var}(X) = 2n$
2. 若 $X_1 \sim \chi^2(n_1), X_2 \sim \chi^2(n_2)$, X_1, X_2 相互独立，则 $X_1 + X_2 \sim \chi^2(n_1 + n_2)$

χ^2 分布举例

```r
library(ggplot2)
x <- seq(0, 25, by=0.1)
chisq_df <- data.frame(x=c(), pd=c(), dof=c())
for (dof in c(2,3,5,10,15)) {
    chisq_df <- rbind(chisq_df, data.frame(x=x, pd=dchisq(x, df=dof), dof=dof))
}
chisq_df$dof <- as.factor(chisq_df$dof)
plot <- ggplot(chisq_df, aes(x=x, y=pd, color=dof)) + geom_line()
print(plot + labs(x="y", y="概率密度", color="自由度"))
```



随着自由度增大，χ^2 分布趋近于正态分布。

上分位数

上 α 分位点 z_α

设 X 是随机变量，$0 < \alpha < 1$，称满足 $P(X > z_\alpha) = \alpha$ 的点 z_α 为 X 的上 α 分位点.

满足 $P\{\chi^2 > \chi^2_\alpha(n)\} = \int_{\chi^2_\alpha(n)}^{\infty} f_{\chi^2(n)}(y)dy = \alpha$ 的点 $\chi^2_\alpha(n)$ 称为 $\chi^2(n)$ 分布的上 α 分位数。

例

$\chi^2(n)$ 分布的上 α 分位数

$$P\{\chi^2(10) > 18.307\} = 0.05 \implies \chi^2_{0.05}(10) = 18.307$$

上分位数

上 α 分位点 z_α

设 X 是随机变量，$0 < \alpha < 1$，称满足 $P(X > z_\alpha) = \alpha$ 的点 z_α 为 X 的上 α 分位点.

满足 $P\{\chi^2 > \chi^2_\alpha(n)\} = \int_{\chi^2_\alpha(n)}^{\infty} f_{\chi^2(n)}(y)dy = \alpha$ 的点 $\chi^2_\alpha(n)$ 称为 $\chi^2(n)$ 分布的上 α 分位数。

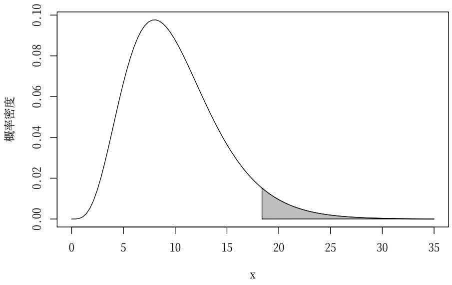
例

$\chi^2(n)$ 分布的上 α 分位数

$$P\{\chi^2(10) > 18.307\} = 0.05 \implies \chi^2_{0.05}(10) = 18.307$$

χ^2 分布的上分位数

```
x <- seq(0,35,length.out=100)
alpha.5 <- qchisq(1 - 0.05, df=10)
plot(x, dchisq(x, df=10), type='l', ylab="概率密度")
sel.x <- x[alpha.5 < x]
polygon(c(sel.x[1], sel.x), c(0, dchisq(sel.x, df=10)), col="gray")
```



正态总体样本方差的分布

设总体 $X \sim N(\mu, \sigma^2)$，σ^2 已知，样本为 $(X_1, \cdots, X_n)$，则 $\frac{n-1}{\sigma^2}$ 与样本方差 S^2 的乘积

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

F 分布

定义

设随机变量 $X \sim \chi^2(n), Y \sim \chi^2(m)$，且 X, Y 相互独立，

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

称 F 服从第一自由度为 n, 第二自由度为 m 的 F 分布.

$$f_F(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{n}{m}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1}\left(1+\frac{n}{m}x\right)^{-\frac{m+n}{2}}$$

F 分布密度函数的推导要领

1. 考察 $Z = \frac{X}{Y}$ 的概率密度函数
2. 乘以系数，$F = \frac{m}{n} Z$

定义

设随机变量 $X \sim \chi^2(n), Y \sim \chi^2(m)$，且 X, Y 相互独立，

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

称 F 服从第一自由度为 n, 第二自由度为 m 的 F 分布.

$$f_F(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{n}{m}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1}\left(1+\frac{n}{m}x\right)^{-\frac{m+n}{2}}$$

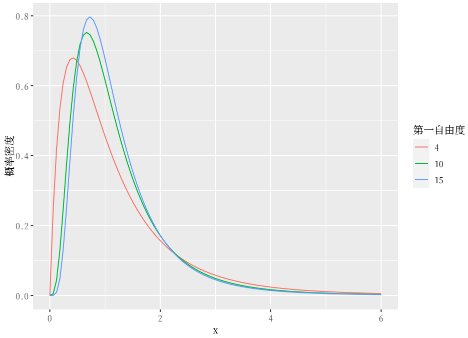
F 分布密度函数的推导要领

1. 考察 $Z = \frac{X}{Y}$ 的概率密度函数
2. 乘以系数，$F = \frac{m}{n} Z$

图示

第二自由度取 10。

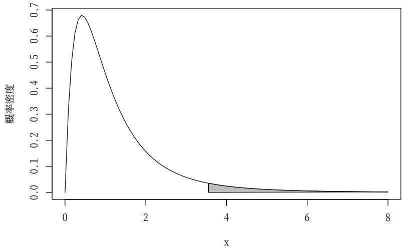
```
x <- seq(0, 6, length.out=100)
F_df <- data.frame(x=c(), pd=c(), n=c())
for (n in c(4, 10, 15)) {
    F_df <- rbind(F_df, data.frame(x=x, pd=df(x, df1=n, df2=10), n=n))
}
F_df$n <- as.factor(F_df$n)
plot <- ggplot(F_df, aes(x=x, y=pd, color=n)) + geom_line()
print(plot + labs(y="概率密度", color="第一自由度"))
```



性质

1. 若 $F \sim F(n,m)$，则 $\frac{1}{F} \sim F(m,n)$
2. $F_{1-\alpha}(n,m) = \frac{1}{F_\alpha(m,n)}$
3. 满足 $P\{F > F_\alpha(n,m)\} = \alpha$ 的 $F_\alpha(n,m)$ 称为 $F(n,m)$ 分布的上 α 分位数。

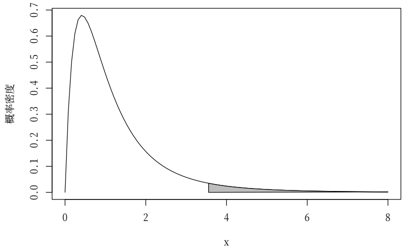
```
x <- seq(0,8,length.out=100)
alpha.5 <- qf(1 - 0.05, df1=4, df2=10)
plot(x, df(x, df1=4, df2=10), type='l', ylab="概率密度")
sel.x <- x[alpha.5 < x]
polygon(c(sel.x[1], sel.x), c(0, df(sel.x, df1=4, df2=10)), col="gray")
```



性质

1. 若 $F \sim F(n,m)$，则 $\frac{1}{F} \sim F(m,n)$
2. $F_{1-\alpha}(n,m) = \frac{1}{F_\alpha(m,n)}$
3. 满足 $P\{F > F_\alpha(n,m)\} = \alpha$ 的 $F_\alpha(n,m)$ 称为 $F(n,m)$ 分布的上 α 分位数。

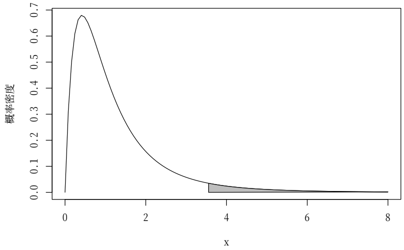
```r
x <- seq(0,8,length.out=100)
alpha.5 <- qf(1 - 0.05, df1=4, df2=10)
plot(x, df(x, df1=4, df2=10), type='l', ylab="概率密度")
sel.x <- x[alpha.5 < x]
polygon(c(sel.x[1], sel.x), c(0, df(sel.x, df1=4, df2=10)), col="gray")
```



性质

1. 若 $F \sim F(n,m)$，则 $\frac{1}{F} \sim F(m,n)$
2. $F_{1-\alpha}(n,m) = \frac{1}{F_\alpha(m,n)}$
3. 满足 $P\{F > F_\alpha(n,m)\} = \alpha$ 的 $F_\alpha(n,m)$ 称为 $F(n,m)$ 分布的上 α 分位数。

```r
x <- seq(0,8,length.out=100)
alpha.5 <- qf(1 - 0.05, df1=4, df2=10)
plot(x, df(x, df1=4, df2=10), type='l', ylab="概率密度")
sel.x <- x[alpha.5 < x]
polygon(c(sel.x[1], sel.x), c(0, df(sel.x, df1=4, df2=10)), col="gray")
```



两个正态总体比较

又设总体 $X' \sim N(\mu', \sigma'^2)$，样本为 $(X_1', \cdots, X_n')$，且样本 $(X_1', \cdots, X_n')$ 与 $(X_1, \cdots, X_n)$ 相互独立。则

$$\frac{S^2/S'^2}{\sigma^2/\sigma'^2} \sim F(n-1, n'-1)$$

t 分布

定义

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

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

为服从自由度为 n 的 t 分布，即（Student 分布，学生氏分布）.

$$f_T(t) = t f_{T^2}(t^2) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}, t \in (-\infty, \infty)$$

推导要领

1. $T^2 = \frac{X^2/1}{Y/n}$，所以 $T^2 \sim F(1, n)$
2. $T = \sqrt{T^2}$ 推得 t 分布的概率密度 $f_T(t)$

定义

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

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

为服从自由度为 n 的 t 分布，即（Student 分布，学生氏分布）.

$$f_T(t) = t f_{T^2}(t^2) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}, t \in (-\infty, \infty)$$

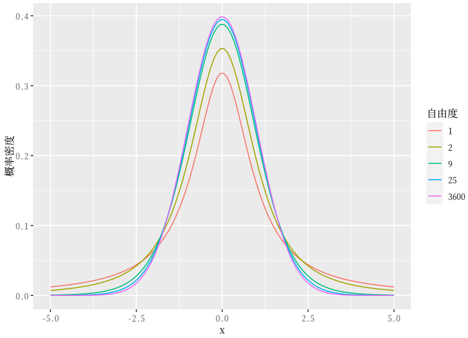
推导要领

1. $T^2 = \frac{X^2/1}{Y/n}$，所以 $T^2 \sim F(1,n)$
2. $T = \sqrt{T^2}$ 推得 t 分布的概率密度 $f_T(t)$

图示

$n \to \infty$ 时为标准正态分布。

```r
x <- seq(-5, 5, length.out=100)
t_df <- data.frame(x=c(), pd=c(), n=c())
for (n in c(1, 2, 9, 25, 3600)) {
    t_df <- rbind(t_df, data.frame(x=x, pd=dt(x, df=n), n=n))
}
t_df$n <- as.factor(t_df$n)
plot <- ggplot(t_df, aes(x=x, y=pd, color=n)) + geom_line()
print(plot + labs(y="概率密度", color="自由度"))
```



t 分布的性质

1. $f_n(t)$ 是偶函数

$$f_n(t) \xrightarrow{n\to\infty} \phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

2. 满足 $P(T > t_\alpha) = \alpha$ 的点 t_α 称为 t 分布的上 α 分位数.
3. $P(T > t_\alpha) = \alpha \implies -t_\alpha = t_{1-\alpha}$
4. 满足 $P(|T| > t_{\alpha/2}) = \alpha$ 的点 $t_{\alpha/2}$ 称为 t 分布的双侧 α 分位数.
5. $P(T > t_{\alpha/2}) = \alpha/2 \implies P(|T| > t_{\alpha/2}) = \alpha$

正态总体样本均值与方差取商的分布

设总体 $X \sim N(\mu, \sigma^2)$，μ 已知，样本为 $(X_1, \cdots, X_n)$，

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

t 分布的性质

1. $f_n(t)$ 是偶函数

$$f_n(t) \xrightarrow{n\to\infty} \phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

2. 满足 $P(T > t_\alpha) = \alpha$ 的点 t_α 称为 t 分布的上 α 分位数.
3. $P(T > t_\alpha) = \alpha \implies -t_\alpha = t_{1-\alpha}$
4. 满足 $P(|T| > t_{\alpha/2}) = \alpha$ 的点 $t_{\alpha/2}$ 称为 t 分布的双侧 α 分位数.
5. $P(T > t_{\alpha/2}) = \alpha/2 \implies P(|T| > t_{\alpha/2}) = \alpha$

正态总体样本均值与方差取商的分布

设总体 $X \sim N(\mu, \sigma^2)$，μ 已知，样本为 $(X_1, \cdots, X_n)$，

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

t 分布的性质

1. $f_n(t)$ 是偶函数

$$f_n(t) \xrightarrow{n\to\infty} \phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

2. 满足 $P(T > t_\alpha) = \alpha$ 的点 t_α 称为 t 分布的上 α 分位数.
3. $P(T > t_\alpha) = \alpha \implies -t_\alpha = t_{1-\alpha}$
4. 满足 $P(|T| > t_{\alpha/2}) = \alpha$ 的点 $t_{\alpha/2}$ 称为 t 分布的双侧 α 分位数.
5. $P(T > t_{\alpha/2}) = \alpha/2 \implies P(|T| > t_{\alpha/2}) = \alpha$

正态总体样本均值与方差取商的分布

设总体 $X \sim N(\mu, \sigma^2)$，μ 已知，样本为 $(X_1, \cdots, X_n)$，

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

t 分布的性质

1. $f_n(t)$ 是偶函数

$$f_n(t) \xrightarrow{n\to\infty} \phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

2. 满足 $P(T > t_\alpha) = \alpha$ 的点 t_α 称为 t 分布的上 α 分位数.
3. $P(T > t_\alpha) = \alpha \implies -t_\alpha = t_{1-\alpha}$
4. 满足 $P(|T| > t_{\alpha/2}) = \alpha$ 的点 $t_{\alpha/2}$ 称为 t 分布的双侧 α 分位数.
5. $P(T > t_{\alpha/2}) = \alpha/2 \implies P(|T| > t_{\alpha/2}) = \alpha$

正态总体样本均值与方差取商的分布

设总体 $X \sim N(\mu, \sigma^2)$，μ 已知，样本为 $(X_1, \cdots, X_n)$，

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