北京航空航天大學
经济管理学院
BEIHANG UNIVERSITY
SCHOOL OF ECONOMICS AND MANAGEMENT

第一讲：样本及抽样分布

康雁飞

数量经济与商务统计系

Outline

Outline

1 概率论与数理统计的区别

2 总体与样本

3 统计量及常用统计量

4 三大抽样分布

5 小结

6 作业

- 概率论：研究随机现象
  - 随机变量及其概率分布全面地描述了随机现象的统计规律性。
  - 在概率论中，通常假定概率分布是已知的，一切计算及其推理均基于这个已知的分布进行。
- 然而，当我们研究并解决实际问题时，情况往往并非如此。

举例

- 某公司要采购一批产品，每件产品不是合格品就是不合格品，但该批产品总有一个不合格品率 p。
- 若从该批产品中随机抽取一件，用 X 表示这一件产品是否合格（不合格记为 $X=1$；合格记为 $X=0$），则 X 服从两点分布 $B(1,p)$，但 p 未知。
- p 的大小决定了该批产品的质量，直接影响采购行为的经济效益。

问题

1. p 的大小如何?
2. p 大概落在什么范围内?
3. 能否认为 p 满足设定要求（如不超过 0.05)?

举例

- 假设一家工厂生产某种电子产品，该产品的重量是关键参数之一。
- 假设产品重量 $X \sim N(\mu, \sigma^2)$。
- 随机抽取 100 个产品，这些产品是在同样的生产条件下生产的。
- 工厂想要利用这些数据来估计生产的平均重量，如何估计？范围如何？能否认为满足工厂生成要求？

数理统计学

- 一门以数据为基础的学科，有很强的应用性：以概率论为理论基础，根据试验或观察得到的数据，来研究随机现象，对研究对象的客观规律性作出合理的估计和判断。
- 任务
  1. 如何获得样本?
  2. 利用样本，从而对事物的某些未知方面进行分析、推断并作出一定的决策。

数理统计学

- **重点**：统计推断（1-9 周）
  1. 样本及抽样分布（第五章）
  2. 参数估计（第六章）：点估计、区间估计、估计的优良性质等
  3. 假设检验（第七章）：基本思想、单（双）正态总体参数假设检验、其他分布参数的假设检验、拟合优度检验等
  4. 方差分析（第八章）：多组总体位置参数的假设检验问题
- **学习要求**：熟悉掌握数理统计的基本理论与思想，并掌握常用的包括点估计、区间估计和假设检验等基本统计推断方法。

Outline

总体与个体

“样本推断总体”是数理统计学的显著特征

- 若规定灯泡寿命低于 1000 小时者为次品，如何确定次品率?
- 由于灯泡寿命试验是破坏性试验，不可能把整批灯泡逐一检测，只能抽取一部分灯泡作为样本进行检验，通过这部分灯泡的寿命数据来推断整批灯泡的次品率。
- 以部分样本的信息来推断总体的信息，就是数理统计学研究的问题之一。

总体与个体

- 总体（Population）：研究对象的全体（本质是一个分布，其数量指标为服从该分布的随机变量）
- 个体（Individual）：总体中的成员
- 总体的容量：总体中包含的个体数
- 有限总体：容量有限的总体
- 无限总体：容量无限的总体，通常将容量非常大的有限总体按无限总体处理

总体与个体：举例

我们要考察北航一年级 2000 名男生的身高情况：

- 2000 人构成该问题的总体，是一个有限总体；而每一个学生即为一个个体。
- 事实上每个学生有许多特征：年龄、身高、体重、籍贯等，而该问题中我们只关心学生的身高（指标 X）如何，其他指标不关心。因此，每个学生的身高数据就是个体，而所有身高全体即为总体。
- 对于不同的学生有不同的身高数据，所有的取值（2000 个学生的身高数据）构成一个分布 $F(x)$，因此 X 可以看成一个随机变量，也称 X 为总体, 服从分布 $F(x)$。

总体与个体：举例

我们想了解北京的空气质量情况，因此关注每天的 PM2.5 值，此时：

- 总体是北京上空一定范围内的空气的 PM2.5 值，可能的取值无限多；
- 这是一个无限总体；
- 个体即为北京某一可能的 PM2.5 取值。

总体与个体：举例

考察全国正在使用的某型号灯泡的寿命时:

- 总体是全国正在使用的该型号灯泡的寿命;
- 可能的观察值的个数很多，可认为无限总体;
- 个体为该型号每个灯泡的寿命。

统计推断的意义和问题

- 磁带的一个质量指标是一卷磁带（20m）上的伤痕数。每卷磁带都有一个伤痕数，全部磁带的伤痕数构成一个总体（无限总体）。其中相当一部分是 0，也有 1、2、3 等，但多于 8 个的伤痕数少见。
- 研究表明：一卷磁带上的伤痕数 X 服从泊松分布 $P(\lambda)$，但分布的参数 λ 却未知。显然，λ 的大小决定了一批产品的质量，直接影响生产方的经济效益。总体参数：描述总体分布特征。
- 这里的总体分布类型明确，但总体含有未知参数 λ，因此总体不是一个特定的泊松分布。
- 数理统计的任务：确定 λ，即确定最终的总体分布。

样本

- 抽样（Sampling）：从总体 X 中抽取有限个个体对总体进行观察的取值过程；
- 随机样本（Random Sample）：从总体 X 中随机抽取 n 个个体，记为
$$X_1, X_2, \cdots, X_n$$
称为总体的一个样本容量（样本量，Sample size）为 n 的随机样本；
- 样本观测值（Observations）：$x_1, x_2, \cdots, x_n$。

样本

1. 样本具有随机性：由于样本是从总体中随机抽取的，抽取前无法预知它们的数值。因此，样本是随机变量，用大写字母 X 表示。
2. 样本观测值是一次抽样的具体实现：样本在抽取后，经观测就有确定的观测值。因此，样本又是一组数值，此时用小写字母 x 表示。

思考：为什么要抽样?

简单随机样本

- 从总体中抽取样本有不同的抽样方法，为了能由样本对总体作出较可靠的推断，就希望样本能很好地代表总体。
- 这需要对抽样方法提出一些要求，最常用的是“简单随机抽样”。

简单随机样本

简单随机抽样（Simple Random Sampling）

- 样本具有随机性（代表性）：总体中每一个个体都有同等机会抽中，即每一样品 X_i 与总体 X 同分布（identically distributed）。
- 样本要有独立性：样本中每一样品的取值不影响其他样品的取值，即 $X_1, X_2, \cdots, X_n$ 相互独立（independent）。

简单随机样本

- 简单随机样本（independently identically distributed samples）：用简单随机抽样方法得到的样本，也称为 i.i.d. 样本。
- 多数统计推断都是以简单随机样本为基础的。

简单随机样本

设总体 X 具有分布函数 $F(x), X_1, X_2, \cdots, X_n$ 为取自该总体的容量为 n 的样本, 则

- 样本联合分布函数:

$$F\left(x_1, \ldots, x_n\right)=\prod_{i=1}^n F\left(x_i\right).$$

- 样本的联合密度函数:

$$f\left(x_1, x_2, \cdots, x_n\right)=\prod_{i=1}^n f\left(x_i\right).$$

如何得到简单随机样本呢?

- 对于有限总体，通常采用有放回简单随机抽样（Simple Random Sampling with replacement)。
- 但当总体容量很大的时候, 有放回抽样有时很不方便, 因此在实际中当总体容量比较大时通常将不放回抽样（Sampling without replacement）所得到的样本近似当做简单随机样本来处理。
- 对于无限总体, 一般采用不放回抽样。

简单随机样本：R 实现

以下是一个咱们班级学生的名单。

```r
library(readxl)
students <- read_excel("data/L1/2024-应用统计学-名单.xlsx", skip = 1)
students
```

```
## # A tibble: 46 x 3
##     学号     姓名                 账号
##     <chr>    <chr>                <chr>
##   1 23080103 陈祉安               23080103
##   2 22377337 麦克然江·买买提       22377337
##   3 22377336 安凯尔江·艾克拜尔     22377336
##   4 22377335 郑昊哲               22377335
```

简单随机样本：R 实现

- 采用简单随机抽样抽出 3 个学生组成一个随机样本：sample()。

```r
sample(students$姓名, 3, replace = TRUE)
```

```
## [1] "李鑫"   "邓磊"   "袁一心"
```

练习

对该数据，利用 R 完成：

1. 分别有放回和无放回抽取 10 名学生的姓名组成一个随机样本，输出学生姓名。
2. 分别有放回和无放回抽取 10 名学生的姓名和学号组成一个随机样本，同时输出学生姓名和学号。

总体与样本

- 样本来自总体，因此样本中含有总体各方面的信息，但是这些信息较为分散，有时显得杂乱无章。
- 思考：如何将样本中有关总体的信息集中起来反映总体的各种特征呢?

Outline

统计量与抽样分布

设 $X_1, X_2, \cdots, X_n$ 为取自某总体的样本, 若样本函数 $T = T(X_1, X_2, \cdots, X_n)$ 中不含有任何未知参数，则称 T 为统计量 (statistic)。统计量的分布称为抽样分布。

统计量与抽样分布

- 统计量是随机变量的函数，也是一个随机变量（$\sum_{i=1}^{n} X_i$ 和 $\sum_{i=1}^{n} X_i^2$ 都是统计量）。
- 若 $x_1, x_2, \cdots, x_n$ 为相当于 $X_1, X_2, \cdots, X_n$ 的观测值，则称 $T(x_1, x_2, \cdots, x_n)$ 为 $T(X_1, X_2, \cdots, X_n)$ 的观测值。
- 尽管统计量不依赖于未知参数，但是其抽样分布一般还是依赖于未知参数的。

思考

假设总体 $X \sim N(\mu, \sigma^2)$，请问

$$g(X_1, X_2, \cdots, X_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

是否为统计量?

常用统计量

- 样本均值

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- 样本方差

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

- 样本标准差

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

- 样本矩（k 阶原点矩和 k 阶中心矩）

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

常用统计量：样本偏度

- 当总体关于分布中心对称时，用样本均值和标准差刻画样本特征很有代表性，但不对称时，就显得不够。
- 样本偏度（反映样本数据与对称性的偏离程度和偏离方向）

$$\hat{\beta}_S = \frac{B_3}{B_2^{3/2}}.$$

- 如果数据完全对称（例：4，7，8，9，12），$B_3 = 0$。
- 如果 $\hat{\beta}_S$ 明显大于 0，表示样本的右尾长，说明数据中有几个较大的数，这时反映总体分布右偏。
- 反之（例：1，4，7，8，9），说明总体分布左偏。

常用统计量：样本峰度

- 样本峰度

$$\hat{\beta}_K = \frac{B_4}{B_2^2} - 3.$$

- $\hat{\beta}_K$ 反映的是总体分布密度曲线在其峰值附近的陡峭程度和尾部粗细。
- $\hat{\beta}_K > 0$ 说明总体分布在峰值附近比正态分布陡峭，尾部更细，尖顶；反之，平顶。
- 样本偏度和样本峰度的定义和计算在不同软件中有少许不同。

样本均值的性质

- 样本均值与样本所有数据的偏差之和为 0，即 $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$。
- 数据观测值与均值的偏差平方和最小，即 $\bar{x} = \text{argmin}_c \sum (x_i - c)^2$。
- 为什么上述成立?

样本均值 $\bar{X}$ 的抽样分布

1. 若总体分布为 $N(\mu, \sigma^2)$，则 $\bar{X}$ 的精确分布为 $N(\mu, \sigma^2/n)$。
2. 若总体分布未知或不是正态分布，但 $E(X) = \mu, Var(X) = \sigma^2$，则由中心极限定理，$n$ 较大时 $\bar{X} \sim AN\left(\mu, \sigma^2/n\right)$（asymptotically normal）。

练习

利用 R 验证总体分别为正态分布、均匀分布、指数分布时样本均值的抽样分布。

请查看这里并思考。

其它性质

- 设总体 X 具有二阶矩，即 $E(X)=\mu$, $Var(X)=\sigma^2$ 存在，则样本均值 $\bar{X}$ 和样本方差 S^2 满足:
$$E(\bar{X})=\mu,\ Var(\bar{X})=\sigma^2/n,\ E(S^2)=\sigma^2.$$
- 无论总体的分布形式，样本均值的期望和总体均值相同，其方差是总体方差的 $1/n$（样本容量越大，样本均值的方差越小)。
- 样本方差的期望等于总体方差。

次序统计量及其分布

设 $X_1, X_2, \cdots, X_n$ 是取自总体 X 的样本，$X_{(i)}$ 称为该样本的第 i 个次序统计量，它的取值是将样本观测值由小到大排列后得到的第 i 个观测值。其中

- $X_{(1)} = \min\{X_1, X_2, \cdots, X_n\}$ 称为该样本的**最小次序统计量**。
- $X_{(n)} = \max\{X_1, X_2, \cdots, X_n\}$ 为该样本的**最大次序统计量**。
- 在一个样本中，$X_1, X_2, \cdots, X_n$ 是独立同分布的，而次序统计量 $\left(X_{(1)}, X_{(2)}, \cdots, X_{(n)}\right)$ 则既不独立, 分布也不相同。

连续总体次序统计量的分布

设总体 X 的密度函数为 $p(x)$, 分布函数为 $F(x)$, $X_1, X_2, \cdots, X_n$ 为样本, 则第 k 个次序统计量 $X_{(k)}$ 的密度函数为

$$p_k(x) = \frac{n!}{(k-1)!(n-k)!}(F(x))^{k-1}(1-F(x))^{n-k}p(x).$$

连续总体次序统计量的分布

- 我们可以先考虑 $X_{(k)}$ 的累积分布函数 $F_{(k)}(x)$。根据定义，$F_{(k)}(x)$ 表示有 k 个随机变量小于等于 x，也就是：

$$F_{(k)}(x) = P(X_{(k)} \leq x) = \sum_{i=k}^{n} C_n^i F(x)^i (1 - F(x))^{n-i}$$

- 我们可以选择任意 i 个随机变量小于等于 x，而剩下的 $n - i$ 个随机变量必须大于 x。这样的选择方案共有 C_n^i 种，每种方案的概率为 $F(x)^i(1 - F(x))^{n-i}$。因此，$F_{(k)}(x)$ 就是所有选出 k 个随机变量的方案的概率之和。

连续总体次序统计量的分布

将分布函数求导

$$
\begin{aligned}
p_k(x) &= \sum_{i=k}^{n} iC_n^i F(x)^{i-1}(1-F(x))^{n-i}p(x) - \sum_{i=k}^{n}(n-i)C_n^i F(x)^i(1-F(x))^{n-i-1}p(x) \\
&= \sum_{i=k}^{n} \frac{n!}{(i-1)!(n-i)!}F(x)^{i-1}(1-F(x))^{n-i}p(x) - \sum_{i=k}^{n} \frac{n!}{i!(n-i-1)!}F(x)^{i}(1-F(x))^{n-i-1}p(x) \\
&= \sum_{i=k}^{n} \frac{n!}{(i-1)!(n-i)!}F(x)^{i-1}(1-F(x))^{n-i}p(x) - \sum_{i=k+1}^{n} \frac{n!}{(i-1)!(n-i)!}F(x)^{i-1}(1-F(x))^{n-i}p(x) \\
&= \frac{n!}{(k-1)!(n-k)!}(F(x))^{k-1}(1-F(x))^{n-k}p(x).
\end{aligned}
$$

最大和最小次序统计量

1. 请写出最大和最小次序统计量的密度函数。
2. 对均匀分布 $U[0,1]$ 的样本，写出次序统计量 $X_{(k)}$ 的密度函数，并说明它是 $Beta(k, n-k+1)$。
3. 请查看这里并思考。

经验分布函数

- 经验分布函数：总体分布函数 F $(\boldsymbol{x})$ 相应的统计量

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\left(X_i \le x\right).$$

- 利用次序统计量, 将 $F_n(x)$ 重新表示为:

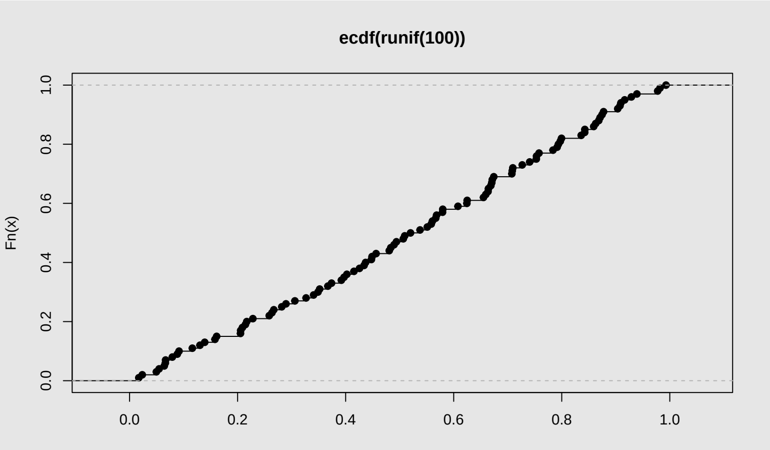
$$F_n(x) = \begin{cases} 0, & x < x_{(1)} \\ \frac{k}{n}, & x_{(k)} \le x < x_{(k+1)}, k = 1, 2, \cdots, n-1 \\ 1, & x \ge x_{(n)} \end{cases}$$

- 格里汶科 (Glivebko) 定理：对于任一实数 x，当 $n \to \infty$ 时 $F_n(x)$ 以概率一致收敛于分布函数 $F(x)$, 即

$$P\left(\lim_{n\to\infty} \sup_x \left|F_n(x) - F(x)\right| = 0\right) = 1.$$

经验分布函数：R 实现

```
plot(ecdf(runif(100)))
```



Outline

三大抽样分布

很多统计推断基于正态分布假设，以标准正态变量为基石构造的三个著名统计量被广泛应用，它们有明确背景和明确密度函数表达式，在统计中常被称为”三大抽样分布”。

1. χ^2 分布 (卡方分布)
2. F 分布
3. t 分布

χ^2 分布 (卡方分布)

χ^2 分布

设 $X_1, X_2, \ldots, X_n$ 独立同分布于标准正态分布 $N(0, 1)$, 则 $\chi^2 = X_1^2 + \ldots + X_n^2$ 的分布称为自由度为 n 的 χ^2 分布, 记为 $\chi^2 \sim \chi^2(n)$.

χ^2 分布

- χ^2 分布的密度函数:
$$p(y) = \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{n/2}} y^{n/2-1} \mathrm{e}^{-\frac{y}{2}} \quad (y > 0).$$
其中伽玛函数: $\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx.$
- $\chi^2(n)$ 是特殊的伽玛分布 $Ga(n/2, 1/2)$。
- 若 $\chi^2 \sim \chi^2(n)$, 则 $E\left(\chi^2\right) = n, Var\left(\chi^2\right) = 2n$。

χ^2 分布的期望

$$E(X) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\int_0^{\infty} y^{\frac{n}{2}} e^{-\frac{y}{2}} dy.$$

做变量代换 $z = \frac{y}{2}$，得

$$E(X) = \frac{2}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\int_0^{\infty} (2z)^{\frac{n}{2}} e^{-z} dz.$$

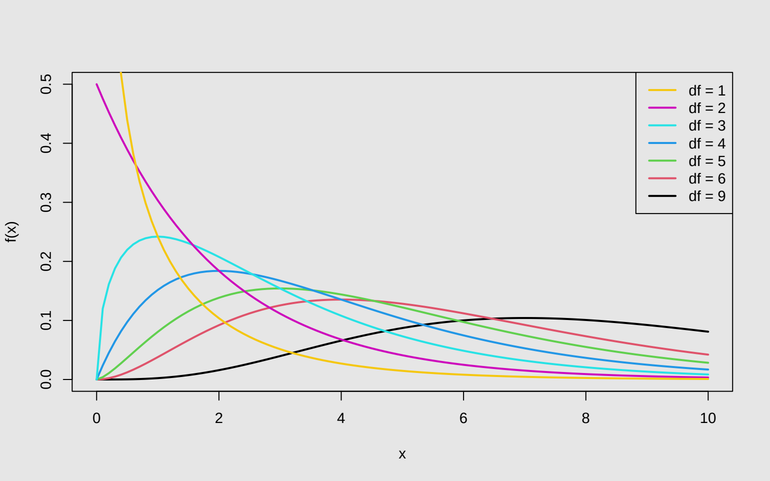
化简可得

$$E(X) = \frac{2^{\frac{n}{2}+1}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\int_0^{\infty} z^{\frac{n}{2}} e^{-z} dz.$$

根据 $\Gamma(\frac{n}{2}+1) = \frac{n}{2}\Gamma(\frac{n}{2})$，有

$$E(X) = n.$$

χ^2 分布



χ^2 分布

- χ^2 分布具有可加性，如果 $\chi_1^2 \sim \chi^2(n_1)$，$\chi_2^2 \sim \chi^2(n_2)$，且二者互相独立，那么
$$\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2).$$
- 当随机变量 $\chi^2 \sim \chi^2(n)$ 时, 对给定 $\alpha(0 < \alpha < 1)$, 称满足 $P\left(\chi^2 \le \chi_{1-\alpha}^2(n)\right) = 1 - \alpha$ 的 $\chi_{1-\alpha}^2(n)$ 是自由度为 n 的卡方分布的 $1 - \alpha$ 分位数。
- 在 R 中，可通过 `qchisq(p，df)` 来求得。

χ^2 分布的分位数

```
qchisq(0.95, 3)
```

```
## [1] 7.814728
```

例题

设 $X_1, X_2, \cdots, X_n$ 是来自 $N\left(\mu, \sigma^2\right)$ 的样本，其中 μ 是已知常数，求统计量 $T = \sum_{i=1}^{n} (X_i - \mu)^2$ 的分布。

思路:

1. $T/\sigma^2 \sim \chi^2(n)$.
2. 得到 T 的密度函数:
$$p(t) = \frac{1}{(2\sigma^2)^{n/2}\,\Gamma(n/2)} \mathrm{e}^{-\frac{t}{2\sigma^2}} t^{\frac{n}{2}-1}, \quad t > 0.$$
3. $T \sim Ga(\frac{n}{2}, \frac{1}{2\sigma^2})$.

χ^2 分布的应用

假设 $X_1, X_2, \cdots, X_n$ 是来自 $N\left(\mu, \sigma^2\right)$ 的样本，其样本均值和样本方差分别为

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

则以下成立：

1. $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
2. $\bar{X}$ 与 S^2 独立，且

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

证明

构造

$$
\begin{aligned}
W = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 &= \sum_{i=1}^{n} \left(\frac{(X_i - \bar{X}) + (\bar{X} - \mu)}{\sigma} \right)^2 \\
&= \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} + \frac{n(\bar{X} - \mu)^2}{\sigma^2} \\
&= \frac{(n-1)S^2}{\sigma^2} + \frac{n(\bar{X} - \mu)^2}{\sigma^2}.
\end{aligned}
$$

因此，

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

t 分布

t 分布

设随机变量 X_1 与 X_2 独立, 且 $X_1 \sim N(0,1), X_2 \sim \chi^2(n)$, 则称 $t = \dfrac{X_1}{\sqrt{X_2/n}}$ 的分布为自由度为 n 的 t 分布, 记为 $t \sim t(n)$。

t 分布

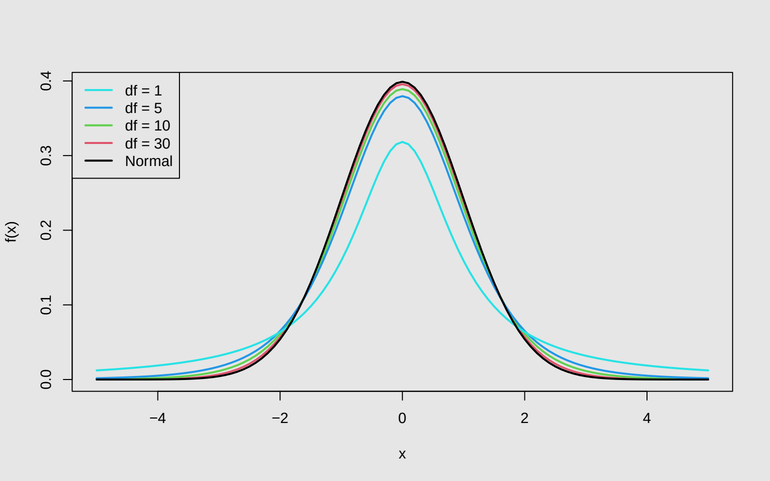
- t 分布的密度函数:

$$p(y) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{y^2}{n}\right)^{-\frac{n+1}{2}}$$

- 若 $t \sim t(n)$, 则
  1. $E(t) = 0, \quad n > 1$
  2. $Var(t) = \frac{n}{n-2}, \quad n > 2$

思考: t 分布和正态分布的区别和联系?

t 分布



t 分布的分位数

1. 当随机变量 $t \sim t(n)$ 时，称满足 $P(t \le t_{1-\alpha}(n)) = 1-\alpha$ 的 $t_{1-\alpha}(n)$ 是自由度为 n 的 t 分布的 $1-\alpha$ 分位数。
2. 分位数 $t_{1-\alpha}(n)$ 可以从附表 4 中查到。
3. 譬如 $n = 10, \alpha = 0.05$，那么从附表 4 上查得 $t_{1-0.05}(10) = t_{0.95}(10) = 1.812$。
4. 由于 t 分布的密度函数关于 0 对称，故其分位数间有如下关系 $t_\alpha(n) = -t_{1-\alpha}(n)$.

t 分布的分位数

```r
qt(0.95, 10)
```

```
## [1] 1.812461
```

t 分布的应用

假设 $X_1, X_2, \cdots, X_n$ 是来自 $N\left(\mu, \sigma^2\right)$ 的样本，其样本均值和样本方差分别为

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

则以下成立：

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

F 分布

设 $X_1 \sim \chi^2(m), X_2 \sim \chi^2(n)$, X_1 与 X_2 独立, 则称 $F = (X_1/m)/(X_2/n)$ 的分布是自由度为 m 与 n 的 F 分布, 记为 $F \sim F(m, n)$, 其中

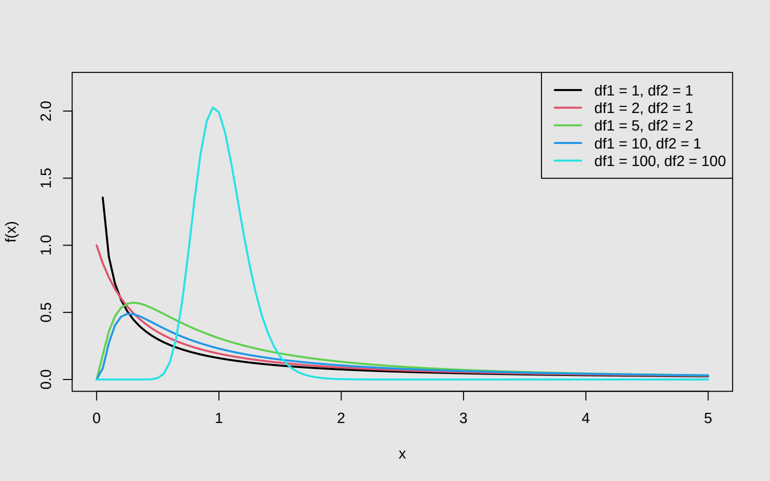
1. m 称为分子自由度
2. n 称为分母自由度

F 分布

F 分布的密度函数：

$$p(y)=\frac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{m}{n}\right)^{m/2}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}y^{\frac{m}{2}-1}\left(1+\frac{m}{n}y\right)^{-\frac{m+n}{2}}$$

F 分布



df1 = 1, df2 = 1
df1 = 2, df2 = 1
df1 = 5, df2 = 2
df1 = 10, df2 = 1
df1 = 100, df2 = 100
f(x)
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5
x

F 分布的分位数

当随机变量 $F \sim F(m, n)$ 时, 对给定 $\alpha(0 < \alpha < 1)$, 称满足 $P(F \leq F_{1-\alpha}(m, n)) = 1 - \alpha$ 的 $F_{1-\alpha}(m, n)$ 是自由度为 m 与 n 的 F 分布的 $1 - \alpha$ 分位数 (附表 5)。

F 分布的性质

1. 若 $F \sim F(m, n)$, 则 $1/F \sim F(n, m)$。
2. $F_\alpha(n, m) = \frac{1}{F_{1-\alpha}(m,n)}$。

F 分布的分位数

```r
qf(0.95, 5, 5)
```

```
## [1] 5.050329
```

F 分布的应用

假设 $X_1, X_2, \cdots, X_m$ 是来自 $N\left(\mu_1, \sigma_1^2\right)$ 的样本, $Y_1, Y_2, \cdots, Y_n$ 是来自 $N\left(\mu_2, \sigma_2^2\right)$ 的样本, 且此两样本相互独立，记

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i, \quad S_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(X_i - \bar{X}\right)^2$$
$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

则以下成立:

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

总结：四大分布之间的关系

假设 X_1 和 X_2 互相独立，则

1. $X_1 \sim N\left(\mu, \sigma^2\right), X_2 \sim N\left(\mu, \sigma^2\right) \to X_1 + X_2 \to N\left(2\mu, 2\sigma^2\right)$
2. $X \sim N(0,1) \to X^2 \sim \chi^2(1)$
3. $X_1 \sim \chi^2\left(n_1\right), X_2 \sim \chi^2\left(n_2\right) \to X_1 + X_2 \sim \chi^2\left(n_1 + n_2\right)$
4. $X_1 \sim N(0,1), X_2 \sim \chi^2(n) \to \frac{X_1}{\sqrt{X_2/n}} \sim t(n)$
5. $X_1 \sim \chi^2\left(n_1\right), X_2 \sim \chi^2\left(n_2\right) \quad \to \quad \frac{X_1/n_1}{X_2/n_2} \sim F\left(n_1, n_2\right)$

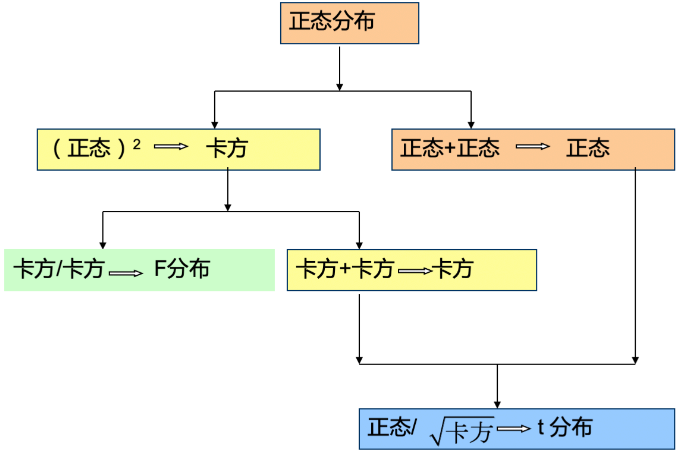
练习

假设 X_1, X_2, X_3, X_4, X_5 为来自总体 $X \sim N\left(0, \sigma^2\right)$ 的一个简单随机样本，记

$$Y_1 = a\frac{X_1 - X_2}{\sqrt{X_3^2 + X_4^2 + X_5^2}}, Y_2 = b\frac{(X_1 - X_2 + X_3)^2}{X_4^2 + X_5^2}$$

那么 a, b 取何值时，Y_1 服从 t 分布，$\boldsymbol{Y}_2$ 服从 F 分布?

正态分布族谱



Outline

1. 概率论与数理统计的区别
2. 总体与样本
3. 统计量及常用统计量
4. 三大抽样分布
5. 小结
6. 作业

小结

- 了解数理统计的基本内容、基本概念（总体、样本、抽样、简单随机样本、参数、统计量等）
- 熟悉掌握常用的统计量的定义、计算、性质及其抽样分布
- 理解并熟悉掌握三大抽样分布（χ^2 分布、t 分布、F 分布）的定义、性质、分布形状及其之间的关系
- 理解并熟悉掌握正态总体样本均值和样本方差的抽样分布
- 初步了解 R，能够运用 R 进行模拟验证及其正态分布、χ^2 分布、t 分布、F 分布的分位数

Outline

作业

1. 总体与样本（习题 5.1)：1-3、5
2. 统计量及其分布（习题 5.3）1、4-5、9-10、15-18；选做：23、25、28、34
3. 三大抽样分布（习题 5.4)：8-11、19；选做：21-22

请于 SPOC 提交。