统计学与 R 语言

第 13 讲　概率分布

张敬信

2022 年 4 月 12 日

哈尔滨商业大学

一. 概率分布概念

随机变量：表示随机现象一组结果的变量：

- 抛 10 次硬币，出现正面的次数（离散）
- 调查了 100 个人的身高数据是随机变量身高的数据（连续，取值实轴某区间）

概率分布：用数学语言表示的某类随机现象的规律性，同一种概率分布，也不是都相同，这是由不同参数值决定和区分的

- 抛 10 次硬币出现正面的次数，表现出二项分布规律性，参数是出现正面的概率 p
- 100 人的身高可能对称地分布在 175cm 附近，离得越远人数越少，表现出一种正态分布规律性，参数是均值 μ 和标准差 σ

1. 离散分布

随机变量取值有限或可列，表示其分布规律：

(1) 列出离散型随机变量 X 的所有可能取值

(2) 列出随机变量取这些值的概率

表格表示：

$X = x_i$	x_1	x_2	$\cdots$	x_n
$P(X = x_i) = p_i$	p_1	p_2	$\cdots$	p_n

函数表示：

$$P(X = x_i) = p_i$$

其中，$p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1.$

这也相当于是**概率密度函数**，累积起来就是**概率分布函数**：

$$F(x) = P(X < x) = \sum_{x_i \leq x} p_i$$

表示取值不超过 x 的累积概率。

注：累积概率，累积完全部取值之后，都等于 $1.$

2. 连续分布

- 取值是某一区间或整个实数轴上的任意一个值
- 它取任何一个特定的值的概率都等于 0
- 不能列出每一个值及其相应的概率，只能考虑它取某一区间值的概率
- 用概率密度函数和概率分布函数来描述

连续情形，单个点的概率是 0, 考虑包含点的任意小区间段的概率才有意义，也就是**概率密度函数**:

$$f(x) = \lim_{\delta \to 0} \frac{P(X \in [x-\delta, x+\delta])}{2\delta}$$

也相当于是随机变量 X 取值为 x 的概率，只是换了一种有意义的定义方式。

概率分布函数，仍是取值不超过 x 的累积概率，就需要积分了:

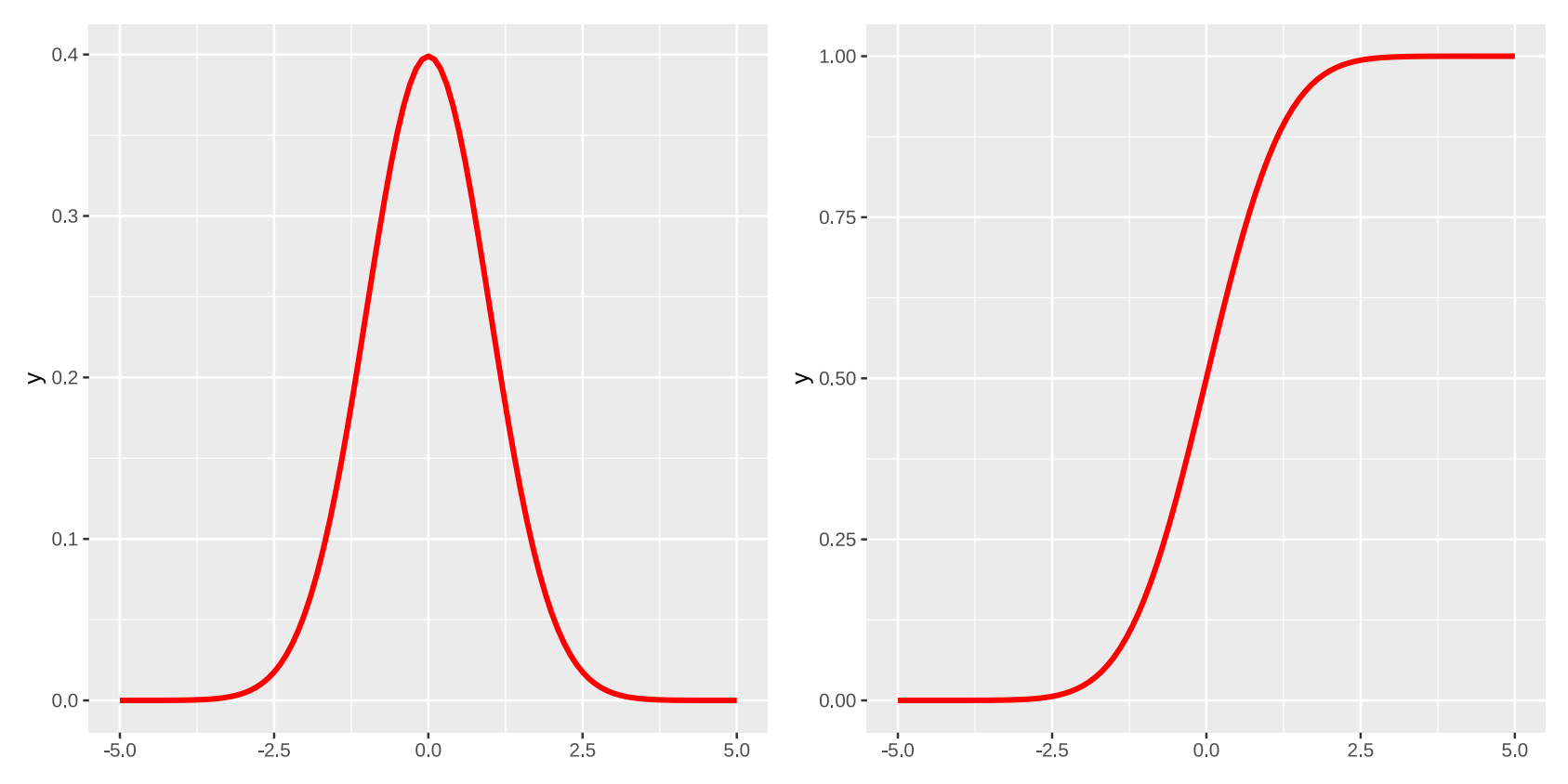
$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\mathrm{dt}$$

注 1：累积概率，累积完全部取值之后，都等于 1.

注 2：概率分布函数也叫累积分布函数，满足 $F'(x) = f(x)$.

```r
library(patchwork)
p1 = ggplot() +
  geom_function(fun = dnorm, color = "red", size = 1.2) +
  xlim(-5, 5)
p2 = ggplot() +
  geom_function(fun = pnorm, color = "red", size = 1.2) +
  xlim(-5, 5)
p1 | p2
```

- 标准正态分布的概率密度函数与概率分布函数：



3. 概率函数的 R 实现

- R 中常用的概率函数有密度函数、分布函数、分位数函数、生成随机数函数，其前缀表示分别为：
    - d = 密度函数（density)
    - p = 分布函数（distribution)
    - q = 分位数函数（quantile)
    - r = 生成随机数（random)
- 上述 4 个字母 + 分布缩写，就构成通常的概率函数。
- **分位数：**分位数是概率分布函数的反函数，p 分位数，是位于 p 位置的数，即比它小的数占比是 p, 比它大的数占比是 $1-p$. 中位数即 50% 分位数。

表 2: 常用概率分布及缩写

分布名称	缩写	参数及默认值
二项分布	binom	size, prob
多项分布	multinom	size, prob
负二项分布	nbinom	size, prob
几何分布	geom	prob
超几何分布	hyper	m, n, k
泊松分布	pois	lambda
均匀分布	unif	min=0, max=1
指数分布	exp	rate=1
正态分布	norm	mean=0, sd=1
对数正态分布	lnorm	meanlog=0, stdlog=1

表 3: 常用概率分布及缩写（续表）

分布名称	缩写	参数及默认值
t 分布	t	df
卡方分布	chisq	df
F 分布	f	df1, df2
Wilcoxon 符号秩分布	signrank	n
Wilcoxon 秩和分布	wilcox	m, n
柯西分布	cauchy	location=0, scale=1
Logistic 分布	logis	location=0, scale=1
Weibull 分布	weibull	shape, scale=1
Gamma 分布	gamma	shape, scale=1
Beta 分布	beta	shape1, shape2

二. 常见离散分布

- **0-1 分布**：1 次伯努利试验，成功概率 p, 失败概率 $1-p$, 成功次数服从 0-1 分布

$$P(X=k)=p^k(1-p)^{1-k}, \qquad k=0,\,1$$

- **二项分布**：n 次伯努利试验，成功概率 p, 失败概率 $1-p$，成功次数服从二项分布，记为 $\mathrm{B}(n,p)$

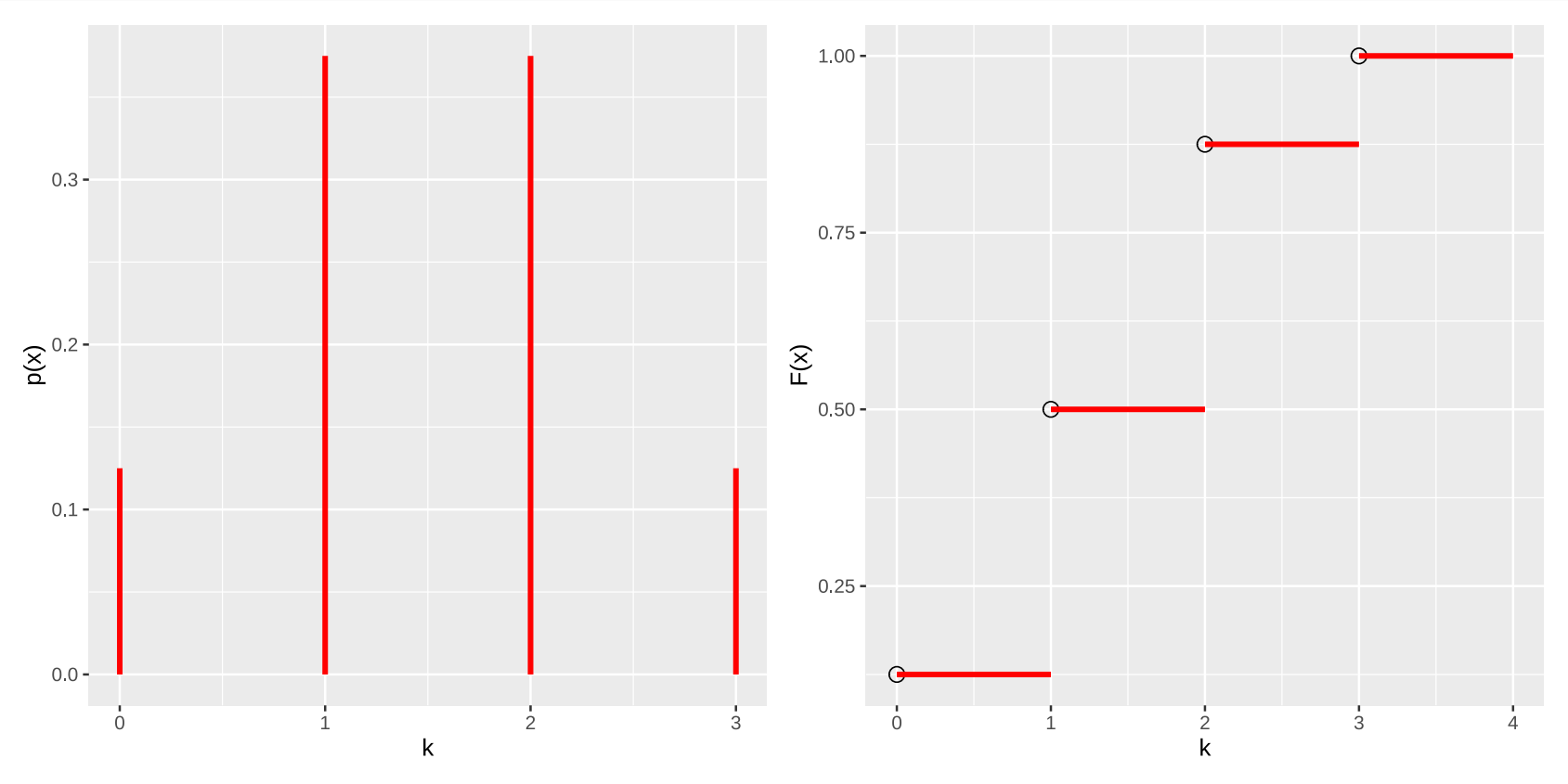
$$P(X=k)=C_n^k p^k(1-p)^{n-k}, \qquad k=0,\,1,\,\cdots,\,n$$

注 1：$0-1$ 分布，是二项分布 $n=1$ 的特例。

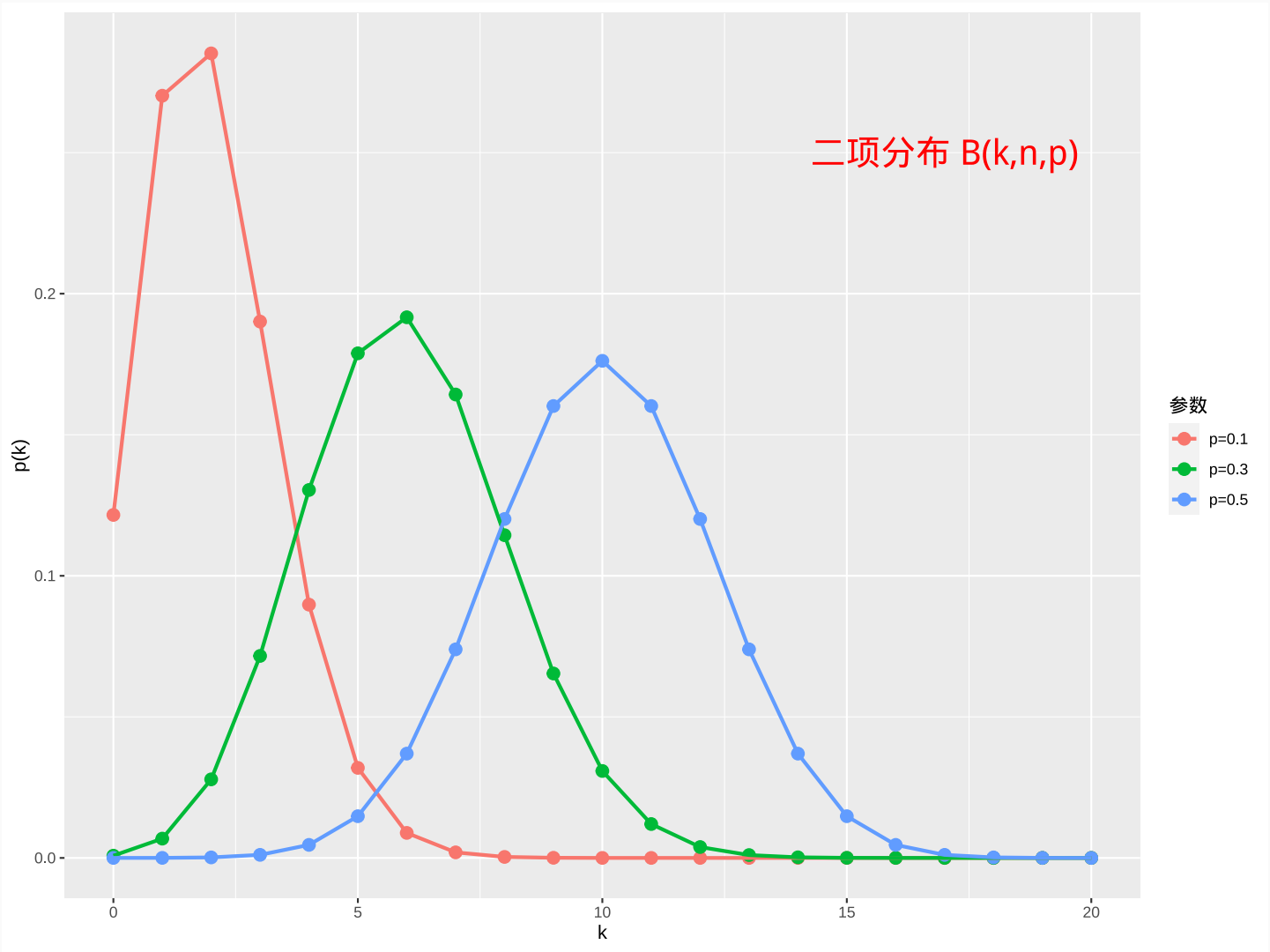
注 2：多项分布，是二项分布从两个结果到多个结果的推广，每个结果的概率确定。

```r
df = tibble(k = 0:3,
       `p(x)` = dbinom(k, 3, 0.5),
       `F(x)` = pbinom(k, 3, 0.5))
p1 = ggplot(df) +
  geom_segment(aes(k, 0, xend = k, yend = `p(x)`),
               size = 1.2, color = "red") +
  labs(y = "p(x)")
p2 = ggplot(df, aes(k, `F(x)`)) +
  geom_point(shape = 1, size = 3, color = "black") +
  geom_segment(aes(xend = k + 1, yend = `F(x)`),
               size = 1.2, color = "red")
p1 | p2
```

- 二项分布 $\mathrm{B}(3, 0.5)$ 的概率密度函数与概率分布函数：



```r
tibble(
  k = 0:20,
  `p=0.1` = dbinom(k, size = 20, 0.1),
  `p=0.3` = dbinom(k, size = 20, 0.3),
  `p=0.5` = dbinom(k, size = 20, 0.5)
) %>%
  pivot_longer(-k, names_to = " 参数",
               values_to = "p(k)") %>%
  ggplot(aes(k, `p(k)`, color = 参数)) +
  geom_point(size = 3) +
  geom_line(size = 1) +
  annotate(geom = "text", x = 17, y = 0.25,
           label = " 二项分布 B(k,n,p)",
           size = 7, color = "red")
```


二项分布 B(k,n,p)
p(k)
0.2
0.1
0.0
0
5
10
15
20
k
参数
p=0.1
p=0.3
p=0.5

已知 Bob 罚球命中率为 60%.

- Bob 罚球 12 次，正好投中 10 次的概率是多少?

```
dbinom(x = 10, size = 12, prob = 0.6)
#> [1] 0.0639
```

- Bob 罚球 10 次，他投中 8 次以上的概率是多少?

```
1 - pbinom(8, size = 10, prob = 0.6)
#> [1] 0.0464
```

- Bob 罚球 10 次，0.8 分位数是罚球几次? ($P(x) > 0.8, x = ?$)

```
qbinom(0.2, size = 10, prob = 0.6)
#> [1] 5
```

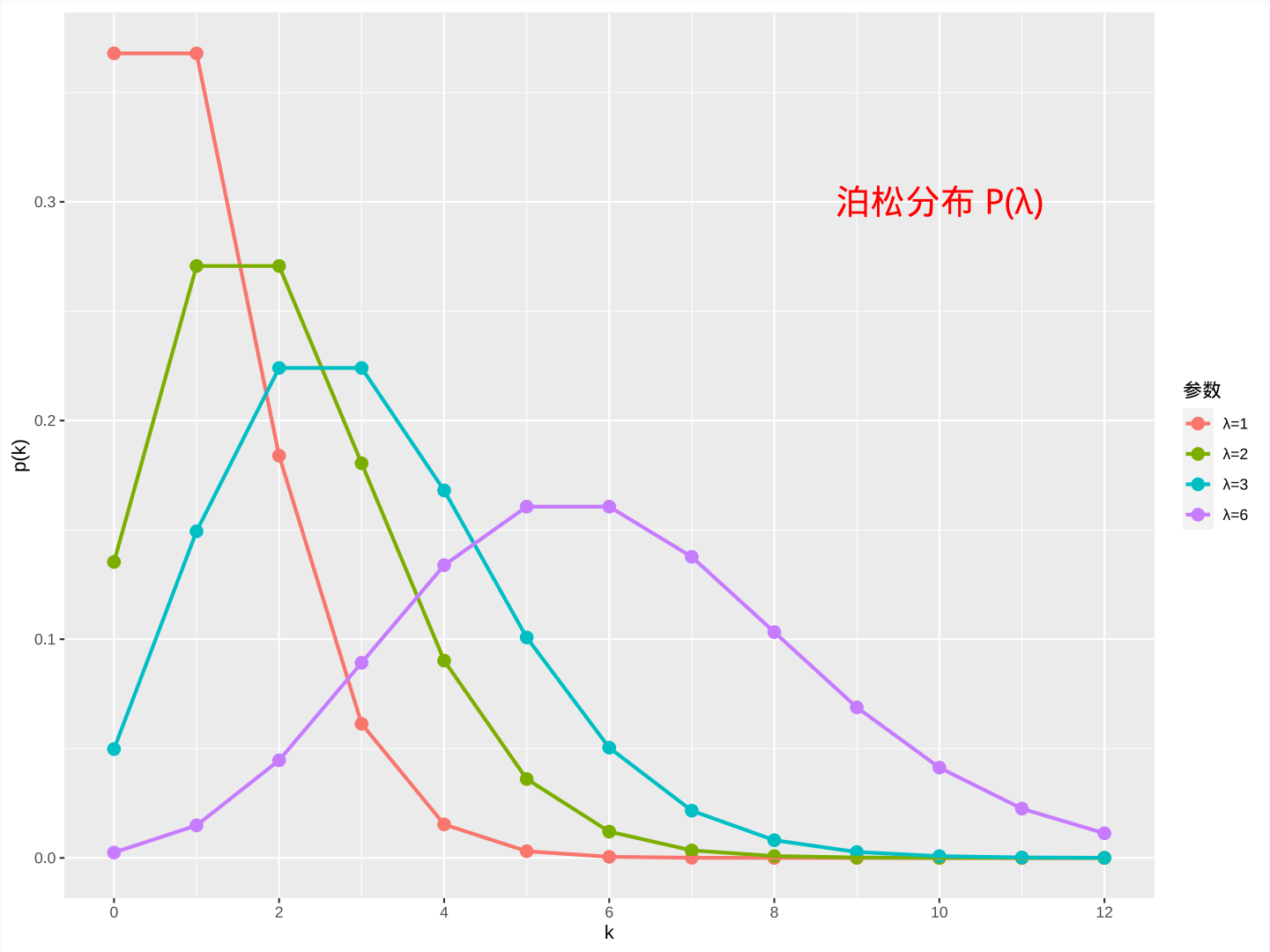
- **泊松分布**：以固定的平均瞬时速率 λ, 随机且独立地发生，单位时间内发生的次数服从泊松分布，记为 $\mathsf{P}(\lambda)$

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad \lambda > 0,\ k = 0,\ 1,\ 2 \cdots$$

注：当 n 很大，p 很小时，可以用泊松分布近似二项分布。

- 某网站平均每小时有 10 笔交易，在某一小时内，该网站交易量超过 8 的概率是多少?

```
1 - ppois(q = 8, lambda = 10)
#> [1] 0.667
```


泊松分布 P(λ)
p(k)
0.3
0.2
0.1
0.0
0
2
4
6
8
10
12
k
参数
λ=1
λ=2
λ=3
λ=6

- **超几何分布**

N 件产品，包含 M 件次品，不放回抽取 n 件，抽到的次品数服从超几何分布：

$$P(X=k)=\frac{C_M^k C_{N-M}^{n-k}}{C_N^n},\qquad 0\le k\le n\le N,\, k\le M$$

- **几何分布**

成功率为 p 的伯努利试验，出现首次成功时的试验次数, 服从几何分布：

$$P(X=k)=(1-p)^{k-1}p, \qquad k=1,\,2,\,\cdots$$

几何分布具有**无记忆性**：前 m 次试验未成功，继续等待到首次成功的试验次数仍服从同样的几何分布。

- **帕斯卡分布**

成功率为 p 的伯努利试验，出现第 r 次成功时的试验次数, 服从帕斯卡分布：

$$P(X=k)=C_{k-1}^{r-1}p^r(1-p)^{k-r},\qquad k=r,\,r+1,\cdots$$

- **负二项分布**

将帕斯卡分布中的 r 从正整数，推广到任意非负实数，就得到负二项分布，记为 $\mathrm{NB}(r,p)$。

三. 常见连续分布

- **均匀分布**

取值在 $[a, b]$ 区间，每个值被选择的可能性都一样：

$$f(x) = \frac{1}{b-a}, \qquad a \leq x \leq b$$

- **指数分布**

直到下一个事件（成功，失败，到达等）发生的等待时间，服从指数分布，记为 $\mathrm{Exp}(\lambda)$:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

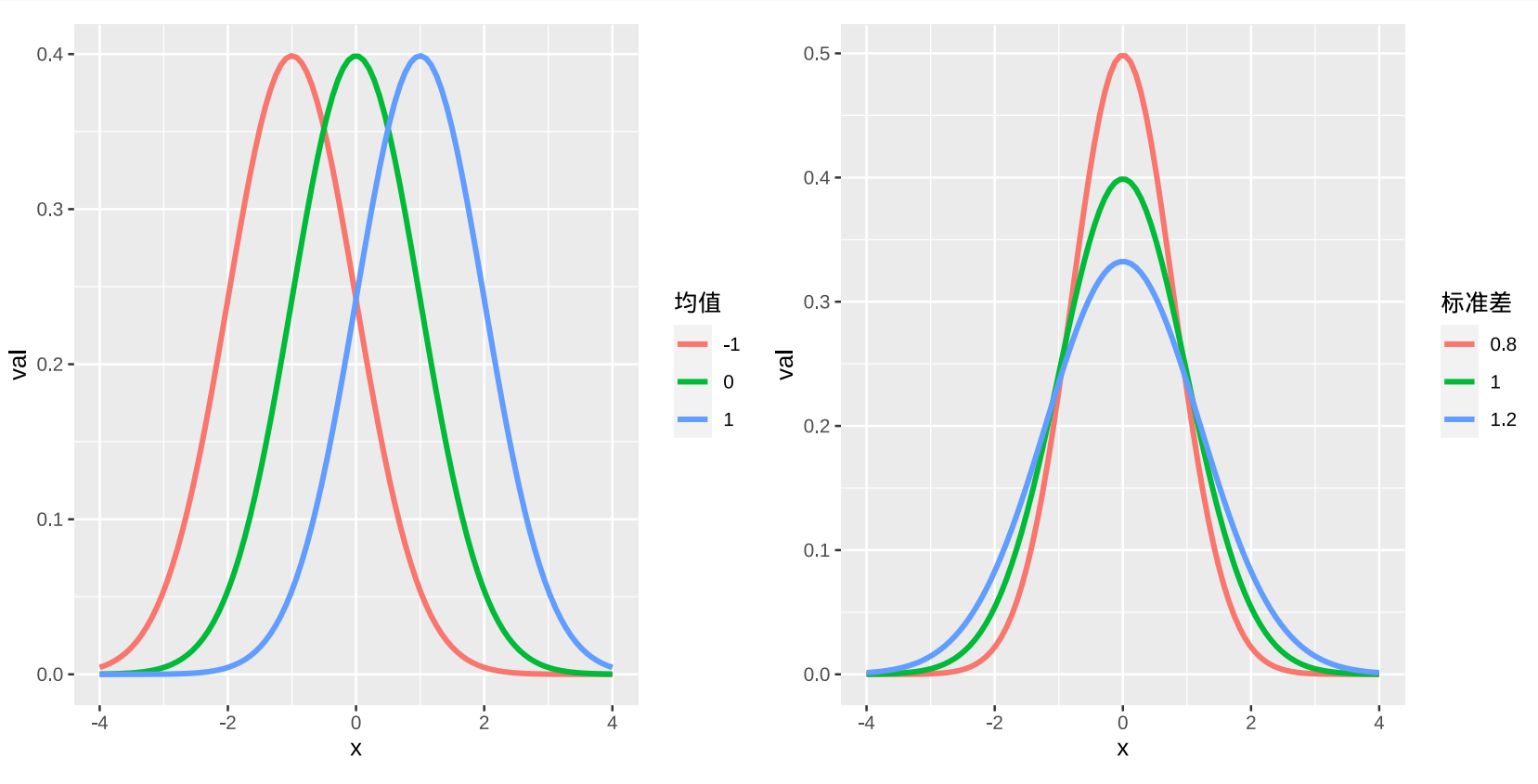
指数分布满足无记忆，即 $P(X > s + t | X > s) = P(X > t)$.

指数分布的另一种描述：风险函数在任意时刻都为常数的分布。

- **正态分布**

若影响某一数值指标的随机因素很多，而每个因素所起的作用都不太大，则数值指标服从正态分布，记为 $\mathsf{N}(\mu, \sigma)$:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty$$


0.4
0.3
0.2
0.1
0.0
val
-4
-2
0
2
4
x
均值
-1
0
1
0.5
0.4
0.3
0.2
0.1
0.0
val
-4
-2
0
2
4
x
标准差
0.8
1
1.2

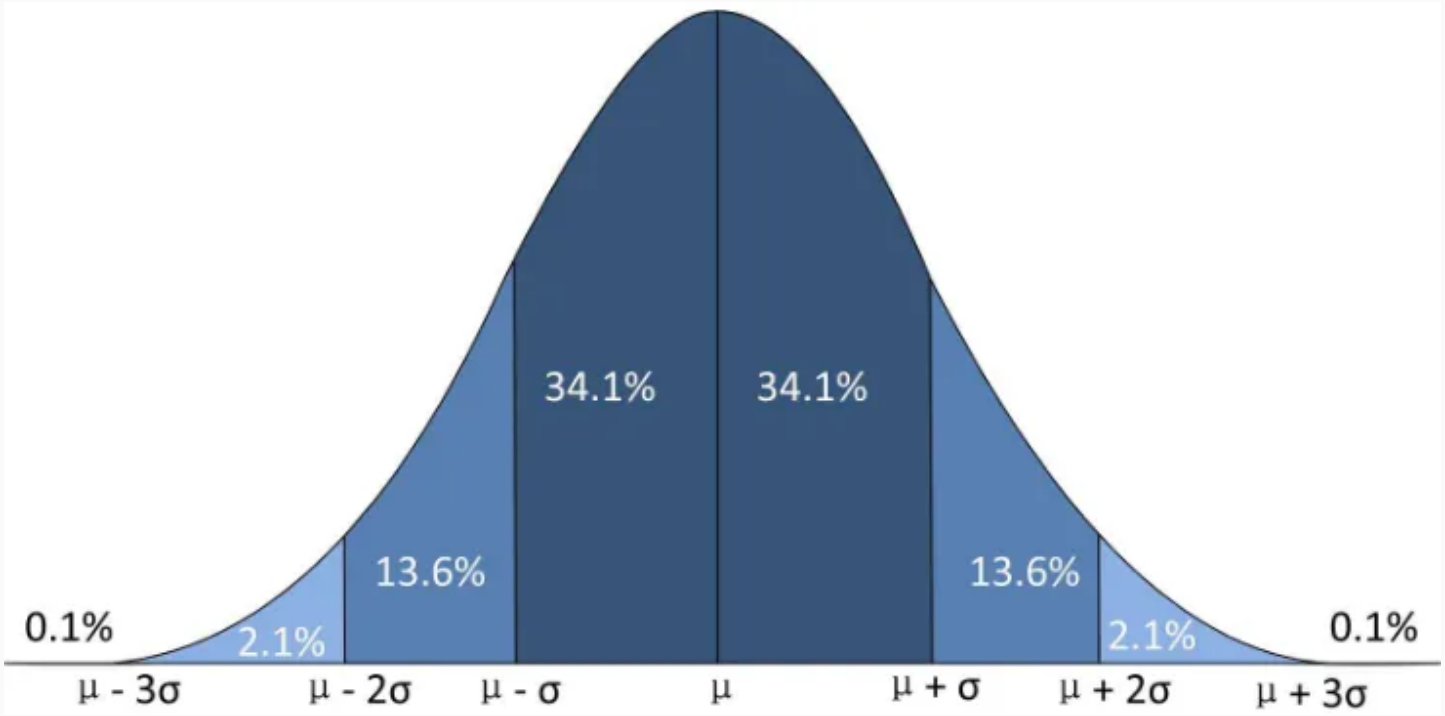


图 1: 正态分布下的面积规律

- **标准正态分布：**均值为 0，标准差为 1 的正态分布

相当于把正态分布的规律简化了，因为它的标准差是 1，对应的横轴上的数值 1 2 直接就是 1 倍标准差、2 倍标准差。所以，利用标准正态分布来说明面积规律就更简单了，可以直接说，以 0 为中心，在 ± 2 的范围内面积约为 95.4%；即当横坐标的值等于 ± 1.96 时，对应的两侧面积之和约为 0.05。

四. 三大抽样分布

设 $X_1, X_2, \cdots, X_n$ 是从总体 X 中抽取的容量为 n 的一个样本，如果由此样本构造一个函数 $T(X_1, X_2, \cdots, X_n)$，不依赖于任何未知参数，则该函数是一个**统计量**。比如，样本均值、样本方差等。

统计量是样本的函数，是统计推断的基础。

样本统计量（作为随机变量）的理论上的概率分布，叫做**抽样分布**。

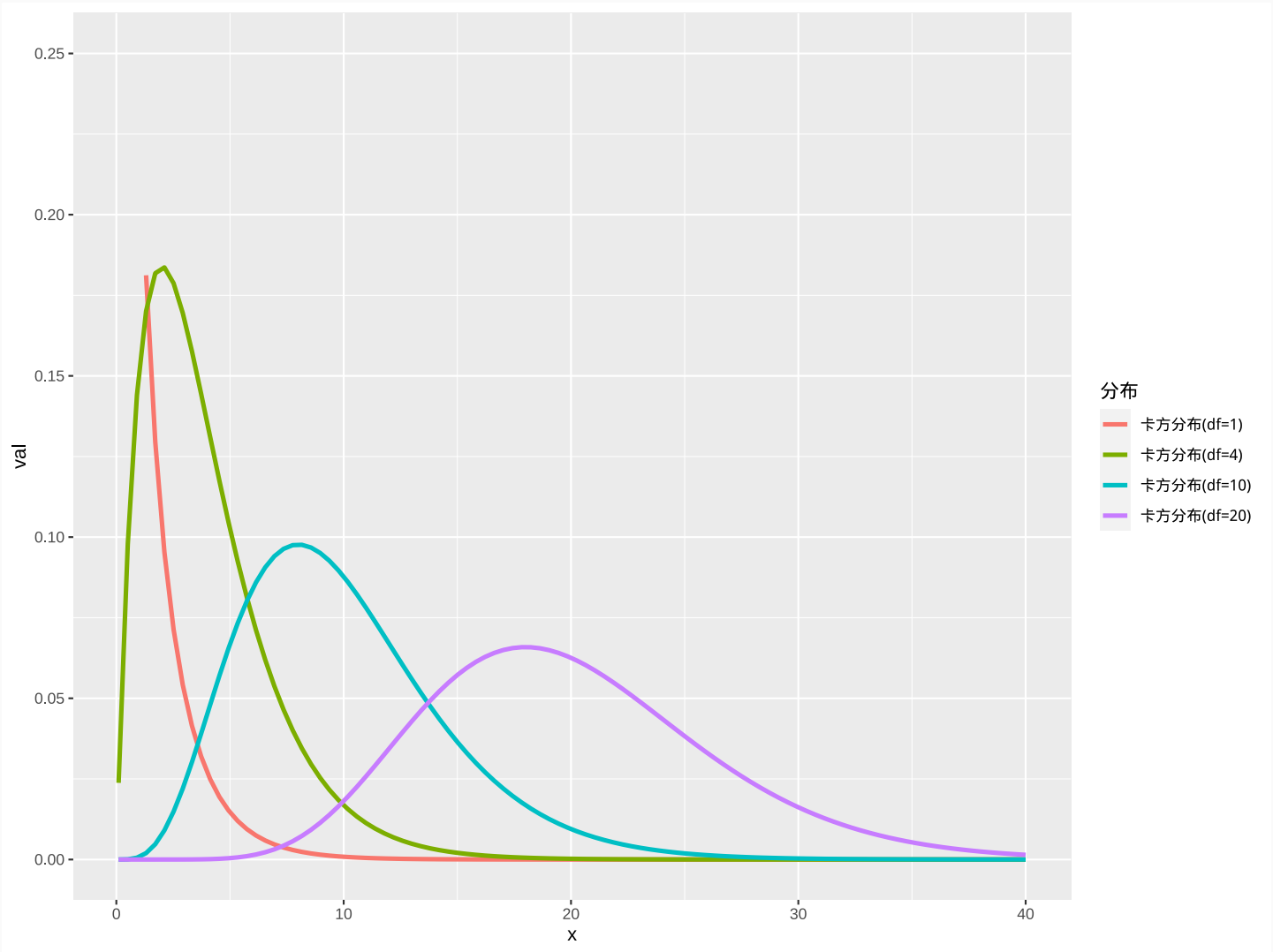
抽样分布的结果来自容量相同的所有可能样本，提供了样本统计量长远而稳定的信息，是进行推断的理论基础。

1. χ^2 分布

设 $X_1, X_2, \cdots, X_n$ 是从总体 $N(0,1)$ 的样本，则统计量

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

服从自由度为 n 的 χ^2 分布，记为 $\chi^2(n)$.


0.25
0.20
0.15
0.10
0.05
0.00
val
0
10
20
30
40
x
分布
卡方分布(df=1)
卡方分布(df=4)
卡方分布(df=10)
卡方分布(df=20)

2. t 分布

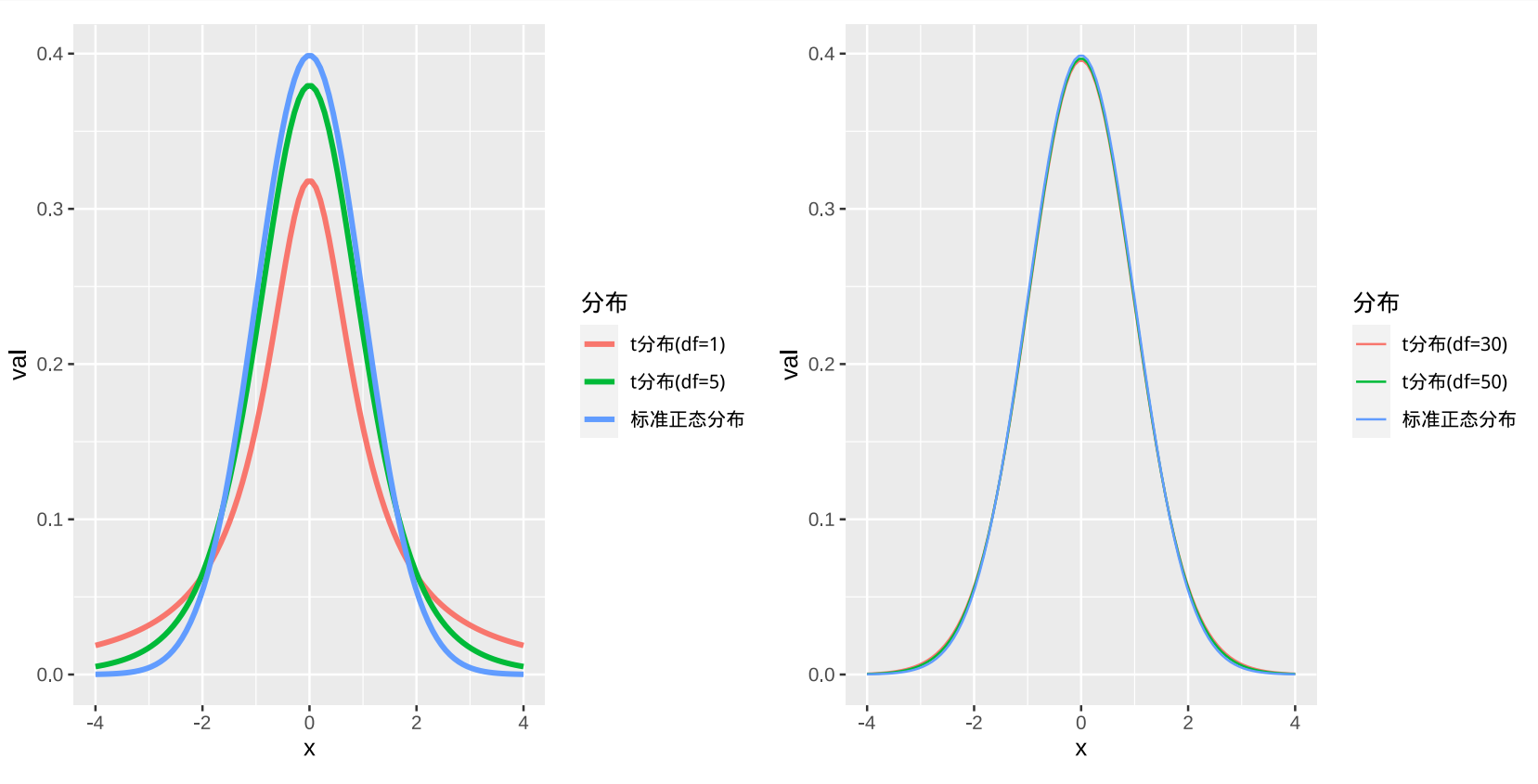
设 $X \sim N(0,1)$, $Y \sim \chi^2(n)$, 且 X, Y 相互独立，则随机变量

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

服从自由度为 n 的 t 分布，记为 $t(n)$.

t 分布可以看作小样本时候的正态分布，当样本量大了之后完全可以用标准正态分布来代替。

t 分布形状与其自由度有关，自由度越小 t 分布与标准正态分布偏离越大；当自由度足够大时（30），t 分布十分接近标准正态分布；50 的时候差别微乎其微。


0.4
0.3
0.2
0.1
0.0
val
-4
-2
0
2
4
x
分布
t分布(df=1)
t分布(df=5)
标准正态分布
0.4
0.3
0.2
0.1
0.0
val
-4
-2
0
2
4
x
分布
t分布(df=30)
t分布(df=50)
标准正态分布

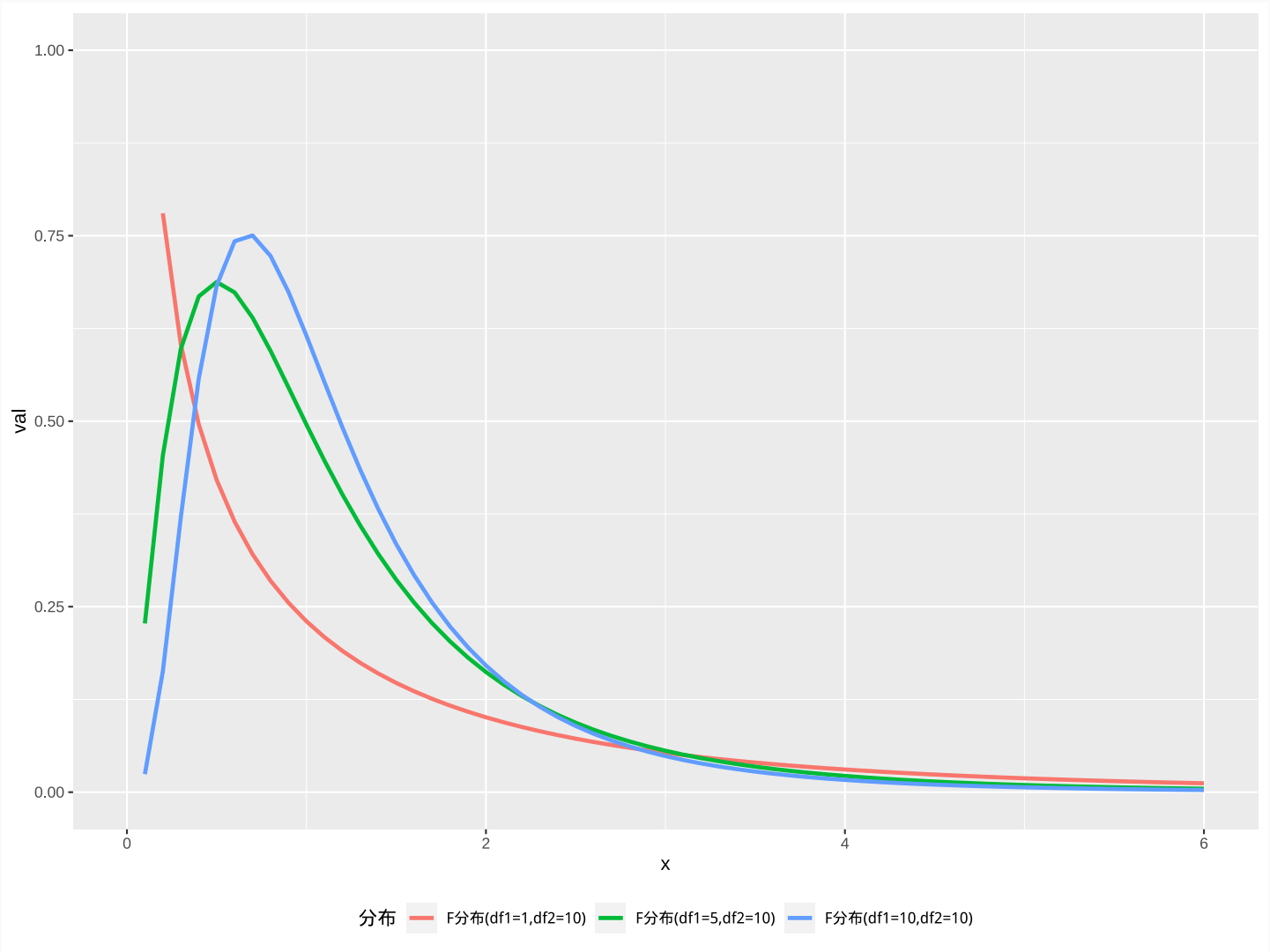
3. F 分布

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

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

服从自由度为 (n_1, n_2) 的 F 分布，记为 $F(n_1, n_2)$.

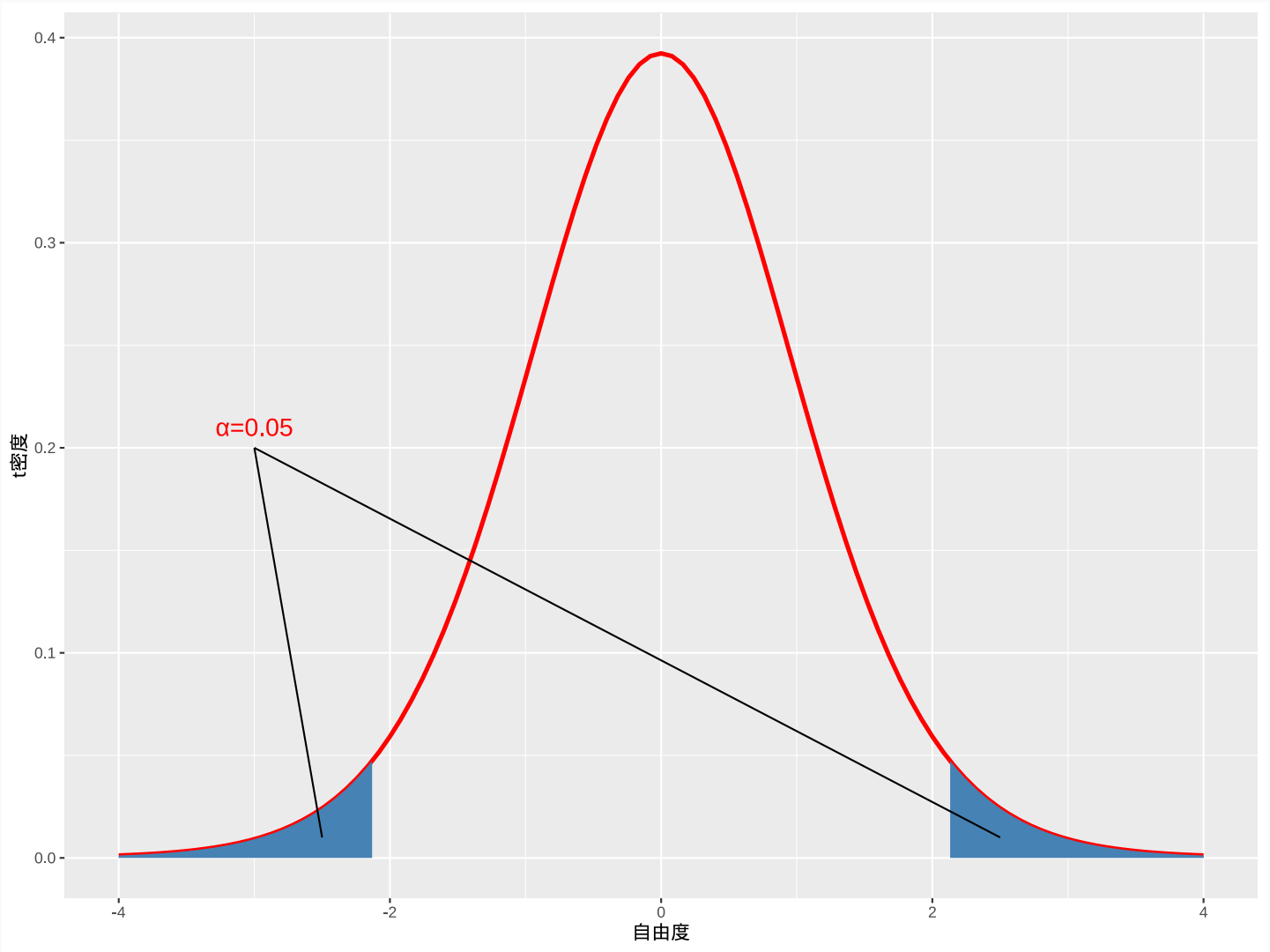
正态分布和 t 分布主要与均值的分布有关，在推断总体均值时比较有用；而 F 分布是与方差有关的分布，可用于分析两个方差是否相等、方差是否等于某一具体值等。


1.00
0.75
0.50
0.25
0.00
val
0
2
4
6
x
分布
F分布(df1=1,df2=10)
F分布(df1=5,df2=10)
F分布(df1=10,df2=10)

三大分布的 0.05 置信尾部

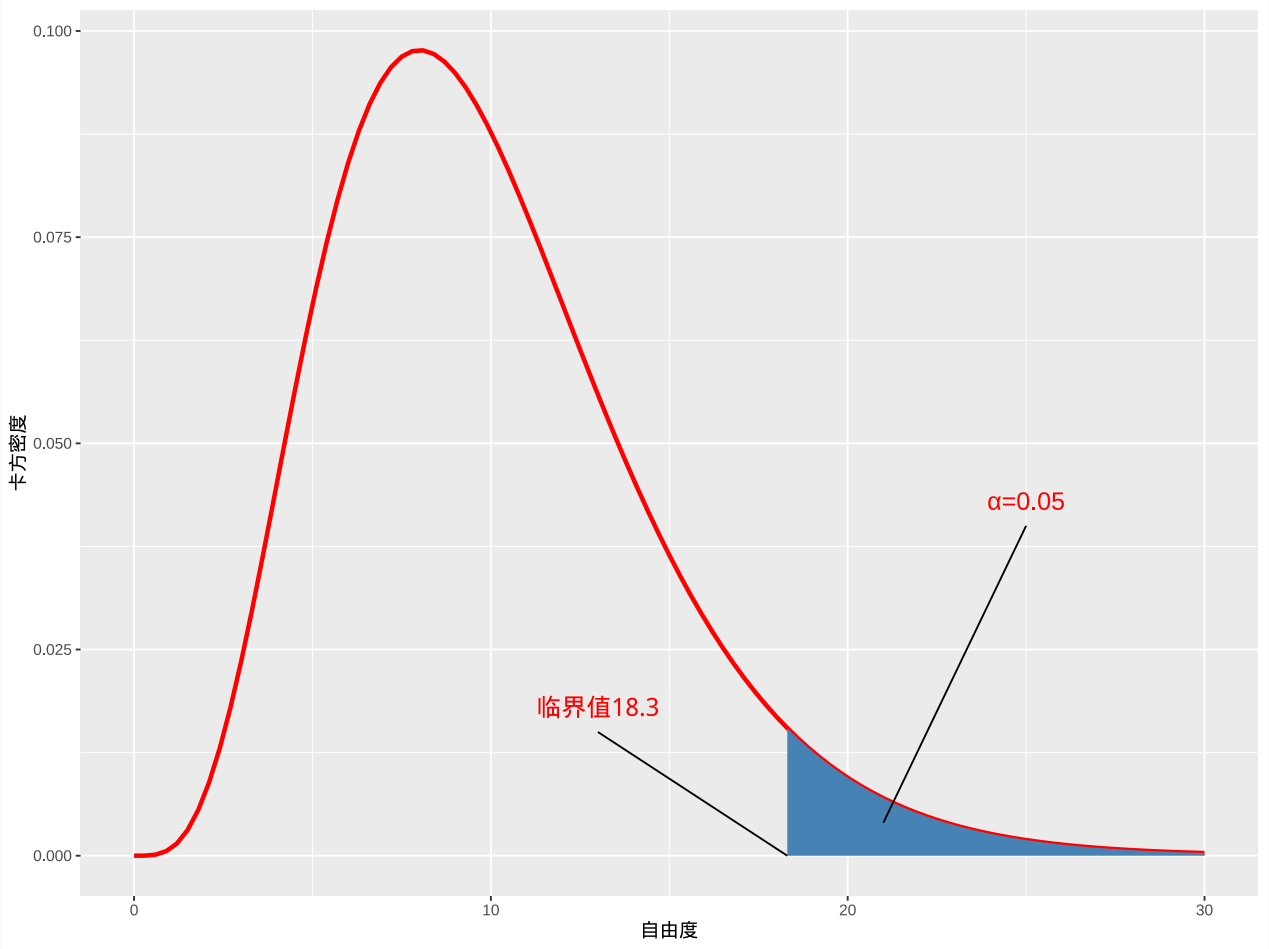
```
(t1 = qt(0.05/2, 15))
#> [1] -2.13
(t2 = qt(1-0.05/2, 15))
#> [1] 2.13

ggplot() +
  stat_function(fun = ~ dt(.x, 15), xlim = c(-4, 4),
                color = "red", size = 1.2) +
  stat_function(fun = ~ dt(.x, 15), xlim = c(t2, 4),
                geom = "area", fill = "steelblue") +
  labs(x = " 自由度", y = "t 密度") +
  geom_segment(aes(x = c(-2.5, 2.5), y = 0.01,
                   xend = -3, yend = 0.2)) +
  annotate(geom = "text", x = -3, y = 0.21,
           label = "α=0.05", size = 5, color = "red")
```


0.4
0.3
0.2
0.1
0.0
t密度
α=0.05
-4
-2
0
2
4
自由度

```r
(chi = qchisq(1 - 0.05, 10))
#> [1] 18.3
```

```r
ggplot() +
  stat_function(fun = ~ dchisq(.x, 10), xlim = c(0, 30),
                color = "red", size = 1.2) +
  stat_function(fun = ~ dchisq(.x, 10), xlim = c(chi, 30),
                geom = "area", fill = "steelblue") +
  labs(x = " 自由度", y = " 卡方密度") +
  geom_segment(aes(c(chi,21), c(0,0.004),
                   xend = c(13,25), yend = c(0.015,0.04))) +
  annotate(geom = "text", x = c(13,25), y = c(0.018,0.043),
           label = c(str_glue(" 临界值 {round(chi,1)}"),
                     "α=0.05"), size = 5, color = "red")
```



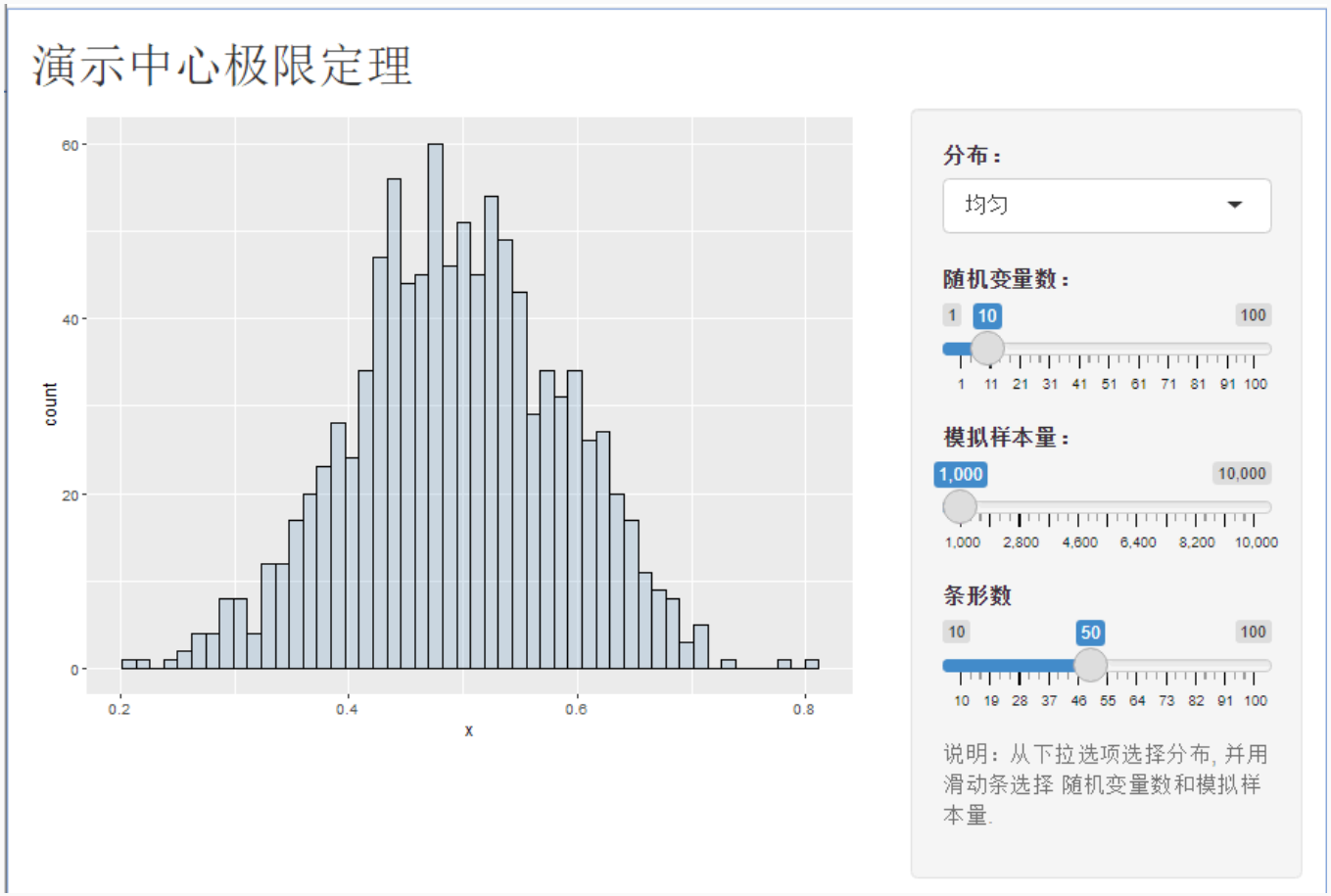
练习： F 分布 0.05 置信尾部?

总结

- t 分布主要是与均值有关的抽样分布，常用于两个均值是否相等（差值是否等于 0)、回归系数是否为 0 的统计检验；
- F 分布是与方差有关的抽样分布，常用于方差齐性检验（两个方差之比)、方差分析（组间方差与组内方差之比)、回归模型检验（模型方差与残差方差之比)；
- χ^2 分布是与方差有关的抽样分布，但在实际中常用于描述分类资料的实际频数与理论频数之间的抽样误差（小样本需要做连续校正)。

中心极限定理

设 $X_1, \cdots, X_n$ 为任意期望为 μ, 方差为 σ^2（有限）分布的抽样，则当 n 足够大时，$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 近似服从 $N\left(\mu, \frac{\sigma^2}{n}\right)$.



本篇主要参阅 (张敬信, 2022), (冯国双, 2018), (贾俊平, 2018), 以及包文档，模板感谢 (黄湘云, 2021), (谢益辉, 2021).

参考文献

冯国双 (2018). **白话统计**. **电子工业出版社, 北京**, 1 edition.

张敬信 (2022). *R* **语言编程：基于** *tidyverse*. **人民邮电出版社, 北京**.

谢益辉 (2021). *rmarkdown: Dynamic Documents for R*.

贾俊平 (2018). **统计学**. **中国人民大学出版社, 北京**, 7 edition.

黄湘云 (2021). *Github: R-Markdown-Template*.