05 CHAPTER

正态分布及其相关分布

1

第5-2章 正态分布及其相关分布

1. 正态分布
2. 二项分布的正态近似
3. 卡夫分布
4. t-分布
5. F-分布

2/33

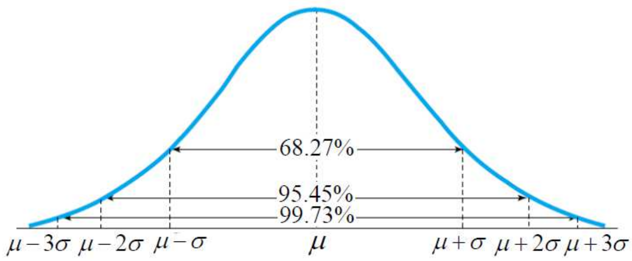
2

1. 正态分布

[定义 8-1] 正态分布(normal distribution) $X \sim N(\mu, \sigma^2)$

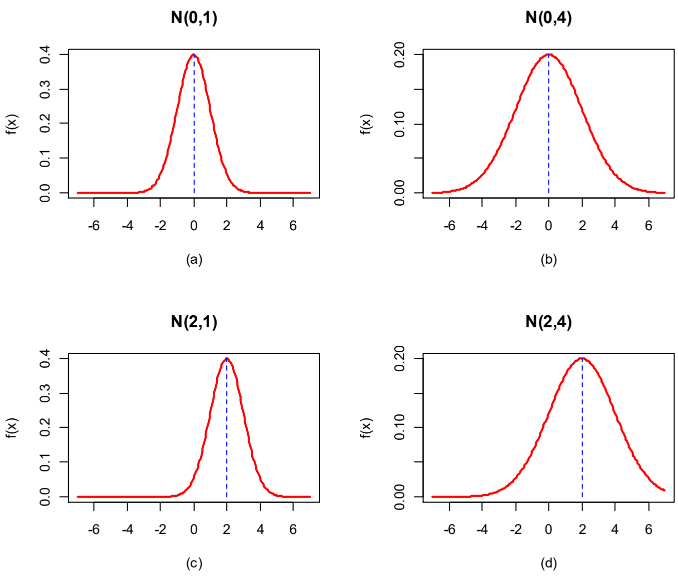
：以期望值为对称中心，由标准差来决定其散布情况的钟形模样分布

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], -\infty < x < \infty$$




```
# 正态分布的R函数

# 概率密度函数 (mean=期望, sd=标准差)
# 若省略mean, sd，就按标准正态分布计算 (mean = 0, sd = 1)
dnorm(x, mean = 0, sd = 1)
# Excel函数 = NORM.DIST(x, mean, sd, FALSE)

# 累积分布函数 (q=分位数)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)
# Excel函数 = NORM.DIST(x, mean, sd, TRUE)

# 分位数 (p=累积概率)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE)
# Excel函数 = NORM.INV(p, mean, sd)

# 正态随机变量 (n=随机数的个数)
rnorm(n, mean = 0, sd = 1)
# Excel函数 = NORM.INV(RAND( ), mean, sd) ⇒ 生成1个随机数
```

[例 8-1] 有关N(0,1), N(0,2²), N(2,1), N(2,2²)的概率密度函数



5/33

5

1. 正态分布

[定义 8-2] 标准正态(standard normal distribution)

: 期望为0，标准差为1的正态分布 $Z \sim N(0,1)$

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), -\infty < z < \infty$$

- CDF $\Phi(y) \equiv P(Z < y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz$
- 分位数 $z_p \equiv \Phi^{-1}(p) = \{y \mid \Phi(y) = p\}$
- MGF $m_Z(t) = E(e^{tZ}) = \int_{-\infty}^{\infty} \frac{e^{tz}}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(z^2 - 2tz + t^2) + \frac{t^2}{2}\right] dz$$

$$= \exp(\frac{t^2}{2}) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(z-t)^2\right] dz = \exp(\frac{t^2}{2})$$

6/33

6

1. 正态分布

- 正态分布的 MGF $X \sim N(\mu, \sigma^2)$

$$m_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} \frac{e^{tx}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$$

$$z = \frac{x-\mu}{\sigma} \Rightarrow x = \mu + \sigma z \Rightarrow dx = \sigma dz$$

$$m_X(t) = e^{\mu t} \int_{-\infty}^{\infty} \frac{e^{\sigma t z}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{z^2}{2}\right) \sigma dz$$

$$= e^{\mu t} \int_{-\infty}^{\infty} e^{\sigma t z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz$$

$$= e^{\mu t} E(e^{\sigma t Z}) = e^{\mu t} m_Z(\sigma t) = \exp(\mu t + \frac{\sigma^2 t^2}{2})$$

$$X = \mu + \sigma Z$$

$$m_X(t) = E(e^{tX}) = E(e^{t(\mu+\sigma Z)}) = e^{\mu t} E(e^{\sigma t Z})$$

$$= e^{\mu t} m_Z(\sigma t) = \exp(\mu t + \frac{\sigma^2 t^2}{2})$$

7/33

7

1. 正态分布

[定理 8-1] 正态分布的线性变换

$$X \sim N(\mu, \sigma^2),\ Y = a + bX \Rightarrow Y \sim N(a + b\mu, b^2\sigma^2)$$

[证明] $m_X(t) = \exp(\mu t + \frac{\sigma^2 t^2}{2})$

$$m_Y(t) = E(e^{tY}) = E(e^{t(a+bX)}) = e^{at} E(e^{btX})$$

$$= e^{at} m_X(bt) = e^{at} \exp\left[\mu bt + \frac{\sigma^2 (bt)^2}{2}\right] = \exp\left[(a+\mu b)t + \frac{(b\sigma)^2 t^2}{2}\right]$$

[推论 8-1] 正态分布的标准化(standardization)

$$X \sim N(\mu, \sigma^2),\ Z = \frac{X-\mu}{\sigma} \Rightarrow Z \sim N(0,1)$$

[证明] $E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma} E(X-\mu) = \frac{1}{\sigma}(\mu - \mu) = 0$

$$Var(Z) = Var\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2} Var(X-\mu) = \frac{\sigma^2}{\sigma^2} = 1$$

8/33

8

1. 正态分布

- 正态分布的概率计算 → 使用标准正态分布

$$P(a<X<b)=P(\frac{a-\mu}{\sigma}<Z<\frac{b-\mu}{\sigma})=\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})$$

[例 8-2] 当X~N(2,4)时, 求P(-1<X<4)

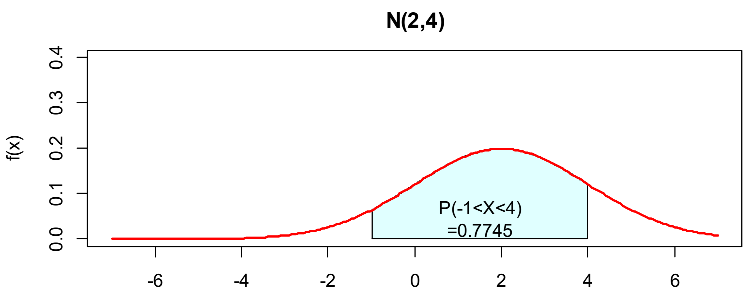
$$P(-1<X<4)=P(\frac{-1-2}{2}<Z<\frac{4-2}{2})=P(-1.5<Z<1)=\Phi(1)-\Phi(-1.5)$$

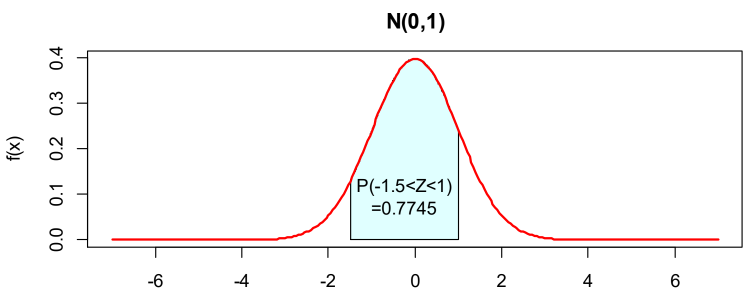
[表 A-1] → $\Phi(1)\equiv P(Z\le 1)\doteq 0.8413,\ \Phi(1.5)\equiv P(Z<1.5)\doteq 0.9332$

$$\Phi(-1.5)\equiv P(Z\le -1.5)=1-\Phi(1.5)\doteq 1-0.9332=0.0668$$

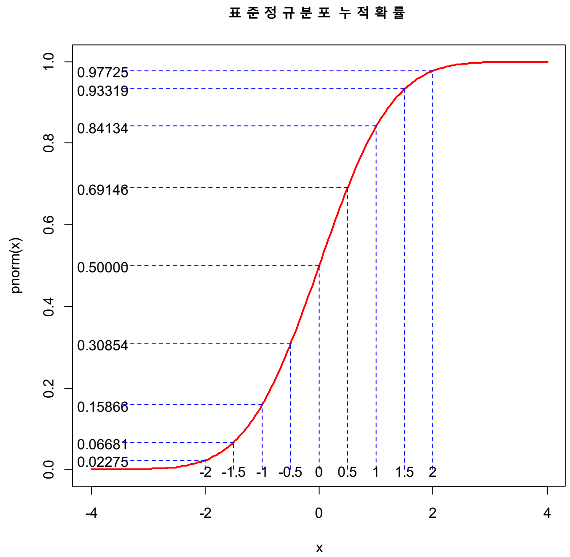
$$P(-1<X<4)\doteq 0.8413-0.0668=0.7745$$

```
# 使用pnorm(x, mean, sd)
pnorm(4, 2, 2) - pnorm(-1, 2, 2); pnorm(1)-pnorm(-1.5)
[1] 0.7745375      [1] 0.7745375
```

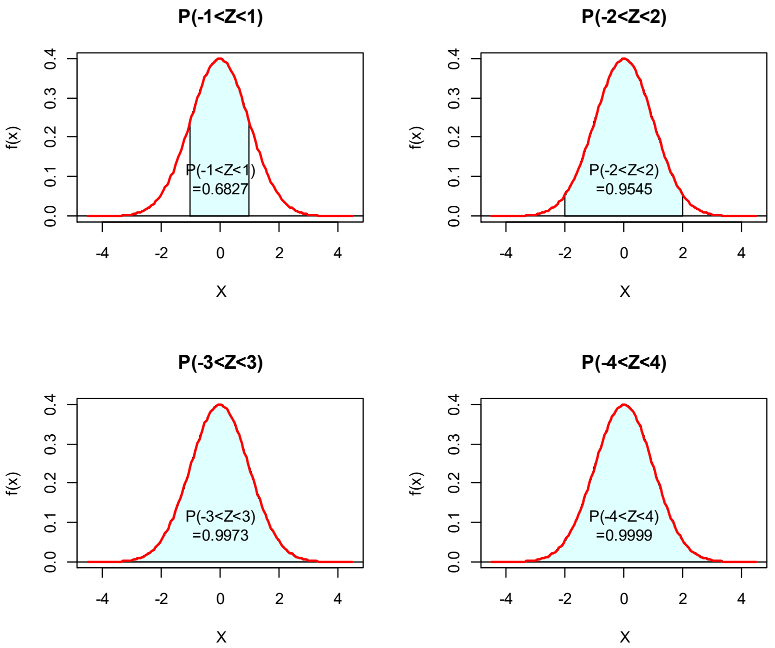





[例 8-3] 请填写<附录 B-1>的标准正态累积分布表

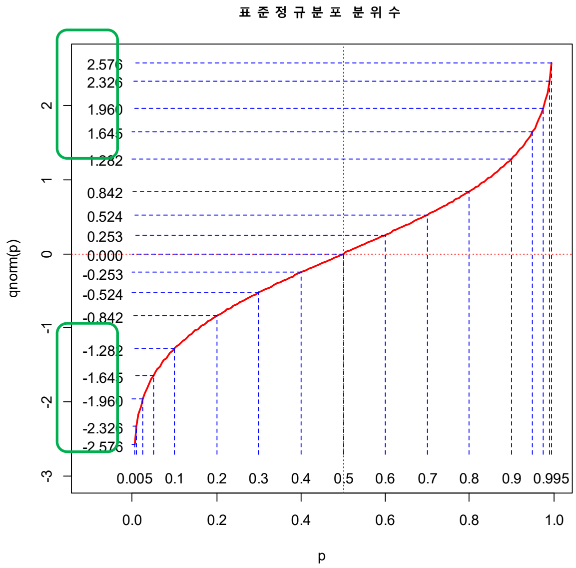


11/33



12/33

[例 8-4] 请填写累积概率分别为0.5%, 1.0%, 2.5%, 5.0%, 10%~90%, 95%, 97.5%, 99%, 99.5%时，标准正态分布的分位数表。



1. 正态分布

[例 8-5] 20多岁男性的身高遵循均值为175，方差为64的正态分布时，求20多岁男性中身高介于180和185之间的比例。

$X \sim N(175, 8^2)$

$$P(180 < X < 185) = P(\frac{180-175}{8} < Z < \frac{185-175}{8})$$
$$= P(0.625 < Z < 1.25) = \Phi(1.25) - \Phi(0.625)$$
$$\Phi(1.25) \doteq 0.8944$$
$$\Phi(0.62) \doteq 0.7324,\ \Phi(0.63) \doteq 0.7357 \quad \Rightarrow \quad \Phi(0.625) \doteq 0.7341$$
$$\Rightarrow \quad P(180 < X < 185) \doteq 0.8944 - 0.7341 = 0.1603$$

```
# 使用pnorm(x, mean, sd)
pnorm(185, 175, 8) - pnorm(180, 175, 8)
[1] 0.1603358
```

1. 正态分布

[例 8-6] 某公司A型号为Q发动机的寿命遵循均值为10年，标准差为1.5年的正态分布。型号Q发动机的无偿保修比率维持在5%内 → 无偿保修期最多几年？

$$X \sim N(10, 1.5^2) \qquad P(X < m) \le 0.05$$

$$P(X < m) = P(Z < \frac{m-10}{1.5}) \le 0.05$$

$$\Rightarrow \frac{m-10}{1.5} \le \Phi^{-1}(0.05) \doteq -1.645$$

$$\Rightarrow m \le -1.645 \times 1.5 + 10 = 7.5325$$

```
# 使用qnorm(p, mean, sd)
qnorm(0.05, 10, 1.5)
[1] 7.53272

10+qnorm(0.05)*1.5
[1] 7.53272
```

15/33

15

1. 正态分布

[定理 8-2] 正态分布的加法法则1

$$X \sim N(\mu_1, \sigma_1^2),\ Y \sim N(\mu_2, \sigma_2^2)\ \& X \perp Y$$
$$\Rightarrow X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

[证明] $m_X(t) = \exp(\mu_1 t + \frac{\sigma_1^2 t^2}{2}),\ m_Y(t) = \exp(\mu_2 t + \frac{\sigma_2^2 t^2}{2})$

$$\Rightarrow m_{X+Y}(t) = m_X(t) m_Y(t) = \exp\left[(\mu_1 + \mu_2)t + \frac{(\sigma_1^2 + \sigma_2^2)t^2}{2}\right]$$

$$\Rightarrow X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

- X与Y不独立的情况

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 + 2\sigma_{XY})$$

16/33

16

1. 正态分布

[例 8-7] 某螺栓直径的均值为20mm，标准差为0.3mm，符合正态分布。
螺丝直径的均值为21.5mm，标准差为0.4mm，符合正态分布。
求该螺栓与螺丝不匹配的概率?

$$X \sim N(20,\ 0.3^2),\ Y \sim N(21.5,\ 0.4^2)$$
$$\Rightarrow Y - X \sim N(1.5,\ 0.4^2 + 0.3^2)$$
$$P(X > Y) = P(Y - X < 0)$$
$$= P(Z < \frac{0 - 1.5}{\sqrt{0.25}}) = \Phi(-3) \doteq 1 - 0.9987 = 0.0013$$

```
# 使用pnorm(p, mean, sd)
pnorm(0, 21.5-20, sqrt(0.4^2 +0.3^2))
[1] 0.001349898
```

17/33

17

1. 正态分布

[推论 8-2] 正态分布的加法法则2

$$X_i \overset{ind}{\sim} N(\mu_i, \sigma_i^2),\ \ i = 1, 2, \ldots, n\ \ \Rightarrow Y \equiv \sum_{i=1}^{n} X_i \sim N(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)$$

[证明] $m_{X_i}(t) = \exp(\mu_i t + \frac{\sigma_i^2 t^2}{2})$

$$\Rightarrow m_Y(t) = \prod_{i=1}^{n} m_{X_i}(t) = \prod_{i=1}^{n} \exp(\mu_i t + \frac{\sigma_i^2 t^2}{2}) = \exp\left(\sum_{i=1}^{n} \mu_i t + \frac{\sum_{i=1}^{n} \sigma_i^2 t^2}{2}\right)$$

[例 8-8] 长度符合N(12.5, 4)分布的四根木棒相互进行连接时，求整天长度介于40和60之间的概率。

$$X \sim N(50,\ 4 \times 4)$$
$$P(40 < X < 60) = P(\frac{40 - 50}{\sqrt{16}} < Z < \frac{60 - 50}{\sqrt{16}}) = \Phi(2.5) - \Phi(-2.5)$$
$$\doteq 0.9938 - (1 - 0.9938) = 0.9876$$

```
# 使用pnorm(x, mean, sd)
pnorm(60, 50, 4)-pnorm(40, 50, 4)
[1] 0.9875807
```

18/33

18

2. 二项分布与正态近似

[定理 8-3] 二项分布的正态近似

$$X \sim B(n,p) \Rightarrow Z \equiv \frac{X-np}{\sqrt{np(1-p)}} \xrightarrow{d} N(0,1)$$

[例 8-9] 不良率为p=0.2的工艺流程，从中抽取n=25个产品进行检查，求发现不良品低于4个的概率。

$$X \sim B(25,\, 0.2) \quad \Rightarrow np(1-p) = 25\times 0.2\times 0.8 = 4$$

$$P(X \le 4) = \sum_{x=0}^{4} \binom{25}{x} (0.2)^x (0.8)^{25-x} \doteq 0.4207$$

$$P(X \le 4) = P(\frac{X-5}{2} \le \frac{4-5}{2}) = P(Z \le -0.5) \doteq 0.3085$$

- 连续性校正(continuity correction)

$$P(X \le 4) \simeq P(X < 4.5) = P(Z \le \frac{4.5-5}{2}) = P(Z \le -0.25) \doteq 0.4013$$

19/33

19

2. 二项分布的正态近似

[例 8-9] 不良率为为p=0.2的工艺流程，从中抽取n=25个产品进行检查，求发现不良品低于4个的概率。(承上)

```
# 正确计算 → 使用pbinom(x, n, p)
pbinom(4, 25, 0.2)
[1] 0.4206743

# 正态近似 → 使用pnorm(x, mean, sd)
pnorm(4, 5, 2)
[1] 0.3085375

# 连续性校正
pnorm(4.5, 5, 2)
[1] 0.4012937
```

20/33

20

2. 二项分布的正态近似

[例 8-10] 掷骰子100次，求出现点数大于4的次数介于40和45之间的概率。

$X \sim B(100, 0.5)$

$$P(40 \le X \le 45) = \sum_{x=40}^{45} \binom{100}{x} (\frac{1}{2})^x (\frac{1}{2})^{100-x} \doteq 0.1665$$

$$P(40 \le X \le 45) \simeq P(39.5 < X < 45.5) = P(\frac{39.5-50}{5} \le Z \le \frac{45.5-50}{5})$$
$$= P(-2.1 \le Z \le -0.9) = [1-\Phi(0.9)] - [1-\Phi(2.1)]$$
$$\doteq (1-0.8159) - (1-0.9821) = 0.1662$$

```
# 正确计算
pbinom(45, 100, 0.5)-pbinom(39, 100, 0.5)
[1] 0.1665007
# 正态近似 (连续性校正)
pnorm(-0.9)-pnorm(-2.1)
[1] 0.1661957
```

21/33

21

3. 卡方分布

[定义 8-3] 卡方分布(chi-square distributiion)

$$Z_1, Z_2, \ldots, Z_n \overset{iid}{\sim} N(0,1) \ \Rightarrow \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$$

- 概率密度函数 $f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}, \ x > 0$

  $\chi^2(\nu) \equiv \Gamma(\alpha = \frac{\nu}{2}, \theta = 2)$
- 动差生成函数 $m(t) = (1-\theta t)^{-\alpha} = (1-2t)^{-\nu/2}$
- 期望 $E(X) = \alpha\theta = (\nu/2) \times 2 = \nu$
- 方差 $Var(X) = \alpha\theta^2 = (\nu/2) \times 2^2 = 2\nu$

22/33

22

3. 卡方分布

卡方分布R函数

卡方分布概率密度函数 (df=自由度, ncp=非中心参数(未使用))

```
dchisq(x, df, ncp = 0)
# Excel 函数 = CHISQ.DIST(x, df, FALSE)
```

卡方分布累积分布函数 F(x)

```
pchisq(x, df, ncp = 0, lower.tail = TRUE)
# Excel 函数 = CHISQ.DIST(x, df, TRUE)
```

卡方分布分位数 (p=累积概率)

```
qchisq(p, df, ncp = 0, lower.tail = TRUE)
# Excel 函数 = CHISQ.INV(p, df)
```

卡方分布随机变量 (n=随机数的个数)

```
rchisq(n, df, ncp = 0)
# Excel 函数 = CHISQ.INV(RAND( ), df) ⇒ 生成一个随机数
```

23/33

23

3. 卡方分布

[定理 8-4] 卡方分布的加法法则

$$X_1, X_2, \ldots, X_k \text{ are indep. } \& \ X_i \sim \chi^2(\nu_i)$$

$$\Rightarrow \sum_{i=1}^{k} X_i \sim \chi^2(\nu_1 + \nu_2 + \cdots + \nu_k)$$

[证明] $Y = X_1 + X_2 + \cdots + X_k$ $\quad E(e^{tX_i}) = (1-\theta t)^{-\nu_i}$

$$m_Y(t) = E(e^{tX_1})E(e^{tX_2})\cdots E(e^{tX_k}) = (1-\theta t)^{-(\nu_1+\nu_2+\cdots+\nu_k)}$$

- 自由度(degree of freedom)

  包括平方和的'独立项'个数

- 样本方差的自由度 $\quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

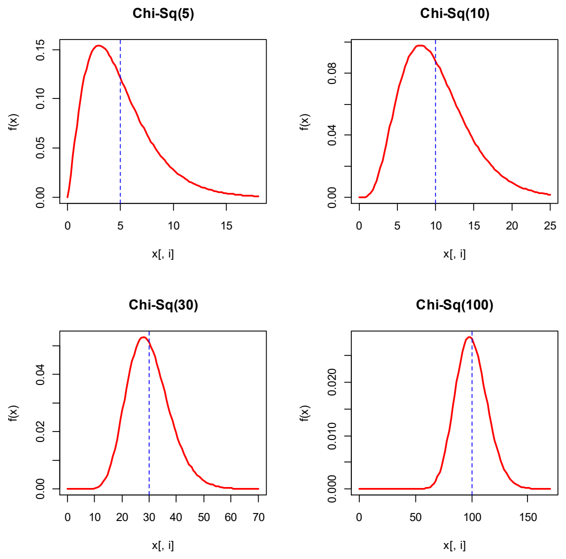
$$\sum_{i=1}^{n}(X_i - \bar{X}) = \sum_{i=1}^{n} X_i - n\bar{X} = 0 \quad \Rightarrow df = n-1$$

24/33

24

[例 8-11] 填写卡方概率分布的分位数[表]　$P(\chi^2 > \chi^2_{1-\alpha;\nu}) = \alpha$

[例 8-12] 自由度为5, 10, 30, 100的卡方分布的概率密度函数



25/33

25

3. 卡方分布

[例 8-13] 从符合标准正态分布的总体中随机抽取n个样本

$$Z_1, Z_2, \ldots, Z_n \Rightarrow P\left(\sum_{i=1}^{n} Z_i^2 \le a\right) = 0.95$$ ➔ 常数a的值?

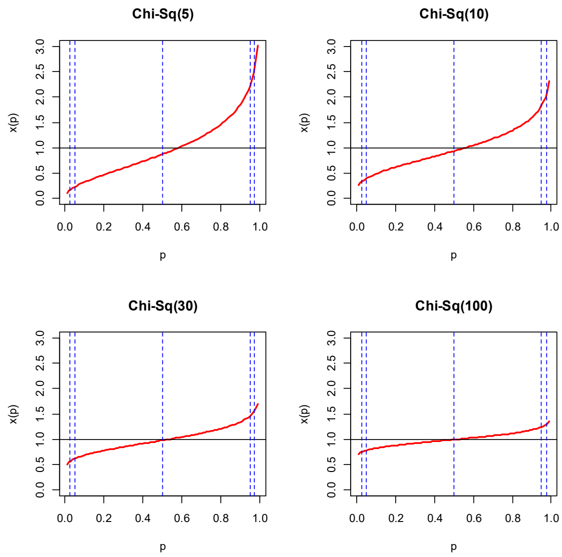
```
# 输入自由度，累积概率
nu <- c(5, 10, 30, 100)
p <- 0.95

# 分位数 (常数a)
qchisq(p, nu)
[1]  11.07050  18.30704  43.77297 124.34211

# 期望对比95%分位数的比率
qchisq(p, nu) / nu
[1] 2.214100 1.830704 1.459099 1.243421
```

☞ 由此可知，自由度越大，期望对比95% 分位数的比率越接近1，越向中心集中。

26/33

26

[例 8-14] 请绘制自由度为5, 10, 30, 100的卡方分布的分位数/期望曲线并比较。



27/33

27

4. t-分布

[定义 8-4] t-分布

: 符合标准正态分布的随机变量Z，与Z相互独立的符合自由度为ν的卡方分布的随机变量Y，则有

$$T \equiv \frac{Z}{\sqrt{Y/\nu}} \sim t(\nu)$$

- 应用

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\dfrac{S}{\sigma}} = \frac{\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2(n-1)}}} = \frac{Z}{\sqrt{Y/(n-1)}} \sim t(n-1)$$

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

28/33

28

4. t-分布

# t-分布的概率密度函数 f(x) (df=自由度， ncp=非中心参数(未使用))
dt(x, df, ncp)
# Excel 函数 = T.DIST(x, df, FALSE)

# t-分布的累积分布函数F(x)
pt(x, df, ncp, lower.tail = TRUE)
# Excel 函数 = T.DIST(x, df, TRUE)

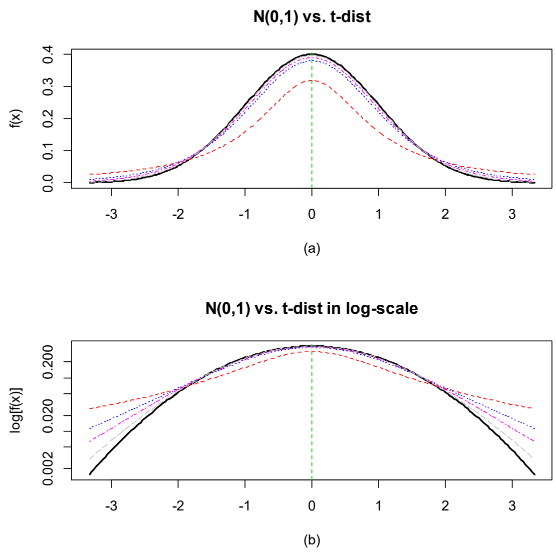
# t-分布的分位数(p=累积概率)
qt(p, df, ncp, lower.tail = TRUE)
# Excel 函数 = T.INV(p, df)

# t-分布的随机变量 (n=随机数的个数)
rt(n, df, ncp)
# Excel 函数 = T.INV(RAND( ), df) ⇒ 生成1个随机数

29/33

29

[例 8-15] 请绘制标准正态分布和自由度为1, 5, 10, 30的t-分布的概率密度函数，并比较区间(-1, 1), (-2, 2), (-3, 3)的概率。

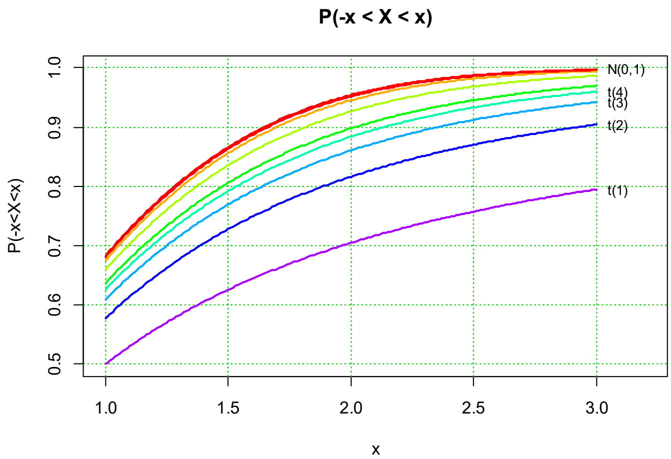


30/33

30

[例 8-16] 绘制t-分布分位数[表]　$P(T > t_{1-\alpha;\nu}) = \alpha$

[例 8-17] 在标准正态分布和t-分布上标示有关1到3的x的趋中概率(ν=1, 2, 3, 4, 5, 10, 30)



31/33

31

5. F-分布

[定义 8-5] F-分布

$$U \sim \chi^2(\nu_1) \perp V \sim \chi^2(\nu_2) \Rightarrow F \equiv \frac{U/\nu_1}{V/\nu_2} \sim F(\nu_1,\nu_2)$$

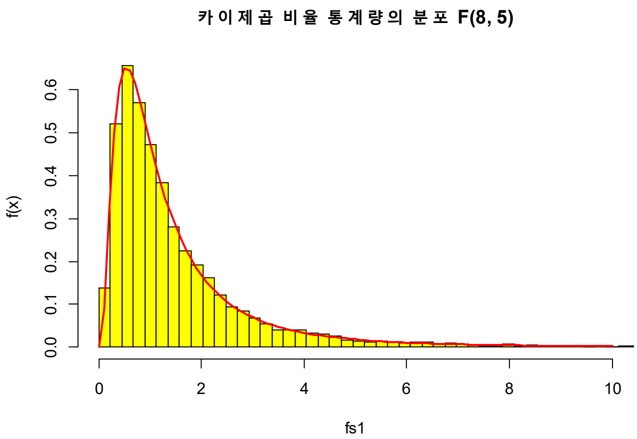
$$\alpha \equiv P(F_{\nu_1,\nu_2} < F_{\alpha;(\nu_1,\nu_2)}) = P\left(F_{\nu_2,\nu_1} > \frac{1}{F_{\alpha;(\nu_1,\nu_2)}}\right) \equiv P\left(F_{\nu_2,\nu_1} > F_{1-\alpha;(\nu_2,\nu_1)}\right)$$

```
# F-分布的R 函数
# F-分布的概率密度函数 (df1, df2 =自由度, ncp=非中心参数(不使用))
df(x, df1, df2, ncp)
# Excel 函数 = F.DIST(x, df1, df2, FALSE)
# F-分布的累积分布函数 F(x)
pf(x, df1, df2, ncp, lower.tail = TRUE)
# Excel 函数 = F.DIST(x, df1, df2, TRUE)
# F-分布的分位数 (p=累积概率)
qf(p, df1, df2, ncp, lower.tail = TRUE)
# Excel 函数 = F.INV(p, df1, df2)
# F-分布的随机变量 (n=随机数的个数)
rf(n, df1, df2, ncp)
# Excel 函数 = F.INV(RAND( ), df1, df2)
```

32/33

32

[例 8-18] 从自由度分别为8和5的独立**卡方分布**总体中各抽取10,000个样本，求标准化比率并绘制直方图，以及确定[**定义** 8-5]中的F(8,5) 分布。



[例 8-19] 绘制F-分布的分位数[表] $\alpha \equiv P(F_{\nu_1,\nu_2} > F_{1-\alpha;(\nu_1,\nu_2)})$