大数据的统计学基础第一次作业

姓名：蒋贵豪　　学号：B+X9bo

目录

1 2021.09.14 作业

1.1 例 1.3.4 重现

给定一组某学院学生的体测数据，如表 1 所示。其中包含了体重 X_1、腰围 X_2、性别 X_3 和班级 X_4。随机抽取 40 人组成一个容量 40 的多维样本。绘制如下可视化图形：

1). X_1 的直方图（含核密度曲线估计）.

2). X_2 的经验分布图.

3). X_3 的条形图.

4). X_1 和 X_2 的散点图（带拟合直线）.

5). X_1 和 X_2 的二维等高线图.

6). X_1 和 X_2 的 Q-Q 图.

7). X_3 和 X_4 的分组条形图.

8). X_3 和 X_1 的分组箱线图.

9). X_4 和 X_2 的分组箱线图.

表 1: 学生体测数据

编号	体重	腰围	性别	班级	编号	体重	腰围	性别	班级
1	101	25	女	A	21	100	25	女	A
2	119	27	女	A	22	168	37	男	A
3	143	33	男	C	23	143	33	男	B
4	162	35	男	B	24	122	30	男	B
5	98	25	女	A	25	111	28	女	C
6	122	29	男	B	26	167	38	男	B
7	135	32	男	B	27	189	43	男	C
8	144	33	男	B	28	147	33	男	B
9	141	33	男	A	29	99	23	女	C
10	180	42	男	B	30	156	38	男	B
11	135	32	男	C	31	131	30	男	C
12	130	32	女	A	32	101	25	女	C
13	154	34	男	B	33	118	27	女	B
14	88	23	女	A	34	176	40	男	C
15	107	26	女	A	35	133	32	男	C
16	125	27	男	B	36	100	25	女	B
17	114	27	女	A	37	157	36	男	C
18	157	34	男	C	38	97	24	女	A
19	142	33	男	A	39	103	25	女	A
20	155	34	男	C	40	109	26	女	C

R 语言代码如下：

```
library(latex2exp)
library(showtext)

x1 <- c(101,119,143,162,98,122,135,144,141,180,135,130,154,88,107,125,
      114,157,142,155,100,168,143,122,111,167,189,147,99,156,131,101,
      118,176,133,100,157,97,103,109)
x2 <- c(25,27,33,35,25,29,32,33,33,42,32,32,34,23,26,27,27,34,33,34,
      25,37,33,30,28,38,43,33,23,38,30,25,27,40,32,25,36,24,25,26)
```

```
x3 <- c('女','女','男','男','女','男','男','男','男','男','男',
      '女','男','女','女','男','女','男','男','男','女','男',
      '男','男','女','男','男','男','女','男','男','女','女',
      '男','男','女','男','女','女','女')
x4 <- c('A','A','C','B','A','B','B','B','A','B','C','A','B','A',
      'A','B','A','C','A','C','A','A','B','B','C','B','C','B',
      'C','B','C','C','B','C','C','B','C','A','A','C')

DATA1 <- data.frame(x1,x2)
DATA2 <- data.frame(x3,x1)
DATA3 <- data.frame(x3,x2)
DATA4 <- data.frame(x1)
DATA5 <- data.frame(x2)
DATA6 <- data.frame(x3)
DATA7 <- data.frame(x3,x4)
DATA8 <- data.frame(x4,x2)

#x1 体重指数直方图绘制
library(ggplot2)
p1 <- ggplot(DATA4, aes(x = x1, y = ..density..)) +
  geom_histogram(binwidth = 20, fill = "lightblue") +
  geom_density() + xlab(TeX('x_1'))

# 绘制 x2 腰围的经验分布图
p2 <- ggplot(DATA5, aes(x = x2)) + stat_ecdf(aes(x2)) +
  xlab(TeX('x_2')) + ylab(TeX('F(x)'))

#X3 性别的条形图绘制
p3 <- ggplot(DATA6, aes(x = x3)) + geom_bar(stat = "count") +
  xlab(TeX('x_3'))

# 绘制 x2 腰围和 x1 体重的带拟合直线的散点图
p4 <- ggplot(DATA1, aes(x1, x2)) +
  geom_point() + xlab(TeX('x_1')) + ylab(TeX('x_2')) +
  geom_smooth(method = "lm")

## 绘制 x1 体重和 x2 腰围的点图和二维等高线
p5 <- ggplot(DATA1, aes(x = x1, y = x2)) +
  geom_point() + stat_density2d() +
  xlab(TeX('x_1')) + ylab(TeX('x_2'))
```

```
## 绘制 x1 和 x2 的 QQ 图
library(tibble)
m <- qqplot(scale(x1), scale(x2))
```

```
X1 = m$x
X2 = m$y
p6 <- ggplot(tibble(X1), aes(sample = X2)) + stat_qq() +
  stat_qq_line() + xlab(TeX('x_1')) + ylab(TeX('x_2'))

## 绘制 x3、x4 分组条形图
p7 <- ggplot(data = DATA7, mapping = aes(x = x4, fill = x3)) +
  labs(fill = TeX('x_3')) + geom_bar(stat = "count",
                                     width = 0.5, position = 'dodge') +
  geom_text(stat = 'count', aes(label = ..count..),
            color = "black", size = 3.5, position = position_dodge(0.5),
            vjust = -0.5) + xlab(TeX('x_4'))

## 绘制 x1 体重和 x3 性别的点图和箱线图
p8 <- ggplot(DATA2, aes(x = x3, y = x1)) +
  geom_boxplot(aes(x = x3, group = x3),width = .25,
               # 设置箱图填充色，边框色
               fill = 'cornsilk', colour = 'grey60') +
  geom_dotplot(aes(x = x3, group = x3),
               # 以 Y 轴堆叠，宽度，类型
               binaxis = 'y', binwidth = .5, stackdir = 'center',
               # 设置填充色
               fill='red') + xlab(TeX('x_3')) + ylab(TeX('x_1'))

## 绘制 x2 和 x4 的点图和箱线图
p9 <- ggplot(DATA8, aes(x = x4, y = x2)) +
  geom_boxplot(aes(x = x4, group = x4),width = .25,
               # 设置箱图填充色，边框色
               fill = 'cornsilk', colour = 'grey60') +
  geom_dotplot(aes(x = x4, group = x4),
               # 以 Y 轴堆叠，宽度，类型
               binaxis = 'y', binwidth = .5, stackdir = 'center',
               # 设置填充色
               fill='red') + xlab(TeX('x_4')) + ylab(TeX('x_2'))

# 将九张图绘制在同一个画板上
library("gridExtra")
```

```
library(showtext)
grid.arrange(p1, p2, p3, p4, p5, p6, p7, p8, p9, ncol = 3, nrow = 3 )
```

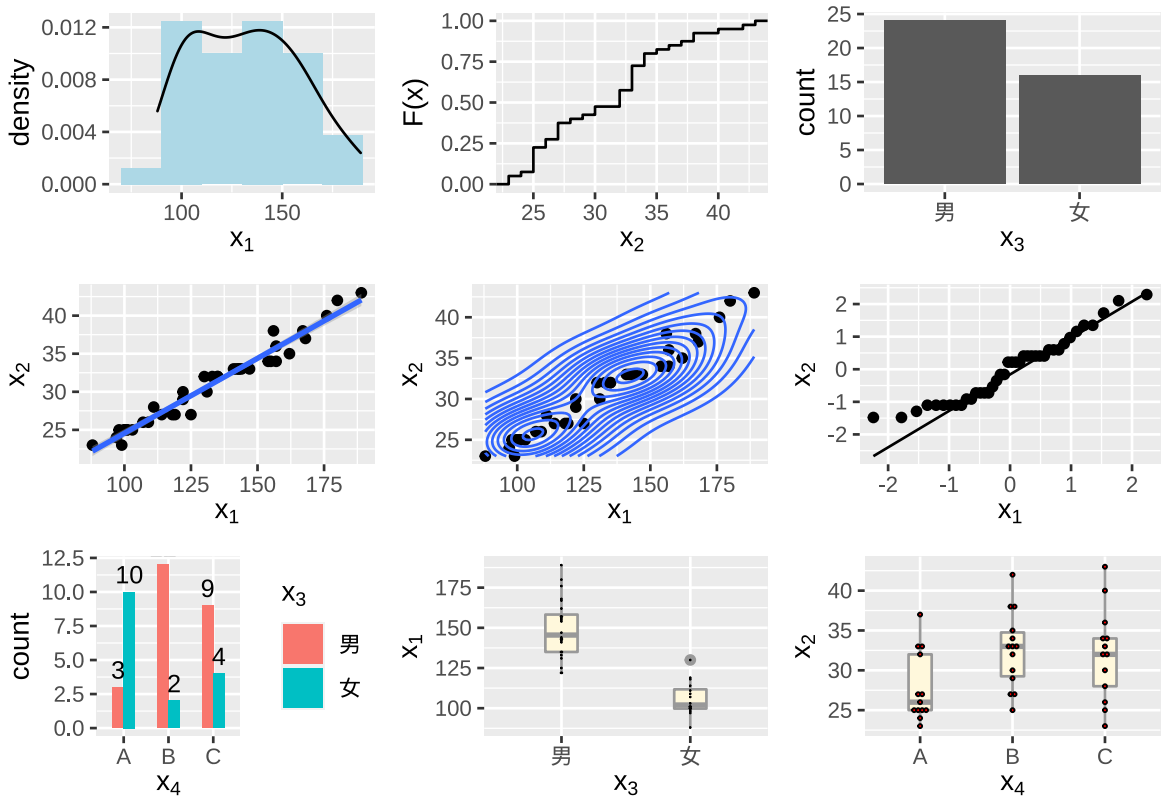

图 1: 高维图展示

1.2 绘制标准正态的经验分布函数与总体分布函数

对一个服从标准正态分布的随机变量 $X \sim N(0,1)$ 随机采样 n 次，绘制 $n = 10$（红色）与 $n = 50$（蓝色）情况下的经验分布函数，黑色为标准正态总体分布函数。R 代码如下：

```
library(latex2exp)
x <- rnorm(10,0,1)
plot(ecdf(x), do.points = FALSE, verticals = TRUE,
     xlim = c(-5,5), main = "", xlab = '', ylab = '',
     col = 'red')
par(new = TRUE)
y <- rnorm(50,0,1)
plot(ecdf(y), do.points = FALSE, verticals = TRUE,
     xlim = c(-5,5), main = "", xlab = '', ylab = '',
     col = 'blue')
par(new = TRUE)
z1 <- seq(-5,5, length.out = 100)
```

```
z2 <- pnorm(z1,0,1)
plot(z1, z2, type = 'l', main = "", xlab = TeX("x"),
     ylab = TeX("F(x)"))
legend(-5,1, col = c("red", "blue"),
       lty = c(1,1), lwd = c(2,2),
       legend = c("n = 10", "n = 50"))
```

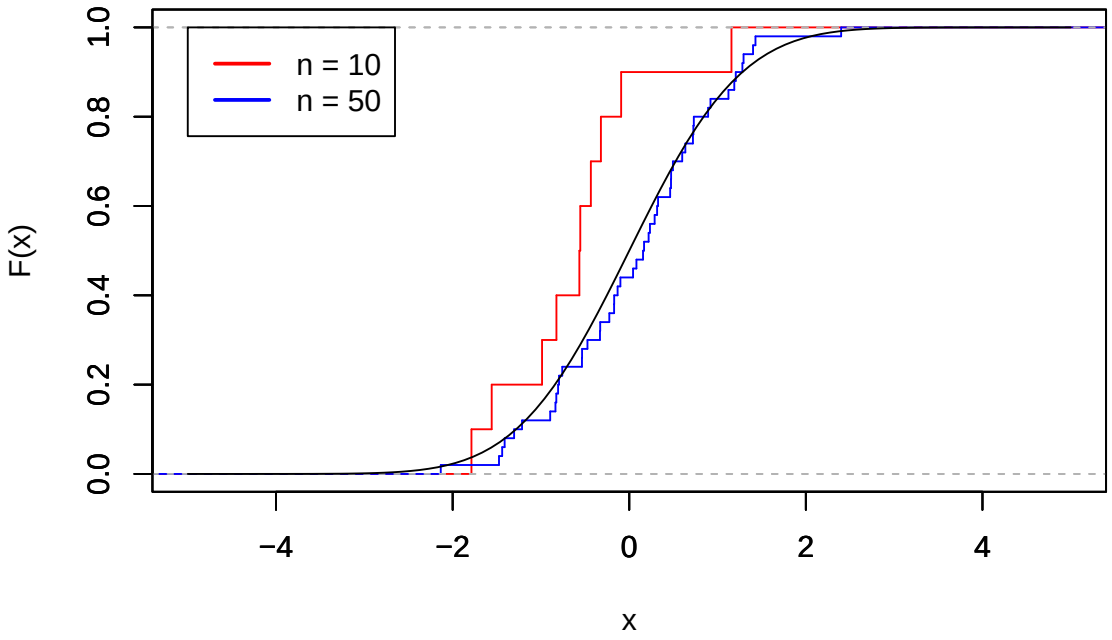

图 2: 标准正态的经验分布函数与总体分布函数

从图 2 中可以看出，当 $n = 10$ 时，经验分布函数与真实分布差距较大。当 $n = 50$ 时，经验分布函数与真实分布已较为接近。随着样本量 n 的增大，经验分布函数越来越接近真实的标准正态分布函数。

1.3 用随机模拟法求标准正态总体 $N(0,1)$ 的样本峰度的分布

样本峰度 γ_2 的定义是 $\gamma_2 = \mu_4/\sigma^4 - 3$，其中，$\mu_4$ 为四阶中心矩，σ^4 为标准差的四次方。我们知道，对于一个服从正态分布的随机变量 $X_i \sim N(0,1)$, 我们可以抽取 n 个样本，计算该随机变量的样本峰度 $\gamma_{2,i}$。然后我们对于 N 个独立同分布于标准正态分布的随机变量 X_i, 我们均计算其样本峰度 $\gamma_{2,i}$。理论上已经证明：对于这 N 个样本峰度 $\gamma_{2,i}$，其渐进分布为 $\gamma_2 \sim AN(0, 24/n)$。我们用如下 R 代码进行此过程的随机模拟，并给出渐进结果图。

```
library(latex2exp)
set.seed(0)
b = vector()
n = 400
for (i in 1:10000){
  s = rnorm(n, mean = 0, sd = 1)
  numerator = sum((s-mean(s))^4)/n
  denominator = (sum((s-mean(s))^2)/n)^2
  b[i] = numerator/denominator-3}
par(mfrow = c(2,2))
plot(density(b), main = "", xlab = TeX("n = 400"), xlim = c(-2,2),
     ylim = c(0,1.8), col = 'red')
par(new = TRUE)
z1 <- seq(-2,2, length.out = 100)
y <- dnorm(z1,0, sqrt(24/n))
Y <- pnorm(z1,0, sqrt(24/n))
plot(z1, y, type = 'l', xlim = c(-2,2), ylim = c(0,1.8), main = "",
     xlab = '', ylab = '', col = 'blue')
plot(ecdf(b), do.points = FALSE, verticals = TRUE,
     xlim = c(-2,2), main = "", xlab = TeX("n = 400"),
     ylab = TeX('F(x)'),col = 'red')
par(new = TRUE)
plot(z1, Y, type = 'l', xlim = c(-2,2), main = "",
     xlab = '', ylab = '', col = 'blue')
set.seed(1)
b = vector()
n = 3600
for (i in 1:10000){
  s = rnorm(n, mean = 0 , sd = 1)
  numerator = sum((s-mean(s))^4)/n
  denominator = (sum((s-mean(s))^2)/n)^2
  b[i] = numerator/denominator-3}
plot(density(b), main = "", xlab = TeX("n = 3600"), xlim = c(-0.5,0.5),
     ylim = c(0,5), col = 'red')
par(new = TRUE)
z1 <- seq(-0.5,0.5, length.out = 100)
y <- dnorm(z1,0, sqrt(24/n))
Y <- pnorm(z1,0, sqrt(24/n))
plot(z1, y, type = 'l', xlim = c(-0.5,0.5), ylim = c(0,5),
     main = "", xlab = '', ylab = '', col = 'blue')
plot(ecdf(b), do.points = FALSE, verticals = TRUE,
```

```
     xlim = c(-0.5,0.5), main = "", xlab = TeX("n = 3600"),
     ylab = TeX('F(x)'), col = 'red')
par(new = TRUE)
plot(z1, Y, type = 'l', xlim = c(-0.5,0.5),
     main = "", xlab = '', ylab = '', col = 'blue')
```

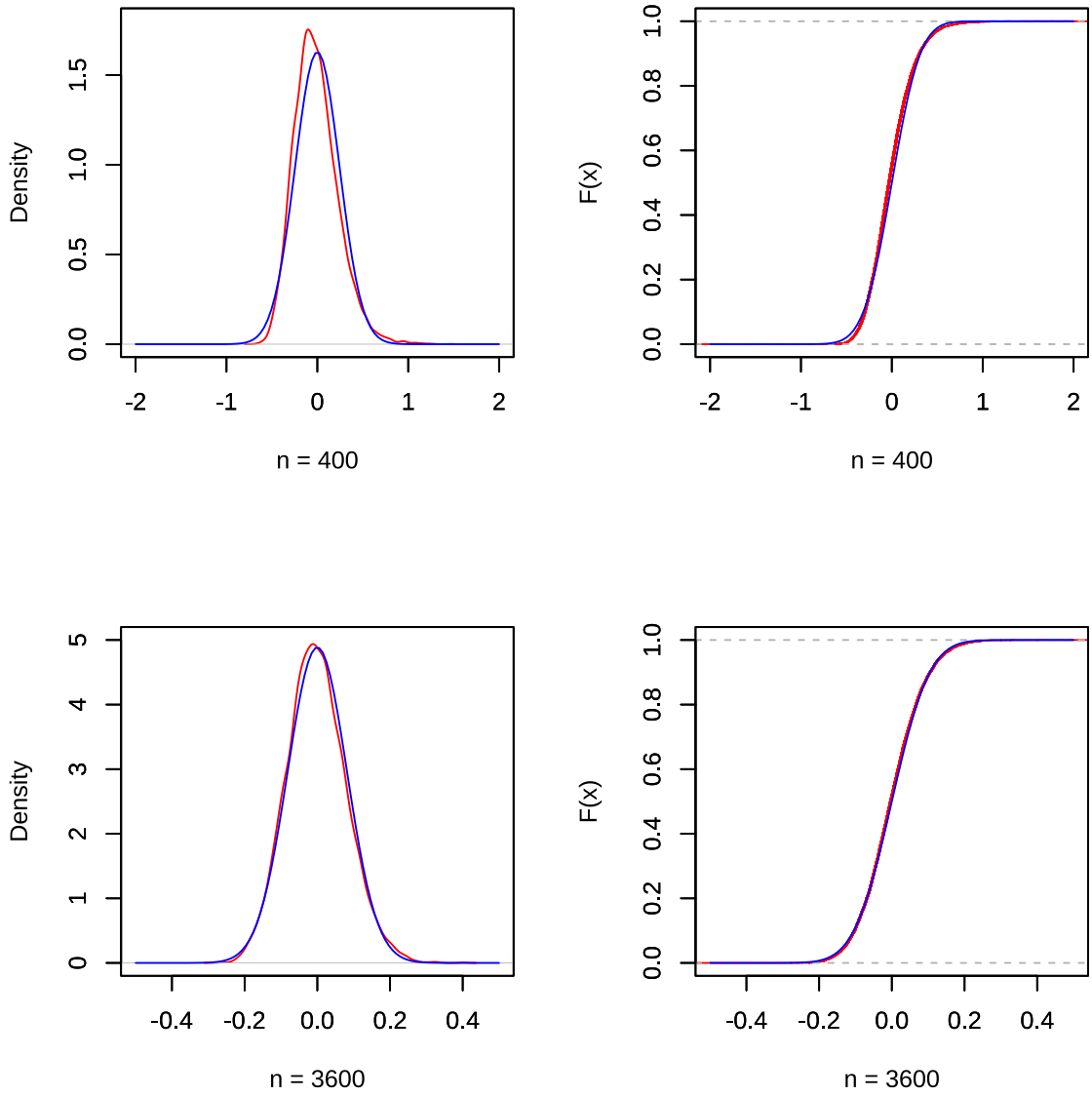

图 3: 样本峰度渐进分布函数与正态函数对比

从图 3 中可以看到，无论是从密度函数还是累计密度函数上看，运用随机模拟的方法，样本峰度 γ_2 的密度函数确实逼近于 $N(0, 24/n)$，并且 $n = 400$ 时，已经有较好的逼近效果，但是仍然有一定的误差。当 $n = 3600$ 时，逼近效果较好。因此 $\gamma_2 \sim AN(0, 24/n)$。

2 2021.09.28 作业

2.1 三个囚犯问题的推广

有三个囚犯 A，B，C 被处以死刑。有一天，政府打算随机赦免其中一人。于是，A 问监狱长自己是不是被释放。但是，政府要求对赦免名单进行保密。于是，监狱长没有告诉 A 是否被释放，而是告诉他 B 被处死了。以 A,B,C 来表示 A，B，C 被释放的事件, $\mathcal{W}$ 表示监狱长告诉 A，B 被处死的事件。

(a) 由于监狱长可以告诉 A 是 B 被处死还是 C 被处死，这两个事件是等概率的。现在我们假设监狱长可以自己给定 B 和 C 被处死的概率 γ 和 $1-\gamma$，如表 2 所示：

表 2: 事件概率表

释放的人	监狱长告诉 A	
A	B 处死	概率为 γ
A	C 处死	概率为 $1-\gamma$
B	C 处死	
C	B 处死	

计算 γ 的函数 $P(A|\mathcal{W})$，并说明当 γ 取何值时，$P(A|\mathcal{W})$ 是小于、等于或大于 $\frac{1}{3}$.

(b) 假设 $\gamma=\frac{1}{2}$，证明：此时假如 A 选择和 C 交换命运，则 A 被释放的概率为 $\frac{2}{3}$.

回答：

(a) 由于

$$P(A|\mathcal{W})=\frac{P(A,\mathcal{W})}{P(\mathcal{W})}$$

其中，

$$P(A,\mathcal{W})=\frac{\gamma}{3}$$

$$\begin{aligned}P(\mathcal{W})&=P(\mathcal{W}|A)P(A)+P(\mathcal{W}|B)P(B)+P(\mathcal{W}|C)P(C)\\&=\gamma\times\tfrac{1}{3}+0\times\tfrac{1}{3}+1\times\tfrac{1}{3}\\&=\tfrac{1+\gamma}{3}\end{aligned}$$

于是，

$$P(A|\mathcal{W})=\frac{\gamma}{1+\gamma}$$

从而，当 $0<\gamma<\frac{1}{2}$ 时，$P(A|\mathcal{W})$ 小于 $\frac{1}{3}$. 当 $\gamma=\frac{1}{2}$ 时，$P(A|\mathcal{W})$ 等于 $\frac{1}{3}$. 当 $\frac{1}{2}<\gamma<1$ 时，$P(A|\mathcal{W})$ 大于 $\frac{1}{3}$.

(b) 由于 $\gamma=\frac{1}{2}$, 我们知道

$$P(A|\mathcal{W})+P(B|\mathcal{W})+P(C|\mathcal{W})=1$$

而

$$P(A|\mathcal{W}) = \frac{1}{3}, P(B|\mathcal{W}) = 0$$

于是

$$P(C|\mathcal{W}) = \frac{2}{3}$$

从而，A 和 C 交换后，释放的概率变为 $\frac{2}{3}$.

2.2 习题 4.13

设 X 和 Y 是具有有限期望的随机变量，

(a) 证明

$$\min_{g(x)} \mathrm{E}(Y - g(X))^2 = \mathrm{E}(Y - \mathrm{E}(Y|X))^2$$

其中 $g(x)$ 取遍所有函数 ($\mathrm{E}(Y|X)$ 有时称作 Y 关于 X 的回归，它表示在已知 X 的条件下对 Y 作出的" 最好" 预测)

(b) 证明

$$\min_{b} \mathrm{E}(X - b)^2 = \mathrm{E}(X - \mathrm{E}(X))^2$$

可以作为 (a) 的一个特例导出.

回答：

(a)

$$\begin{aligned}&\mathrm{E}(Y - g(X))^2 = \mathrm{E}((Y - \mathrm{E}(Y|X)) + (\mathrm{E}(Y|X) - g(X)))^2 \\ &= \mathrm{E}(Y - \mathrm{E}(Y|X))^2 + \mathrm{E}(\mathrm{E}(Y|X) - g(X))^2 + 2\mathrm{E}[(Y - \mathrm{E}(Y|X))(\mathrm{E}(Y|X) - g(X))]\end{aligned}$$

其中，交叉项 $\mathrm{E}[(Y - \mathrm{E}(Y|X))(\mathrm{E}(Y|X) - g(X))]$ 由重期望公式为 0.

于是

$$\mathrm{E}(Y - g(X))^2 = \mathrm{E}(Y - \mathrm{E}(Y|X))^2 + \mathrm{E}(\mathrm{E}(Y|X) - g(X))^2 \geq \mathrm{E}(Y - \mathrm{E}(Y|X))^2$$

等号当且仅当 $g(X) = \mathrm{E}(Y|X)$ 取到，于是：

$$\min_{g(x)} \mathrm{E}(Y - g(X))^2 = \mathrm{E}(Y - \mathrm{E}(Y|X))^2$$

(b) 我们在 (a) 中，取 $g(x)$ 恒为常数 b, 且有 $\mathrm{E}(Y|b) = \mathrm{E}(Y)$, 于是：

$$\min_{b} \mathrm{E}(Y - b)^2 = \mathrm{E}(Y - \mathrm{E}(Y))^2$$

2.3 习题 4.28

设 X 和 Y 是一对独立的标准正态随机变量，

(a) 证明 $X/(X+Y)$ 服从 Cauchy 分布；

(b) 求 $X/|Y|$ 的分布；

(c) 给出 (b) 更为一般的定理.

回答：

(a) 设 $U=X+Y$, $V=X/(X+Y)$, 于是 $x=uv$, $y=u(1-v)$, 从而

$$|J|=\begin{vmatrix} v & u \\ 1-v & -u \end{vmatrix}=|u|$$

由

$$f_{X,Y}(x,y)=\frac{1}{2\pi}\mathrm{e}^{-\frac{x^2+y^2}{2}}$$

我们有：

$$f_{U,V}(u,v)=\frac{1}{2\pi}\mathrm{e}^{-\frac{(uv)^2+(u(1-v))^2}{2}}|u|=\frac{1}{2\pi}\mathrm{e}^{-\frac{u^2(1-2v+2v^2)}{2}}|u|$$

于是：

$$p(v)=\int_{-\infty}^{+\infty}\frac{1}{2\pi}\mathrm{e}^{-\frac{u^2(1-2v+2v^2)}{2}}|u|\,\mathrm{d}u=\int_{0}^{+\infty}\frac{1}{\pi}\mathrm{e}^{-\frac{u^2(1-2v+2v^2)}{2}}u\mathrm{d}u=\int_{0}^{+\infty}\frac{1}{2\pi}\mathrm{e}^{-\frac{z(1-2v+2v^2)}{2}}\mathrm{d}z$$

计算得：

$$p(v)=\frac{1}{\pi(1-2v+2v^2)}=\frac{1}{\pi}\frac{\frac{1}{2}}{\left(\frac{1}{2}\right)^2+\left(v-\frac{1}{2}\right)^2},-\infty<v<+\infty$$

于是，$X/(X+Y)$ 服从 $\lambda=1/2$, $\mu=1/2$ 的 Cauchy 分布.

(b) 设 $U=X/|Y|$, $V=|Y|$, 于是我们有：$x=uv$, $y=\pm v$, 从而

$$|J|=\begin{vmatrix} v & u \\ 0 & \pm 1 \end{vmatrix}=|v|$$

由

$$f_{X,Y}(x,y)=\frac{1}{2\pi}\mathrm{e}^{-\frac{x^2+y^2}{2}}$$

于是 U,V 的联合分布为：

$$f_{U,V}(u,v)=\frac{1}{2\pi}\mathrm{e}^{-\frac{(uv)^2+(v)^2}{2}}|v|+\frac{1}{2\pi}\mathrm{e}^{-\frac{(-uv)^2+(-v)^2}{2}}|v|=\frac{v}{\pi}\mathrm{e}^{-\frac{v^2(1+u^2)}{2}},-\infty<u<+\infty,0<v<+\infty$$

计算得：

$$p(u)=\int_{0}^{+\infty}\frac{1}{\pi}\mathrm{e}^{-\frac{(u^2+1)v^2}{2}}v\mathrm{d}v=\int_{0}^{+\infty}\frac{1}{2\pi}\mathrm{e}^{-\frac{(u^2+1)z}{2}}\mathrm{d}z=\frac{1}{\pi(u^2+1)},-\infty<u<+\infty$$

于是，$X/|Y|$ 服从 $\lambda=1$, $\mu=0$ 的 Cauchy 分布.

(c) 更为一般的定理：设 X 和 Y 是一对独立的标准正态随机变量，则 $X/\sqrt{(aX)^2+(bY)^2}$, $a^2+b^2=1$, 服从 Cauchy 分布.

2.4 习题 4.30

假设 Y 在条件 $X = x$ 下的条件分布是 $N(x, x^2)$, 且 X 的边缘分布是 $(0, 1)$ 区间上的均匀分布. 求 $\mathrm{E}Y$, $\mathrm{Var}Y$, $\mathrm{Cov}(X, Y)$.

回答：

$$\mathrm{E}Y = \mathrm{E}[\mathrm{E}(Y|X)] = \mathrm{E}X = \tfrac{1}{2}$$
$$\mathrm{Var}Y = \mathrm{Var}[\mathrm{E}(Y|X)] + \mathrm{E}[\mathrm{Var}(Y|X)] = \mathrm{Var}X + \mathrm{E}X^2 = \tfrac{1}{12} + \tfrac{1}{3} = \tfrac{5}{12}$$
$$\mathrm{E}XY = \mathrm{E}[\mathrm{E}(XY|X)] = \mathrm{E}[X\mathrm{E}(Y|X)] = \mathrm{E}X^2 = \tfrac{1}{3}$$
$$\mathrm{Cov}(X, Y) = \mathrm{E}XY - \mathrm{E}X\mathrm{E}Y = \tfrac{1}{3} - \tfrac{1}{4} = \tfrac{1}{12}$$

2.5 习题 4.41

证明：任意随机变量与常量都不相关.

回答：

$$\mathrm{Cov}(X, c) = \mathrm{E}cY - \mathrm{E}c\mathrm{E}Y = c\mathrm{E}Y - c\mathrm{E}Y = 0$$

其中，c 为常数。从而，任意随机变量与常量都不相关.

2.6 习题 4.44

证明：对任意随机向量 $(X_1, ..., X_n)$, 有：

$$\mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{Var}X_i + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}(X_i, X_j)$$

回答：

设 $\mu_i = \mathrm{E}X_i$, 则有：

$$\begin{aligned}
\mathrm{Var}(\sum_{i=1}^{n} X_i) =& \mathrm{E}(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i)^2 = \mathrm{E}[\sum_{i=1}^{n} (X_i - \mu_i)]^2 \\
=& \sum_{i=1}^{n} \mathrm{E}(X_i - \mu_i)^2 + 2 \sum_{1 \le i < j \le n} \mathrm{E}(X_i - \mu_i)(X_j - \mu_j) \\
=& \sum_{i=1}^{n} \mathrm{Var}X_i + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}(X_i, X_j)
\end{aligned}$$

2.7 习题 5.29

某工厂生产一种小册子并将其按每箱 100 册的数量打包. 已知小册子重量的均值为 1 盎司，标准差为 0.05 盎司. 厂家希望计算 P(100 本小册子的重量超过 100.4 盎司) 的值，以帮助检测每个箱子中的小册子是否有多. 说说你会怎么样近似地计算这一概率，并指出假设和使用到的定理.

回答：

设 X_i 为第 i 本小册子的重量，则有：$\mathrm{E}X_i = 1$, $\sigma_i = 0.05$，假设 X_i 独立同分布。则有：

$$P(\sum_{i=1}^{100} X_i > 100.4) = P(\bar{X} > 1.004)$$

由

$$\bar{X} \sim N(1, 0.05^2/100)$$

于是：

$$P(\bar{X} > 1.004) = P(Y > (1.004-1)/(0.05/10)) = P(Y > 0.8) = 0.2119$$

用到了棣莫弗-拉普拉斯极限定理。

2.8 习题 5.30

设有两个独立的大小为 n 的随机样本取自方差为 σ^2 的总体，样本均值分别为 $\bar{X}_1$ 和 $\bar{X}_2$, 求 n 使 $P(\left|\bar{X}_1 - \bar{X}_2\right| < \sigma/5) \approx 0.99$, 说明你的结论的合理性.

回答：

由于 $\bar{X}_i \sim N(\mu, \sigma^2/n), i = 1, 2$, 而 $\bar{X}_1$ 和 $\bar{X}_2$ 独立同分布，于是 $\bar{X}_1 - \bar{X}_2 \sim N(0, 2\sigma^2/n)$，故有：

$$\begin{aligned} &P(\left|\bar{X}_1 - \bar{X}_2\right| < \sigma/5) = P(\tfrac{-\sigma/5}{\sqrt{2}\sigma/\sqrt{n}} < \tfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{2}\sigma/\sqrt{n}} < \tfrac{\sigma/5}{\sqrt{2}\sigma/\sqrt{n}}) \\ &= P(-\tfrac{1}{5}\sqrt{\tfrac{n}{2}} < Y < \tfrac{1}{5}\sqrt{\tfrac{n}{2}}) = 0.99 \end{aligned}$$

其中，$Y \sim N(0, 1)$, 于是 $P(Y \geq \frac{1}{5}\sqrt{\frac{n}{2}}) = 0.005$, 查表得 $\frac{1}{5}\sqrt{\frac{n}{2}} = 2.576$, 于是 $n = 50 \times 2.576^2 \approx 332$

2.9 习题 5.34

设随机变量 $X_1, ..., X_n$ 取自均值为 μ、方差为 σ^2 的总体，证明：

$$\mathrm{E}\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = 0, \mathrm{Var}\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = 1$$

即中心极限定理中对 $\bar{X}_n$ 进行标准化后与极限 $N(0, 1)$ 分布有相同的均值和方差.

回答：

由题意我们可以知道：$\mathrm{E}\bar{X}_n = \mu$,$\mathrm{Var}\bar{X}_n = \sigma^2/n$, 于是我们有：

$$\mathrm{E}\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{\sqrt{n}}{\sigma}\mathrm{E}(\bar{X}_n - \mu) = \frac{\sqrt{n}}{\sigma}(\mu - \mu) = 0$$

$$\mathrm{Var}\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{n}{\sigma^2}\mathrm{Var}(\bar{X}_n - \mu) = \frac{n}{\sigma^2}\mathrm{Var}(\bar{X}_n) = \frac{n}{\sigma^2}\frac{\sigma^2}{n} = 1$$

2.10 习题 5.43

补充定理 5.5.24 证明的细节：

定理 5.5.24：设随机变量序列 Y_n 满足：$\sqrt{n}(Y_n-\theta)$ 依分布收敛于 $N(0,\sigma^2)$, 函数 g 在指定的 θ 处满足：$g'(\theta)$ 存在且不为 0，则：

$$\sqrt{n}[g(Y_n)-g(\theta)] \xrightarrow{d} N(0,\sigma^2[g'(\theta)]^2)$$

(a) 证明：如果 $\sqrt{n}(Y_n-\mu)$ 依分布收敛于 $N(0,\sigma^2)$, 则 Y_n 依概率收敛于 μ.

(b) 详述 Slutsky 定理在证明中的作用.

回答：

(a)

$$\lim_{n\to\infty} P\left(|Y_n-\theta|<\varepsilon\right) = \lim_{n\to\infty} P\left(\sqrt{n}\,|Y_n-\theta|<\sqrt{n}\varepsilon\right) = P(|X|<\infty) = 1$$

其中，$X\sim N(0,\sigma^2)$. 因此，$Y_n \xrightarrow{P} \theta$.

(b) 由 Slutsky 定理：

$$g'(\theta)\sqrt{n}(Y_n-\theta) \xrightarrow{d} g'(\theta)X$$

其中：$X\sim N(0,\sigma^2)$，因此，我们有：

$$\sqrt{n}[g(Y_n)-g(\theta)] = g'(\theta)\sqrt{n}(Y_n-\theta)+\Delta \xrightarrow{d} N(0,\sigma^2[g'(\theta)]^2)$$

其中余项 $\Delta \xrightarrow{P} 0$