データ解析基礎（テキストの解答例）　　担当：北野 利一；　2006 年 1 月 17 日

【8.1】選挙の議席数予測

ベルヌーイ試行の繰り返しによる「成功」の総数は，２項分布で表現できる．したがって，成功率 prob=0.4 で，試行数 size=700 とすれば，成功の総数 x は，次のように図示できる（図 8.1）．

```
> x <- 250:330
> plot(x, pbinom(x, size=700, prob=0.4))
> abline(h=c(.025, .975), lty=2)
```

図から判断すれば，累積確率がおよそ 0.025 となる下限は，L = 254 となり，超過確率がおよそ 0.025 となる上限は，U = 305 である．これは，以下のように確認できる．

```
> pbinom(254, size=700, prob=0.4)
[1] 0.02410222
> pbinom(305, size=700, prob=0.4)
[1] 0.9750458
```

以上の事柄を図に依らずに，数式で算出しよう．そのためには，２項分布が正規分布に近似できることを利用すると都合がよい．すなわち，**1）**２項分布の平均は size * prob であり，その分散は size * prob * (1 - prob) であること，また，**2）**正規分布において，累積確率がおよそ 0.025 となる下限は，『平均 − 1.96 * √分散』，超過確率がおよそ 0.025 となる上限は，『平均 ＋ 1.96 * √分散』となることを利用すればよい．したがって，そのような上限および下限は，以下のように算定できる．

```
> 700 * 0.4 - 1.96 * sqrt(700 * 0.4 * 0.6)
[1] 254.5955
> 700 * 0.4 + 1.96 * sqrt(700 * 0.4 * 0.6)
[1] 305.4045
```

ここで，1.96 という値は，累積確率を 0.025 および 0.975 とする標準正規分布のパーセント点から，以下のように算出すればよい．

```
> round(qnorm(0.025), 2)
[1] -1.96
> round(qnorm(0.975), 2)
[1] 1.96
```

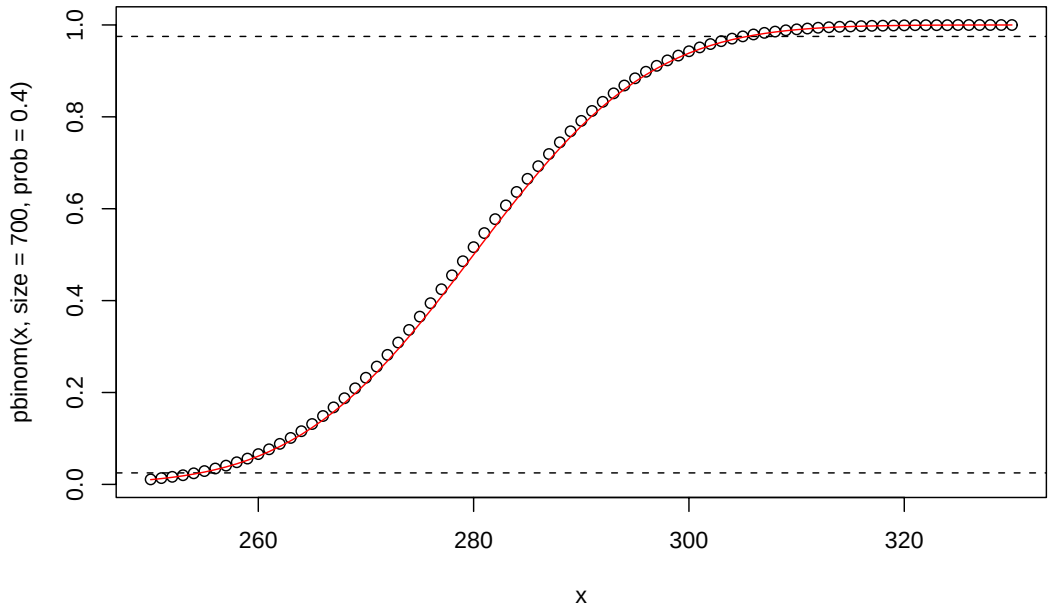

図 8.1 正規分布による２項分布の近似

なお，２項分布が正規分布に近似できることを図示で確認するためには，以下のようにすればよい．

```
> lines(x, pnorm(x, m=700 * 0.4, sd=sqrt(700 * 0.4 * 0.6)), col="red")
```

以上の事柄は，ある政党の７００人の立候補者に対して，当選率を平均的に 0.4 と見た時に，その政党の議席獲得数の分布を表していると解釈することもできる．しかし，現実的には，小選挙区制や比例代表制などのように，選挙制度は複雑であり，２項分布が単純に適用できるとは考えにくい．そもそも，当選を“確率”として扱うことに，当事者である政治家には抵抗があるかもしれない．もっとも，特定の議員個人の当否について，この解析は言及するものではない．当選がほぼ確実と見込まれる熟練議員もいて，当選が読めない激戦区の議員もいる．これらをならして平均の当選率としているのである．時代の寵児となった首相は，このような計算に長けているのかもしれない．

【8.3】野球の打率

今シーズンのヒット数を２項分布で表現する．３割バッター，すなわち，450 打数の内，

```
> 450 * 0.3
[1] 135
```

135 本以上のヒットを打てる確率は，打率 0.28 の打者の場合，

```
> 1 - pbinom(135, size=450, prob=0.28)
[1] 0.1592524
```

約 0.16 であることがわかる．２項分布を正規分布に近似した場合には，次のような値を得る．

```
> 1 - pnorm(135, m=450 * 0.28, sd=sqrt(450 * 0.28 * 0.72))
[1] 0.1723521
```

超過確率が 0.2 となる標準正規分布のパーセント点を求める．

```
> round(qnorm(1 - 0.2), 2)
[1] 0.84
```

これを利用し，２項分布を正規分布に近似して考えれば，打数 size に対して，size * 0.3 本以上のヒットを打てる確率が 0.2 となる（この時，パーセント点は 0.84 である）ためには，打数 size は，

```
size * 0.28 + 0.84 * sqrt(size * 0.28 * 0.72) = size * 0.3
```

を満足する必要がある．上式を整理して，打数 size は次のように得られる．

```
> 0.28 * 0.72 /(0.02/0.84)^2
[1] 355.6224
```

以下のように検算すれば，確かに，３割バッターとなる確率は 0.2 であることが確認できる．

```
> 1 - pbinom(355*0.3, size=355, prob=0.28)
[1] 0.1999426
```

なお，この場合，打率をベルヌーイ試行の成功率とみて，ヒット数を２項分布に従う確率変数と見なしている．このことは，比較的妥当な発想と考える．しかしながら，この適用に問題があることも考察しておくことが重要である．例えば，打率は，年間とおして一定値と扱えるのか？という疑問や，また，連続する打席に相関がないか？（つまり，三振した打席の後は，次もヒットが打てないこともあるし，ヒットの後は，ヒットがでやすい等，１つ前の打席の成績が影響しないか？）というような，統計的独立性が成立するかという疑問など，現実データに対して理論を適用する際に注意して考察すべきである．

【9.2】ねじの直径

標本平均および標本分散は，以下のように算出できる．

```
> nezi <- scan()
1: 1.22  1.24  1.25  1.19  1.17  1.18
7:
Read 6 items
> mean(nezi)
[1] 1.208333
> var(nezi)
[1] 0.001096667
```

上の算出法ではブラックボックスである. 具体的に何をしているのか, わからない人は, 以下を参考にせよ.

```
> sum(nezi)/length(nezi)
[1] 1.208333
> sum((nezi - .Last.value)^2)/(length(nezi) - 1)
[1] 0.001096667
```

【9.4】標本分散の不偏性

別紙を参照．なお，ポイントは，$E(XiXj) = \mu^2$ (i not= j), $= \mu^2 + \sigma^2$ (i = j) を利用するということである．

【9.5】２項分布の応用

３回信号を送って，正しい信号が送れた回数 x の分布は，２項分布に従う．

```
> dbinom(x, size=3, prob=0.9) -> prob.3
> names(prob.3) <- 0:3
> prob.3
    0     1     2     3
0.001 0.027 0.243 0.729
```

上記の表は, 各 x の値に対する確率を表している．本問のシステムでは, 回数の多い信号を送られた信号と見なすので，x >= 2 の場合が，正しい信号を認識したことになる．したがって，信号が正しく伝達される確率は,

```
> sum(prob.3[3:4])
[1] 0.972
```

となる．これは，以下のように算出することと同等である．

```
> 1 - pbinom(1, size=3, prob=0.9)
[1] 0.972
```

したがって，５回信号を送る場合，信号が正しく伝達される確率は，以下のとおり．

```
> 1 - pbinom(2, size=5, prob=0.9)
[1] 0.99144
```

【9.7】交通事故統計

手始めに，必要なデータを入力しよう．

```
> matrix(c(9.7, 4.0, 5.7, 7.8, 8.4,
+        526.6, 508.7, 703.8, 867.2, 621.6),
+ nrow=2, byrow=T, dimnames=list(c("die", "injure"),
+        c("Hokkaido","Tokyo","Osaka","Fukuoka","All.japan"))) -> traffic.acc
> traffic.acc
       Hokkaido Tokyo Osaka Fukuoka All.japan
die         9.7   4.0   5.7     7.8       8.4
injure    526.6 508.7 703.8   867.2     621.6
```

i) 年間の交通事故死亡者数が１０人未満となる確率について, 各都市の値は以下のとおりである. 計算に際して，１９８８年の統計値を平均発生率（lambda）として，ポアソン分布を適用している.

```
> round(ppois(9, lambda=traffic.acc["die","Hokkaido"]), 3)
[1] 0.496
> round(ppois(9, lambda=traffic.acc["die","Tokyo"]), 3)
[1] 0.992
> round(ppois(9, lambda=traffic.acc["die","Osaka"]), 3)
[1] 0.935
> round(ppois(9, lambda=traffic.acc["die","Fukuoka"]), 3)
[1] 0.741
```

ii) １日あたりの交通事故死傷者数が５人未満となる確率についても, 以下のとおりに算出できる. ここでは，年間あたりの統計値を１日あたりに換算し，平均発生率として与えている.

```
> round(ppois(4, lambda=traffic.acc["injure","Hokkaido"]/365), 3)
[1] 0.984
> round(ppois(4, lambda=traffic.acc["injure","Tokyo"]/365), 3)
[1] 0.986
> round(ppois(4, lambda=traffic.acc["injure","Osaka"]/365), 3)
[1] 0.954
> round(ppois(4, lambda=traffic.acc["injure","Fukuoka"]/365), 3)
[1] 0.907
```

【10.1】天秤での計測

１回毎の誤差は，N（0, 0.1）の正規分布に従うので，１０回計測したときの誤差は，中心極限定理（あるいは，独立な確率変数に対する，分散の加法性）により，N（0, 0.1/10），つまり，N（0, 0.01）の正規分布に従うことがわかる．したがって，重さ 100 g の物体を１０回計測したときの標本平均は，N（100, 0.1/10）の正規分布に従うといえる．よって，|X_bar - 100| > 0.3 となる確率は,

```
> pnorm(100 - 0.3, m=100, sd=sqrt(0.01))
[1] 0.001349898
```

となる下側の確率と，次の上側の確率の合計である.

```
> 1 - pnorm(100 + 0.3, m=100, sd=sqrt(0.01))
[1] 0.001349898
```

正規分布は対称な分布であるので，以下のようにすれば，上側と下側を合計する必要はない.

```
> pnorm(100.3, m=100, sd=sqrt(0.01), lower.tail=F) * 2
[1] 0.002699796
```

以上の考察は，次に示すコードを実行して得られる図 10.1 を参考にせよ．なお，確率 0.0027 は微細であるので，注意しないと見えない.

```
> par(mfrow=c(2,1)) -> op
> x <- seq(100-.5, 100+.5, by=0.01)
> plot(x, dnorm(x, m=100, sd=0.1), type="l")
> xl <- seq(100-.5, 100-.3, by=.01)
> polygon(c(xl, rev(xl)), c(dnorm(xl, m=100, sd=0.1), rep(0, length(xl))), col="gray")
> xu <- seq(100+.5, 100+.3, by=-.01)
> polygon(c(xu, rev(xu)), c(dnorm(xu, m=100, sd=0.1), rep(0, length(xl))), col="gray")
> lines(c(99.6, 100.4), c(0,0), lty=3)
> abline(v=100 + c(-.3,.3), lty=3)
> plot(x, pnorm(x, m=100, sd=0.1), type="l")
> abline(h=c(pnorm(100+.3, , m=100, sd=0.1),
+            pnorm(100-.3, , m=100, sd=0.1)), lty=3)
> abline(v=100 + c(-.3,.3), lty=3)
> par(op)
```

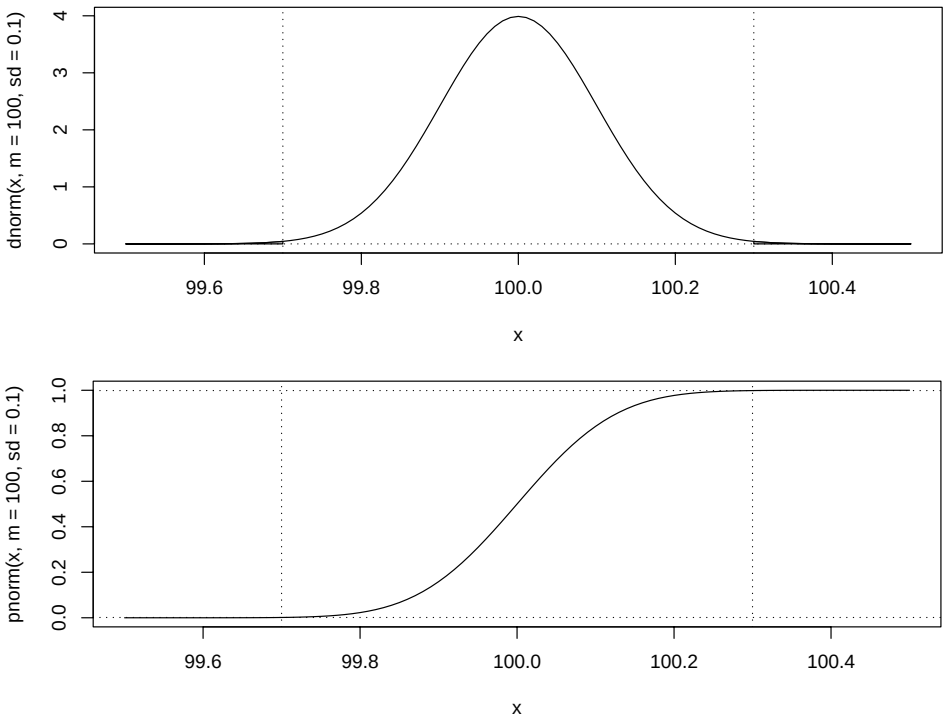

図 10.2 天秤での計測誤差（中心極限定理の適用）

【10.3】正規母集団の標本平均と標本分散

この場合は，母分散が既知であるので，標本平均は，N（4, 15/10）の正規分布に従う．また，標本分散／母分散の比に自由度を乗じたものは，カイ自乗分布に従う．

i) 標本平均が，３と６の間にある確率は，以下のように得られる．

```
> round(pnorm(6, m=4, sd=sqrt(15/10)) -
+       pnorm(3, m=4, sd=sqrt(15/10)), 3)
[1] 0.742
```

ii) カイ自乗分布（自由度 `df = 10 - 1 = 9`）を用いて，標本分散の累積分布を図示しよう（図 10.3）．

```
> x <- seq(0,35, length=200)
> plot(x, pchisq(9*x/15, df=10 - 1), type="l")
> abline(h=0.95, lty=3)
```

ここで，`1 - pchisq(9*x/15, df=9) = 0.05` を満たす x は，図から判断して，およそ 28 あたりである．自由度 `df = 9` のカイ自乗分布において，$\alpha = 0.05$ のパーセント点は，

```
> qchisq(1 - 0.05, df=10 - 1)
[1] 16.91898
```

である（テキストの付表３，p.282 も見よ）．そこで，母分散と自由度により，標本分散としてとりうる値に換算すると，以下のようになる．

```
> qchisq(1 - 0.05, df=10 - 1) * 15/9
[1] 28.19830
```

つまり，超過確率が 0.05 となる標本分散は，28.2 である．念のため，以下のように確認する．

```
> round(pchisq(9* 28.2/15, df=9, lower.tail=F), 3)
[1] 0.05
```

注意：カイ自乗分布の数学的表現を，テキストでは紹介していない（＊少しオドロクベキコトことかもしれない！なぜなら，ポアソン分布や指数分布などの数学的表現式は，テキストに明記されていることと対

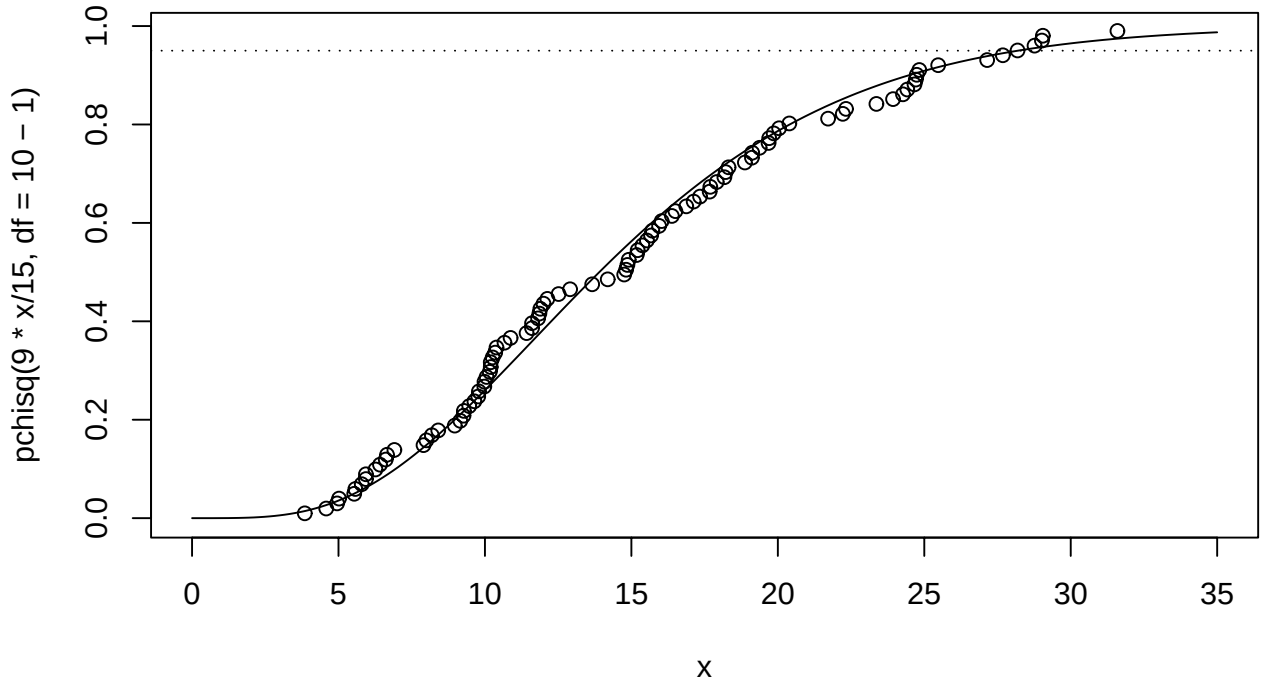

図 10.3 正規分布を母集団とする標本分散の分布（カイ自乗分布）

照的である）．それは，このレベルでは，その表現式を知る必要も特にないためである（なぜなら，ポアソン分布などは，具体的な現象の確率分布であり，他方，カイ自乗分布などは，推定値の分布であることが背景となっている．つまり，推定値の分布については，幾つかの代表的な有意水準 α に対するパーセント点がわかればよいからである）．カイ自乗のパーセント点を用いずに，以下のような算出も可能である．

```
> uniroot(function(x) 0.95 - pchisq(9*x/15, df=10-1), c(25, 30))$root
[1] 28.1983
```

シミュレーションによる確認（図 10.3 に重ねて表示）：

```
> sample.vars <- numeric(100)
> for (j in 1:100) {sample <- rnorm(10, m=4, sd=sqrt(15))
+                   sample.vars[j] <- var(sample)}
> points(sort(sample.vars), 1:100/101)
> sort(sample.vars)[94:100]
[1] 27.14169 27.68399 28.18199 28.76643 29.01191 29.05104 31.59696
```

【10.6】２つの標本の標本分散の比

この場合の標本分散の比 s1^2/s2^2 に対して，s1^2/s2^2 * σ2^2/σ1^2 は，自由度を `df1 = 10 - 1 = 9` と `df2 = 8 - 1 = 7` とするエフ分布に従う．そこで，標本分散の比の分布を図示しよう（図 10.6）．

```
> x <- seq(0, 3.5, length=100)
> plot(x, pf(x*4/3, df1=9, df2=7), type="l")
> abline(h=0.95, lty=3)
```

この場合のエフ分布について，α = 0.05 のパーセント点は，

```
> qf(0.05, df1=9, df2=7, lower.tail=F) # same as: qf(1 - 0.05, df1=9, df2=7)
[1] 3.676675
```

となる（テキストの付表４，p.284 も見よ）．母分散の比を用いることにより，超過確率が 0.05 となる標本分散比は，以下のように，2.76 程度とわかる．

```
> qf(0.05, df1=9, df2=7, lower.tail=F) * 3/4
[1] 2.757506
```

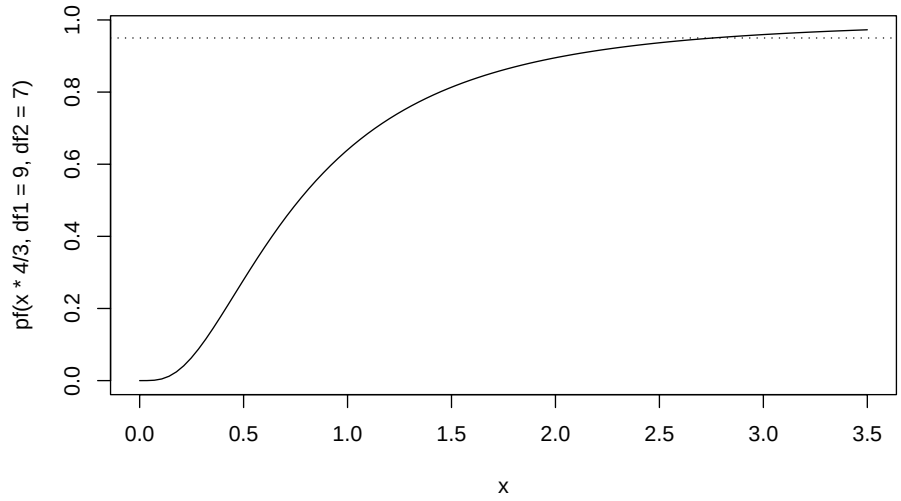

図 10.6 正規分布を母集団とする標本分散の比の分布（エフ分布）

【10.9】パーセント点の関係

i) a) 定義式によれば，自由度１のカイ自乗は，標準正規分布に従う変量を自乗したものであり，それゆえ，正規分布の両側の確率の和が，カイ自乗の超過確率に等しくなることは，明白である．以下は，確認のための一例である．

```
> pchisq(qnorm(1:19/20)^2, df=1, lower=F)
 [1] 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1
```

b) ティ分布に従う量を自乗すれば，分子は正規分布に従う量であり，分母は標本分散である．また，共に真値は母分散である．また，分子の自由度１であり，分母の自由度は，標本サイズ－１，すなわち，ティ分布の自由度となる．a) での関係と対比するとよい．以下は，確認のための一例である．

```
> pf(qt(1:19/20, df=3)^2, df1=1, df2=3, lower=F)
 [1] 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1
```

c) この式は，ティ分布の自由度が増加すれば，ティ分布は正規分布に近似できるという意味である．

```
> pt(qnorm(.975), df=5^(0:5))
[1] 0.8498262 0.9463536 0.9693815 0.9738875 0.9747780 0.9749556
```

ii) 有為水準 $\alpha = 0.05$ および自由度 $k = 120$ の時，諸量は次のとおりである．

```
> # a)
> qnorm(0.05/2)
[1] -1.959964
> .Last.value^2
[1] 3.841459
> qchisq(1 - 0.05, df=1)
[1] 3.841459

> # b)
> qt(0.05/2, df=120)
[1] -1.979930
> .Last.value^2
[1] 3.920124
> qf(1-0.05, df1=1, df2=120)
[1] 3.920124

> # c)
> qt(0.05, df=120)
[1] -1.657651
> qnorm(0.05)
[1] -1.644854
```

【11.7】東京と大阪の最低気温

```
> temp.tokyo <- scan()
1:  21.8  22.4  22.7  24.5  25.9  24.9  24.8  25.3  25.2  24.6
11:
Read 10 items
> mean(temp.tokyo)
[1] 24.21
> var(temp.tokyo)
[1] 1.938778
```

i) 母分散が未知であるので，母平均の信頼区間は，（X_bar + t_0.025 * s/sqrt(n), X_bar + t_0.975 * s/sqrt(n)）として得られる．t_0.025 および t_0.975 は，ティ分布から求める．

```
> qt(c(0.025, 0.975), df=9)
[1] -2.262157  2.262157
```

分散が既知であれば，上記の係数の代わりに，正規分布を用いて，

```
> qnorm(c(0.025, 0.975))
[1] -1.959964  1.959964
```

を用いることになる．以上により，東京の最低気温の平均の推定は，有意水準を５％ととれば，

```
> round(mean(temp.tokyo) + qt(c(0.025, 0.975), df=9) *
+   sqrt(var(temp.tokyo)/10), 2)
[1] 23.21 25.21
```

となる．なお，有意水準を１％とすれば，以下のとおり．

```
> round(mean(temp.tokyo) + qt(c(0.005, 0.995), df=9) *
+   sqrt(var(temp.tokyo)/10), 2)
[1] 22.78 25.64
```

ii) カイ自乗分布を用いれば，母分散の信頼区間は，（(n-1)/chisq_0.975 * s^2, (n-1)/chisq_0.0025 * s^2）として得られる．chisq_0.025 および chisq_0.975 は，カイ自乗分布から求める．

```
> qchisq(c(0.975, 0.025), df=9)
[1] 19.022768  2.700389
```

したがって，分散の信頼区間は，有意水準を５％とすれば，

```
> round(9*var(temp.tokyo)/qchisq(c(0.995, 0.005), df=9), 2)
[1]  0.74 10.06
```

となり，また，有意水準を１％とすれば，以下のとおり．

```
> round(9*var(temp.tokyo)/qchisq(c(0.975, 0.025), df=9), 2)
[1] 0.92 6.46
```

同時期の大阪での最低気温は，以下のとおりである．

```
> temp.osaka <- scan()
1:  22.1  25.3  23.3 25.2 25.3  24.9  24.9  24.9 24.9 24.0
11:
Read 10 items
> mean(temp.osaka)
[1] 24.48
> var(temp.osaka)
[1] 1.095111
```

iii) 大阪の最低気温の分散が，東京の最低気温の分散に等しいと仮定できるとき，最低気温の差の信頼区間は，（X_bar - Y_bar + t_0.025 * s * sqrt(1/n + 1/m), X_bar - Y_bar + t_0.975 * s * sqrt(1/n + 1/m)）となる．こ

の時，注意すべきは，t_0.025 および t_0.975 を算出する際のティ分布の自由度が n + m - 2 であることと，標本分散 s^2 が，次式で示す合併分散であることである．

$$s^2 = ((n-1) * Var(X) + (m-1) * Var(Y))/(n + m - 2)$$

したがって，この場合，有意水準を５％とした時の信頼区間は，次のようになる．

```
> s2.pool <- 9*(var(temp.tokyo) + var(temp.osaka))/(10*2 - 2)
> s2.pool
[1] 1.516944
> round(mean(temp.tokyo) - mean(temp.osaka) + qt(c(0.025, 0.975), df=18) *
+ sqrt(s2.pool * 2/10), 2)
[1] -1.43  0.89
```

なお，信頼区間の中にゼロが含まれることから，東京の最低気温と大阪のそれとの違いには，有為な差が見られるとはいえないことがわかる．これは，検定として以下のように示すことも可能である．

```
> t.test(temp.tokyo, temp.osaka, var.eq=T)

    Two Sample t-test

data:  temp.tokyo and temp.osaka
t = -0.4902, df = 18, p-value = 0.6299
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.4272036  0.8872036
sample estimates:
mean of x mean of y
    24.21     24.48
```

検定結果の出力レポートには，信頼区間も報告されていることに注意．また，個々の平均値も報告されている．ところで，上記のデータを箱ひげ図により図示すれば，図 11.7 (a) を得る．

```
> boxplot(temp.tokyo, temp.osaka, names=c("Tokyo","Osaka"))
```

また，等分散性については，エフ分布で検定できる．

```
> var.test(temp.tokyo, temp.osaka)

    F test to compare two variances

data:  temp.tokyo and temp.osaka
F = 1.7704, num df = 9, denom df = 9, p-value = 0.4077
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4397407 7.1275946
sample estimates:
ratio of variances
          1.770394
```

図によれば，バラツキは異なるようにみえるが，この例題では標本数が少ないため，バラツキが異なることを有為に示すことができない．したがって，本問において，等分散の仮定のもとでの検討は許容できる．また，東京と大阪の最低気温を時系列として表示したものが，図 11.7 (b) である．

```
> plot(temp.tokyo, type="o")
> lines(temp.osaka, type="o", col="red")
> legend(8, 23, c("Tokyo","Osaka"), lty=rep(1,2), pch=rep(1,2), col=c("black","red"))
```

以下のようにすれば，東京と大阪の最低気温の違いを日数として比較することができる．

```
> sum(temp.tokyo > temp.osaka)
[1] 4
```

東京の最低気温が大阪のそれより高い場合は４日，その逆は６日である．このことからも，両者に有為な差があるといえない．また，東京と大阪の最低気温の差は，図 11.7 (c) と表される．

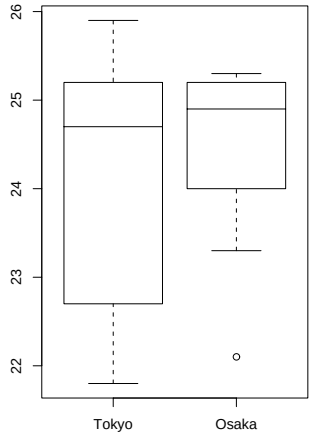

図 11.7 (a) 箱ヒゲ図

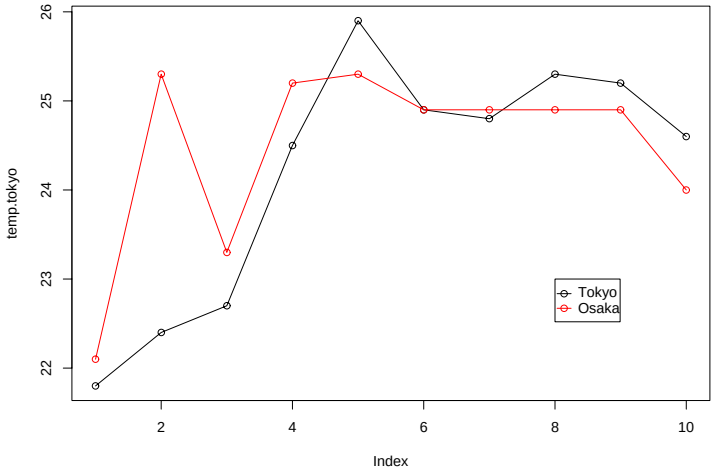

図 11.7 (b) 東京および大阪の最低気温の時系列グラフ

```
> plot(temp.diff <- temp.tokyo - temp.osaka, type="b", pch=4); abline(h=0, lty=2)
```

少し外れているように見える．両地点の差の検定は以下のように検討することができる．

```
> t.test(temp.tokyo, temp.osaka, paired=T)

    Paired t-test

data:  temp.tokyo and temp.osaka
t = -0.8267, df = 9, p-value = 0.4298
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.0088559  0.4688559
sample estimates:
mean of the differences
                  -0.27
```

このことは，以下のように確認しながら実行することも可能である．

```
> mean(temp.diff) + qt(c(0.025,0.975), df=9) * sqrt(var(temp.diff)/10)
[1] -1.0088559  0.4688559
```

また，以下のように図示して，そのバラツキの正規性を確認できる（図 11.7 (d) を参照）．やはり，二日目の両地点の差は，正規分布から外れているようである．

```
> qqline(temp.diff, col="magenta")
```

以上の考察で主張したいことは，本問のような簡単な資料でもいろんな視点で解析できることである．

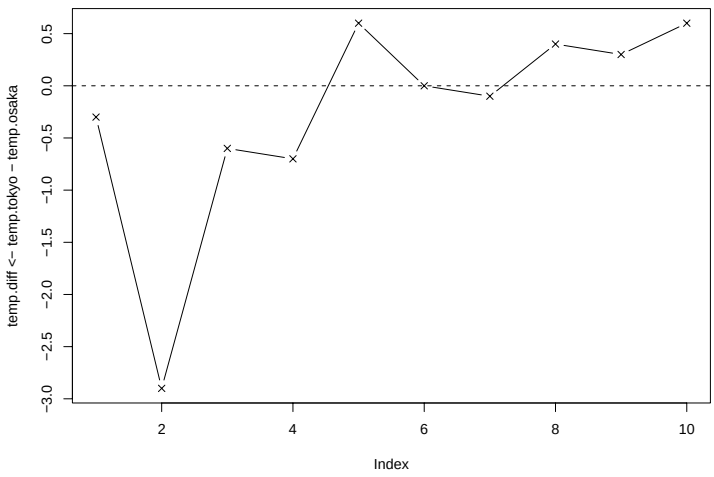

図 11.7 (c) 東京および大阪の最低気温の差についての時系列グラフ

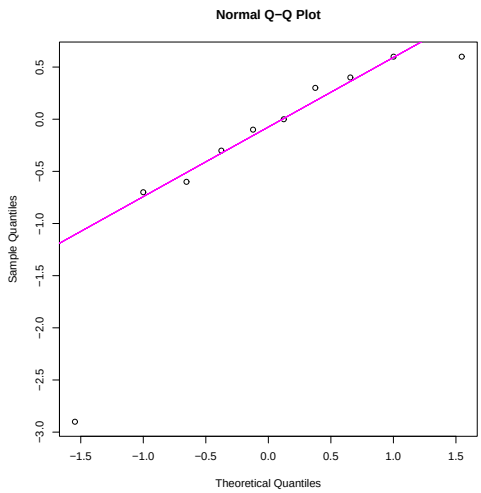

図 11.7 (d) 正規性の確認

以下は，北野担当の出題ではないが，統計計算環境 R の使用例として，解答例を作成してみた．

【1.5】 炭酸ガス

統計計算環境 R の特徴として，著名なデータがソフトウェアに添付していることである．ここでは，ハワイの Mauna Loa 観測所にて，1959 年から 1997 年までに観測された炭酸ガス濃度（単位：ppm）の月平均値データを用いる．出典は以下のとおり．

Keeling, C.D., T.P. Whorf, M. Whalen, and J. van der Plicht (1995): Interannual extremes in the rate of rise of atmospheric carbon dioxide since 1980, Nature, Vol. 375, pp.666-670. (see also, `ftp://cdiac.esd.ornl.gov/pub/maunaloa-co2/`)

単純に，時系列のグラフとして，このデータを図示すれば，図 1.5 (a) のようになる．

```
> data(co2)
> plot(co2); axis(1, seq(1965, 1995, by=5))
```

かなり明確な周期をもって変動しながら, 増加傾向のトレンドを有することがわかる. 要約統計量を見る．

```
> summary(co2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  313.2   323.5   335.2   337.1   350.3   366.8
```

最小値と最大値を参考に，年毎の１２ヶ月の変化を図示したものが，図 1.5 (b) である．周期的な変化として，春（５月から６月にかけて）に年最大をとり，秋（９月から１０月にかけて）に年最小をとることがわかる．

```
> plot(c(1,12), c(310,370), type="n", axes=F, xlab="month", ylab="co2")
> axis(1, 1:12, month.abb)
> axis(2)
> for (j in 1:39) lines(1:12, co2[1:12 + 12*(j - 1)])
```

では，視点を変えて，月毎の経年変化を見てみよう（図 1.5 (c) を参照）．各月において，経年的に単調増加なトレンドが見られる．

```
> monthplot(co2)
```

月毎の変化を表した本図に，年毎の変化を表したグラフを重ねて表示することも可能である．

```
> for (j in 1:39) lines(1:12, co2[1:12 + 12*(j - 1)], col="red") # again!
```

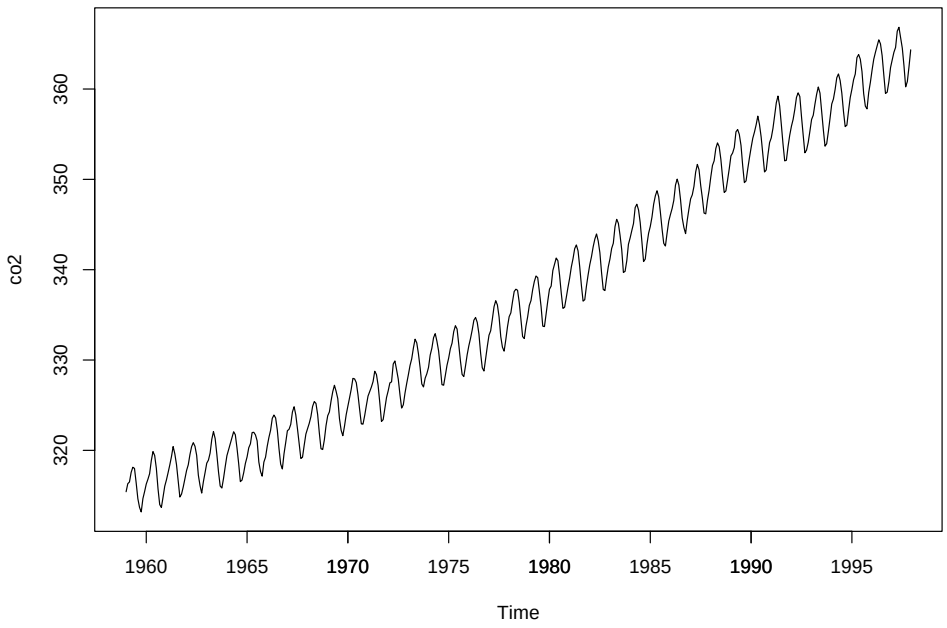

図 1.5 (a) Mauna Loa 観測所における炭酸ガス濃度（単位：ppm）の変化

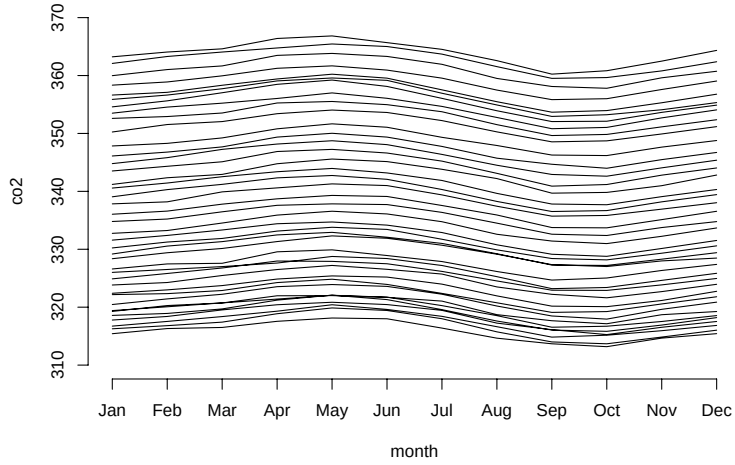

図 1.5 (b) 各年の季節変化

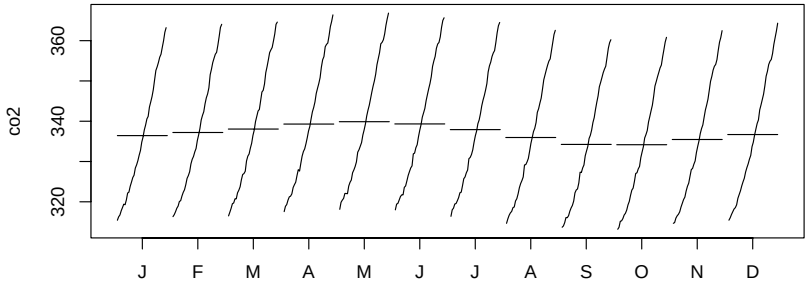

図 1.5 (c) 月毎の経年変化

また，樹木の年輪のように，月毎の変化曲線が互いに沿うように描かれており，月毎には変動しながら，経年的には着実に上昇していることがわかる．ただし，各曲線間のひらきは不規則であり，概略的には，１２ヶ月のの変動であるものの，その他の周期変動などが含まれており，複雑であることがわかる．季節変動とトレンドの分離には，より高度な手法が求められる．例えば，以下のような解析手法もある（が，現時点の学習レベルでは，その詳細を知る必要はない）．

```
> plot(stl(co2, s.window="periodic"))
> ?stl
```

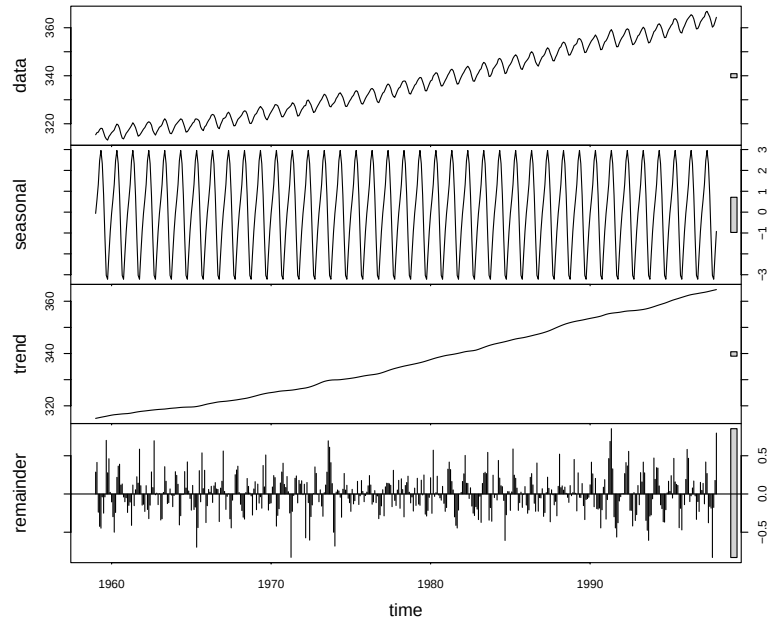

図 1.5 (d) 季節変動，トレンド，残差の各成分に分解

【3.1】得票率

```
> vote <- scan()
1: 41.4  76.3  59.2  51.8  52.5  53.2  62.4  55.0  57.7  63.2  37.5  48.5  32.4
14: 20.5  47.9  68.9  68.5  52.5  63.3  58.8  59.7  48.4  40.7  51.0  50.9  34.3
27: 25.8  32.1  34.4  55.1  60.3  57.0  45.6  54.2  55.1  55.7  70.3  61.8  47.6
40: 42.5  71.3  55.2  65.2  42.9  54.7  62.0  48.2  45.2
49:
Read 48 items
> home <- scan()
1: 52.8  71.2  72.6  63.7  81.3  81.8  70.9  74.0  73.2  72.9  66.7  65.7  43.7
14: 55.5  79.6  85.7  75.3  80.5  73.0  77.0  77.5  69.2  60.0  78.2  79.5  61.8
27: 49.6  59.6  72.1  71.0  76.3  72.8  71.8  60.7  67.0  71.8  71.2  68.3  68.5
40: 54.8  76.0  65.8  69.4  66.9  69.7  71.2  59.6  62.4
49:
Read 48 items
> pref <- c(
+ "HOKKAIDO", "AOMORI", "IWATE", "MIYAGI", "AKITA", "YAMAGATA", "FUKUSHIMA",
+ "IBARAKI", "TOCHIGI", "GUNMA", "SAITAMA", "CHIBA", "TOKYO", "KANAGAWA",
+ "NIIGATA", "TOYAMA", "ISHIKAWA", "FUKUI", "YAMANASHI", "NAGANO", "GIHU",
+ "SHIZUOKA", "AICHI", "MIE", "SHIGA", "KYOTO", "OSAKA", "HYOGO", "NARA",
+ "WAKAYAMA", "TOTTORI", "SHIMANE", "OKAYAMA", "HIROSHIMA", "YAMAGUCHI",
+ "TOKUSHIMA", "KAGAWA", "EHIME", "KOCHI", "FUKUOKA", "SAGA", "NAGASAKI",
+ "KUMAMOTO", "OITA", "MIYAZAKI", "KAGOSHIMA", "OKINAWA", "ALL.japan")
> elect <- data.frame(home, vote)
> row.names(elect) <- pref

> plot(elect)
> cor(elect)
          home      vote
home 1.0000000 0.6419464
vote 0.6419464 1.0000000
> abline(lm(vote ~ home, data=elect), col="red")
```

自民得票率と持ち家比率は，図 3.1 からも相関係数からも，互いに相関関係があるとみてとれる．なお，図中の赤線は，回帰直線である．個人的見解であるが，１９８３年当時，持ち家者の保守的な考えと，自

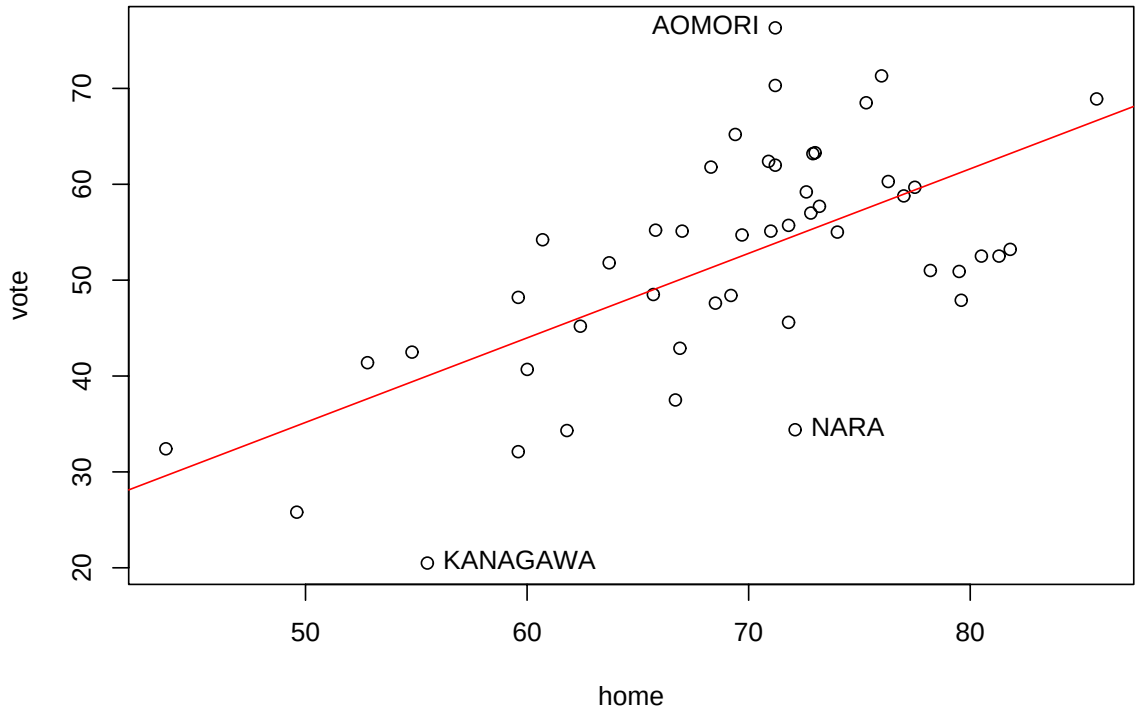

図 3.1 持ち家率と自民党投票率の関係

民党の安定した政策と一致する点がなんらかあったと思われる．なお，回帰直線から，やや外れた所にあるデータは，以下のように判別することができる．

```
> identify(elect, lab=row.names(elect))
[1]  2 14 29
```

また，持ち家率の小さい方から５つの都道府県を示すと，以下のように得られる（表計算ソフトでいえば，home での並び替えと考えよ）．

```
> elect[order(elect$home),][1:5,]
         home vote
TOKYO    43.7 32.4
OSAKA    49.6 25.8
HOKKAIDO 52.8 41.4
FUKUOKA  54.8 42.5
KANAGAWA 55.5 20.5
```

【3.2】タバコと肺がん

Ａ氏は，１）『（たとえ，タバコと肺がんの相関関係を示されたにせよ，）その因果関係が示されたわけではない』，あるいは，２）『（統計的証拠としては，一般的な傾向が示されたにすぎない．）ひとはひと，じぶんはじぶん！（じぶんは，その一般的傾向に対して，外れ値の位置にいると思う．つまり，タバコを吸うと肺がんに罹るという傾向に対して，タバコを吸わなくても肺がんに罹る人もいれば，タバコを吸っていても肺がんに罹らない人もいる）』というような言い訳をいうかもしれない（言い訳は，他にもあるだろう）．さて，あなたなら，論理にもとづく統計学の有用性をＡ氏にどのように説明しますか？

【3.3】社会的リスク

対象とするデータを以下のように入力する．

```
> woman <- 1:30
> college <- scan()
1:   1  5  2  3  6  7 15  8  4 11 10 14 18 13 22
16: 24 16 19 30  9 25 17 26 23 12 20 28 21 27 29
31:
Read 30 items
> company <- scan()
1:   8  3  1  4  2  5 11  7 15  9  6 13 10 22 12
16: 14 18 19 17 22 16 24 21 20 28 30 25 26 27 29
31:
Read 30 items
> prof <- scan()
1:  20  1  4  2  6  3 12 17  8  5 18 13 23 26 29
16: 15 16  9 10 11 30  7 27 19 14 21 28 24 22 25
31:
Read 30 items
```

以下のようにスピアマンとケンドールの順位相関係数から，例えば，女性有権者団体と大学教授らの有識者との社会リスクに対する考え方について，相関はあると判断できる．

```
> cor.test(woman, prof, method="spearman")

    Spearman's rank correlation rho

data:  woman and prof
S = 1828, p-value = 0.0006945
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.5933259
```

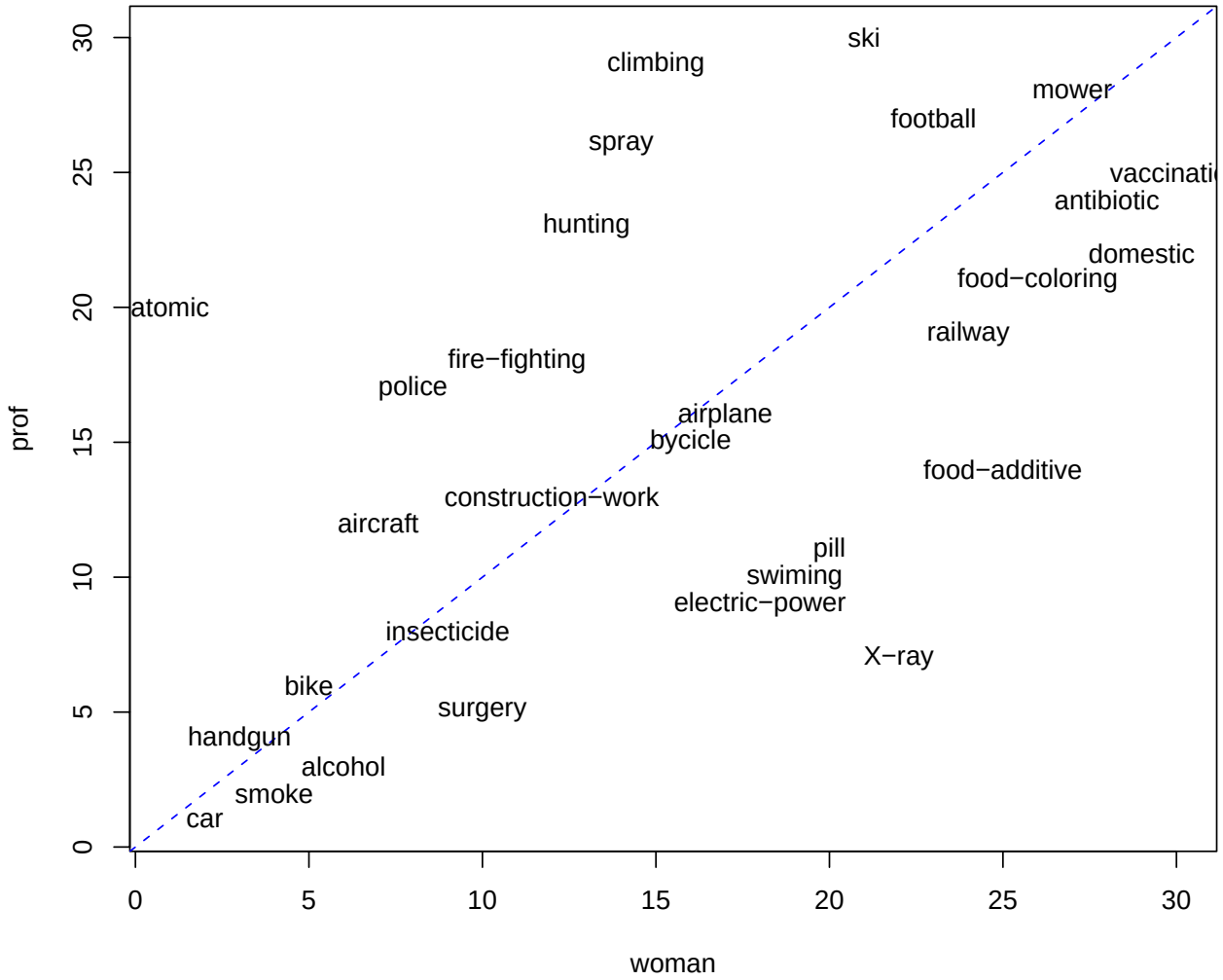

図 3.3 社会リスクについての考え方の違い（女性有権者団体と大学教授ら）

```
> cor.test(woman, prof, method="kendall")

    Kendall's rank correlation tau

data:  woman and prof
T = 313, p-value = 0.0004875
alternative hypothesis: true tau is not equal to 0
```

女性有権者団体と大学教授らの有識者との社会リスクに対する考え方について, 図 3.3 として順位の散布図を表現することもできる．特にスピアマンの相関については，図 3.3 から確認しやすい．

```
> activities <- c(
+ "atomic", "car", "handgun", "smoke", "bike", "alcohol", "aircraft", "police",
+ "insecticide", "surgery", "fire-fighting", "construction-work", "hunting",
+ "spray", "climbing", "bycicle", "airplane", "electric-power", "swiming", "pill",
+ "ski", "X-ray", "football", "railway", "food-additive", "food-coloring", "mower",
+ "antibiotic", "domestic", "vaccination")
```

```
> plot(woman, prof, type="n")
> abline(0, 1, lty=2, col="blue")
> text(woman, prof, activities)
```

授業の学習レベルから逸脱するが，3 つのグループを同時に比較する（この場合, 相関の比較というより, 順位付けの類似性をみる）ことも，ケンドールの一致係数 W を導入すれば可能である．興味のある人は，例えば，以下のテキストなどを参考にせよ．

Siegel, S. (1956): Nonparametric Statistics: for the behavioral science, McGraw-Hill.（藤本 **熙** 監訳による邦訳あり）