村上　明
pp. 77-82

4.2 Tests for two independent vectors

二つの数値ベクトルを比較する場合、対応あり/なしの区別が必要

- 対応なしの例：ある作品の語の頻度表から 100 語をランダムに選び、別の作品の語の頻度表からランダムに選んだ 100 語と比較する
- 対応ありの例：決められた 100 語について、ある作品と別の作品の頻度を比較する

本節ではまず対応なしを、その後対応ありを扱う。

4.2.1 Are the distributions the same?

分布が同一であるかどうかを検定する

- Figure 4.1 で二つの山があったのは、semantically opaque な語と semantically transparent な語の差と考えられる
- データフレーム ver の中にそれを記した列がある
- 違いは Figure4.5 を見ればわかる
- 以下の手順により、Figure4.5 を作成できる

```
ver$Frequency = log(ver$Frequency)
ver.transp = ver[ver$SemanticClass == "transparent",]$Frequency
ver.opaque = ver[ver$SemanticClass == "opaque",]$Frequency
ver.transp.d = density(ver.transp)
ver.opaque.d = density(ver.opaque)
xlimit = range(ver.transp.d$x, ver.opaque.d$x)
ylimit = range(ver.transp.d$y, ver.opaque.d$y)
plot(ver.transp.d, lty = 1, col = "black", xlab = "frequency", ylab = "density", xlim = xlimit,
ylim = ylimit, main = "")
lines(ver.opaque.d, col = "darkgrey")
```

- 二つの山が出たのが偶然でないことを示すために two-sample Kolmogorov-Smirnov test を用いる

```
ks.test(jitter(ver.transp), jitter(ver.opaque))
```

→p 値が小さい
→transparent と opaque の二つの subset にわけることは正しい

4.2.2 Are the means the same?

平均の差の検定

- Chapter2 で使った ratings の頻度を simple/complex をグルーピングファクターにして plant/animal でわける

```
bwplot(Frequency ~ Class | Complex, data = ratings)
```

→simplex 側を見ると、animal の方が plant よりも頻度が高いのではないかと思える
→箱ひげ図により symmetrical（＝正規分布を仮定できる）であることがわかるので、two-sample の t 検定を用いる

```
simplex = ratings[ratings$Complex == "simplex",]
freqAnimals = simplex[simplex$Class == "animal",]$Frequency
```

freqPlants = simplex[simplex$Class == "plant",]$Frequency
t.test(freqAnimals, freqPlants)

※t.test 関数に数値ベクトルを二つ入れれば独立した二群の検定になる
※ウェルチの検定を用いている。これは二群の分散が異なっていても、それを補正してくれる。普通は自由度は整数だが、補正の結果そうではなくなっているのがわかる

・ 95%の信頼性区間は自動的に出るが、例えば 99%を出したければ以下のようにする

t.test(freqAnimals, freqPlants, conf.level = 0.99)

・ 独立した二群の検定に t 検定を用いることができるのは、正規分布が仮定できるデータのみ
・ そうではない場合は wilcox.test を用いる

wilcox.test(ver.opaque, ver.transp)

→Kolmogorov-Simirnov test を用いた際と結果は同じ
→二つの異なった分布を持つデータを扱っている

Chapter1 の verbs を再びここで用いる
tapply(verbs$LengthOfTheme, verbs$AnimacyOfRec, mean)

・二群に差があるかどうかをウェルチの t 検定で見る
t.test(LengthOfTheme ~ AnimacyOfRec, data = verbs)

→有意差あり

4.2.3 Are the variances the same?
分散が同一かどうかの検定
1. まずはデータの準備

x = rnorm(50, mean = 0, sd = 2)
y = rnorm(30, mean = 1, sd = 1)

2. var.test( )を用いて検定する

var.test(x, y)

※表中の F 値は x と y の分散の比率。以下と同じ。

var(x)/var(y)

自由度は観測値の数マイナス 1 なので、p 値もそれを用いて計算できる

*2 * (1 - pf(var(x)/var(y), 49, 29))*

この検定は正規分布が仮定できるデータにのみ適用できる。
ノンパラメトリックの場合は ansari.test や mood.test を用いる