Corpus Statistics
Baayen (2008; pp. 71-74)
2008/11/11

Toru Hitomi

4.1 Tests for single vectors

4.1.1 Distribution tests

どのように分布が自分の持つデータを特徴づけるかを知ることは有用！

e.g. 多くの統計分析において、データの正規性が仮定されている。
→自分のデータが正規分布に近似しているかを確かめる（正規性の検定）必要がある。

- Baayen and Lieber (1997)のデータ：オランダ語の接頭辞（*ver-*）の頻度分布の特徴を把握する。
  まず：
  データセット（ver）の頻度情報をもとに、確率密度分布を表示する。

```
> plot(density(ver$Frequency))
```

  →かなり形が歪んでいることがわかる。
  頻度を指数化して、ある程度ゆがみを矯正

```
> ver$Frequency = log(ver$Frequency)
> plot(density(ver$Frequency))
```

  Pattern 1. Q-Q プロットを作成し、正規分布と比較する！
  ① qqnorm()を用いて、乱数の場合と、実際の観察データの Q-Q プロットを作成し、グラフを比較する。

```
> qqnorm(length(ver$Frequency), 4, 3))
> abline(v=qnorm(0.025), col="grey")
> abline(h=qnorm(0.025, 4, 3), col="grey")

> abline(v=qnorm(0.975), col="grey")
> abline(h=qnorm(0.975, 4, 3), col="grey")
> qqnorm(ver$Frequency)
```

  Pattern 2. Shapiro-Wilk test（シャピロ - ウィルク検定）で検定する！
  ① shapiro.test()を用いて、統計量（W 値）を算出し、有意確率を見る。

```
> shapiro.test(ver$Frequency)
```

```
Shapiro-Wilk normality test
data:  ver$Frequency
W = 0.9022, p-value < 2.2e-16
```

p 値が 0.00000000000000022 よりも小さいので、帰無仮説が棄却
→ver$Frequency の分布は正規分布ではない！！

Toru Hitomi

Pattern 3.　**Kolmogorov-Smirnov one-sample test（コルモゴロフ - スミルノフ検定）**で検定する！

① ks.test()を用いて、統計量（D 値）を算出し、有意確率を見る。

```
> ks.test(ver$Frequency, "pnorm", mean(ver$Frequency), sd(ver$Frequency))
```

＊構造：ks.test(観察データ, 比べたい確率関数, corresponding parameters)

```
        One-sample Kolmogorov-Smirnov test

data:  ver$Frequency
D = 0.1493, p-value < 2.2e-16
alternative hypothesis: two-sided

Warning message:
In ks.test(ver$Frequency, "pnorm", mean(ver$Frequency),
sd(ver$Frequency)) :
cannot compute correct p-values with ties
```

Shapiro-Wilk 検定と同様の結果

KS 検定では、データ内に重複するデータ（**ties**）を想定していない。

② jitter()を用いて、ノイズをデータに発生させて、重複データをなくす。

```
> ks.test(jitter(ver$Frequency), "pnorm", mean(ver$Frequency),
+ sd(ver$Frequency))
        One-sample Kolmogorov-Smirnov test

data:  jitter(ver$Frequency)
D = 0.1493, p-value < 2.2e-16
alternative hypothesis: two-sided
```

・ テキスト（Baayen, 2008）のイントロ部分での単語の頻度が確率的に等しいかどうかを知りたい！（**χ^2 test、カイ 2 乗検定**）

① データセットを作る。

```
> intro=c(75, 68, 45, 40, 39, 39, 38, 33, 24, 24)
> names(intro)=c("the","to","of","you","is","a","and","in","that","data")
> intro
```

the	to	of	you	is	a	and	in	that	data
75	68	45	40	39	39	38	33	24	24

② chisq.test()を用いて、統計量（χ^2 値）を算出し、有意確率を見る。

```
> chisq.test(intro)

        Chi-squared test for given probabilities

data:  intro
X-squared = 59.7294, df = 9, p-value = 1.512e-09
```

p 値 = 0.000000001512
→帰無仮説が棄却＝イントロの単語の頻度確率は等しくない！