전산통계

Chapter 2. 표준균일분포의 검정

- 예) $a = 3,\ c = 1,\ x_0 = 3, m = 8$일 때, 난수의 생성결과

$$x_n = 3 \cdot x_{n-1} + 1\ (mod\ 8)$$

3, 2, 7, 6, 3, 2, 7, 6, ...

주기가 4인 수열

```
> m <- 8
> a <- 3
> c <- 1
> x[1] <- 3
> for(i in 2:10) x[i] <- (a*x[i-1]+c)%%m
> x
 [1] 3 2 7 6 3 2 7 6 3 2
```

카이제곱분포를 이용한 적합도 검정

- H_0 : 생성된 난수들은 표준균일분포를 따른다.
- 단계 1) 구단 [0, 1]을 k등분하여 k개의 분할을 구한다.

  - $\left[0, \frac{1}{k}\right)$, $\left[\frac{1}{k}, \frac{2}{k}\right)$, ⋯ , $[\frac{k-1}{k}, 1)$

- 단계 2) $u_1,\ u_2,\ u_3,\ \cdots\ ,\ u_k$의 자료를 분할된 구간으로 정리하여 각 구간의 관측도수 $n_i,\ i = 1,2,\cdots,k$를 구한다. $(n_1 + n_1 + \cdots + n_k = N)$
- 단계 3) 표준균일분포에서 각 구간에 속하는 기대도수는 $\frac{N}{k}$이므로 다음의 식을 계산한다.

  - $W = \frac{k}{N}\sum_{i=1}^{k}(n_i\ -\frac{N}{k})^2$

- 단계 4) 난수가 표준균일분포를 따른다면 W는 자유도가 $k-1$인 카이제곱분포를 따른다. 즉, $W < \chi_\alpha^2(k-1)$이면 표준균일분포를 따른다고 한다.

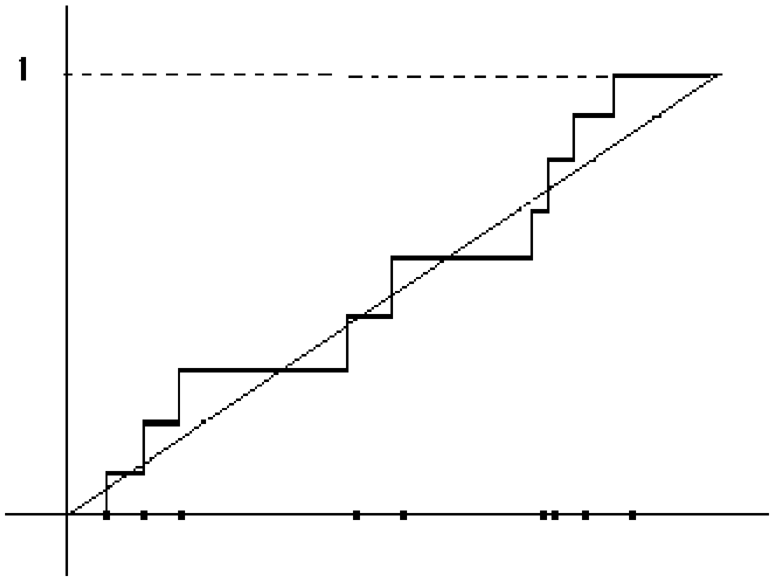
카이제곱분포를 이용한 적합도 검정

```
library(MASS)
data(survey)
str(survey)

x <- table(survey$W.Hnd)
x
chisq.test(x, p = c(0.3, 0.7))
```

콜모고로프-스미르노프(Kolmogorov-Smirnov) 적합도 검정

- 표본값 $u_1,\ u_2,\ u_3,\ \cdots,\ u_N$의 경험분포함수 (empirical distribution function)
  - $F_N(u) = \frac{1}{N}\sum_{i=1}^{N} I_{[u_i,u)}(u)\ \ I_{[u_i,u)}(u) = \begin{cases} 1 & if\ \ u \geq u_i \\ 0 & if\ \ u < u_i \end{cases}$
  - $F_N(u)$는 u값 이하가 되는 u_i의 개수를 N으로 나눈 값이며 계단함수의 형태를 가진다.



- $H_0 : F_N(u) = F_0(u)$
- 글리벵코-칸탈리 정리 (Glivenko-Cantelli theorem)
  - 임의의 양의 실수 ε에 대하여 $\lim_{N\to\infty} P(max_u| F_N(u) - F_0(u)| > \varepsilon) = 0$
  - 귀무가설이 맞다는 가정에서, N이 충분히 큰 경우에 모든 u값에 대하여 $|F_N(u) - F_0(u)|$의 값이 매우 작은 값 ε보다 작을 확률이 1에 가까워진다.
- 콜모고로프-스미르노프 검정통계량
  - $D_N = max_u|F_N(u) = F_0(u)|$
  - 분포무관 성질
- 난수의 경우, 귀무가설의 분포가 표준균일분포이므로 $F_0(u) = u$가 된다.

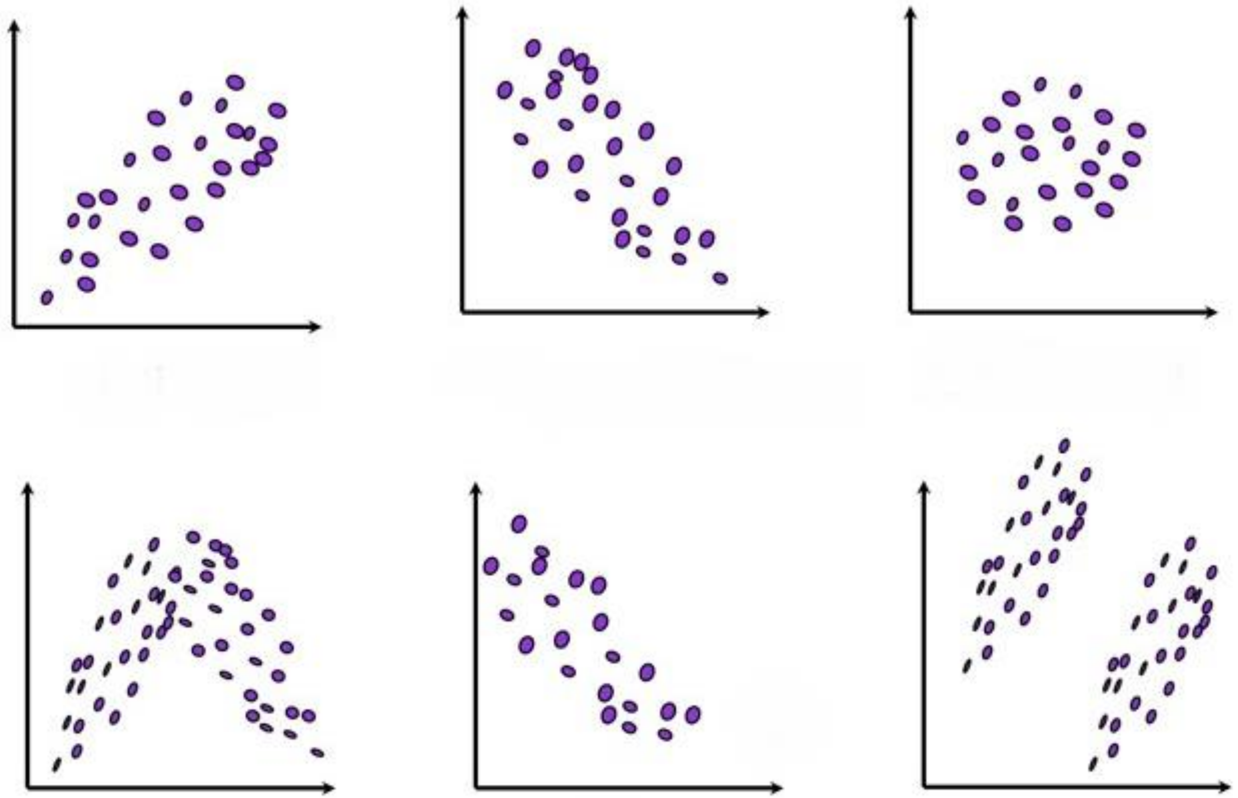
- ## K-S test as a goodness-of-fit test
- x <- rnorm(100,0.5,1)
- u <- runif(50,0,1)
- summary(x)
- summary(u)

- x11()
- par(mfrow=c(1,2), mar=c(2,2,2,2), cex=1.2)
- hist(x,freq=FALSE)
- hist(u,freq=FALSE)

- # k-s test
- ks.test(x,u)
- ks.test(x,"pnorm",0.5,1)
- ks.test(u,"punif")

난수의 독립성 검정

- 그래프를 통한 주관적 판정



- 런 검정 (run test)
  - 난수값들의 증감형태를 가지고 통계적 분포를 통해 판정
  - 런 : 같은 성질을 가진 연속된 수들의 집합
  - 런의 길이
  - Up and down
  - Run above and below the mean

- Up and down
  - 변화점을 이용해 런을 구분

$$0.34\ 0.51\ 0.68\ 0.44\ 0.18\ 0.93\ 0.87\ 0.57\ 0.73\ 0.92$$

  - 귀무가설이 맞다는 가정하에 N개의 수에 대한 런의 개수를 R이라 하면

    $E(R) = \frac{2N-1}{3}, \quad Var(R) = \frac{16N-29}{90}$이고, N이 큰 경우 근사적으로 정규분포를 따른다.

  - $Z = \frac{R-(2N-1)/3}{\sqrt{(16N-29)/90}}$

- Run above and below the mean
  - 자료의 평균을 중심으로 사용

0.34 0.51 0.68 0.44 0.18 0.93 0.87 0.57 0.73 0.92 (평균 0.617)

  - 평균보다 큰 수의 개수를 m, 작은 수의 개수를 n이라 하면, 런의 개수 R에 대하여 귀무가설이 맞다는 가정 아래

$$P(R=2k)=\frac{2\binom{m-1}{k-1}\binom{n-1}{k-1}}{\binom{N}{m}},\ P(R=2k+1)=\frac{\binom{m-1}{k}\binom{n-1}{k-1}+\binom{m-1}{k-1}\binom{n-1}{k}}{\binom{N}{m}}$$

- Run above and below the mean
  - 자료의 평균을 중심으로 사용

0.34 0.51 0.68 0.44 0.18 0.93 0.87 0.57 0.73 0.92 (평균 0.617)

  - 평균보다 큰 수의 개수를 m, 작은 수의 개수를 n이라 하면, 런의 개수 R에 대하여 귀무가설이 맞다는 가정 아래

$$P(R=2k)=\frac{2\binom{m-1}{k-1}\binom{n-1}{k-1}}{\binom{N}{m}}, P(R=2k+1)=\frac{\binom{m-1}{k}\binom{n-1}{k-1}+\binom{m-1}{k-1}\binom{n}{k}}{\binom{N}{m}}$$

- Run above and below the mean
  - m과 n의 값이 큰 경우에는 정규분포에 근사시켜 검정 가능
  - $Z = \frac{R-E(R)}{\sqrt{Var(R)}}, \quad E(R) = \frac{2mn+1}{N}, \quad Var(R) = \frac{2mn(2mn-N)}{N^2(N-1)}$

```
library(tseries)

x <- factor(sign(rnorm(100)))
runs.test(x)
```