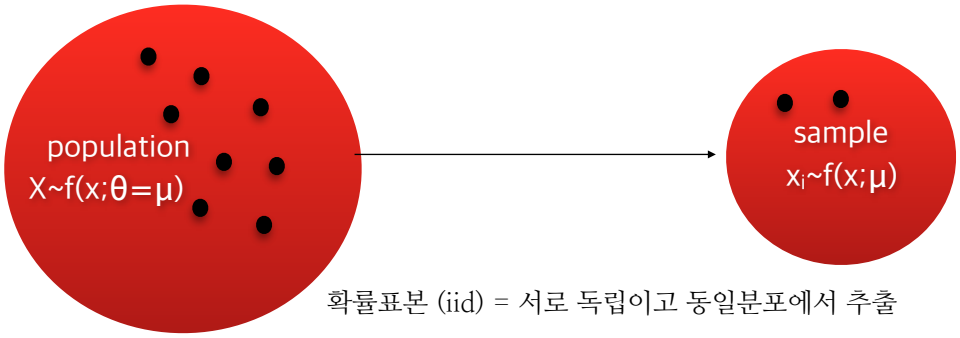
모평균 추론 개념



확률표본 (iid) = 서로 독립이고 동일분포에서 추출 $x_i \sim f(x;\mu)$

데이터 $x_1 = 12.5, x_1 = 1.5, x_3 = 3, ..., x_n = 7.5$

- 모집단 개체 확률변수의 분포는 모수가 평균 μ인 분포를 갖는다. – 모집단의 분포도 모른다.
- 표본 데이터의 관측치는 측정형 값이다. (예) 일주일 공부시간

MVUE for $\theta = \mu$

- 모수, 모집단 평균의 최소분산불편추정량은 표본 평균이다. $\hat{\theta} = \bar{x} = \frac{\sum x_i}{n}$
- 추정값, 표본평균의 평균과 분산 : $E(\bar{x}) = \mu, V(\bar{x}) = \frac{\sigma^2}{n}$: 모집단 분산 σ^2 도 모른다.
- μ 처럼 관심 모수는 아니지만 모집단의 분산도 모르는 값이다. 이를 ancillary parameter 보조 모수라 함 – 당연히 이에 대한 추정값도 필요하 – 모집단 분산의 MVUE = $\hat{\sigma}^2 = s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$

MVUE 샘플링분포

- $\sum x_i \sim Normal(\mu, \frac{\sigma^2}{n})$ by 중심극한 정리, 모집단 분산을 알지 못하면 표본분산 s^2 으로 추정
- 소표본 : T-정리 : 1) $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$ 2) $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$ (모집단 정규분포 가정 필요) 3) $(\bar{x}, s^2)$ 은 서로 독립이다. 그러므로 4) $\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n-1)$

- 피봇 통계량 : $\frac{\overline{x}-\mu}{s/\sqrt{n}} \sim t(n-1)$ (소표본일 경우 모집단 정규분포 가정 필요)

데이터 검증

표본 평균을 사용하므로 1) 치우침 (물론 대표본 이론에 의해 표본크기가 충분히 크면 문제 없지만) 2) 극단값 존재, 이 두 문제에 대해서는 추론 전에 검증되어야 한다.

예제 데이터 : BASEBALL.csv

OBA: (hits + bases on balls + hit by pitcher) / (at bats + bases on balls + hit by pitcher)

EBP: (total bases − hits) / at bats

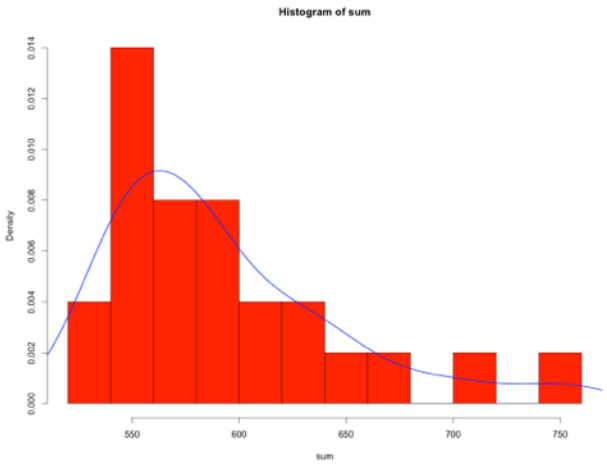
sum=OBA+EBP

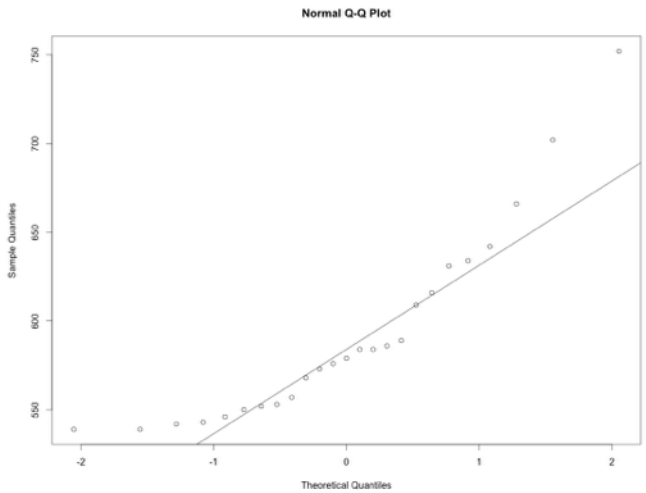
1) 치우침 skewness

(1) 진단

- 시각적 진단 : Q-Q plot, 히스토그램

hist(sum,,breaks=10, col="red",prob=T) lines(density(sum), col="blue", lwd=2) qqnorm(sum);qqline(sum)	우로 치우친 형태





• 정규성 검정 : A-D 통계량, K-S 통계량

library(nortest) ad.test(sum) shapiro.test(sum)	귀무가설 : 데이터는 정규분포를 따른다. 대립가설 : 데이터는 정규분포를 따르지 않는다 (결론적으로 무슨 분포인지는 모른다)

```
> ad.test(sum)

        Anderson-Darling normality test

data:  sum
A = 1.1395, p-value = 0.004498
```

```
> shapiro.test(sum)

        Shapiro-Wilk normality test

data:  sum
W = 0.8536, p-value = 0.002059
```

*) 두 통계량 모두 유의확률이 유의수준 5%보다 작으므로 귀무가설이 기각되어 데이터는 정규분포를 따른다고 할 수 없다.

a) Shapiro Wilk W-통계량

- $W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$, 상수 a_i는 분산행렬을 이용하여 구함

b) Kolmogorov D-통계량

- $D = \sup_x |F_n(x) - F(x)|$

c) Anderson-Darling AD 통계량

- $A^2 = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 [F(x)(1 - F(x))]^{-1} dF(x)$

(2) 해결 : 정규변환 Normality Transformation

예제 데이터는 우로 치우친 형태를 가지고 있고, 정규분포를 따르지 않으므로 적절한 정규변환이 필요하다.

ad.test(sqrt(sum[-1])) - Babe Ruth (첫 관측치 제외)

W = 0.9066, p-value = 0.02977

ad.test(log(sum[-1]))

A = 0.7164, p-value = 0.05342

로그변환 후 정규분포를 따르게 된다.

$$Y^* = \begin{cases} Y^3, \textit{left} \\ Y^2, \textit{mild left} \\ \sqrt{Y}, \textit{mild right} \\ \ln(Y), \textit{right} \\ 1/Y, \textit{severe right} \end{cases}$$

2) 극단 값

(1) 진단

- 평균, 표준편차 이용 : 실증적 법칙 ($\bar{x} \pm 3s$ 구간 99.7% 포함) – 이미 치우침이 있는 경우에는 적절하지 않음 – Chevyshev's Inequality : $P(|X-\mu| \le k\sigma) \ge 1-\frac{1}{k^2}$

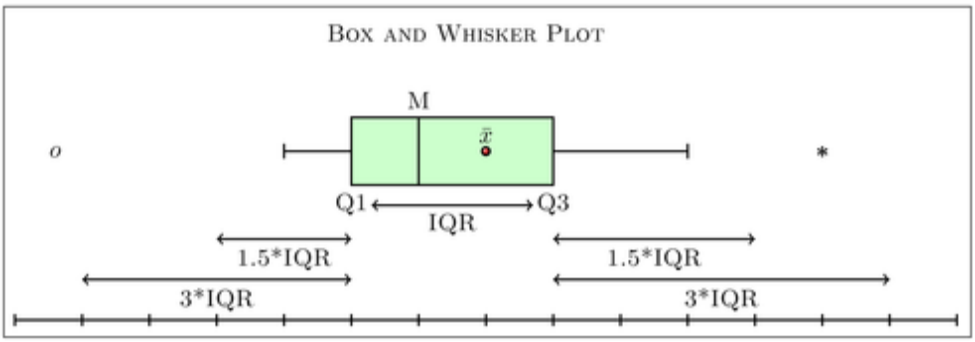
- 상자-수염 Box whisker plot 이용 :

  (min, Q1, 중앙값, Q3, max) 5개 순서통계량

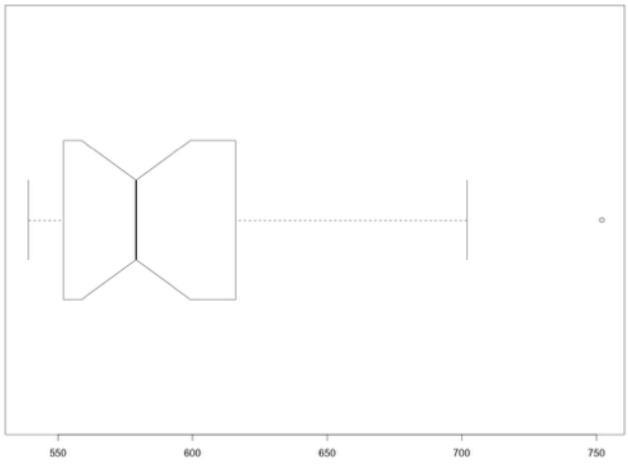
  사분위 범위 Inter Quartile Range : $IQR = Q_3 - Q_1$

  mild outlier – $(Q_1 - 1.5IQR, Q_3 + 1.5IQR)$

  severe outlier – $(Q_1 - 3*IQR, Q_3 + 3*IQR)$



fivenum(sum) IQR(sum) boxplot(sum,horizontal = T, notch=T) boxplot(sum,horizontal = T, notch=T)$out	[1] 539 552 579 616 752 IQR(sum) [1] 64 [1] 752 (babe ruth) 한 개의 이상치를 갖는다.



정규변환, 이상치 제외 전 95% 신뢰구간

〉 t.test(sum)

(570.2 614.8)

로그변환, 이상치 제외 후 95% 신뢰구간

〉 t.test(log(sum[-1]))

(6.34, 6.40) –〉 (e^6.34, e^6.40)

(566.8, 601.8)

=〉 구간 폭도 줄어들고 하향 조정 되었음