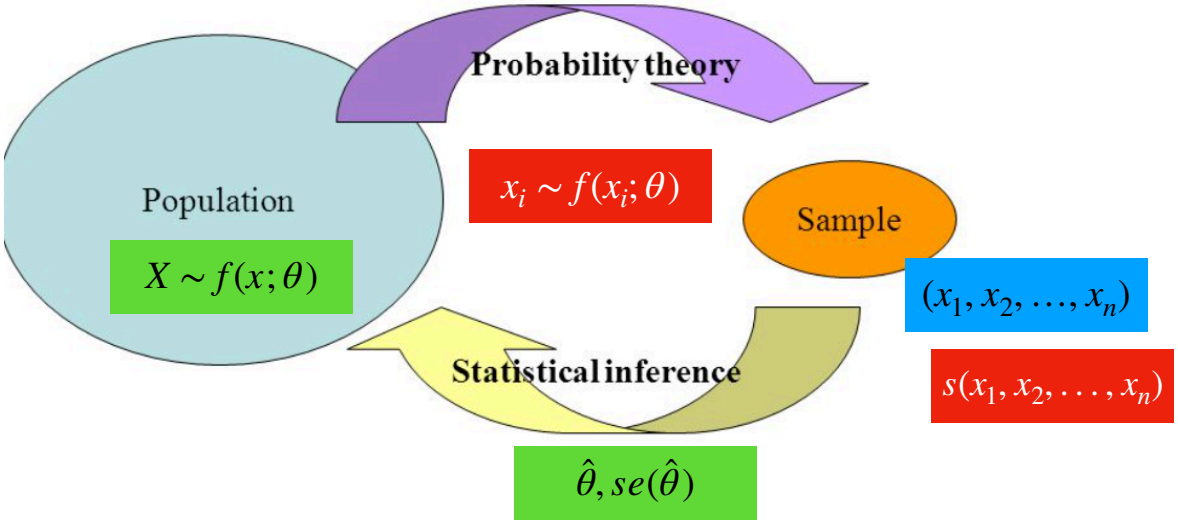
개념

모집단과 모수

확률표본 관측치(데이터)는 모집단의 관심특성을 알아보기 위하여 수집된다. 모집단의 관심특성은 모집단 확률분포함수 $f(x)$와 대표값인 모수(θ)으로 요약된다. 확률표본 $(x_1, x_2, \ldots, x_n) \sim f(x;\theta)$



대표 모수

- 데이터 중앙위치 : 모집단 평균 μ, 비율 p [*]비율은 모집단 확률변수가 (0, 1) 이진형 결과만 있음
- 데이터 흩어짐 : 분산 σ^2

모집단의 대표값 모수는 확률표본 데이터로부터 계산된 통계량 $s(x_1, x_2, \ldots, x_n)$을 이용하여 정보를 얻는다. 확률표본으로 계산된 통계량 중 추정에 사용되는 통계량은 추정량 $\hat{\theta}$이라 하고 (가장 좋은 추정량을 MVUE 최소분산불편추정량) 통계적 가설검정에 사용되면 검정통계량이라 하며 가설 검정의 경우 추정량의 확률분포함수(샘플링 분포) $f(\hat{\theta})$가 필요하다.

통계량 $s(x_1, x_2, \ldots, x_n)$

확률표본으로부터 계산된 값을 통계량이라 하고 통계량은 확률표본의 함수이다.

모수에 대한 가장 좋은 점추정치는 MVUE (불편 추정량 $E(s = \hat{\theta}) = \theta$ 중 추정분산 $V(\hat{\theta})$이 가장 적은 통계량)이다. 점추정량 MVUE의 확률분포함수를 이용하여 모수에 대한 신뢰구간을 구하거나 통계적 가설(모수의 설정값)을 검정한다.

$$\frac{\hat{\theta} - E(\hat{\theta})}{s(\hat{\theta})} \sim Dist.???$$

모수가 모집단 평균인 경우 $\theta = \mu$ MVUE 표본평균은 $\hat{\theta} = \bar{x}$이다. 표본평균은 중심극한 정리에 의해 샘플링 분포는 정규분포에 근사한다. $\dfrac{\bar{x} - \mu}{s(\bar{x}) = s/\sqrt{n}}$

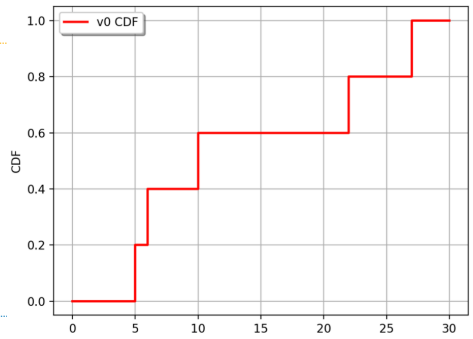
순서통계량 order statistic

정의

데이터 관측값 $(x_1, x_2, \ldots, x_n)$을 크기 순으로 정렬한 통계량으로 $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$

- 최대값 maximum value : 데이터 관측값 중 가장 큰 값 : $x_{(n)}$
- 최소값 minimum value : 데이터 관측값 중 가장 작은 값 : $x_{(1)}$
- 범위 range : 최대값과 최소값의 차이 : $R = x_{(n)} - x_{(1)}$

(예제 데이터) 1 0 0 4 3 3 2 1 2 2 데이터 크기 $n = 10$ $(x_1, x_2, \ldots, x_{10})$
순서통계량 $(x_{(1)} = 0, x_{(2)} = 0, \ldots, x_{(10)} = 4)$
최소값 $x_{(0)} = 0$, 최소값 $x_{(10)} = 4$



순서통계량 분포

확률변수 X 누적확률분포함수 정의 : $F(x) = P(X \le x)$

데이터 실증 CDF $(x_{(i)}, F(x_{(i)}))$

최소값 $x_{(1)}$ 분포함수

누적분포함수 : $F(x_{(1)}) = P(X_{(1)} \le x_{(1)}) = P(all\ X_i \ge x_{(1)}) = 1 - [1 - F(x_{(1)})]^n$

확률분포함수 : $f(x_{(1)}) = n[1 - F(x)]^{n-1} f(x)$

최소값 $x_{(n)}$ 분포함수

누적분포함수 : $F(x_{(n)}) = P(X_{(n)} \le x_{(n)}) = P(all\ X_i \le x_{(n)}) = F(x)^n$

확률분포함수 : $f(x_{(n)}) = nF(x)^{n-1} f(x)$

$x_{(k)}$ 분포함수

확률분포함수 : $f(x_{(k)}) = k\binom{n}{k} f(x)[F(x)]^{k-1}[1 - F(x)]^{n-k}$

중앙값 median $\tilde{x}$

데이터 관측값을 크기 순서에서 나열한 순서통계량의 순서 가운데 있는 관측값

- Median Depth =중위값 위치, $MD = \frac{n+1}{2}$: 정수가 아닌 경우 양쪽 정수 값 순서통계량의 평균
- $F(\tilde{x}) = 0.5$

표본크기가 10인 경우 $MD = \frac{10+1}{2} = 5.5$ 중앙값 $\tilde{X} = \frac{x_{(5)} + x_{(6)}}{2} = (2+2)/2 = 2$

사분위값 Quartile

- 사분위값 quartile : 데이터 크기 순 25%(일 first 사분위,), 50%(이사분위, 중위값), 75%(삼사분위) 값,
- $F(Q_1) = 0.25, F(Q_3) = 0.75$
- 사분위 위치 Quartile Depth, $QD = \frac{\llcorner MD \lrcorner + 1}{2}$ (ㄴMDㄱ는 MD를 넘지 않는 최대 정수, ㄴ9.5ㄱ 인 경우 9)

표본크기가 10인 경우 $MD = 5.5, QD = \frac{5+1}{2} = 3$이므로 $Q_1 = x_{(3)} = 1, Q_3 = x_{(8)} = 3$

백분위값 percentile quantile

- $F(\alpha\ percentile) = \alpha$
- 20-분위값 percentile : 데이터를 크기순으로 정렬했을 때 데이터의 20%는 그 값보다 작고 80%는 그 값보다 큰 그 값을 20%-분위값이라 한다. 25% -백분위값이 제1사분위값이다.
- 보간법 (예 : 표본크기 n=22이면, 80% 백분위 위치는 0.8*22=17.6 위치, $x_{(17)}$, $x_{(18)}$ 활용한 보간법으로 (0.4:0.6) 배분하여 구한다.

표본크기가 10인 경우 80% 분위값은 $x_{(8)} = 3$이다.

중앙위치

확률표본 데이터(관측값 $(x_1, x_2, \ldots, x_n)$)의 중앙위치 center location measure에 대한 정edian보이며 관측값 절대 크기의 중앙과 순서의 중앙으로 나뉜다. 순서의 중앙은 앞에서 살펴본 중앙값 median이다.

중앙값 median

계산식 : $\widetilde{X} = x_{(md)}$, 중앙위치 median depth = $\dfrac{n+1}{2}$

- [장점] 절대 크기로 인한 영향이 적어 (치우침의 영향으로 인한 왜곡 없음) 중앙 위치 통계량으로 가장 적절하다.
- 그러나 중앙값 샘플링 확률분포함수(모수 구간추정 및 가설 검정에 반드시 필요)를 구하기 어려워 모수 추론이 불가능 - 비모수추론 distribution free test을 한다.
- 중앙 위치 척도는 크기(치우침, 이상치)의 영향이 적은 중앙값이 적절하나 모집단 평균 추론을 위하여는 중심극한정리에 의해 샘플링 분포가 알려진 표본평균을 주로 활용한다.

평균 mean

(기호) μ : 모집단 평균(모수) , $\bar{x}$: 표본평균(통계량)

계산식 : $\mu = \dfrac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$

- [장점] 표본평균 샘플링 확률분포함수를 알 수 있어(중심극한정리) 신뢰구간 및 가설검정 추론이 가능하다. -집단 평균 비교 가능
- [단점] 좌우 대칭이 아닌 치우침 데이터는 중앙 위치의 왜곡이 발생할 수 있다.
- 그러므로 데이터의 분포를 좌우 대칭(정규변환)으로 만든 후 평균을 이용하는 것이 적절하다.
- 통계량 표본평균의 평균은 $E(\bar{x}) = \mu$, (추정)분산은 $V(\bar{X}) = \dfrac{\sigma^2}{n}$, 통계량의 표준편차인 표준오차 (standard error) $s(\bar{x}) = \dfrac{\sigma}{\sqrt{n}}$이다.

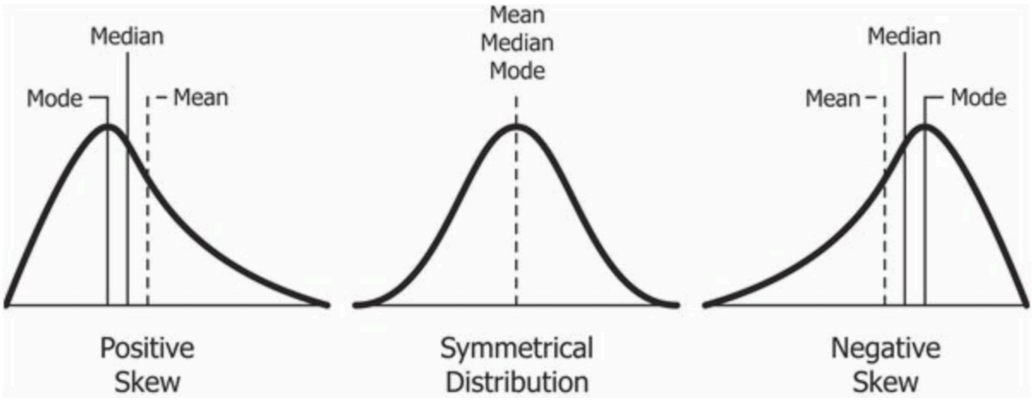
Central Limit Theorem : 모집단 확률분포함수에 관계없이 표본 크기가 충분히 크면() 표본평균(표본합)은 정규분포에 근사한다. $\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) = z$

치우침

치우침

정규분포나 t-분포와 같이 중앙을 중심으로 데칼코마니 좌우 동일하면 이를 벨모양 bell shaped, 좌우 대칭 symmetric 분포라 한다. 이와 달리 오른쪽 꼬리가 길거나 왼쪽 꼬리가 긴 경우 이를 치우친 skewed 분포라 한다.

- 우로 치우침 positive, right skewed : 오른쪽 꼬리가 길다. (평균>중앙값) 오른쪽 꼬리가 길다는 것은 상대적으로 크기가 큰 관측값이 존재한다는 것이고 그 값이 순서는 1이지만 크기는 매우 크므로 중앙값보다는 평균이 커지게 된다.
- 좌로 치우침 negative, left skewed: 왼쪽 꼬리가 길다. (평균<중앙값)



우로 치우침과 확률분포함수 형태

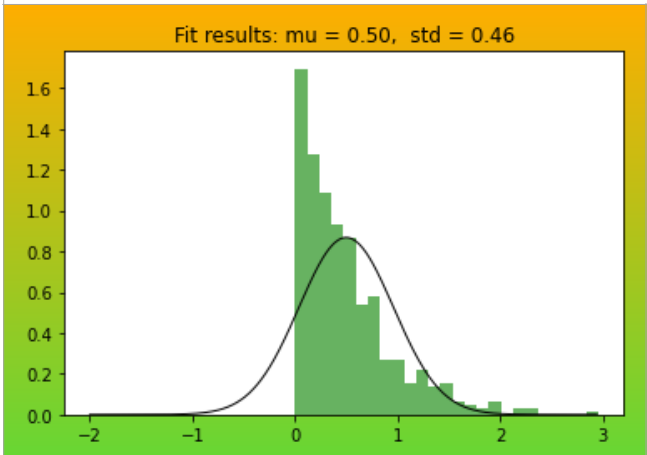
평균이 0.5인 지수분포를 따르는 데이터 500개를 생성 generating하여 히스토그램과 정규분포함수를 적용해 보자. Numpy 모듈의 random 함수 이용하여 생성 후 data 오브젝트에 저장

```
import numpy as np
data=np.random.exponential(1/2,500)
```

```
import scipy.stats as st
import matplotlib.pyplot as plt
mu, sd = norm.fit(data)
plt.hist(data, bins=25, density=True, alpha=0.6, color='g')
x=np.arange(-2,max(data),0.01)
p_norm =st.norm.pdf(x, mu, sd)
plt.plot(x, p_norm, 'k', linewidth=1)
title = "Fit results: mu = %.2f,  std = %.2f" % (mu, sd)
plt.title(title)
plt.show()
```

```
print("Mean=%.3f , Median=%.3f" %(mu,np.quantile(data,0.5)))
```

데이터 히스토그램을 그린 결과 우로 치우친 형태를 가진다. 실선 그래프는 데이터 평균과 표준편차를 따르는 정규분포를 적합한 것이다. 우로 치우친 향태를 가지므로 중앙값 0.331 보다 평균 0.488이 더 크다. 치우침이 클수록 평균과 중앙값의 차이는 커진다.




```
Mean=0.488 , Median=0.331
```

P-P plot

[확률 그림 Probability -Probability Plot] 데이터의 누적분포함수(y-축)와 이론분포(일반적으로 정규분포)의 누적분포함수(x-축)를 산점도에 나타내고 45도 기울기 직선을 그어 점들이 직선상에 있으면 데이터가 이론 데이터 분포에 따른다고 시각적으로 판단한다.

[확률 그림 Qunatile -Quantile Plot] 데이터의 백분위 값(Y-축)과 이론분포(일반적으로 정규분포 혹은 다른 데이터 셋)의 백분위 값(x-축)을 표시하여 P-P 플롯과 동일하게 활용한다.

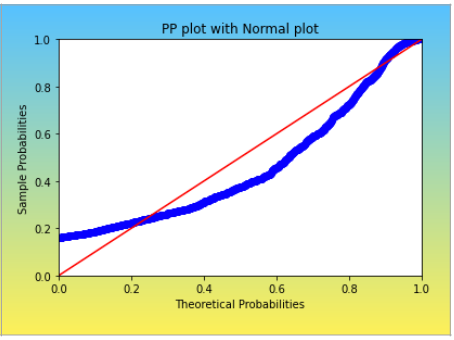
[데이터 백분위 값 구하기]

```
import numpy as np
q=np.linspace(0, 100, 101)
data_q=np.percentile(data, q)
print(data_q[50],np.quantile(data,0.5))
```

```
0.3697005852026205 0.3697005852026205
```

[우로 치우친 분포는 왼쪽 꼬리부분은 데이터 백분위 값이 크고 중앙에 가까워지므로 데이터 백분위 값이 작아지다가 오른쪽 꼬리에 가까워질수록 다시 커진다]

```
pp_x=sm.ProbPlot(data,fit=True,dist=norm)
pp_x.ppplot(line="45")
plt.title('PP plot with Normal plot')
plt.show()
```



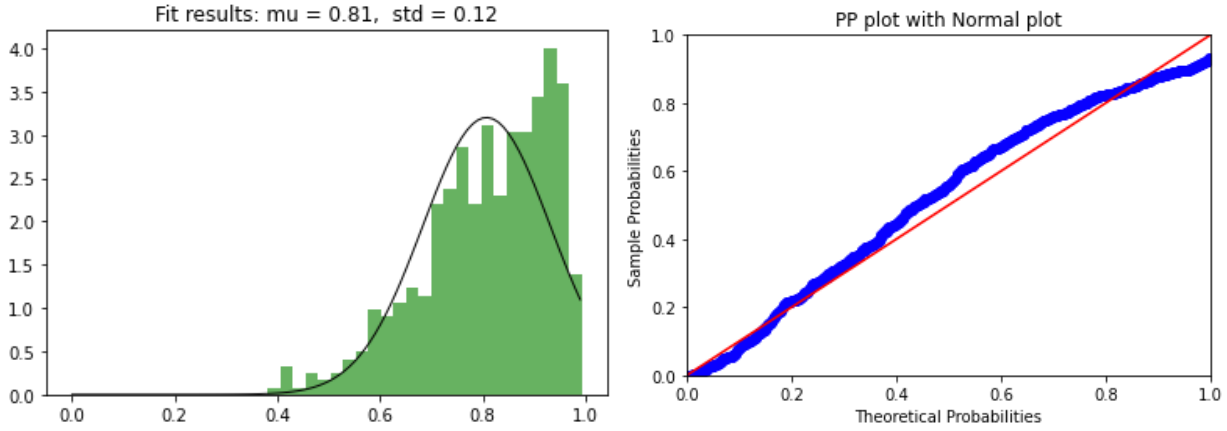
좌로 치우침과 확률분포함수 형태

모수($\alpha = 8, \beta = 2$)인 베타분포를 따르는 데이터 500개를 생성 generating하여 히스토그램과 정규분포함수를 적용해 보자. Numpy 모듈의 random 함수 이용하여 생성 후 data 오브젝트에 저장

```
import numpy as np
data=np.random.beta(8,2,500)
```

```
0.8260151707946077 0.8260151707946077
```

[좌로 치우친 분포는 왼쪽 꼬리부분은 데이터 백분위 값이 작고 중앙에 가까워지므로 데이터 백분위 값이 커지지다가 오른쪽 꼬리에 가까워질수록 다시 작아진다] <-> 우로 치우침과는 반대 향태

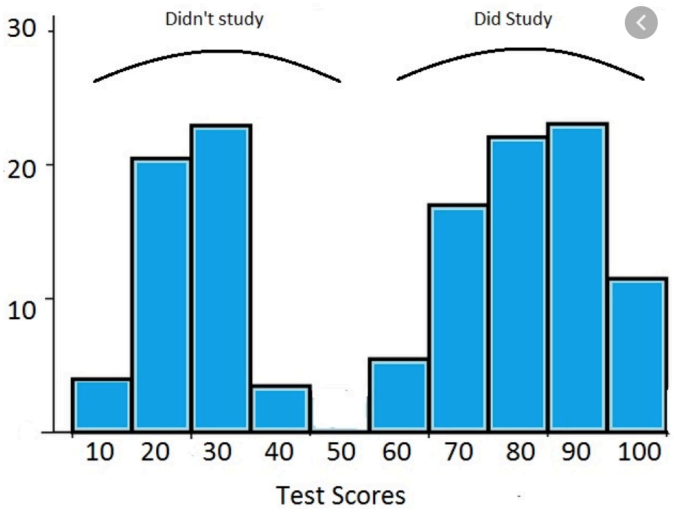


봉우리

봉우리 개수

두 개의 서로 다른 집단 데이터의 히스토그램을 그리는 경우 봉우리가 2개 나타나는 경우가 발생 - 상자 수염 그림으로는 판별 불가

옆의 히스토그램은 시험 공부한 집단과 그렇지 않은 집단의 시험 성적 히스토그램이다. -> 데이터 분리하여 분석



산포(흩어짐)

범위 range

계산식 : $R = x_{(n)} - x_{(1)}$: 데이터 최대값과 최소값의 차이

- 치우침과 이상치에 영향이 커 자주 사용하지 않는다.
- 활용 : Empirical Rule(실증적 법칙)에 의하여 데이터 대부분은 $\mu \pm 3\sigma$ 구간에 들어간다. 그러므로 $R = \frac{\sigma}{6}$이므로 범위를 알면 $\hat{\sigma} = \frac{R}{4}$로 추정한다. 6으로 나누지 않는 이유는 과소 추정을 막기 위함이다. 필드에서 표준편차를 과소 추정하면 표본크기 결정 등에 왜곡이 발생할 수 있기 때문이다.

사분위범위 Inter Quartile Range

계산식 : $IQR = Q_3 - Q_1$: 데이터 최대값과 최소값의 차이

- 치우침, 이상치에 영향을 받지 않아 좋은 산포 척도이나 샘플링 분포를 구하기 어려워 흩어짐 추론에는 사용하지 않는다.
- 활용 : 국내 대학은 수능성적을 공시할 때 입학생의 평균 수능성적만 공시하나 미국의 경우에는 사분위 범위, 즉 상위 25%~하위25% 성적을 공지한다. 서로 다른 집단의 산포를 비교하는데 가장 적절한 척도이다.

분산, 표준편차 Variance, standard deviation $\sigma^2,\ \sigma,\ s^2,\ s$

계산식 : 모집단 분산 $\sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$, 표본분산 $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$, 분산의 양의 제곱근이 표준편차임

- 치우침과 이상치의 영향을 많이 받지만 샘플링 분포가 알려져 있어 추론에 사용된다.
- 그러므로 치우침과 이상치가 있는 경우는 이를 해결한 후 추론을 진행하는 것이 적절하다.

변동계수(coefficient of variation, CV)

표준편차를 평균으로 나눈 값 : $CV = \frac{s}{\bar{x}}$ [참고] $RSE = \frac{se(\bar{x})}{\bar{x}}$

- 측정단위가 서로 다른 자료의 산포를 비교할 때 사용한다.
- [예] 공대생의 일주일 공부시간 평균=10시간, 표준편차=5, 문과생은 평균 5시간, 표준편차=3시간이다. 누가 더 꾸준히 공부하는가? 꾸준하다는 것은 공부 시간의 편차가 작다는 것이다. 그럼 문과생이 더 꾸준한가? 아니다. 공대생의 CV=0.5이고 문과생은 0.6이므로 공대생이 더 꾸준히 공부한다.

상대표준오차 RSE relative standard erroe: 추정량의 표본추출오차를 나타내는 척도로 상대표준오차는 추정값의 표준오차를 추정값 자체로 나누어서 산출함

샘플링 분포

개념

모집단 모수에 대한 구간추정, 가설검정을 위하여 확률표본으로 부터 구한 통계량(MVUE)의 분포를 샘플링분포 $f(\hat{\theta})$라 한다. 모수 θ, MVUE 추정량을 $\hat{\theta}$라 하면 추정량의 평균은 θ(왜냐하면 불편 추정량이므로), 표준편차는 (추정)표준오차인 $s(\hat{\theta})$이다.

$$\frac{\hat{\theta} - E(\theta)}{s(\hat{\theta})} \sim app.\ sampling\ distribution$$

모수와 통계량 & 샘플링분포

모수(θ)	상황	MVUE	표준오차 $s(\hat{\theta})$	근거/대체	샘플링분포
평균 μ	대표본	$\bar{x}$	$s(\bar{x}) = \frac{\sigma}{n}$	중심극한정리 σ는 s로 대체	$\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim z$
	소표본			모집단 정규분포 가정	$\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t(n-1)$
비율 p	대표본	$\hat{p} = \frac{x}{n}$	$s(\bar{x}) = \frac{p(1-p)}{n}$	중심극한정리	$\frac{\hat{p}-p}{p(1-p)/\sqrt{n}} \sim z$
	소표본	$\sum x$	$s(\sum x) = np(1-p)$	이항분포 근거 $\sum x_i \sim B(n,p)$	$p = P(\sum x_i \le x)$
분산 σ^2	-	s^2	$s(s^2) = \frac{2\sigma^4}{n-1}$	모집단 정규분포 가정	$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$

$100(1-\alpha)\,\%$ 신뢰구간

모비율 $P(\hat{p} - z_{(1-\alpha/2)}\frac{\hat{p}(1-\hat{p})}{\sqrt{n}} \le \mu \le \bar{x} + z_{(1-\alpha/2)}\frac{\hat{p}(1-\hat{p})}{\sqrt{n}} = \alpha$

모평균 $P(\bar{x} - z_{(1-\alpha/2)}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + z_{(1-\alpha/2)}\frac{s}{\sqrt{n}}) = \alpha$

모분산 $P(\frac{s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{s^2}{\chi^2_{1-\alpha/2}}) = \alpha$