데이터와 분포 #4

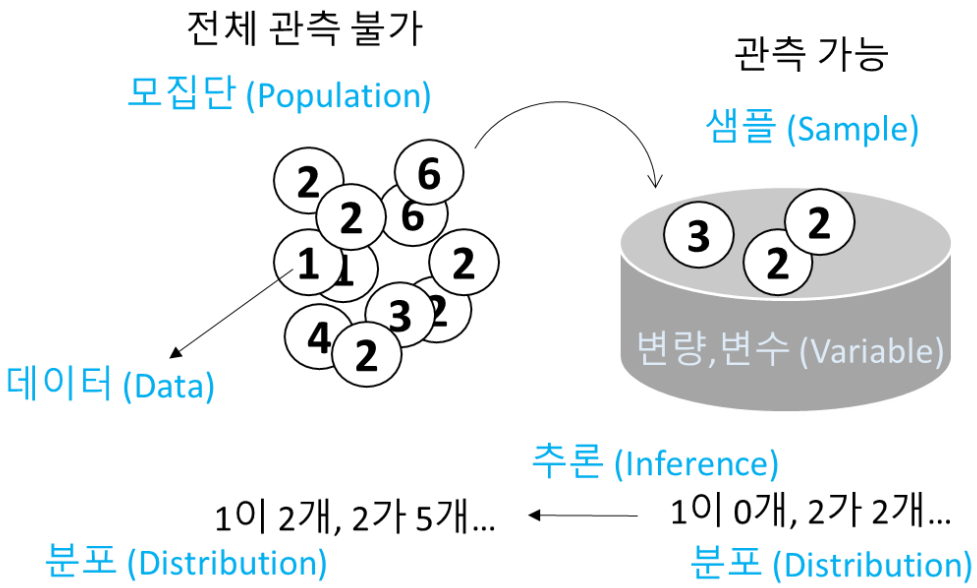
과학기술연합대학원대학교
한국생명공학연구원 스쿨
시스템생명공학전공

haseong@kribb.re.kr
김하성

https://github.com/greendaygh/2019-USTOT-stat

Summary of lecture #1

통계
데이터, 정보
일변량
요약통계량
-Center: mean, median..
-Spread: variance, range..
-Shape: skewness, ..



Summary of lecture #2

이변량 데이터 비교

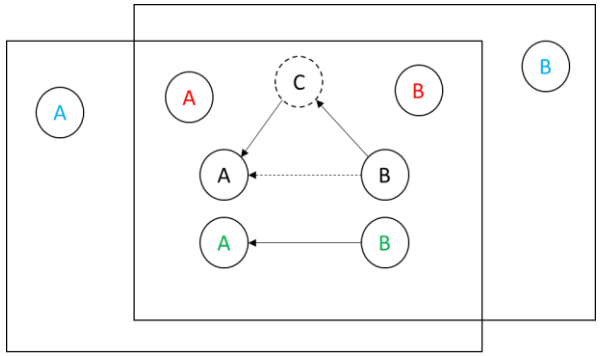
1. Numerical data

   Unpaired data – Similarity with summaries

   Paired data – Relationship (Covariance / correlation / regression)

2. Categorical data

   Paired data – Relationship (chi-squared statistic)



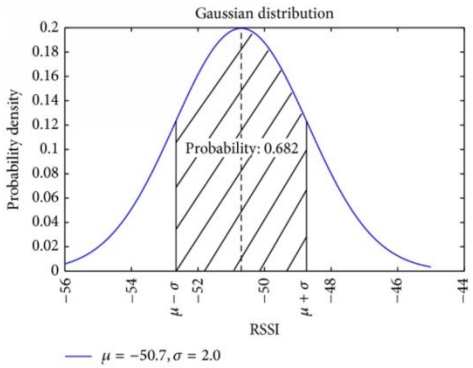
Independence (독립)
Correlation (상관)
Association (연관)
Causation (인과)

Summary of lecture #3

다변량 데이터 비교

1. Boxplot, Scatter plot, Heatmap, etc
2. Population, Random variable, Distribution

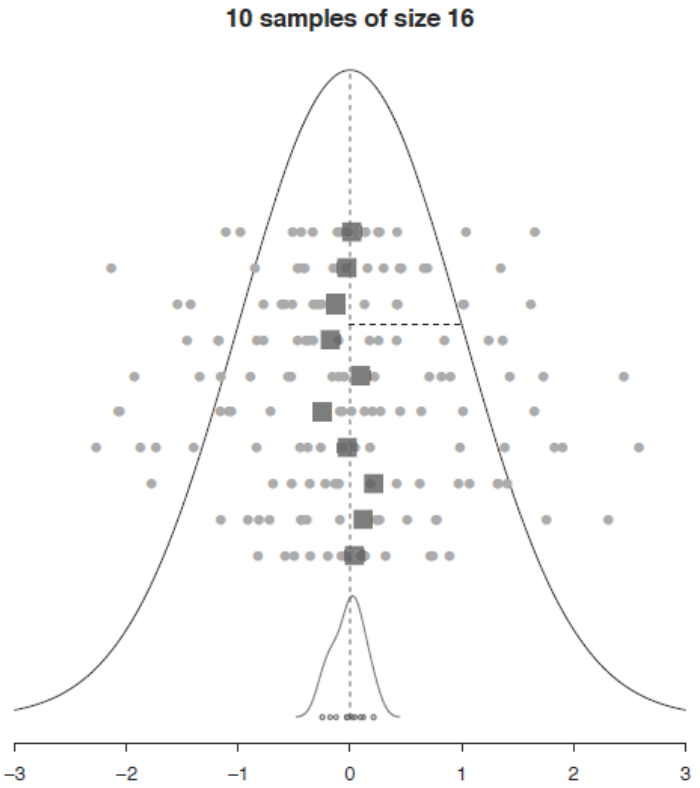
Population	Variable (Observed)	Random variable (Before observation)
Cars93.Price	Cars93.Price	Cars93.Price
15.9	15.9	
33.9	33.9	Cars93.Price == 15.9 ?
29.1	29.1	Cars93.Price == 15.9 probability?
37.7	34.7	P(Cars93.Price=15.9)=
30.0	40.1	P(X=x)=
15.7	13.4	
20.8	18.4	
23.7		
26.3		
34.7		
40.1		
13.4		
….		



Statistical inference

- Population – 랜덤하게 선발된 데이터의 분포로 추정
- parameters – population의 분포를 설명하는 상수
- Samples – 관측 데이터
- Statistics – 샘플의 summary

- Confidence interval
- Significance tests



Simulation

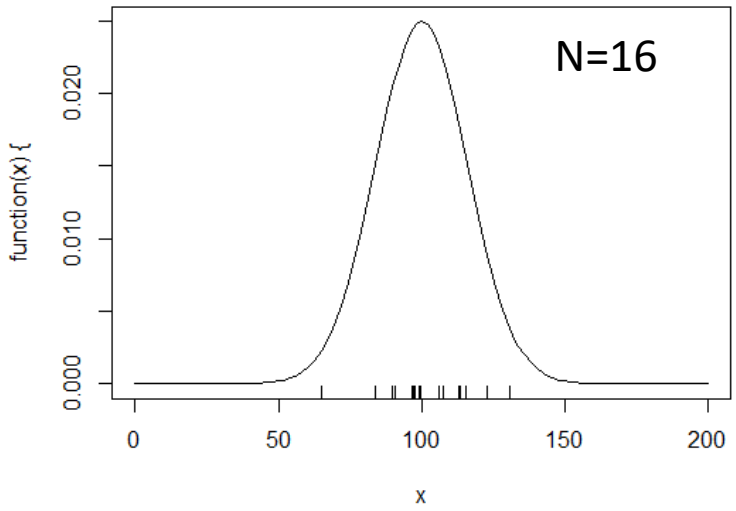
Simulation

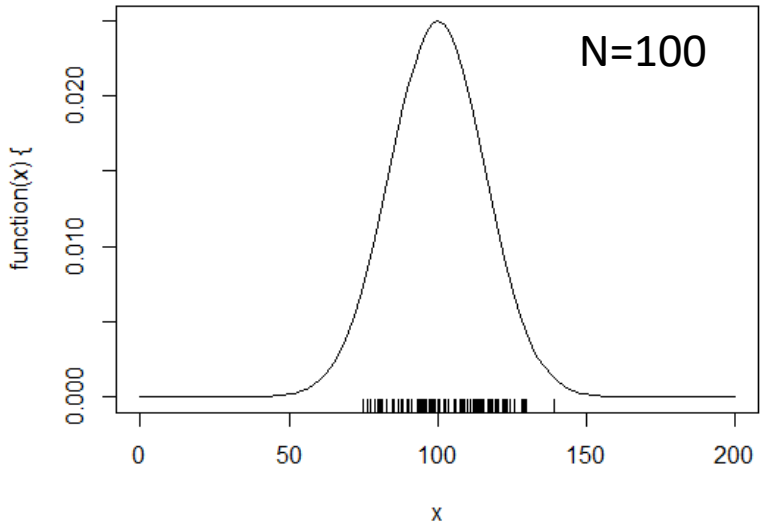
모집단 (모형) 특성을 연구하기 위한 모사 데이터

Specify a probability model and parameters

Draw samples randomly from the probability model

```
> mu <- 100; sigma <- 16
> n <- rnorm(16, mu, sigma)
> n
 [1]  97.12261  64.88136  83.88370 122.74030
 [5] 113.03665  97.66376 130.64473  98.98449
 [9] 115.40776  99.73519  96.78120  89.89863
[13] 113.45250 106.00097  90.73636 107.50135
> mean(n)
[1] 101.7795
```





The central limit theorem

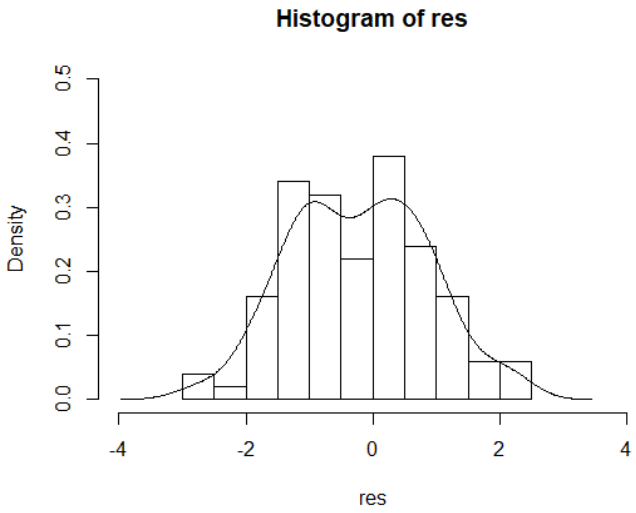
number of samples

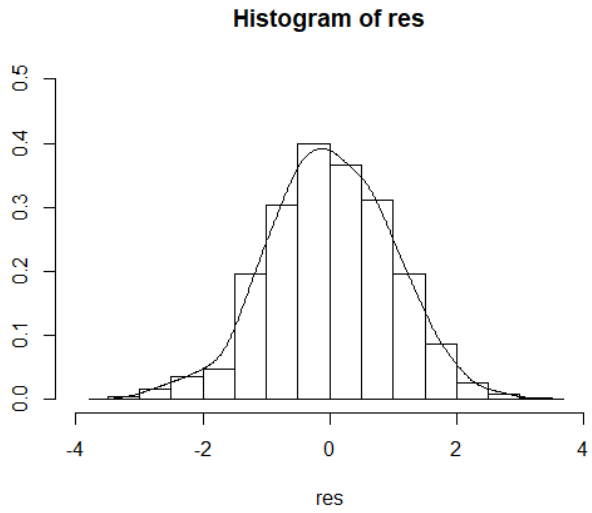
number of repeats

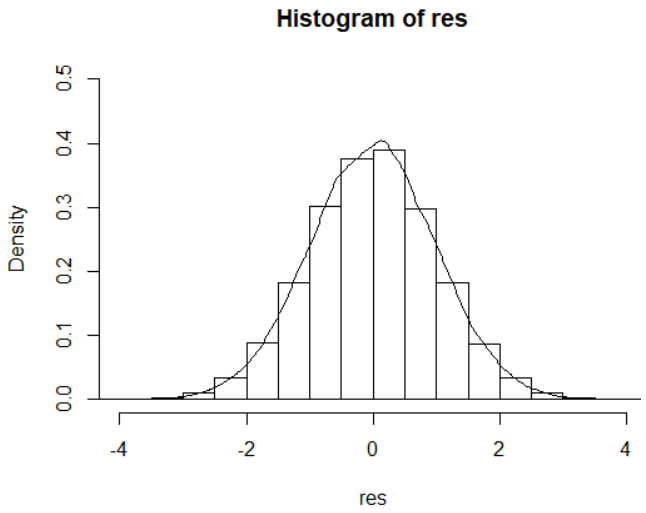
N=7
M=100

N=7
M=1000

N=7
M=100000

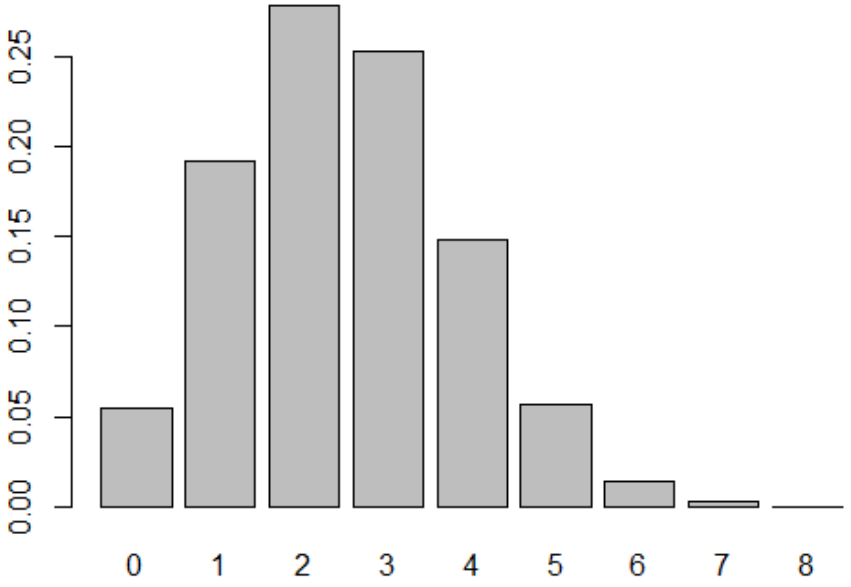






Estimating probabilities

- {1, 2, 3, …, 19, 20}
- Draw 10 numbers from {1, 2, 3, 4, …, 78, 79, 80}
- A player wins if 3 or more numbers are matched



Significance tests

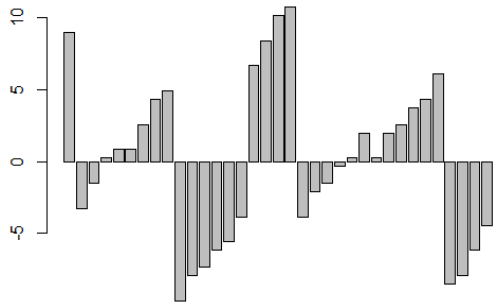
- Q: Does a treatment induce a notable effect?
- Ex: Dose consuming an amount of honey during exercise increase performance?
- 7 people, two groups (Control: 23, 33, 40, Treatment: 19, 22, 25, 26)

	Mean
Control	32
Treatment	23

Diff: 9 Significant? (유의?)

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
[1,]    1    1    1    1    1    1    1    1    1     1     1
[2,]    2    2    2    2    2    3    3    3    3     4     4
[3,]    3    4    5    6    7    4    5    6    7     5     6
```

…



How many of > 9 ?
How many of ≠ 9 ?

H0 vs. H1
H0: C=T

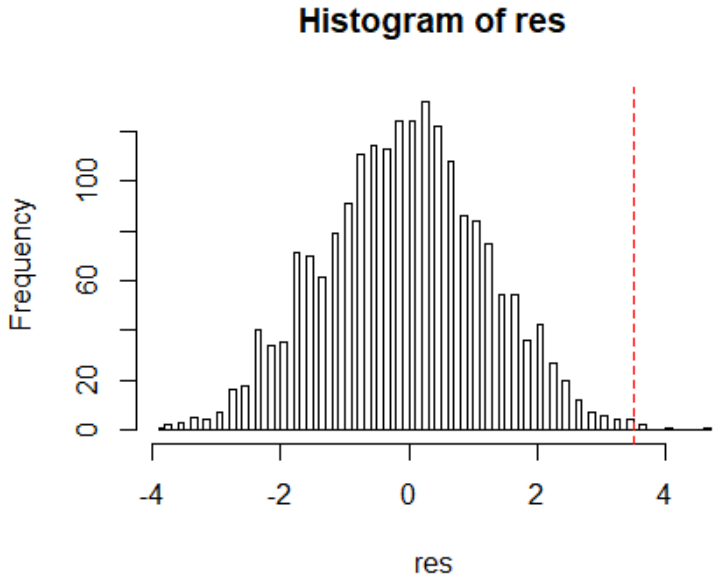
3/35 = 0.085

Significance tests

- Does caffeine make you jittery?
- 10 control vs 10 treatment
- difference: 3.5

```
caf <- c(245, 246, 246, 248, 248, 248, 250, 250, 250, 252)
no_caf <- c(242, 242, 242, 244, 244, 245, 246, 247, 248, 248)
the_data <- stack(list(caffeine=caf, no_caffeine=no_caf))

obs <- mean(caf)-mean(no_caf)
res <- replicate(2000, {
  ind <- sample(1:20, 10, replace=F)
  mean(the_data$values[ind]) - mean(the_data$values[-ind])
})
hist(res, br=100)
abline(v=obs, col="red", lty=2)
```



How well dose the sample estimate population?

*Estimation (추정): To provide an estimator of a population parameter, and its error margin

- ex: A poll asking a random sample of 1003 whether marriages between same-sex couples should be recognized by law as valid. 55% said yes
- a randomly selected person would responding yes with p = 0.55, $\hat{p} - p \approx 0$

*Unbiased estimators : $E(\bar{x}) = \mu, E(\hat{p}) = p, E(s^2) = \sigma^2$

- Where is the most of the data? == What does range contains 95% of the data?

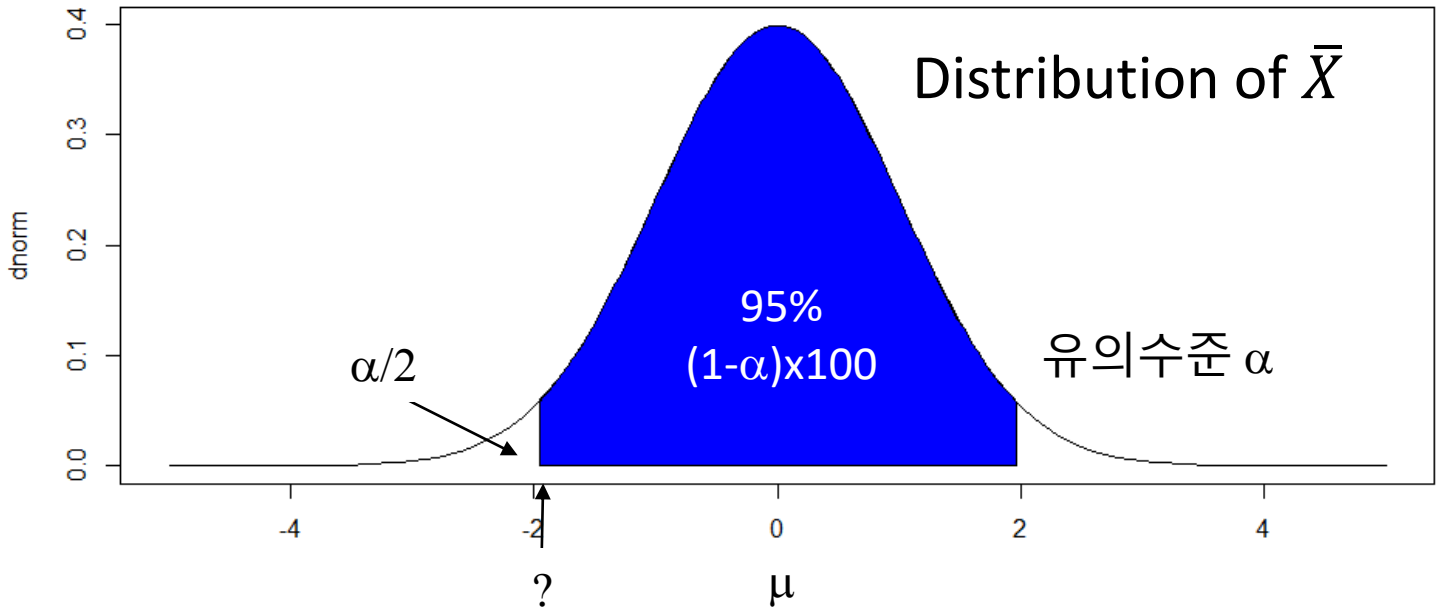
Estimation of population parameter (μ)

- A new process making tires
  - H0: Lifetime X < Lifetime X (new) of a tire
  - std dev = 5000
  - n=100, measure $\bar{X}$ = 38,000km
  - What is the μ of a tire with the new process?

$$\hat{\mu} = \bar{X} = 38{,}000$$

  - How confidence with this?
  - What is the probability that another 100 samples give the same value?

Confidence intervals



? = k * std of $\bar{X}$

k is a value such that p(z < k) = 0.025 in standard normal dist.

1. The probability that μ is within $\pm$ 980 is 0.95
2. $\mu - 980 < \bar{X} < \mu + 980$ ➔ $\mu \in (\bar{X} - 980, \bar{X} + 980)$
3. If one repeats to compute $\bar{X}$ and its 95% CI with n=100, then 95% of the CIs will include μ

Confidence intervals #2

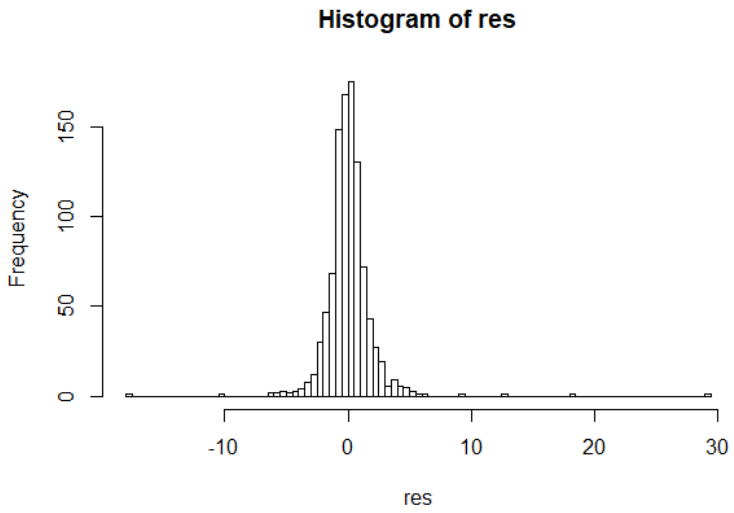
$$\bar{x} = \mu \quad \text{vs.} \quad \bar{x} - \mu = 0$$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim T$$

Estimator of population mean

```
mu <- 100
sigma <- 16
M <- 1000
n <- 4
res <- replicate(M, {
        x <- rnorm(n, mu, sigma)
        (mean(x)-mu) / sd(x)/sqrt(n)
}
```

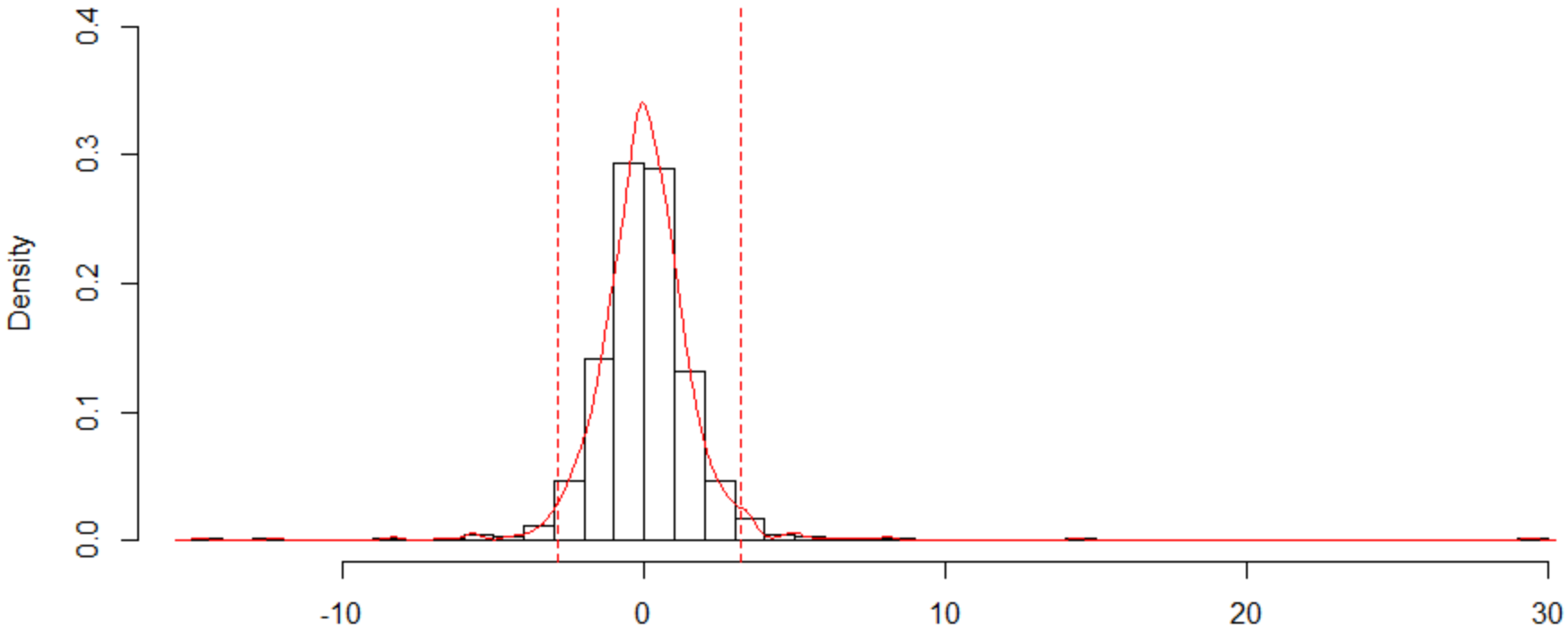


Confidence intervals #3

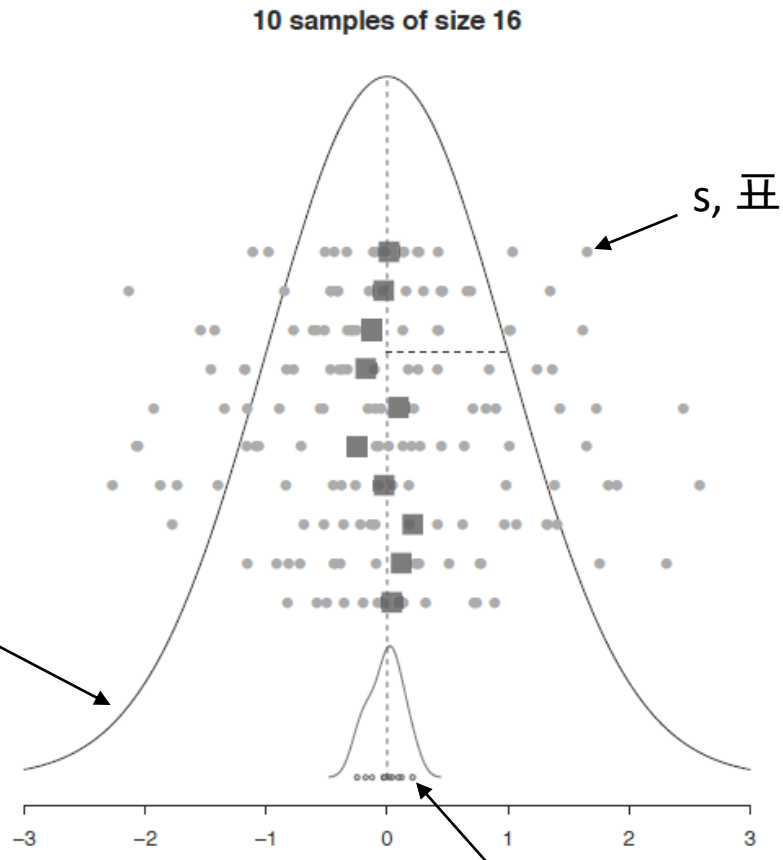
```
> quantile(res, c(0.025, 0.975))
     2.5%     97.5%
-3.031903  3.701658
```

$$-3.03 < \frac{\bar{x} - \mu}{SE} < 3.70$$

$$\bar{x} - 3.70 \cdot SE < \mu < \bar{x} + 3.03 \cdot SE$$



Deviations



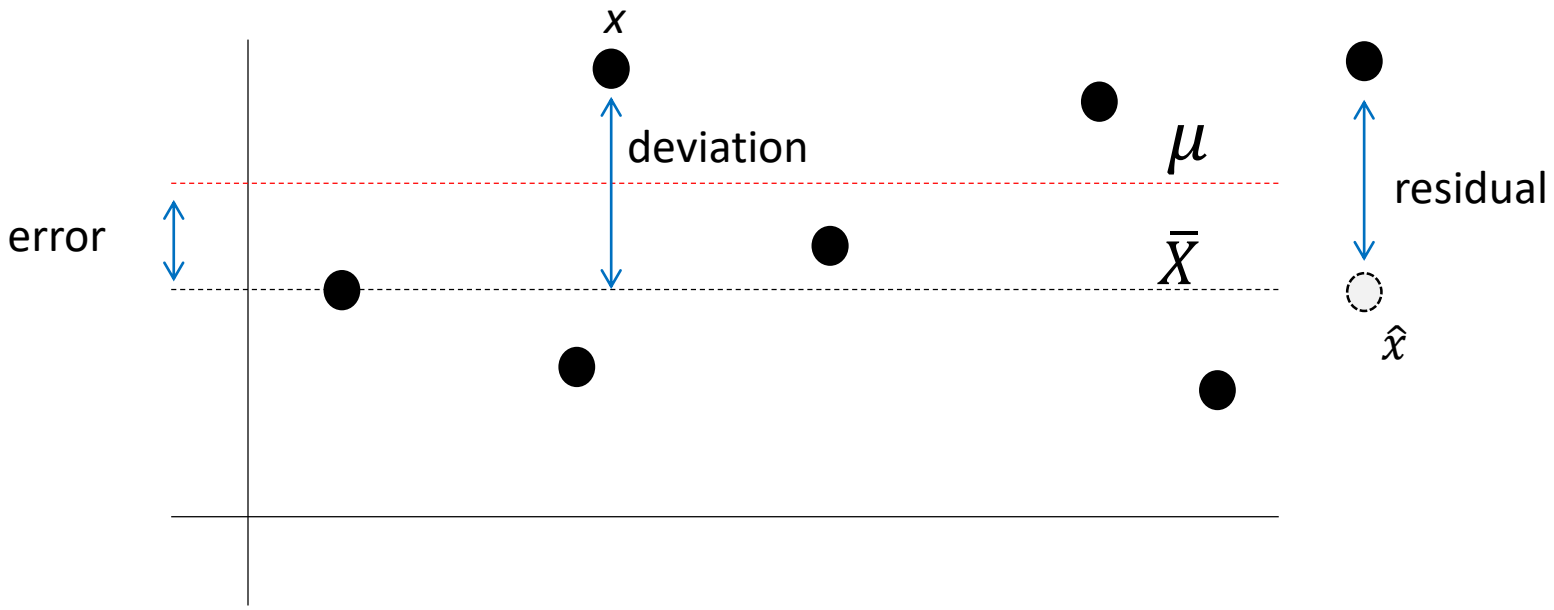
s, 표본표준편차

σ, 표준편차

$s/\sqrt{n}$ (표준오차) 표본평균 표준편자

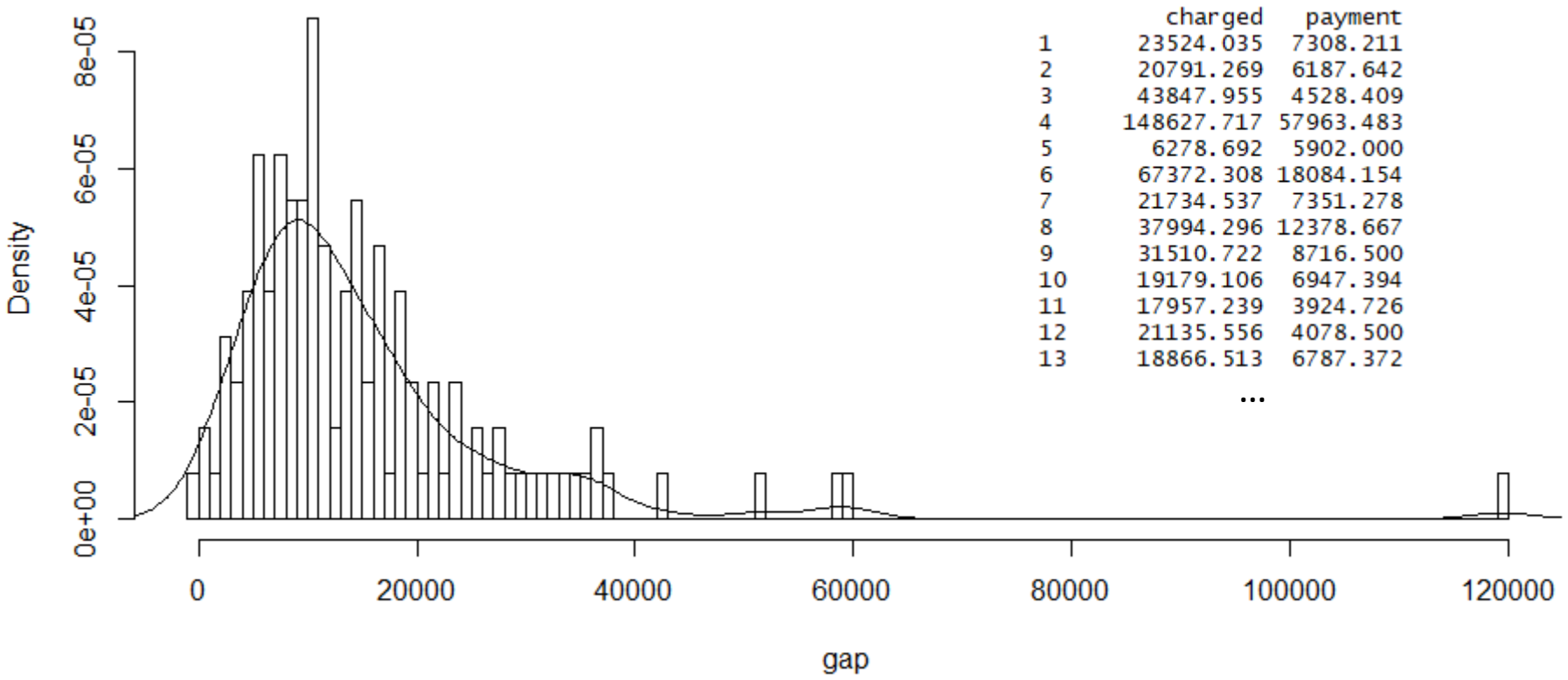
Deviation vs. Error vs. Residual

- 편차 (deviation) – difference between observed data (x) and sample mean ($\bar{X}$)
- 오차 (error) - Diff. between sample mean ($\bar{X}$) and population statistics (mean) (μ)
- 잔차 (residual) - Diff. between observed data (x) and predicted data ($\hat{x}$)



The basic bootstrap #1

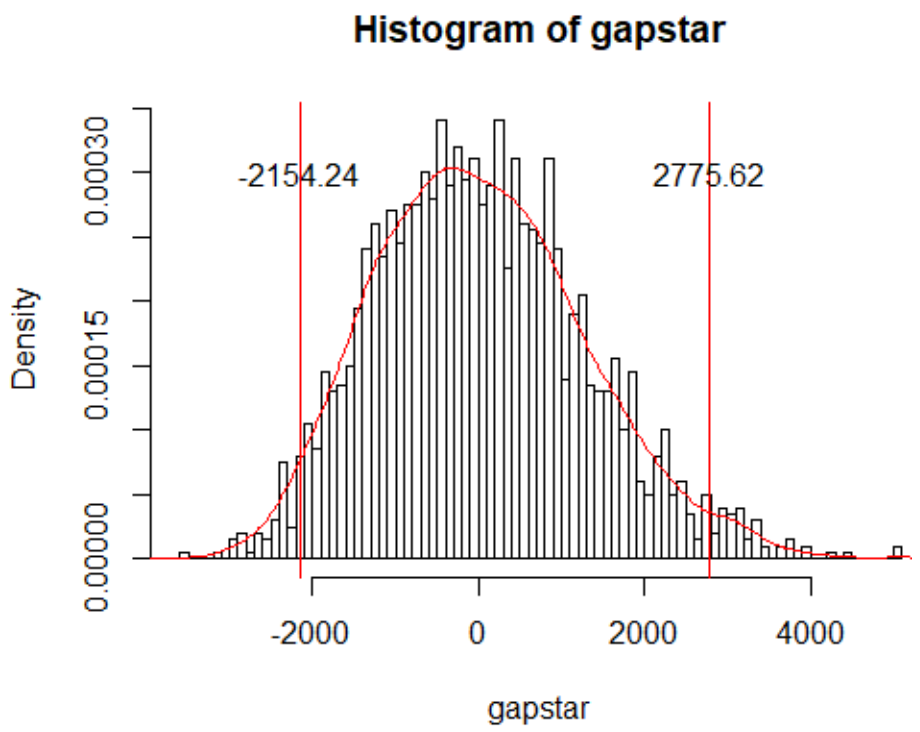
- Computer-intensive method for generating sampling distribution
- mean(Charged – Payment) = 15,445

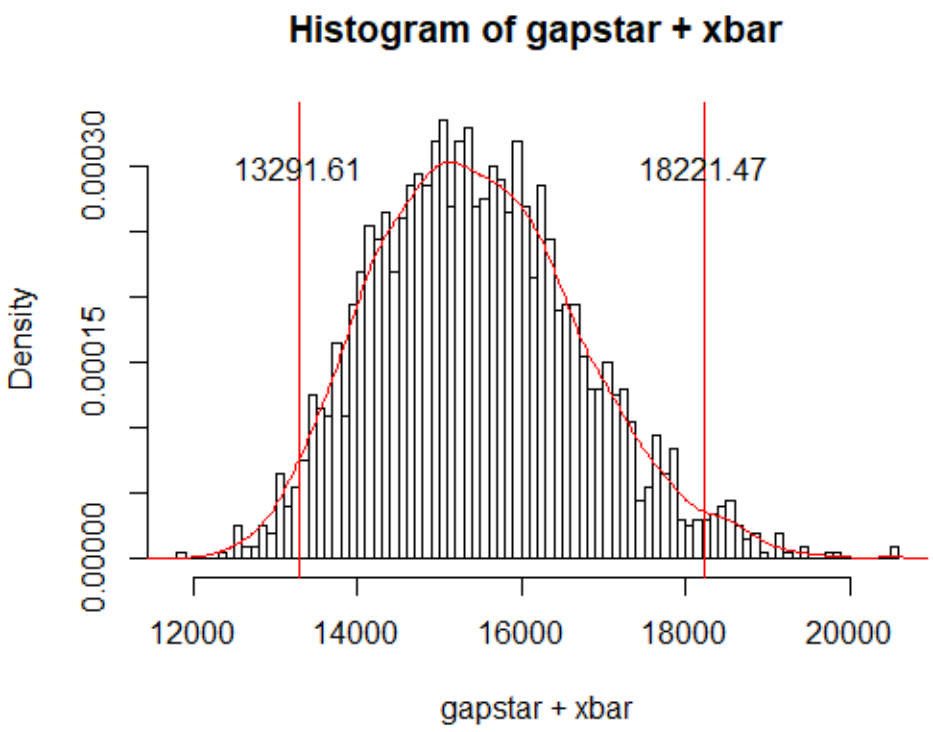


	charged	payment
1	23524.035	7308.211
2	20791.269	6187.642
3	43847.955	4528.409
4	148627.717	57963.483
5	6278.692	5902.000
6	67372.308	18084.154
7	21734.537	7351.278
8	37994.296	12378.667
9	31510.722	8716.500
10	19179.106	6947.394
11	17957.239	3924.726
12	21135.556	4078.500
13	18866.513	6787.372
	...	

The basic bootstrap #2

- 95% confidence interval





15,445 -2154.24 < μ < 15,445 + 2775.62

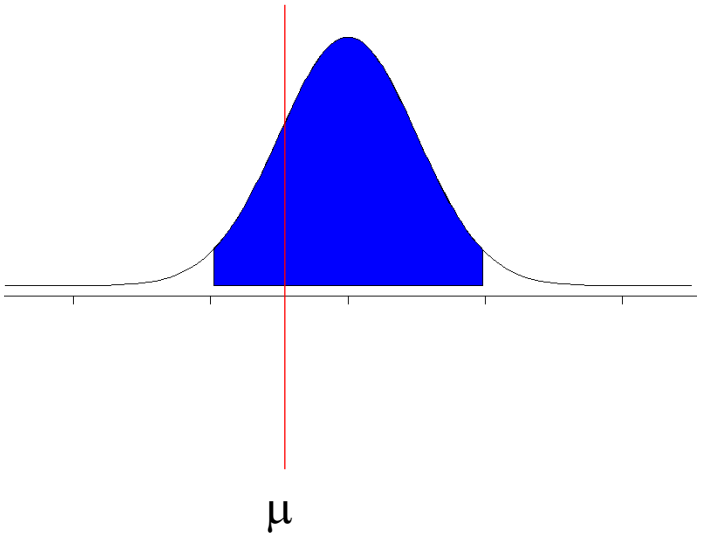
- 99% confidence interval ?

Interpretation of CI

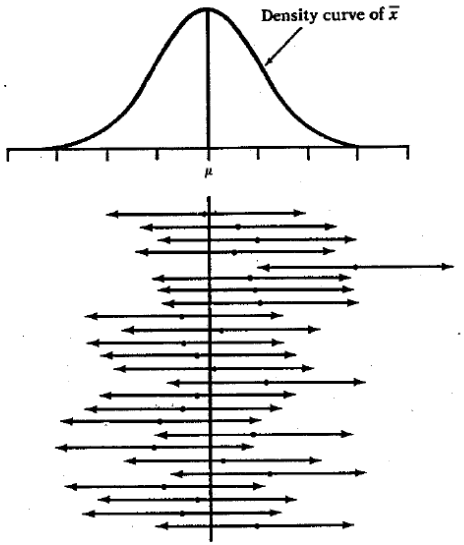
μ 가 CI 에 있을 확률이 0.95 이다

vs

Random sampling을 100번 수행하고 각각의 CI를 구할 때
그들 중 95%가 μ를 포함한다



vs



Bayesian analysis #1

- Conditional probability

$$P(B|A) \text{ vs } P(A|B)$$

- θ: unknown parameter, random
- Random -> Probability

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

$$P(\theta|X) \propto P(X|\theta)P(\theta)$$

$$Posterior\ dist\ \propto likelihood\ \cdot prior\ dist$$

Bayesian analysis #2

$$P(\theta|X) \propto P(X|\theta)P(\theta)$$

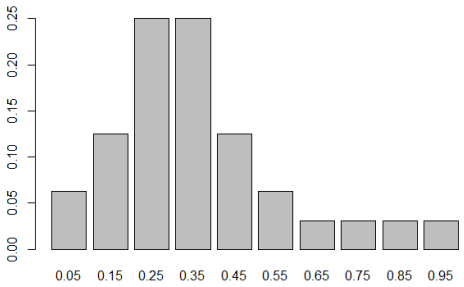
- likelihood: how likely the data is given a specific value for the parameter, can be computed from the model
- Bernoulli trials with *p*, # of success: *s*, total *n* trials

$$P(X|p) = p^s(1-p)^{n-s}$$

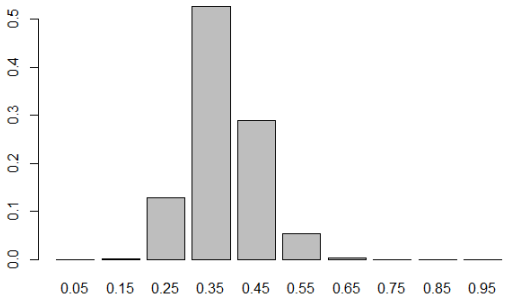
Bayesian analysis #3

- Let p be the proportion of college students getting less than 8 hours of sleep
- Observed 11 students out of 27
- A researcher believes around p=0.3

p=0.407 (11/27)?



prior



posterior

$$P(X|\theta) = p^s(1-p)^{n-s}$$

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

p=0.05일 때 $P(X|\theta) = 0.05^{11}(1-0.05)^{27-11}$

CI for the population mean

$$x_1, x_2, \ldots, x_n \sim N(\mu, \sigma^2)$$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - \mu}{SD(\bar{x})}$$

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\bar{x} - \mu}{SE(\bar{x})} = \frac{observed - expected}{SE}$$

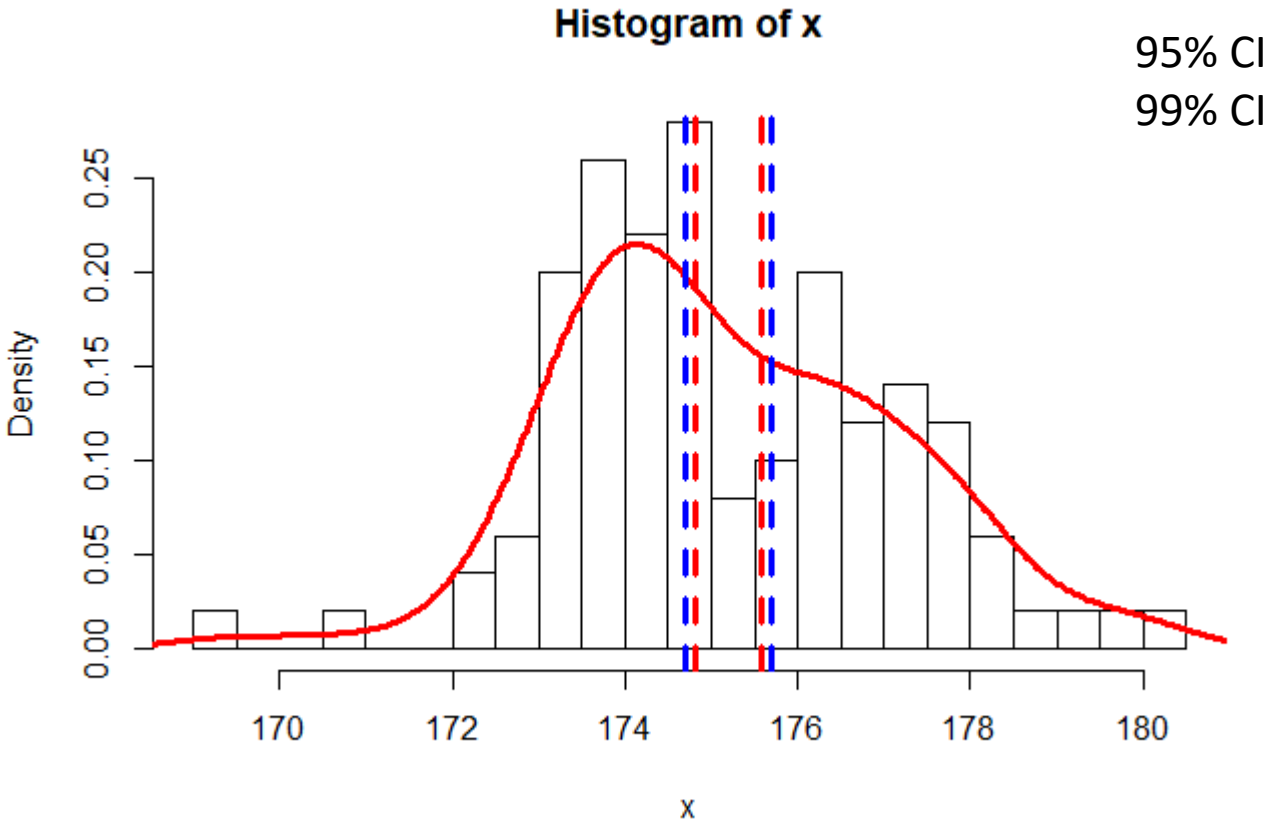
- *n* is large enough, CLT implies the sampling dist. approx. std. normal

$$P\left(-t^* < \frac{\bar{x} - \mu}{SE(\bar{x})} < t^*\right) = 1 - \alpha$$

$$P(\bar{x} - t^* \cdot SE(\bar{x}) < \mu < \bar{x} + t^* \cdot SE(\bar{x})) = 1 - \alpha$$

CI for the population mean

- Average height (mean:175cm, sd:2, n:100)



Significance tests

- Significance test for the mean (t-test)

Significance tests

- Assumes a value for the population parameter and computes a probability based on a sample given that assumption

> 무죄라는 건 죄가 없다는 뜻이 아냐. 죄를 저질렀다는 사실을 증명하지 못했다는 뜻이지.[3][4]
> - 차영우, 드라마 《개과천선》에서.

- H0 : Null hypothesis (no guilty)
- H1 : Alternative hypothesis (guilty) – no used in the trial
- Jury: determination of guilty by proofing a failure of the assumption of H0 (no guilty) to explain the evidence well enough
  - Evidence: statistic, Well enough: probability (p-value)

→under H0 is true, the probability that statistic is observed

→p-value (유의확률)

- Significance test: to calculate p-value

Standard levels of p-values

p-value range	significance stars	common description
$[0, 0.001]$	***	extremely significant
$(0.001, 0.01]$	**	highly significant
$(0.01, 0.05]$	*	statistically significant
$(0.05, 0.1]$	.	could be significant
$(0.1, 1]$		not significant

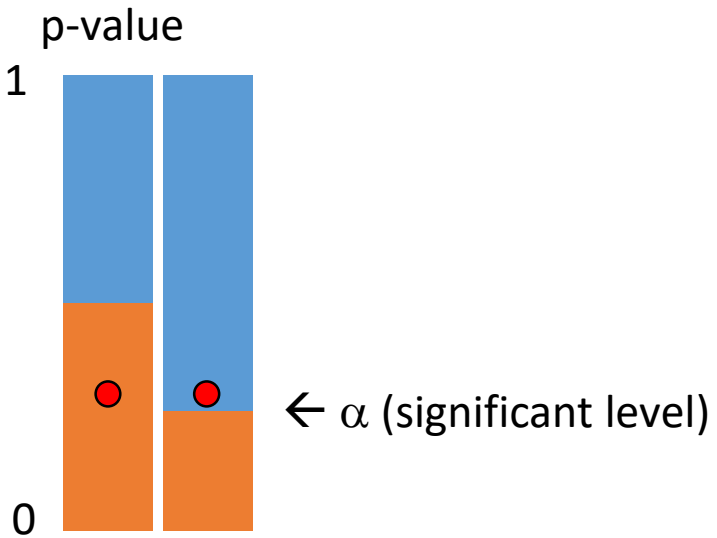
Table 9.1: Level of significance for range of p-values.

Decision of significance test

- We don't actually prove the null to be false or true
- "Reject" and "Accept" by specifying significance level α
- Errors
  - H0 is true but rejected: Type I error (no guilty man is found guilty)
  - H0 is false but accepted: Type II error (guilty man is found no guilty)

Table of error types		Null hypothesis (H_0) is	
		True	False
Decision About Null Hypothesis (H_0)	Fail to reject	Correct inference (True Negative) (Probability = 1 - α)	Type II error (False Negative) (Probability = β)
	Reject	Type I error (False Positive) (Probability = α)	Correct inference (True Positive) (Probability = 1 - β)

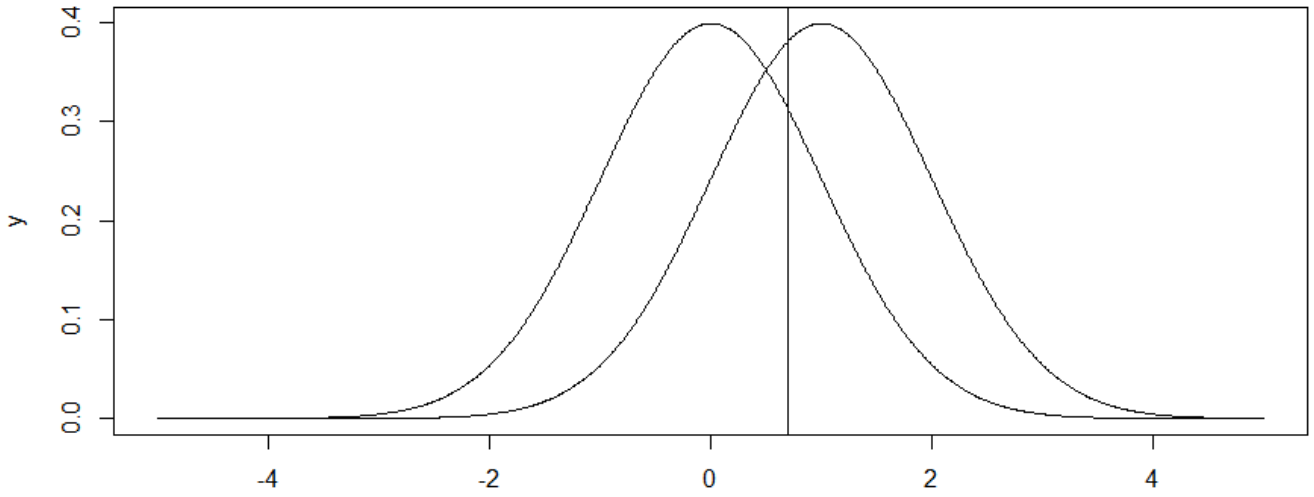
wiki



Ex) Which mean?

- A machine with calibration: N(0, 1)
- No Calibration: N(1, 1)
- Drawn a single number: 0.7
- Decide whether the machine is in calibration or not? (under the strong assumptions of N)

H0: $\mu=0$ vs. H1: $\mu=1$

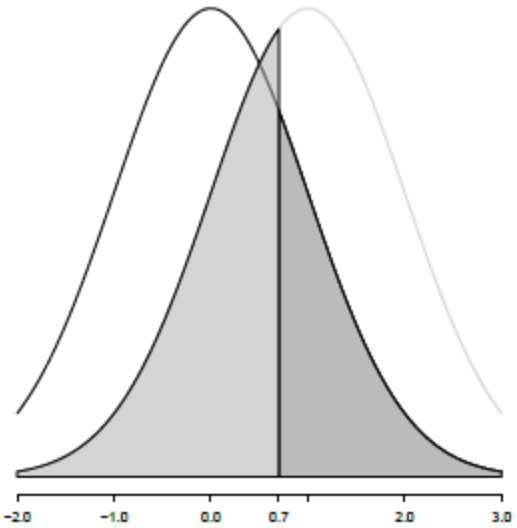


Testing

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{0.7 - 0}{1} = 0.7$$

1-pnorm(0.7, 0, 1) : 0.2419
pnorm(0.7, 1, 1) : 0.382

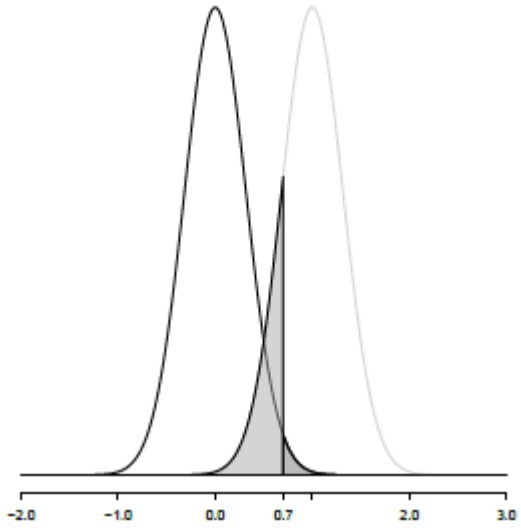


Testing sample means

- 0.7 is a value of sample mean

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$\mathrm{t} = \frac{0.7 - 0}{1/\sqrt{10}} = 0.0134$$



H0: μ=0 vs. H1: μ=1

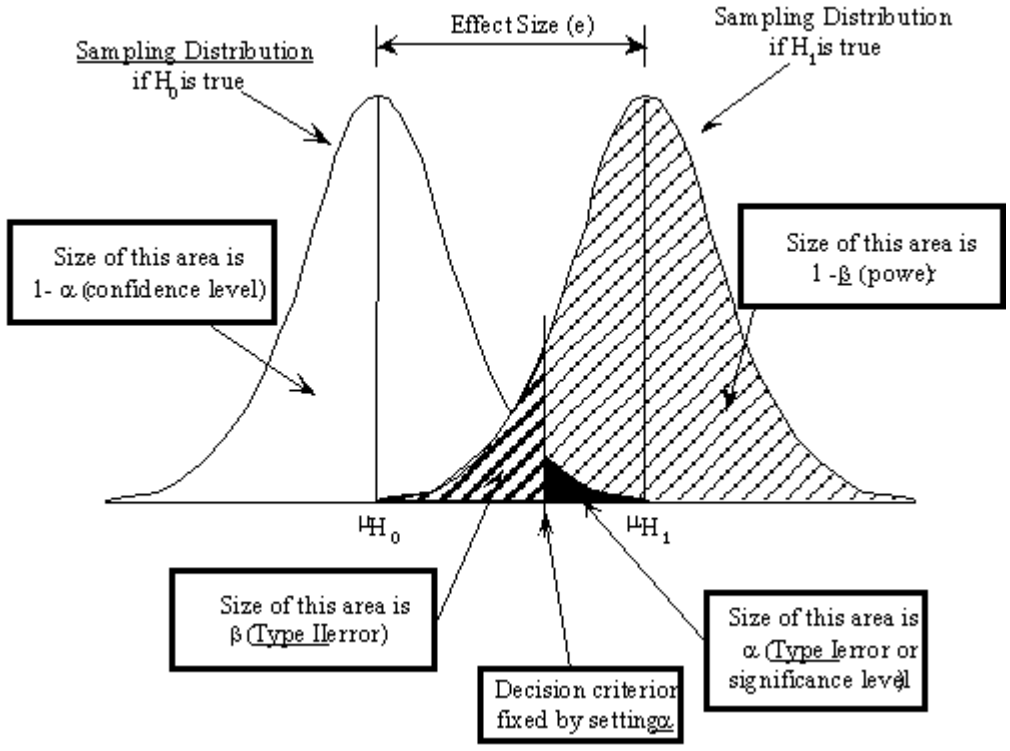
Reject H0

Power

- Errors
  - H0 is true but rejected: Type I error (no guilty man is found guilty)
  - H0 is false but accepted: Type II error (guilty man is found no guilty)

Table of error types		Null hypothesis (H_0) is	
		True	False
Decision About Null Hypothesis (H_0)	Fail to reject	Correct inference (True Negative) (Probability = 1 - α)	Type II error (False Negative) (Probability = β)
	Reject	Type I error (False Positive) (Probability = α)	Correct inference (True Positive) (Probability = 1 - β)

wiki



https://webstat.une.edu.au/unit_materials/c5_inferential_statistics/power.html

Steps for computing a p-value

- Specify some model for the underlying data (assume a family of the population)
- Identify H0 and H1
- Specify a test statistic that discriminates H0 and H1
- Collect the data
- Calculate test statistic
- Calculate p-value

Significance test for a population proportion

$$H_0: p = p_0, \quad p_0 = E(\hat{p}|H_0)$$

$$H_1: p > p_0, p < p_0, p \neq p_0$$

$$z = \frac{\hat{p} - E(\hat{p}|H_0)}{SD(\hat{p}|H_0)}$$

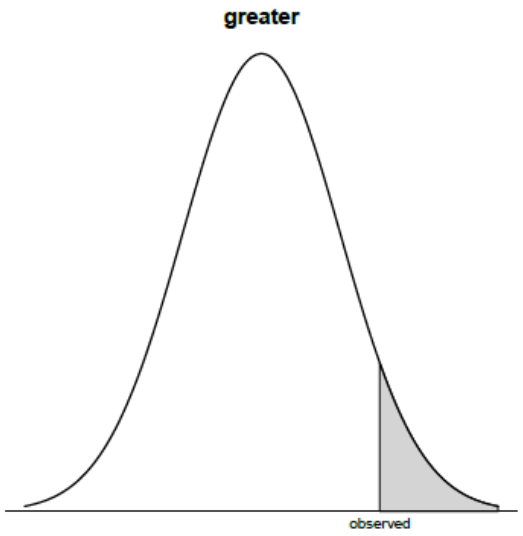
$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

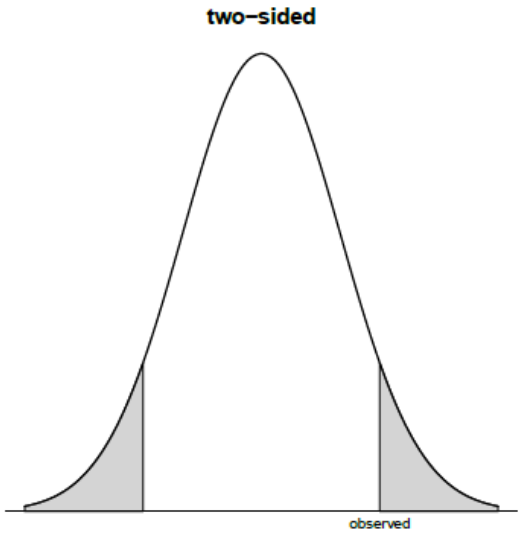
CLT

$$X \sim Bin(n, p)$$

$$z = \frac{X - np}{\sqrt{np(1-p)}} \rightarrow N(0,1)$$

$$p\text{-value} = \begin{cases} \mathrm{P}(\hat{p} \leq \text{observed value} \,|\, H_0) & \text{if } H_A : p < p_0, \\ \mathrm{P}(\hat{p} \geq \text{observed value} \,|\, H_0) & \text{if } H_A : p > p_0, \\ \mathrm{P}(|\hat{p} - p_0| \geq |\text{observed value} - p_0| \,|\, H_0) & \text{if } H_A : p \neq p_0. \end{cases}$$





Ex) Poverty rate

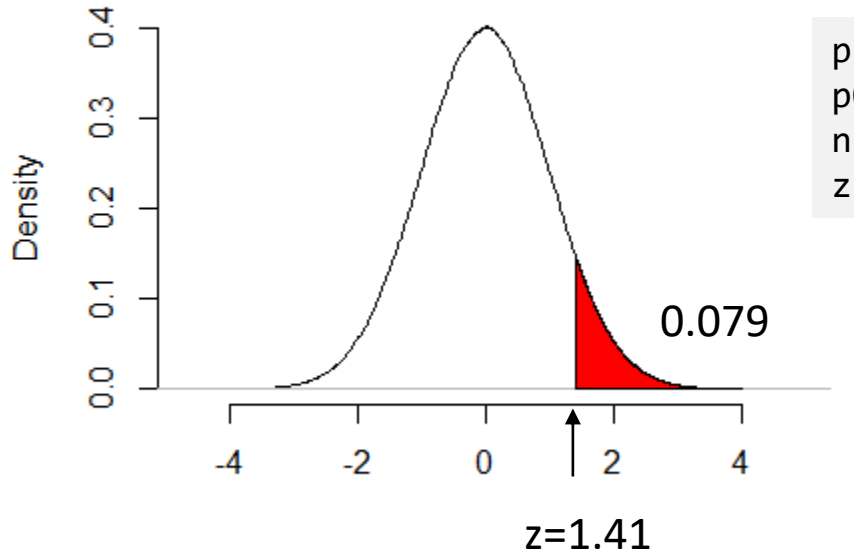
Year 2000, Poverty rate: 11.7

Year 2010, Poverty rate: 15.0

Year 2011, it is 15.13, Does it continue to rise?

n: 150,000, x: 22,695

$$H_0: p = 0.15 \; vs. \; H_a: p > 0.15$$




```
phat <- 22695/150000
p0 <- 0.1500
n <- 150000
z <- (phat - p0)/sqrt(p0*(1-p0)/n)
```

Significance test for the mean (t-tests)

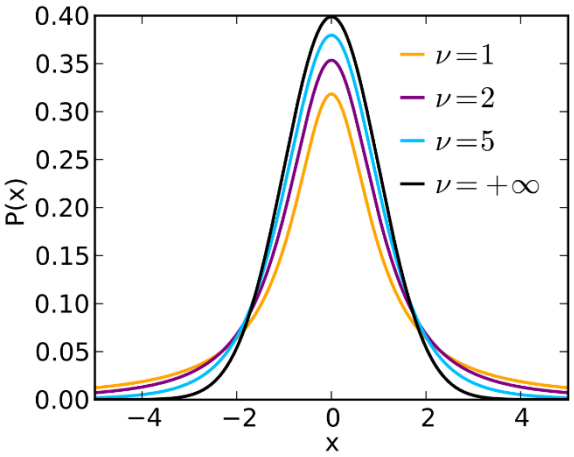
- Unknown mean of a parent population

$$H_0: \mu = \mu_0,$$

$$H_a: \mu > \mu_0, \mu < \mu_0, \mu \neq \mu_0$$

$$T = \frac{\bar{x} - E(\bar{x}|H_0)}{SE(\bar{x}|H_0)}$$

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{observed - expected}{SE}$$



wiki

Ex) SUV gas mileage

The actual mileage of a new SUV = 17 miles / gallon

A consumer group suspects it is lower

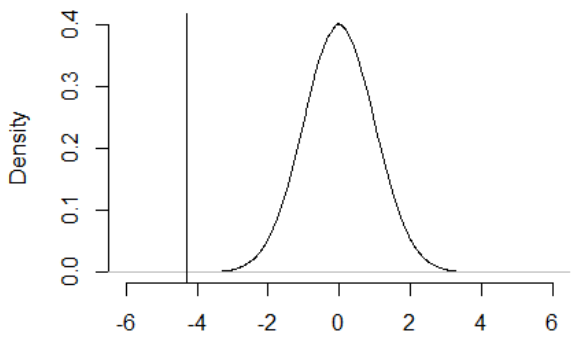
$$H_0: \mu = 17 \; vs. \; H_a: \mu < 17$$

```
mpg <- c(11.4, 13.1, 14.7, 14.7, 15, 15.5, 15.6, 15.9, 16, 16.8)
xbar <- mean(mpg)
s <- sd(mpg)
n <- length(mpg)
pt((xbar-17)/(s/sqrt(n)), df=9, lower.tail=T)
t.test(mpg, mu=17, alternative="less")
```

```
> t.test(mpg, mu=17, alternative="less")

        One Sample t-test

data:  mpg
t = -4.2847, df = 9, p-value = 0.001018
alternative hypothesis: true mean is less than 17
95 percent confidence interval:
     -Inf 15.78127
sample estimates:
mean of x
    14.87
```



Two sample tests of proportion

$$H_0: p_1 = p_2$$

$$H_a: p_1 < p_2, p_1 > p_2, p_1 \neq p_2$$

$$z = \frac{\widehat{p_1} - \widehat{p_2} - E(\widehat{p_1} - \widehat{p_2}|H_0)}{SD(\widehat{p_1} - \widehat{p_2}|H_0)}$$

$$z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\hat{p}\,(1-\hat{p})/\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Two sample tests of center (t-test)

$$H_0: \mu_x = \mu_y$$

$$H_a: \mu_x < \mu_y, \mu_x > \mu_y, \mu_x \neq \mu_y$$

$$T = \frac{(\bar{x} - \bar{y}) - E(\bar{x} - \bar{y}|H_0)}{SE(\bar{x} - \bar{y}|H_0)}$$

$$SE(\bar{x} - \bar{y}) = s_p \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} \qquad s_p = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}}.$$

Steps for computing a p-value

- Specify some model for the underlying data (assume a family of the population)
- Identify H0 and H1
- Specify a test statistic that discriminates H0 and H1
- Collect the data
- Calculate test statistic
- Calculate p-value

Ex) Differing dosage of antigen drug

$$H_0: \mu_x = \mu_y \quad vs. \; H_A: \mu_x \neq \mu_y$$

```
m300 <- c(284, 279, 289, 292, 287, 295, 285, 279, 306, 298)
m600 <- c(298, 307, 297, 279, 291, 335, 299, 300, 306, 291)
t.test(m300, m600)
t.test(m300, m600, var.equal=T)
```

등분산 가정

```
> t.test(m300, m600)

        Welch Two Sample t-test

data:  m300 and m600
t = -2.034, df = 14.509, p-value = 0.06065
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -22.3557409   0.5557409
sample estimates:
mean of x mean of y
    289.4     300.3

> t.test(m300, m600, var.equal=T)

        Two Sample t-test

data:  m300 and m600
t = -2.034, df = 18, p-value = 0.05696
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -22.1584072   0.3584072
sample estimates:
mean of x mean of y
    289.4     300.3
```

Ex) pre- and post tests

A college course is working

$$H_0: \mu_x = \mu_y \quad vs. \; H_A: \mu_x < \mu_y$$

$$H_0: \mu = 0 \quad vs. \; H_A: \mu < 0$$

```
x <- c(5,3,5,7,4,4,7,4,3)
y <- c(2,3,2,4,2,2,3,4,2)
t.test(x, y)

z <- x-y
t.test(z)
t.test(x, y, paired = T)
```

```
> t.test(x, y)

        Welch Two Sample t-test

data:  x and y
t = 3.4641, df = 12.8, p-value = 0.004283
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.7507266 3.2492734
sample estimates:
mean of x mean of y
 4.666667  2.666667
```

```
> t.test(z)

        One Sample t-test

data:  z
t = 4.2426, df = 8, p-value = 0.002827
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.9129392 3.0870608
sample estimates:
mean of x
        2
```

```
> t.test(x, y, paired = T)

        Paired t-test

data:  x and y
t = 4.2426, df = 8, p-value = 0.002827
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9129392 3.0870608
sample estimates:
mean of the differences
                      2
```

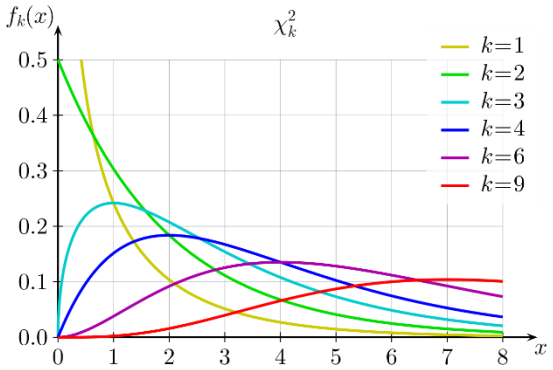
Measures of association

- The chi-squared statistic – a common summary of a table
- summary function

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

f_o: observed
f_e: expected

H0: 두 변수가 독립
H1: 두 변수가 독립이 아님



$$p - value = P(\chi^2 \geq \text{observed value} | H_0)$$

주름과 흡연 관계

3. 한 의학연구가에 의하면 흡연은 눈가에 주름이 지게 하는 요인이 된다고 한다. 이러한 주장이 타당한가를 알아보기 위해 30대 남자 1000명을 랜덤하게 추출하여 조사한 결과 다음과 같은 표를 얻었다. 30대 남자들을 대상으로 볼 때 연구가의 주장이 옳은지 판단하는 연관성을 나타내는 카이제곱 값을 구하라.

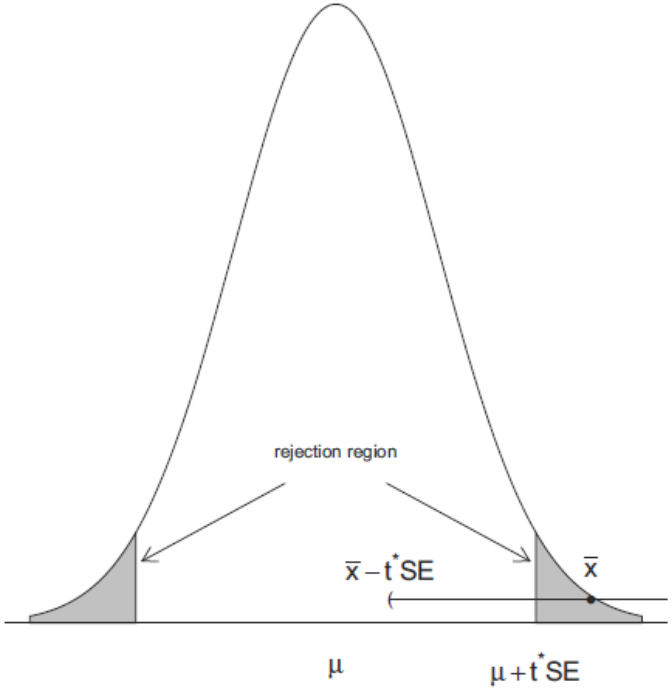
	주름있음	주름없음	
흡연자	186 (124.2)	114 (175.8)	300 (0.3)
비흡연자	228 (289.8)	472 (410.2)	700 (0.7)
	414 (0.414)	586 (0.586)	1000

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \quad = 74.9652$$

p-value < 2.2e-16

Significance tests and confidence intervals

$$H_0: \mu = \mu_0 \quad vs.\ H_1: \mu \neq \mu_0$$



(1- α)*100% 신뢰구간에 μ0 포함되어있지 않으면 양측검정 유의수준 α에서 H0 기각

$$P\left(-t^* < \frac{\bar{x} - \mu}{SE(\bar{x})} < t^*\right) = 1 - \alpha$$

Variable type vs. Tests

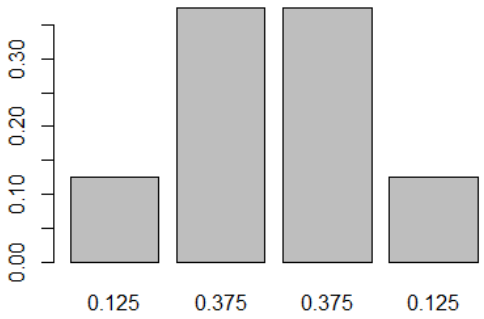
독립변수	종속변수	목적	모수적	비모수적	예
범주형	범주형	두 변수간 관계	카이제곱	Fisher's Exact Test	
범주형 (n=2)	연속형	그룹 평균 비교	독립표본 t-test	Kolmogorov-Smirnov Test Wilcoxon rank-sum test	
			대응표본 t-test	Wilcoxon singed rank test	
범주형 (n>2)	연속형	그룹 평균 비교	분산분석	Kruskal-Wallis test	
연속형	연속형	두 변수간 관계	상관분석	Spearman's correlation	
			회귀분석	Kendall's tau test Stuart's tau test	

숙제 #3 solution

1. 공평한 (p=1/2) 동전 세 개를 동시에 던질 경우 앞면의 개수로 정의된 확률변수에 대한 분포를 알아보고자 한다. 확률변수를 X라 할 때 X의 분포를 구하고 그래프를 그리시오. 또한 기대값과 분산을 구하시오.

1	2	3
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

X = {0, 1, 2, 3}
P(X=0) = 1/8
P(X=1) = 3/8
P(X=2) = 3/8
P(X=3) = 1/8
E[X] = 0*1/8 + 1*3/8 + 2*3/8 + 3*1/8 = 1.5
V[X] = E[X^2] – E[X]^2 = 3 – 2.25 = 0.75



숙제 #3 solution

2. 공평한 (p=1/2) 동전 100개를 동시에 던질 경우 앞면의 개수를 확률변수 X로 정의하고 이에 대한 기대값과 분산을 구하시오.

X = {0, 1, 2, …, 100}
E[X] = np = 100*0.5 = 50
V[X] = np(1-p) = 100*0.5*0.5 = 25

3. 한 야구선수가 평균 3번 타석에서 1번 안타를 친다고 한다. 4번의 연속된 타석에서 모두 안타를 칠 확률을 구하시오.

p = 1/3
X = {0, 1, 2, 3, 4}
P(X=4) =

$$P(X = 4) = \binom{4}{4} p^4 (1-p)^0$$

$$= 1/3^4 = 0.0123456789$$

END

수고하셨습니다