데이터와 분포 #3

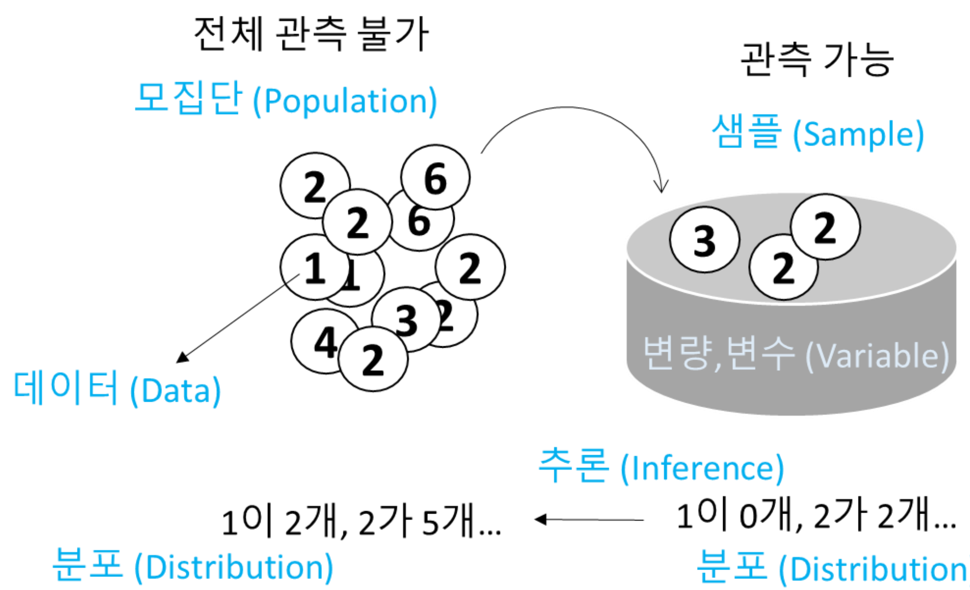
과학기술연합대학원대학교
한국생명공학연구원 스쿨
시스템생명공학전공

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

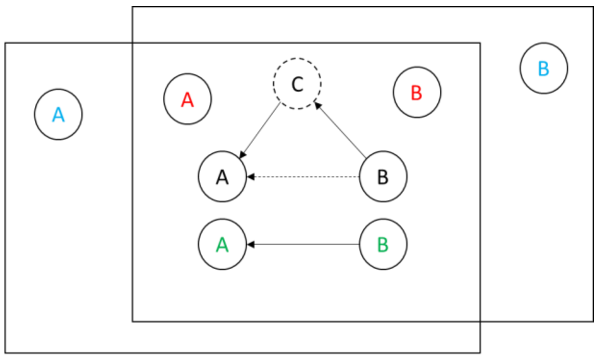
이변량 (변수 2개) 데이터 비교

1. Numerical data
   1. Unpaired data (Independent data) – Similarity with summaries
   2. Paired data – Relationship

      Covariance / correlation / regression

2. Categorical data
   1. Paired data - Relationship

      chi-squared statistic



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

Multivariate data

이변량 분석법 호환

그래프 이용 다변량 데이터 비교

Variables vs. samples

Airquality

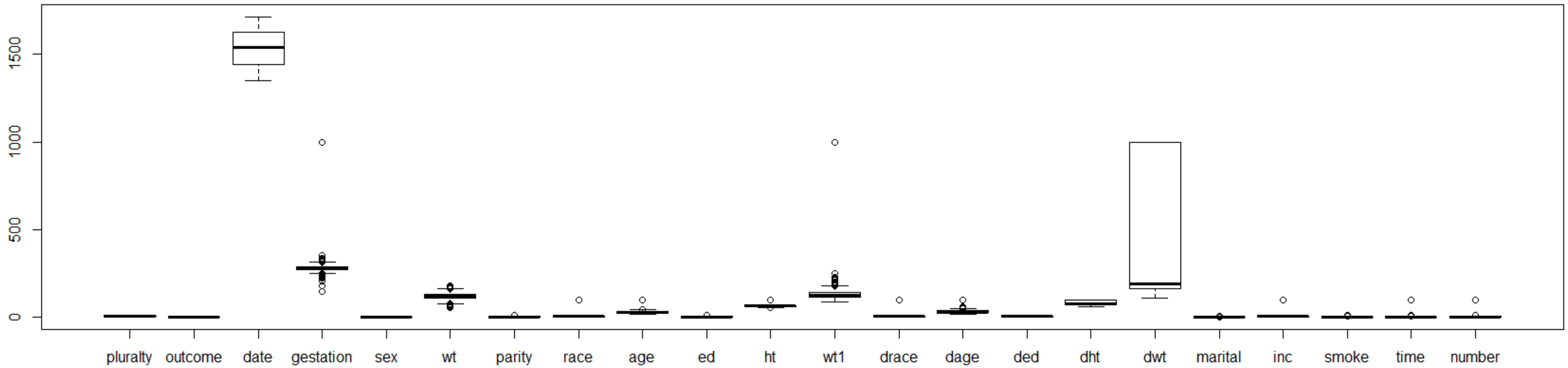
	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6
7	23	299	8.6	65	5	7
8	19	99	13.8	59	5	8
9	8	19	20.1	61	5	9
10	NA	194	8.6	69	5	10
11	7	NA	6.9	74	5	11
12	16	256	9.7	69	5	12
13	11	290	9.2	66	5	13
14	14	274	10.9	68	5	14
15	18	65	13.2	58	5	15
16	14	334	11.5	64	5	16

Babies

	id	pluralty	outcome	date	gestation	sex	wt	parity
1	15	5	1	1411	284	1	120	1
2	20	5	1	1499	282	1	113	2
3	58	5	1	1576	279	1	128	1
4	61	5	1	1504	999	1	123	2
5	72	5	1	1425	282	1	108	1
6	100	5	1	1673	286	1	136	4
7	102	5	1	1449	244	1	138	4
8	129	5	1	1562	245	1	132	2
9	142	5	1	1408	289	1	120	3
10	148	5	1	1568	299	1	143	3
11	164	5	1	1554	351	1	140	2
12	171	5	1	1593	282	1	144	4
13	175	5	1	1491	279	1	141	3
14	183	5	1	1446	281	1	110	5
15	194	5	1	1524	273	1	114	3

row: 샘플
column: 변수

Boxplot for multivariate data



Babies dataset

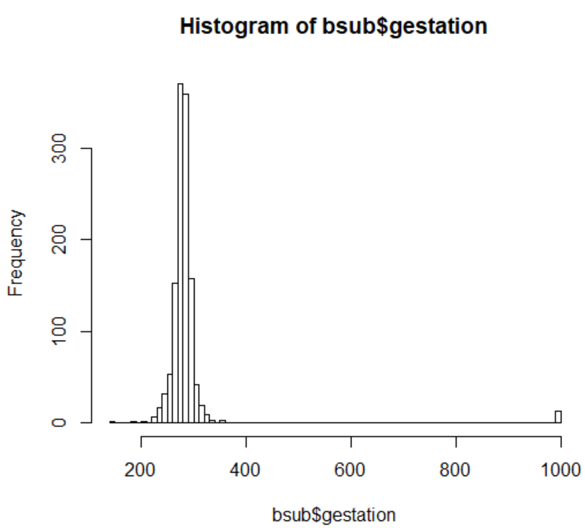
Stat Labs: Mathematical Statistics through Applications Springer-Verlag (2001)

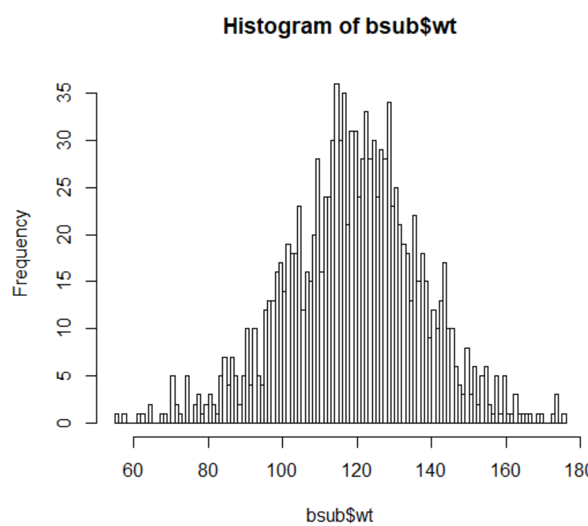
gestation: length of gestation in days
sex: infant's sex 1=male 2=female 9=unknown
wt: birth weight in ounces (999 unknown)
race: mother's race 0-5=white 6=mex 7=black 8=asian 9=mixed 99=unknown
age: mother's age in years at termination of pregnancy, 99=unknown
ht: mother's height in inches to the last completed inch 99=unknown
drace: father's race, coding same as mother's race.
dage: father's age, coding same as mother's age.
dht: father's height, coding same as for mother's height
dwt: father's weight coding same as for mother's weight
smoke: does mother smoke? 0=never, 1= smokes now, 2=until current pregnancy, 3=once did, not now, 9=unknown
time: If mother quit, how long ago? 0=never smoked, 1=still smokes, 2=during current preg, 3=within 1 yr, 4= 1 to 2 years ago, 5= 2 to 3 yr ago, 6= 3 to 4 yrs ago, 7=5 to 9yrs ago, 8=10+yrs ago, 9=quit and don't know, 98=unknown, 99=not asked
number: number of cigs smoked per day for past and current smokers 0=never, 1=1-4,2=5-9, 3=10-14, 4=15-19, 5=20-29, 6=30-39, 7=40-60, 8=60+, 9=smoke but don't know,98=unknown, 99=not asked

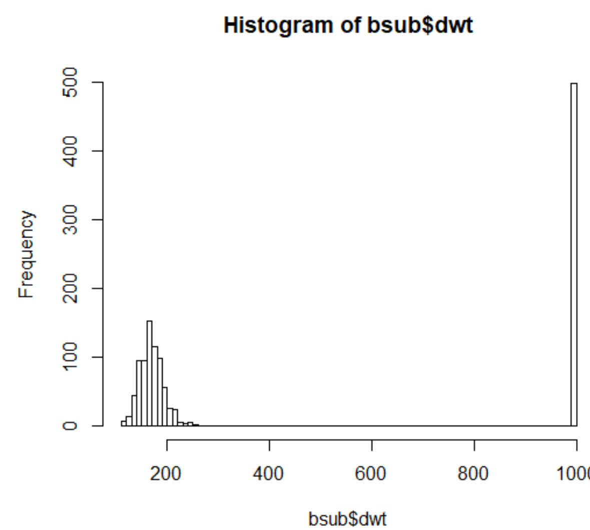
Standardization

$$z_i = \frac{x_i - \bar{x}}{s}$$

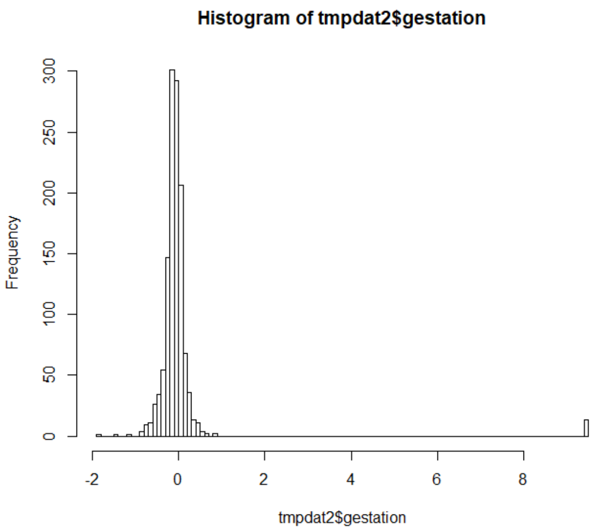
- In babies dataset, standardize gestation and weight using apply and sweep

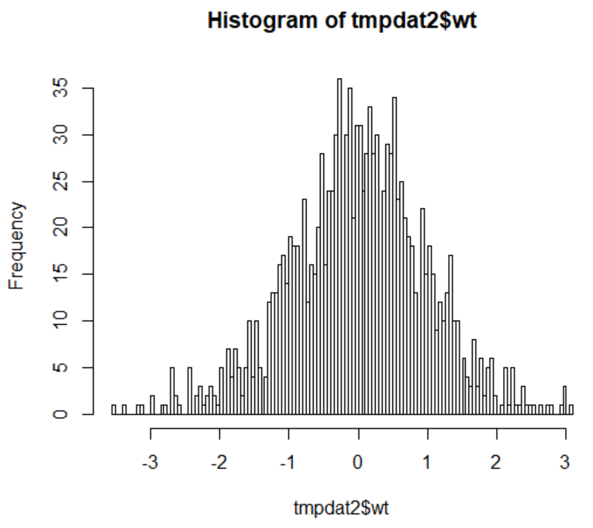


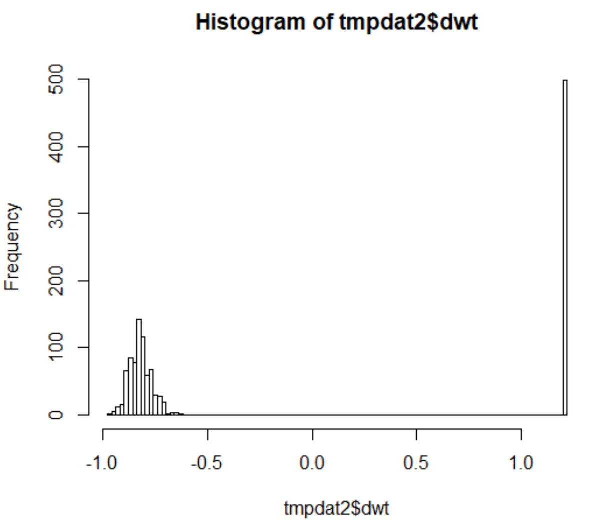


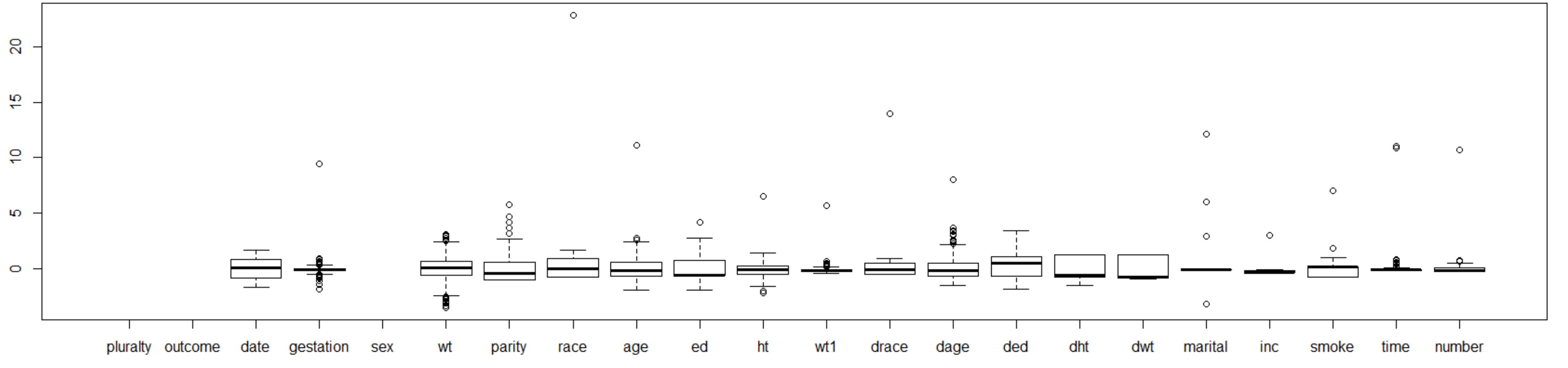


After standardization







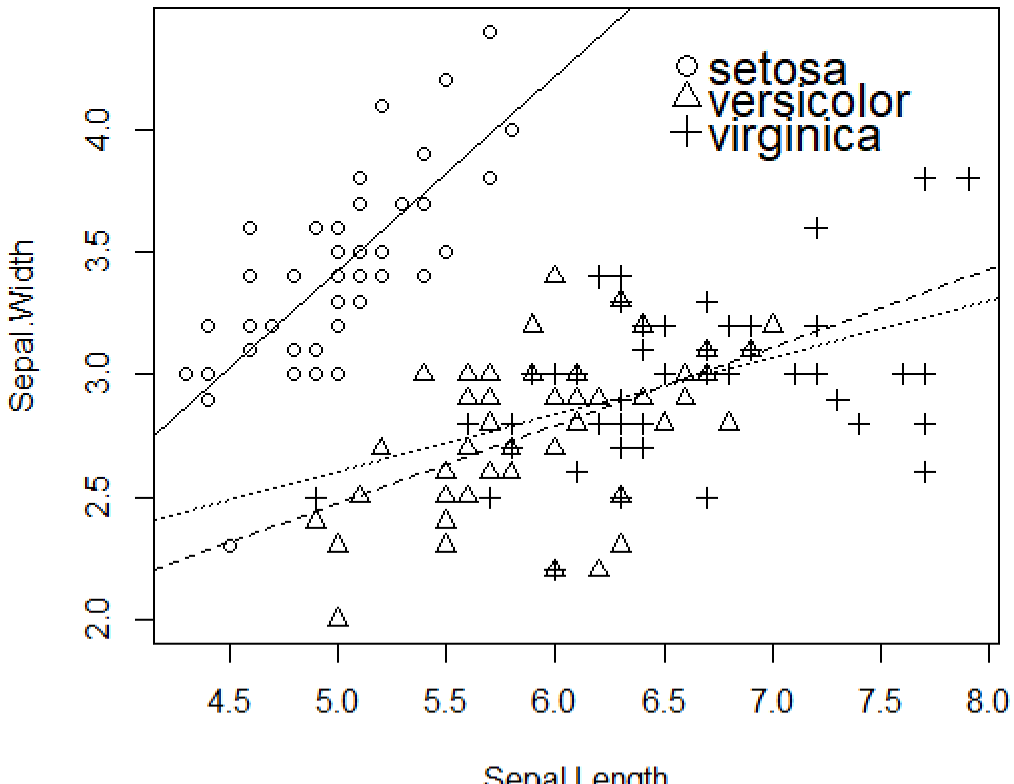


표준화 장점?

Scatterplot for multivariate data

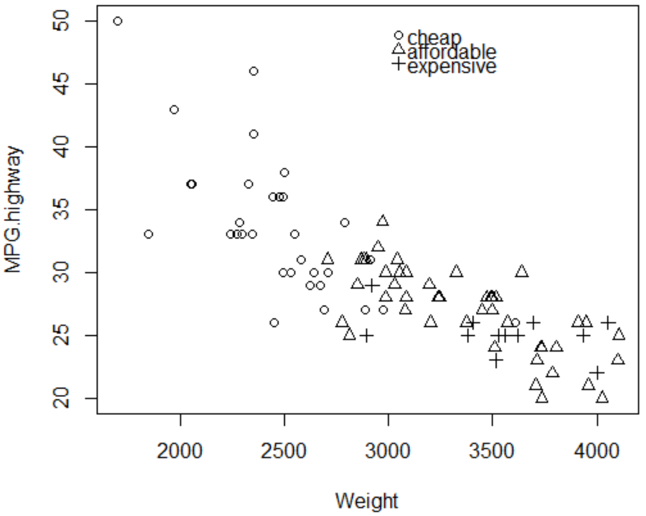
Iris example

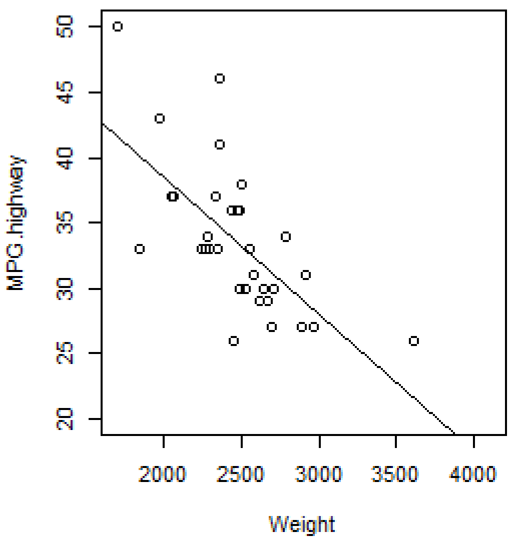
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

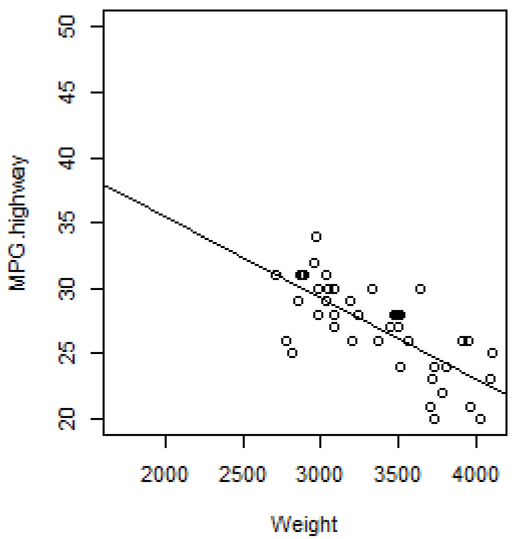


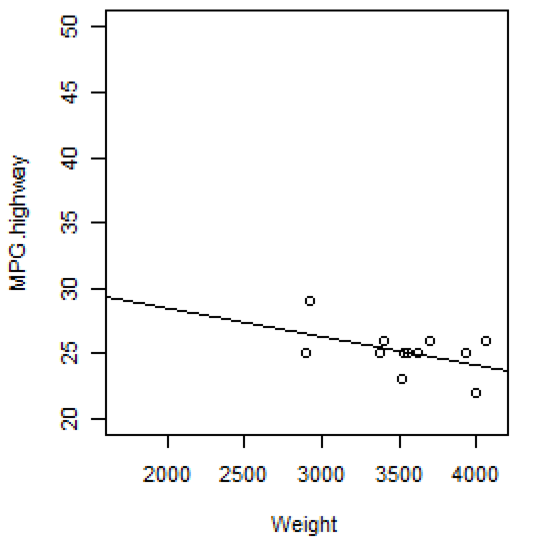
Scatterplot

- Cars93
- Turn numeric variables to categorical variables
- Relationship highway mileage and weight by car price



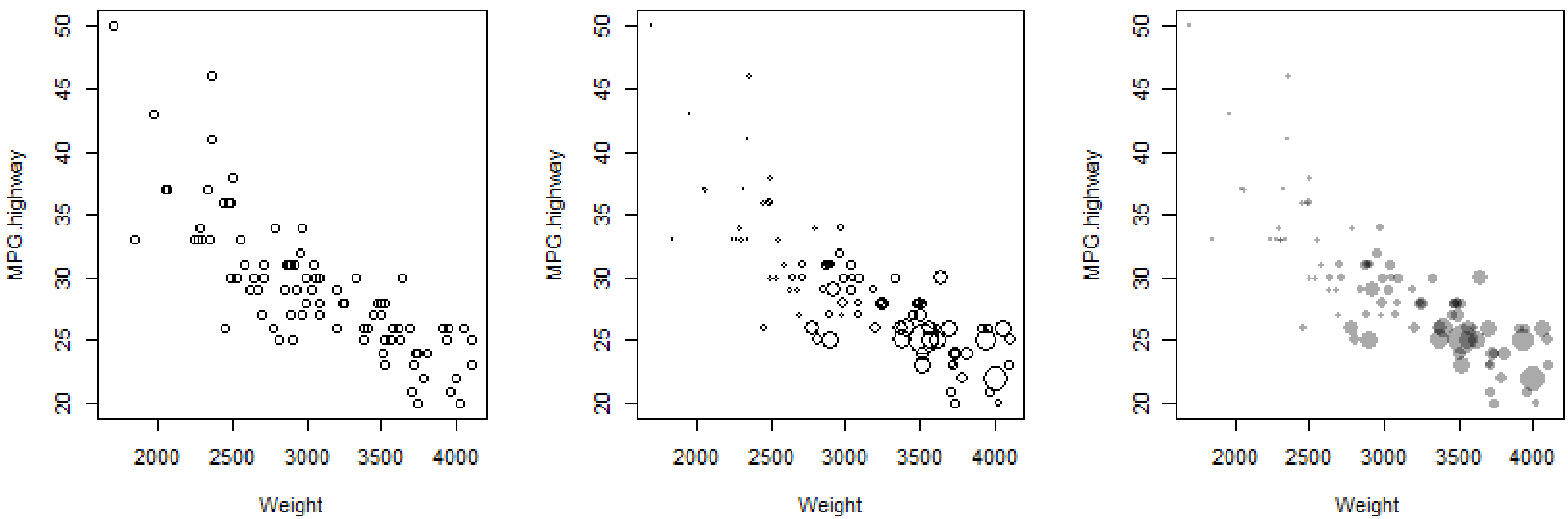






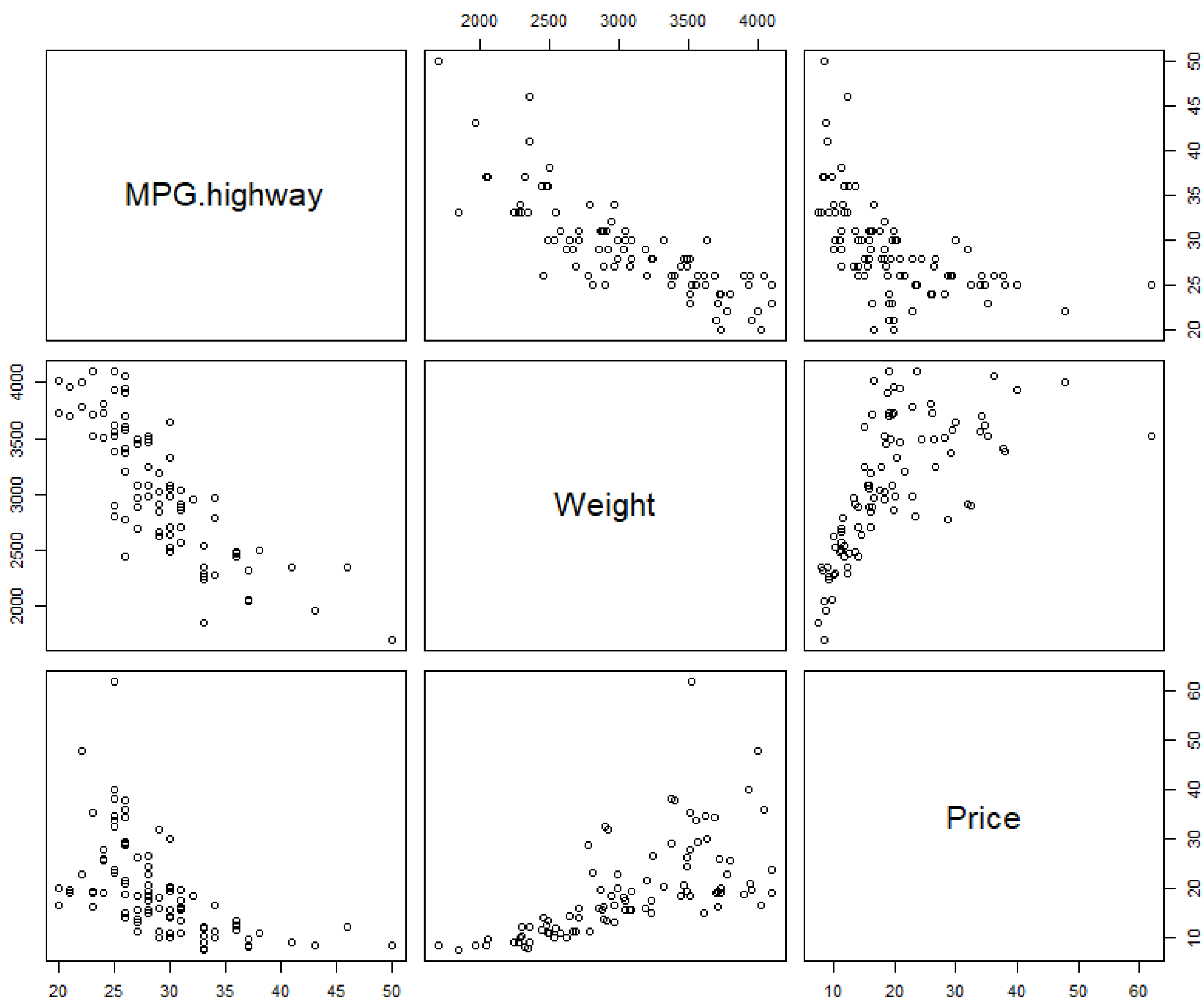
Bubble chart

- Cars93



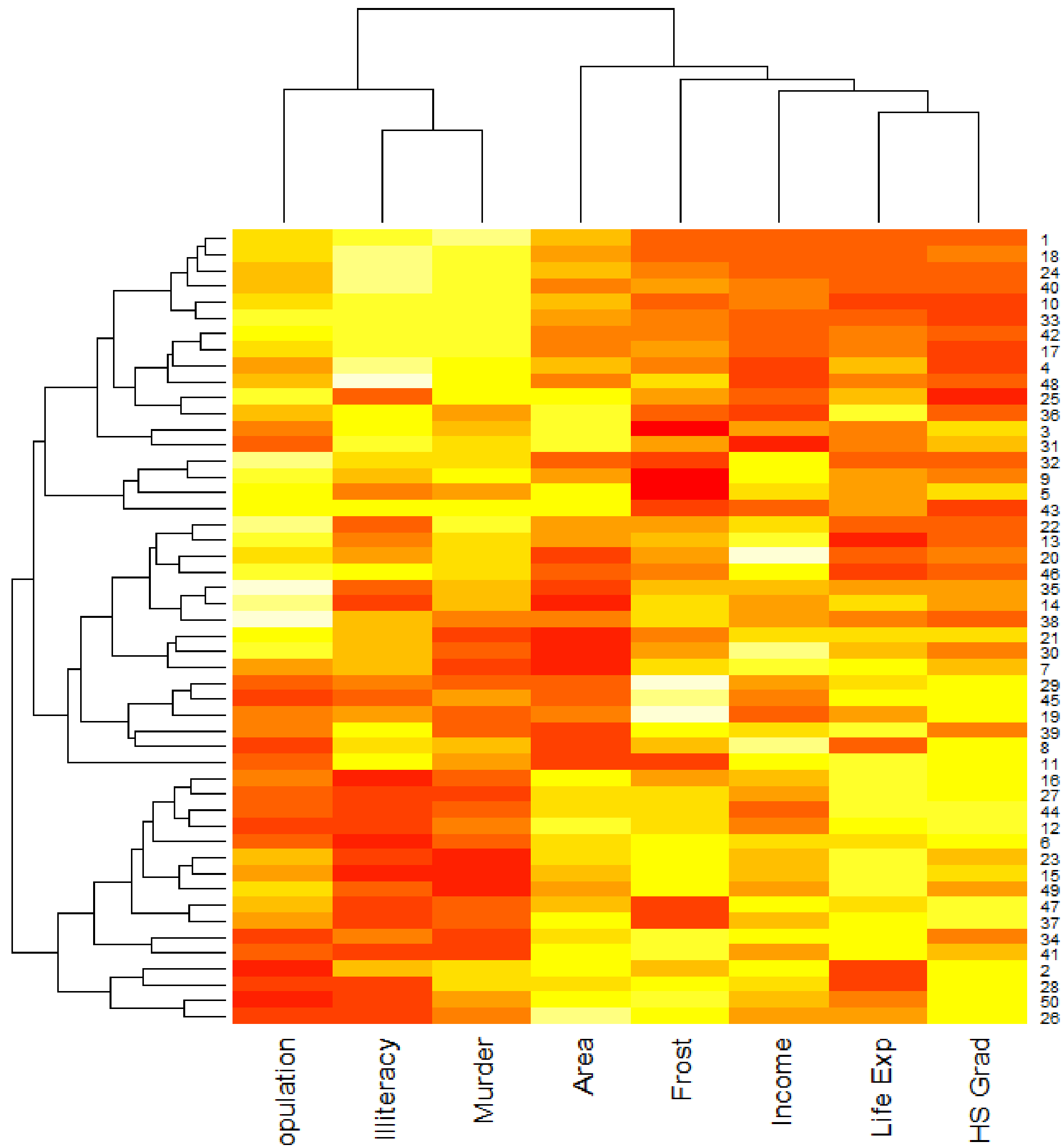
Pairs plots

- Cars93



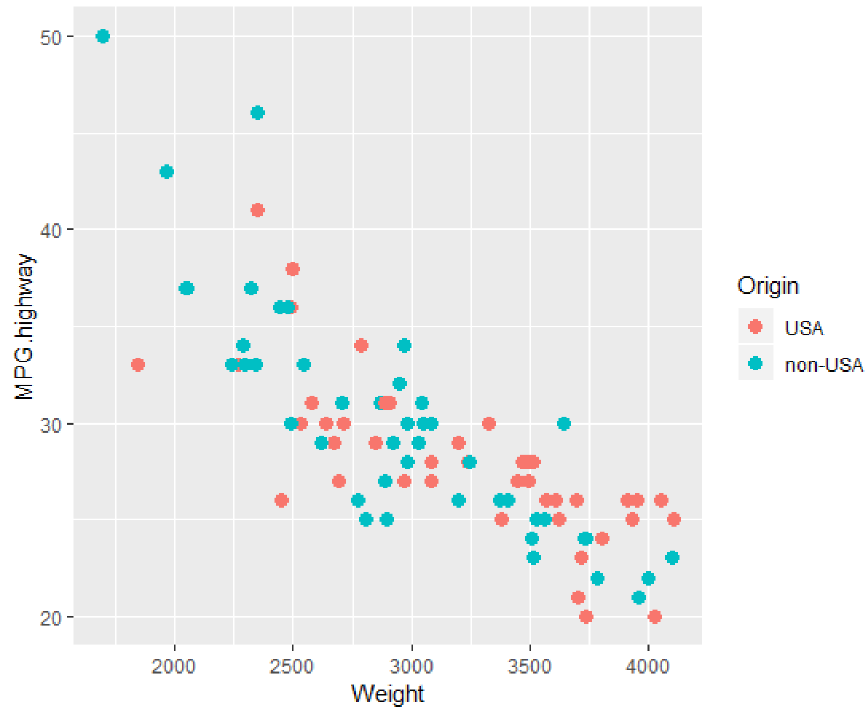
Heatmap

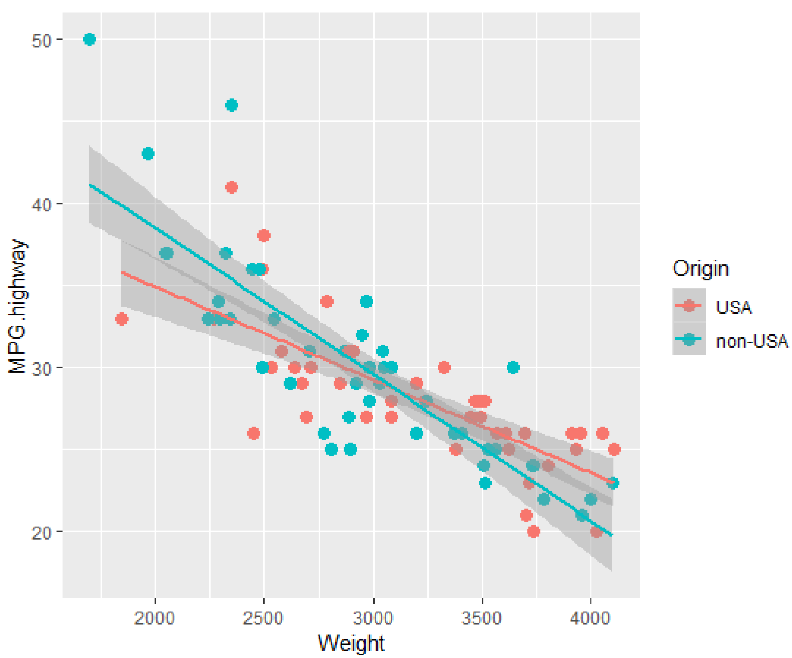
- Popular graphical trick in big data visualization

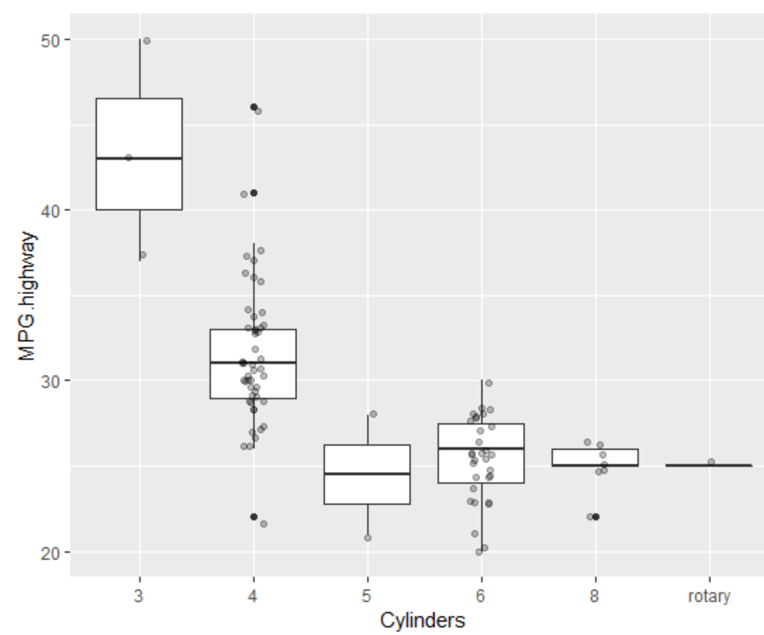


ggplot2

- http://ggplot2.org
- grid graphics engine for R
- grammar of graphics
- Pros – visual appeal, popularity, big data handling, easy to use
- Cons – relatively difficult to learn
- Two main component blocks: aesthetics and geometries



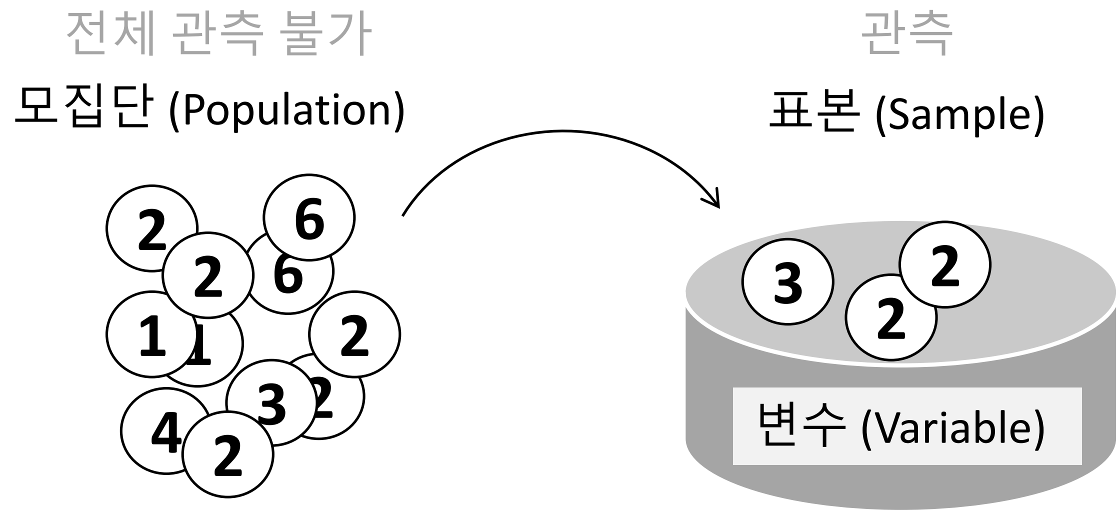




Population

Random variable

Distribution (sample / population)



Q: 대한민국 성별에 따른 흡연 비율?

모집단:

변수:

표본: 256명

데이터:

Populations and random variables

- Variable (변수) - 특정 현상을 이해하기 위한 특징
- Data (데이터) – 특징의 값(들)
- Population (모수) – 변수가 갖는 가능한 모든 값의 범위

- Random variable (확률 변수) – 값을 관측하기 전의 변수
- A data point (데이터) – realization of the random variable
- Probability of the random variable has the one data point
- Distribution of a random variable

Random variable

Population

Cars93.Price
15.9
33.9
29.1
37.7
30.0
15.7
20.8
23.7
26.3
34.7
40.1
13.4
11.4
15.1
15.9
16.3
16.6
18.8
38.0
18.4
15.9
33.9
29.1
37.7
30.0
15.7
20.8
23.7
26.3
38.0
18.4

....

Variable
(Observed)

Cars93.Price
15.9
33.9
29.1
37.7
30.0
15.7
20.8
23.7
26.3
34.7
40.1
13.4
18.4

Random variable
(Before observation)

Cars93.Price

Cars93.Price == 15.9 ?
Cars93.Price == 15.9 probability?
P(Cars93.Price=15.9)=
P(X=x)=

Probability

Definition

$$P(E) = \frac{\#\, of\ events\ in\ E}{\#\, of\ events\ in\ total}$$

Rules

1. $P(E) > 0$
2. $\sum_{All\ possible\ out\ comes} P(E) = 1$
3. $P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$

Examples

- Gender vs. smoking status
  - E1: to select one female, P(E1)
  - E2: to select a heavy smoker, P(E2)
  - P(E1 or E2)

*variable? value?

- Number of heads in two coin tosses
  - (H, T)
  - X: number of heads → X=0 or X=1 or X=2
  - P(X=0) , P(X=1) , P(X=2)

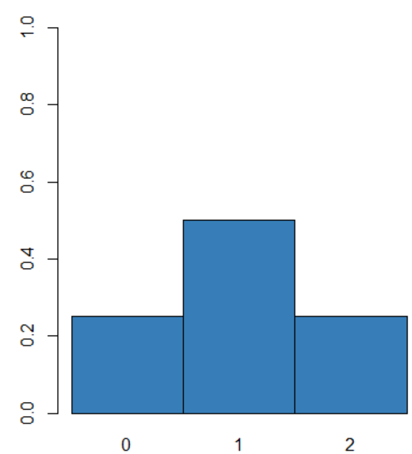
- Picking balls from a bag
  - with/without replacement

Distribution of a random variable

- The probabilities of occurrence of different possible outcomes in an experiment.

- Number of heads in two coin tosses
  - (H, T)
  - X: number of heads → X=0 or X=1 or X=2
  - P(X=0) =1/4 , P(X=1) = 2/4, P(X=2) = 1/4

Coin1	Coin2
H	H
H	T
T	H
T	T



Discrete random variables

- X be a discrete random variable
- Range of X is the set of all x, $P(X=x) > 0$

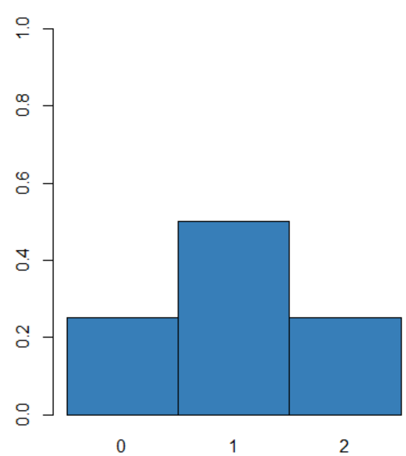
- Number of heads in two coin tosses
  - (H, T)
  - X: number of heads → $X=0$ or $X=1$ or $X=2$
  - $P(X=0) = 1/4$, $P(X=1) = 2/4$, $P(X=2) = 1/4$

- Picking balls from a bag
  - N balls ($R + G = N$), pick a ball twice with replacement
  - X: number of red balls chosen
  - $x=\{0, 1, 2\}$

Specifying a distribution of discrete R.V.

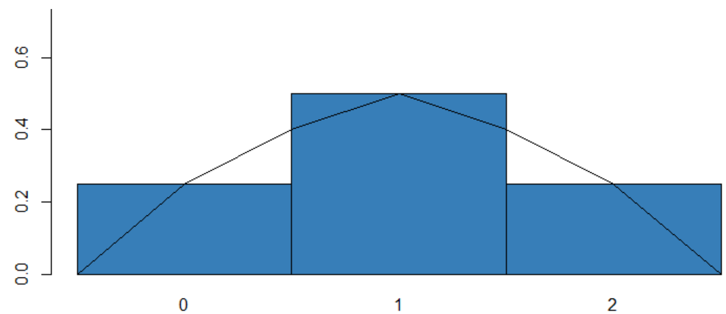
- Specifying the range of value k
- Assigning to each k a number $p_k = \mathrm{P}(X=k)$

such that $p_k > 0$ and $\sum p_k = 1$



Continuous random variables

- Continuum of possible values → new definition of probability
- Density of X



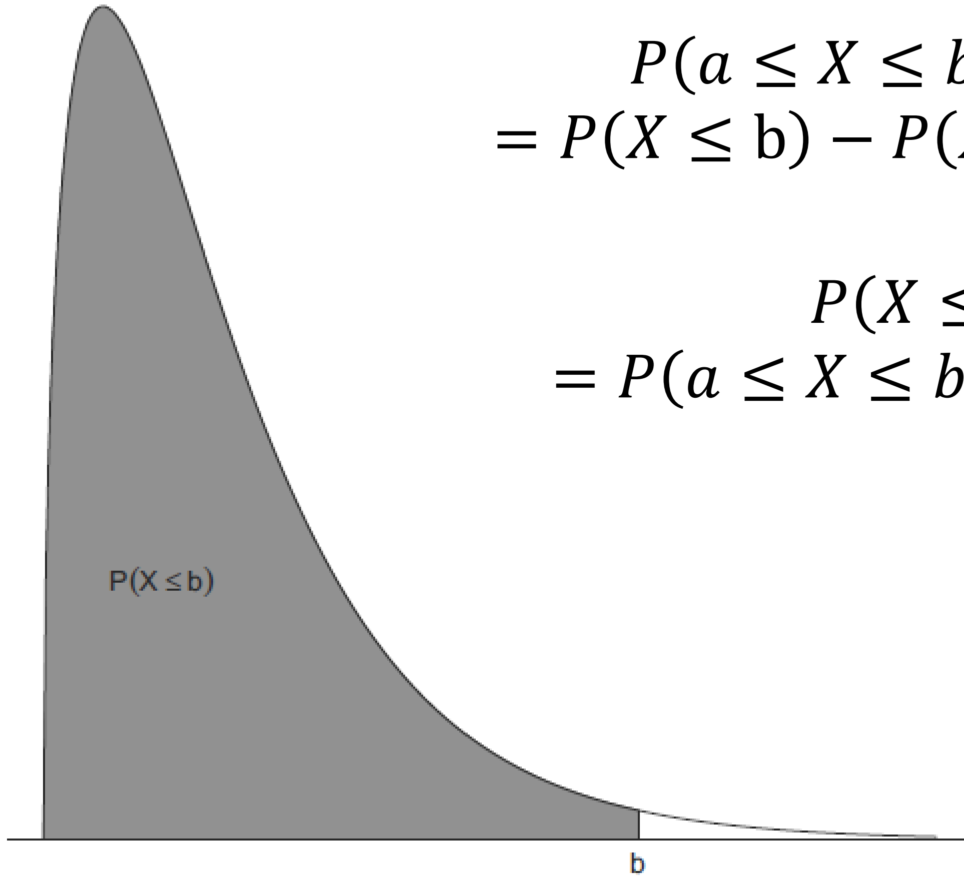
P(X=k)=?

- Probability density function (p.d.f. 확률밀도함수)

$$f(x) = P(a \leq X \leq b) \text{ for any } x$$

- Cumulative distribution function (c.d.f. 누적분포함수)

$$F(x) = P(X \leq x)$$



$$P(a \leq X \leq b) \\ = P(X \leq \mathrm{b}) - P(X \leq \mathrm{a})$$

$$P(X \leq \mathrm{b}) \\ = P(a \leq X \leq b) - P(X \leq \mathrm{a})$$

$\mathsf{P}(X \leq b)$ is defined by the area to left of b under the density of X.

Check points

- Population vs sample
- Random variable
- Probability
- Distribution of random variable
- What’s next??

$$P(X = x)$$

Mean and standard deviation (Discrete)

- Expected value of a random variable (vs. summaries)
- population mean μ
- population standard deviation σ

$$\mu = E(X) = \sum xP(X = x)$$

$$\sigma^2 = VAR(X) = E\left((X-\mu)^2\right)$$
$$= E(X^2) - E(X)^2$$
$$= \sum x^2 P(X = x) - \sum xP(X = x)^2$$

Mean and standard deviation (Continuous)

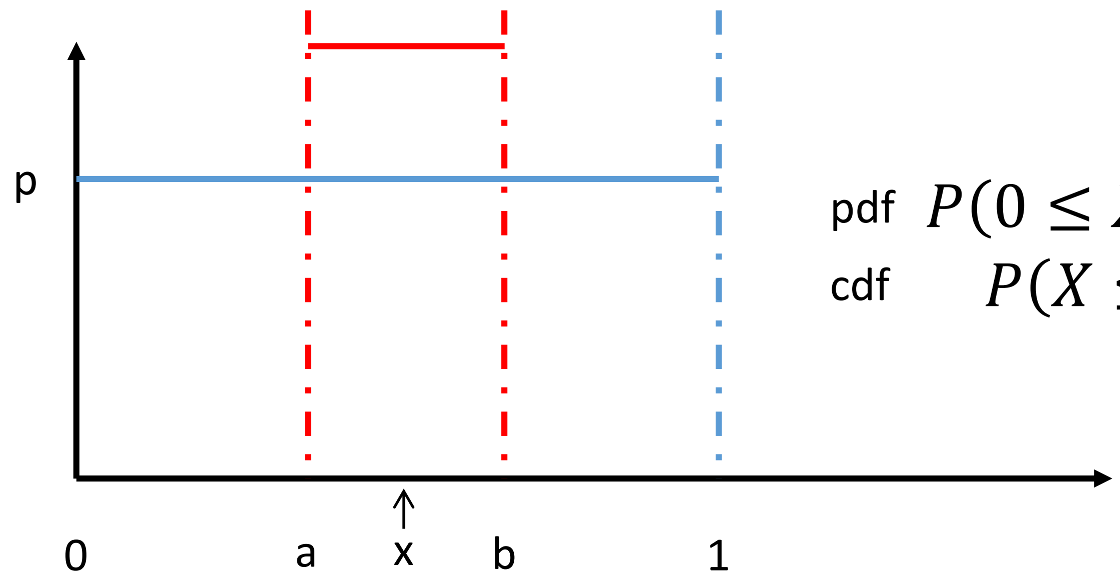
- population mean μ = E(X)
- population standard deviation σ = SD(X)

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

$$\begin{aligned} \sigma^2 &= VAR(X) = E\big((X-\mu)^2\big) \\ &= E(X^2) - E(X)^2 \\ &= \int_{-\infty}^{\infty} x^2 f(x) dx - {\int_{-\infty}^{\infty} x f(x) dx}^{\;2} \end{aligned}$$

Mean and standard deviation (Uniform)

- X has a uniform distribution on [0, 1]



pdf $P(0 \le X \le 1) = 1$

cdf $P(X \le x) = x$

pdf $P(a \le X \le b) = \frac{1}{b-a}$ if $a \le x \le b$ or 0

cdf $P(X \le x) = \frac{x-a}{b-a}$ if $x \le a$ or 0

$$\mu = E(X) = \int_0^1 xf(x)dx = \int_0^1 xdx = \left[\frac{x^2}{2} + c\right]_0^1$$

Example

- 동시 2개 동전 던질 경우 앞면의 개수

  X: number of heads, x = {0, 1, 2},

  P(X=x)=

  E(X)=

  VAR(X)=

Sampling from a population

- 확률변수의 관측값은 모집단의 분포를 설명할 수 있음
- 모집단으로부터 표본을 추출하여 데이터를 관측하고 모집단의 분포를 추론
- 이를 확률변수 Sequence, X_1, X_2, X_3, …, X_n 라 하면

- *Identically distributed*: 각 확률변수가 같은 분포를 가질 경우
- *Independent*: 특정 확률변수의 값이 다른 확률변수의 분포에 영향을 주지 않을 경우

Example)
동전 하나 n 번 던지는 경우, X_i = 앞면이 나오면 1 아니면 0 으로 정의할 때 X_1, X_2, X_3, …, X_n 는 i.i.d sequence라고 함.

Distributions

Sampling distributions

Popular distributions for population

Sampling distributions

- A *statistic* is a numeric value summarizing a random sample

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \ \dots + X_n)$$

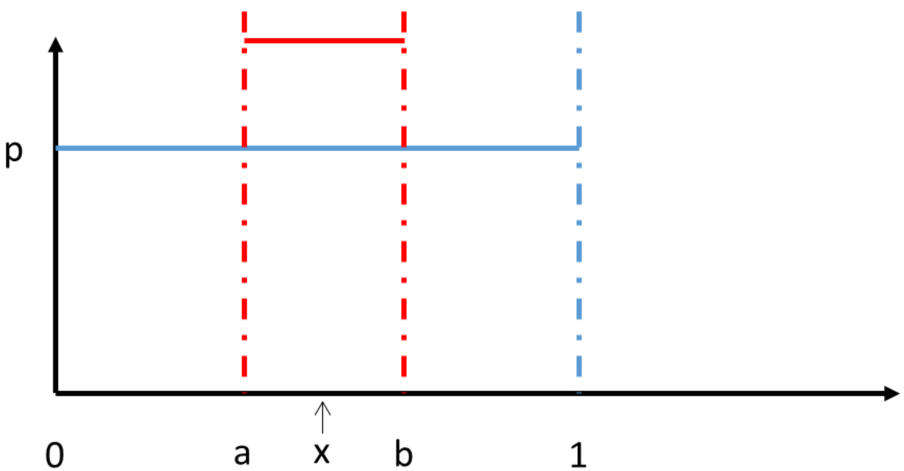
- 통계량이 random sample에 의해 계산된 경우 이 통계량도 Random variable
- 이 때 statistic 의 분포를 sampling distribution 이라 함
- sampling distribution 은 이론적으로 복잡하나 일반적인 경우에 대해서는 잘 알려져 있음

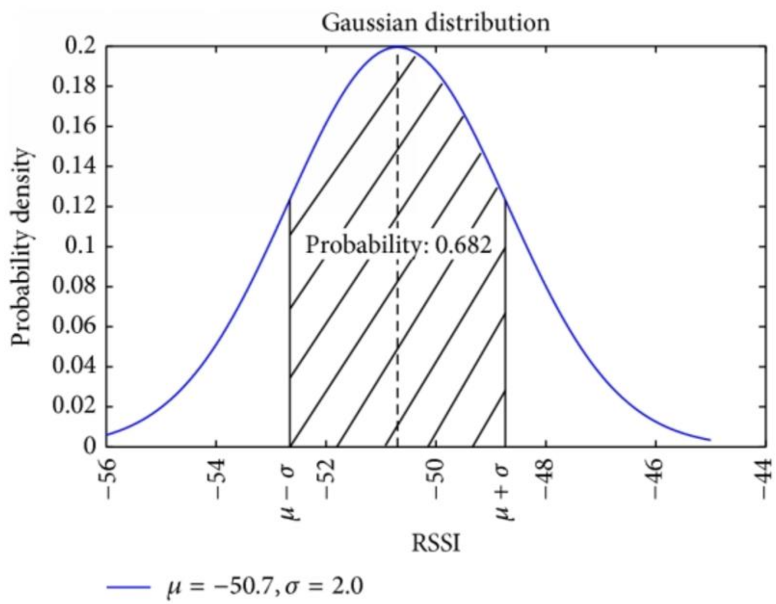
Sample mean & standard deviation

$$E(\bar{X}) = \mu \quad \text{and} \quad \mathrm{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Families of distributions

- 유사한 특성을 가진 분포
- 각 family는 분포를 결정하는 parameter의 function으로 표현
- Ex. uniform distribution





f(x)=1 with parameters 0, 1
mean = 1/2, var = 1/12

f(x)=1 with parameters a, b
mean = (b-a)/2, var = (b-a)^2/12

동전던지기 #1

하나의 동전 던져서 앞면이 나올 확률은? (앞면1, 뒷면0)
확률 변수 X: 동전 던져서 나오는 면 (only two values 1 and 0)

$$P(X = 1) = p$$

동전을 던져서 앞면이 나올 경우의 기대값? 분산?

$$E(X) = 0 \times P(X = 0) + 1 \times \mathrm{P}(\mathrm{X} = 1) = p$$

$$VAR(X) = E(X^2) - E(X)^2 = p - p^2 = p(1 - p)$$

p 계산?

동전던지기 #2

공평한 (p=1/2) 2개의 동전을 던져서 앞면이 둘 다 나올 확률? (앞면1, 뒷면0)
확률 변수 X: 동전 던져서 나오는 앞면의 개수 x=(0,1,2)

$P(X=x)$ P(X=0) =1/4 , P(X=1) = 2/4, P(X=2) = 1/4

두 동전을 던져서 앞면이 나올 개수의 기대값? 분산?

$$E(X) = 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1 \ (np)$$

$$VAR(X) = E(X^2) - E(X)^2 = 1.5 - 1 = 0.5 \ (np(1-p))$$

E(X) =1 의미?

동전던지기 #3

충분히 큰 n개의 동전을 던져서 5개 동전만 앞면이 나올 확률? (앞면1, 뒷면0)
확률 변수 X: 동전 던져서 나오는 앞면의 개수 x=(0,1,2, …, n)

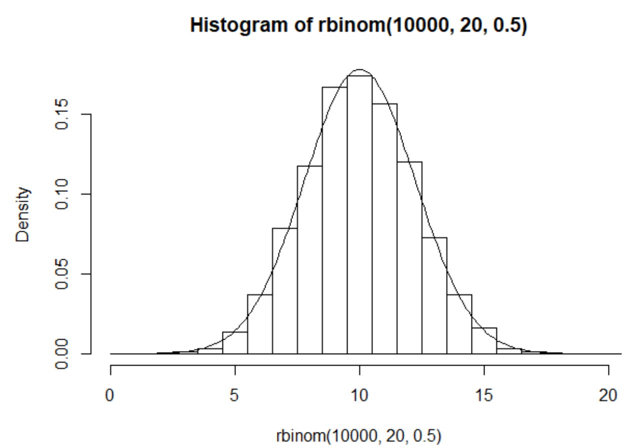
$$P(X = x) = ?$$

n개 동전을 던져서 나올 앞면의 개수의 기대값? 분산?

$$E(X) = np$$

$$VAR(X) = np(1 - p)$$

20개 동전을 던져서
나오는 앞면의 개수



Bernoulli random variables

- X has only two values (0, 1) (success, failure)
- Distribution of X characterized by p = $P(X=1)$

➔ Bernoulli(p)

- E(X)=p, Var(X)=p(1-p),
- Probability mass function

$$f(x)=p \quad \text{if } x=1$$

$$f(x)=1-q \quad \text{if } x=0$$

동전 100번 던지기, 앞면 1, 뒷면 0

```
> x
  [1] 0 0 1 0 0 1 0 0 0 1 1 1 0 0 0 0 0 1 1 0 1 0
 [23] 1 0 0 1 1 0 1 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0
 [45] 0 0 0 1 1 0 1 0 0 1 1 1 0 1 0 0 1 1 0 0 0 1
 [67] 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 1 1 0 1 0
 [89] 0 1 0 0 0 0 0 1 0 0 0 0
> mean(x)
[1] 0.32
> var(x)
[1] 0.219798
```

Binomial random variables

- X counts the number of successes in n Bernoulli trials
- Distribution of X characterized by n and p= $P(X$=1)

➔ Binomial(n, p)

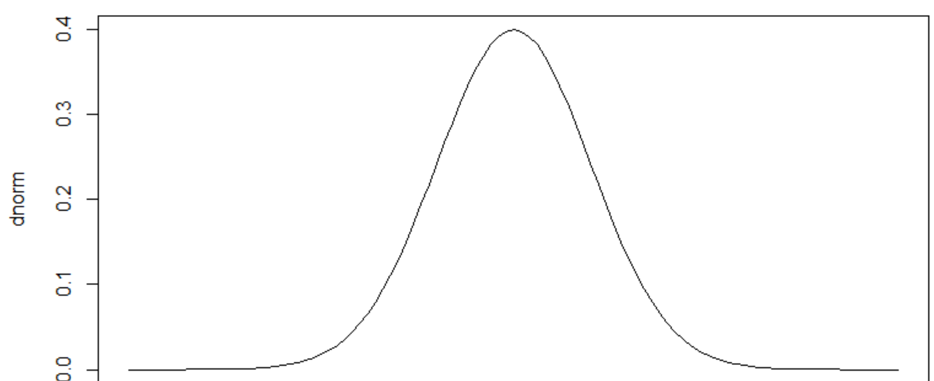
- $\mathrm{E}(X)$=np, Var(X)=np(1-p)
- Probability mass function

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$
$$=\frac{n!}{(n-k)!\,k!}p^k(1-p)^{n-k}$$

Normal random variables

- “Bell-shaped”, continuous ➔ density
- distribution of X characterized by μ and σ
  ➔ Normal(μ, σ^2)
- $\mathrm{E}(X)=\mu$, $\mathrm{Var}(X)=\sigma^2$
- Probability density function

$$f(x|\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



μ

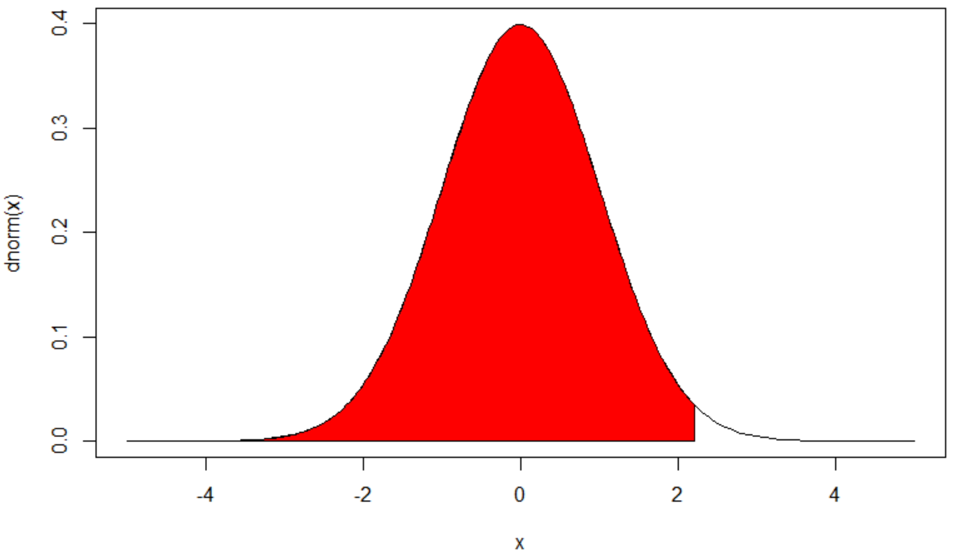
Normal Approximation to the Binomial

- Use a continuous distribution (the normal distribution) to approximate a discrete distribution (the binomial distribution)
- According to the Central Limit Theorem, the the sampling distribution of the sample means becomes approximately normal if the sample size is large enough.
- The factorials in the formula of binomial pmf cause difficulty of computation

More read
https://www.thoughtco.com/normal-approximation-to-the-binomial-distribution-3126589

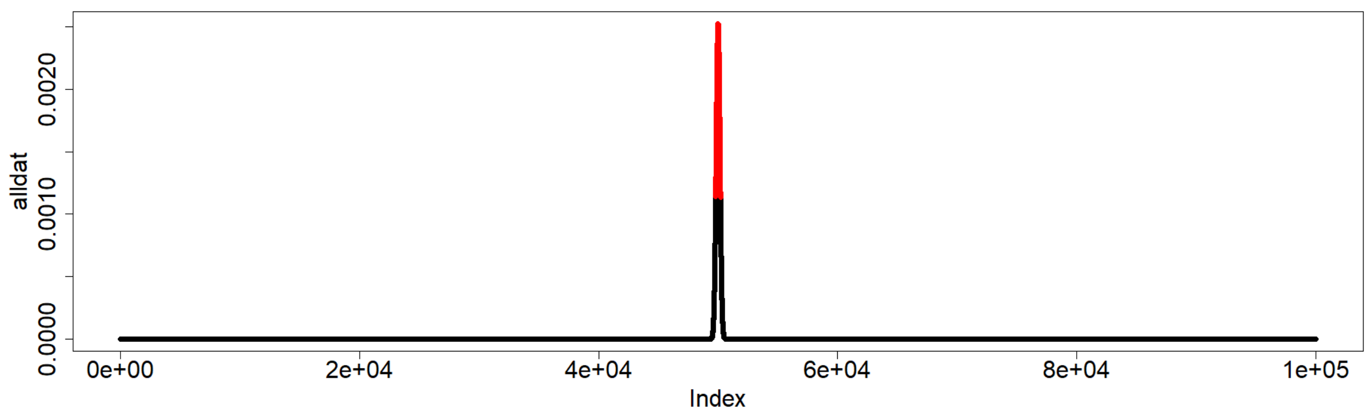
Standard normal distribution

- Z ~ Normal(0,1)
- p(Z ≤ 2.2)
- p(-1 < Z ≤ 2)
- p(Z > 2.5)
- b such that p(-b < Z ≤ b) = 0.90



Ex) coin toss 100000

- if X={H, T} success or failure, p=p(X=H)=1/2
- X: number of heads (a fair coin), X={0, 1, 2, ..., n}
- n = 100,000, p=0.5



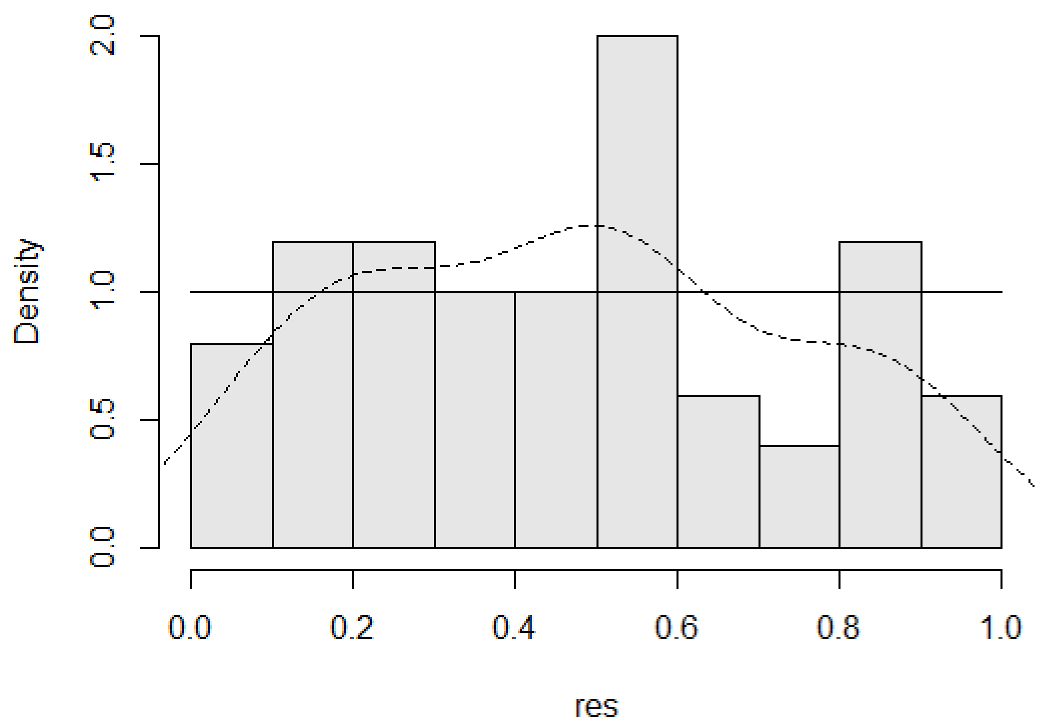
- p(49,800 < X ≤ 50,200)

  = p(X=49,801) + p(X=49,802) + … + p(X=50,200)

  = p(X ≤ 50,200) – p(X ≤ 49,800)

Uniform distribution

- No preferred values over [a, b]
- density: 1/(b-a)
- Uniform(a, b)

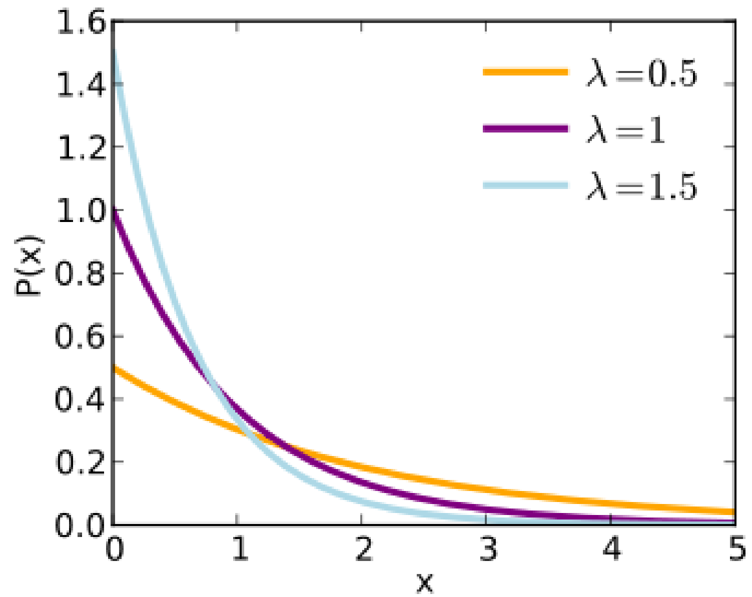


Exponential distribution

- Length of time
- density:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{where } x \ge 0 \\ 0 & \text{where } x < 0 \end{cases}$$

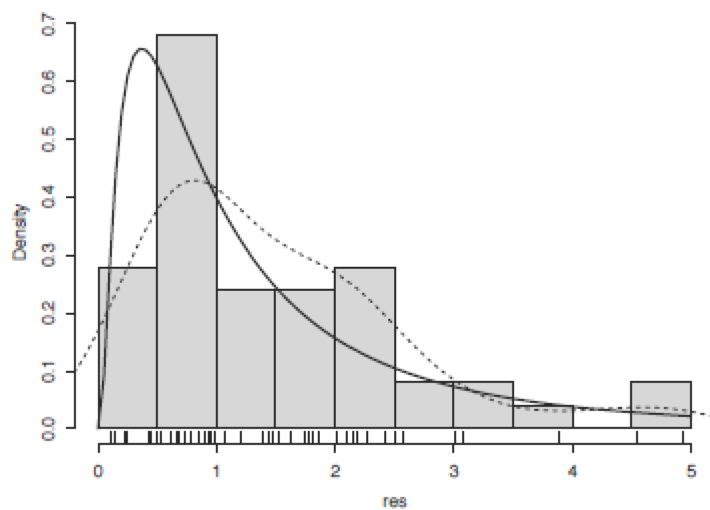
- Exponential(λ)
- vs. poisson



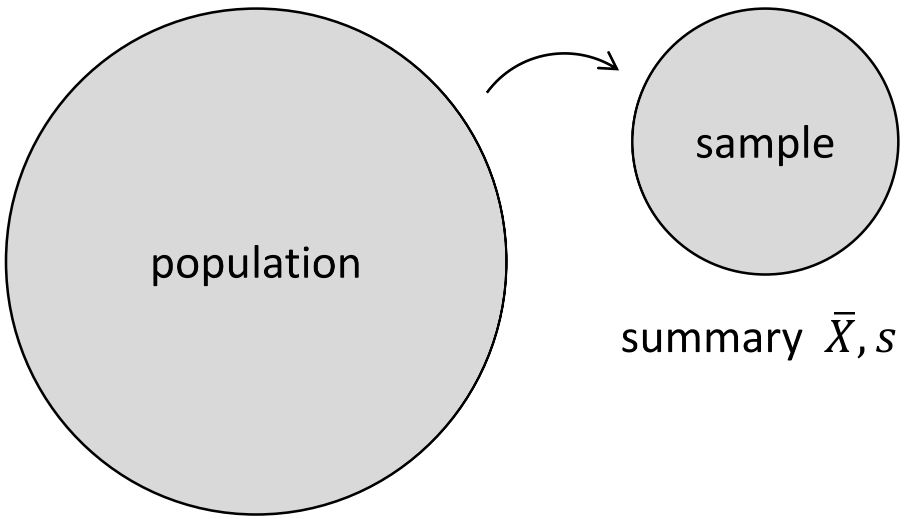
-wiki-

Lognormal distribution

- heavily skewed continuous distribution on positive numbers
- log(X) follows normal distribution

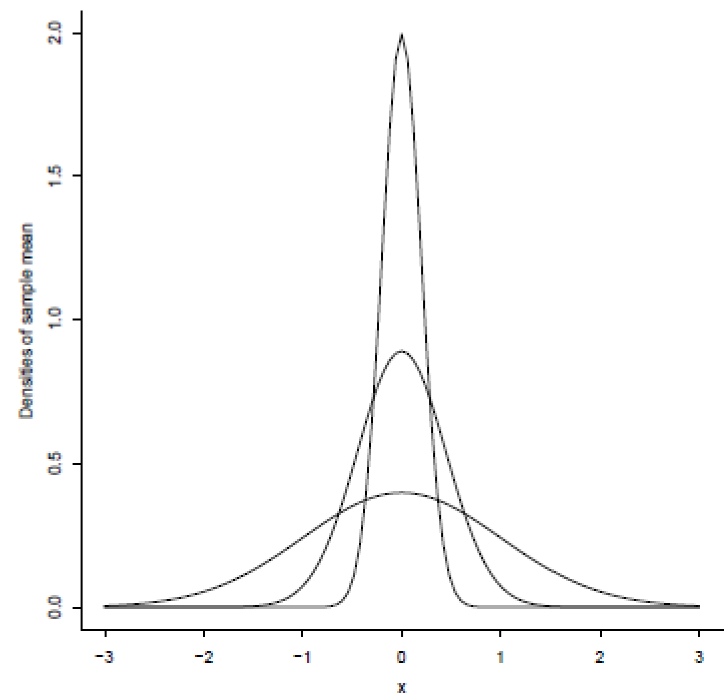


The central limit theorem



summary $\bar{X}, s$

summary μ, σ



$$P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq b\right) \approx P(Z \leq b)$$

숙제 #2 solution (다음시간제출, A4용지 사용, 이름, 학번 명시)

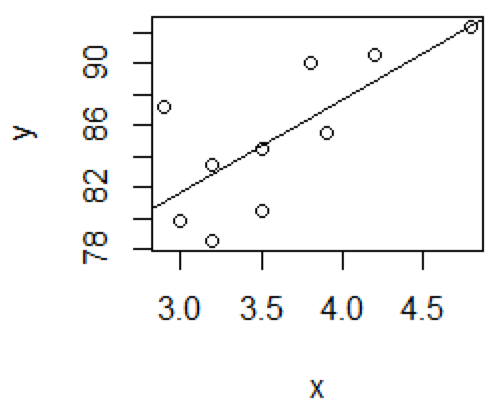
1. 다음 X와 Y로부터 두 변수의 correlation을 계산하고 그 의미를 해석하시오

X	1	2	3	4	5
Y	7	5	3	1	-1

-1, X와 Y가 강한 음의 상관관계가 있음

2. 어느 화학 제품의 공정 수율(Y)을 그 제품을 만들 때 들어가는 원료의 촉매량(X)에 영향을 받는 것으로 알려져 있다. 그 관련성을 알기 위해 다음 데이터를 얻었다. 설명변수 X와 반응 변수 Y 사이에 회귀직선을 적합하고 산점도를 그린 후 회귀직선을 그리시오.

X	3.5	3.9	3.2	4.2	4.8	3.0	3.2	3.5	2.9	3.8
Y	80.5	85.5	83.5	90.5	92.4	79.8	78.5	84.5	87.2	90.0



y=5.965*x + 63.767

숙제 #2 solution (다음시간제출, A4용지 사용, 이름, 학번 명시)

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

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

숙제 #3 (다음시간제출, A4용지 사용, 이름, 학번 명시)

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.

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

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

Next

Statistical inference