Statistics Review

Sample Space

Sample space (표본공간): The sample space ***S*** is a set that contains all possible outcomes from an experiment.

Experiment (실험): Any process or procedure that produces more than one outcome.

Event (사건, 이벤트): A subset of the sample space ***S***.

Discrete Sample Spaces (이산형)

Tossing a coin: S = {H, T}

Tossing a die: S = {1, 2, 3, 4, 5, 6}

Tossing 3 coins:

S = { HHH,
HHT, HTH, THH,
HTT, THT, TTH,
TTT }

Count # people that enter the post office:

S = {0, 1, 2, 3, 4, 5,…}

Continuous Sample Spaces (연속형)

Values between 0 and 130:

$S = \{x \mid 0 < x < 130\}$

Lifetime of a light bulb:

$S = \{x \mid 0 \leq x < \infty\}$

Points on a circle of radius 2:

$S = \{(x,y) \mid x^2 + y^2 = 4\}$

Points inside a circle of radius 2:

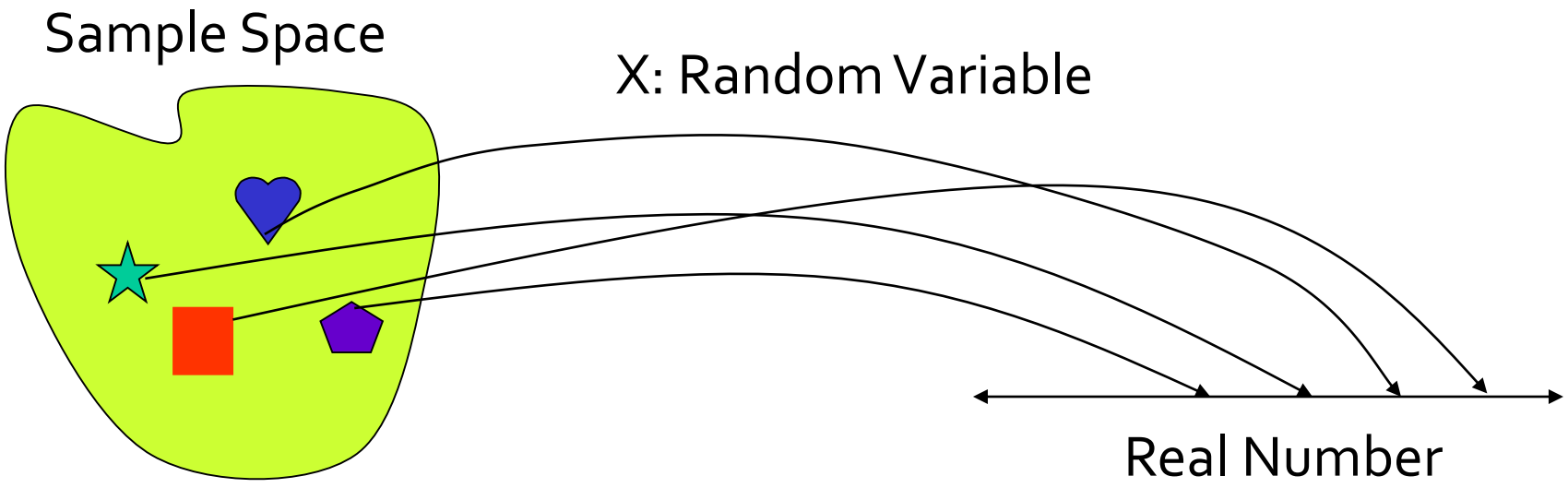
$S = \{(x,y) \mid x^2 + y^2 < 4\}$

Events (사건)

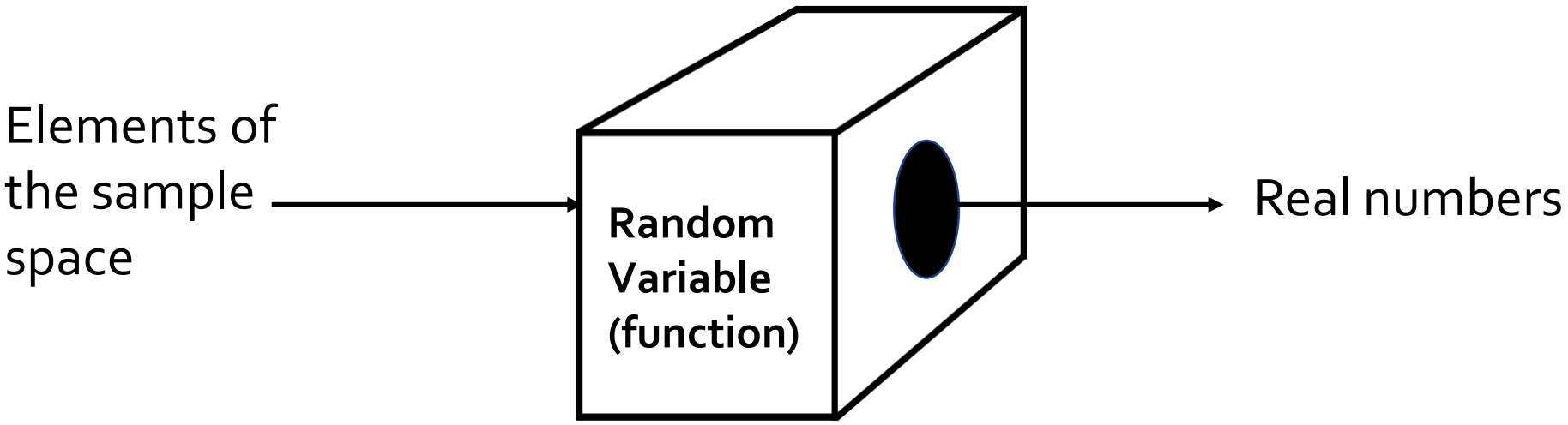
- An **event** is a subset of a sample space.
  - Toss 3 coins: A = {exactly two heads}
    = {HHT, HTH, THH}
  - Light bulb: A = {lasts < 200 hours}
    = $\{x \mid 0 \leq x < 200\}$
  - Roll a die: A = {odd number}
    = {1, 3, 5}

Random Variable (확률변수)

Random variable is a function that assigns real number to each element of the sample space.



Random Variable



Real numbers = f (Elements of the sample space)

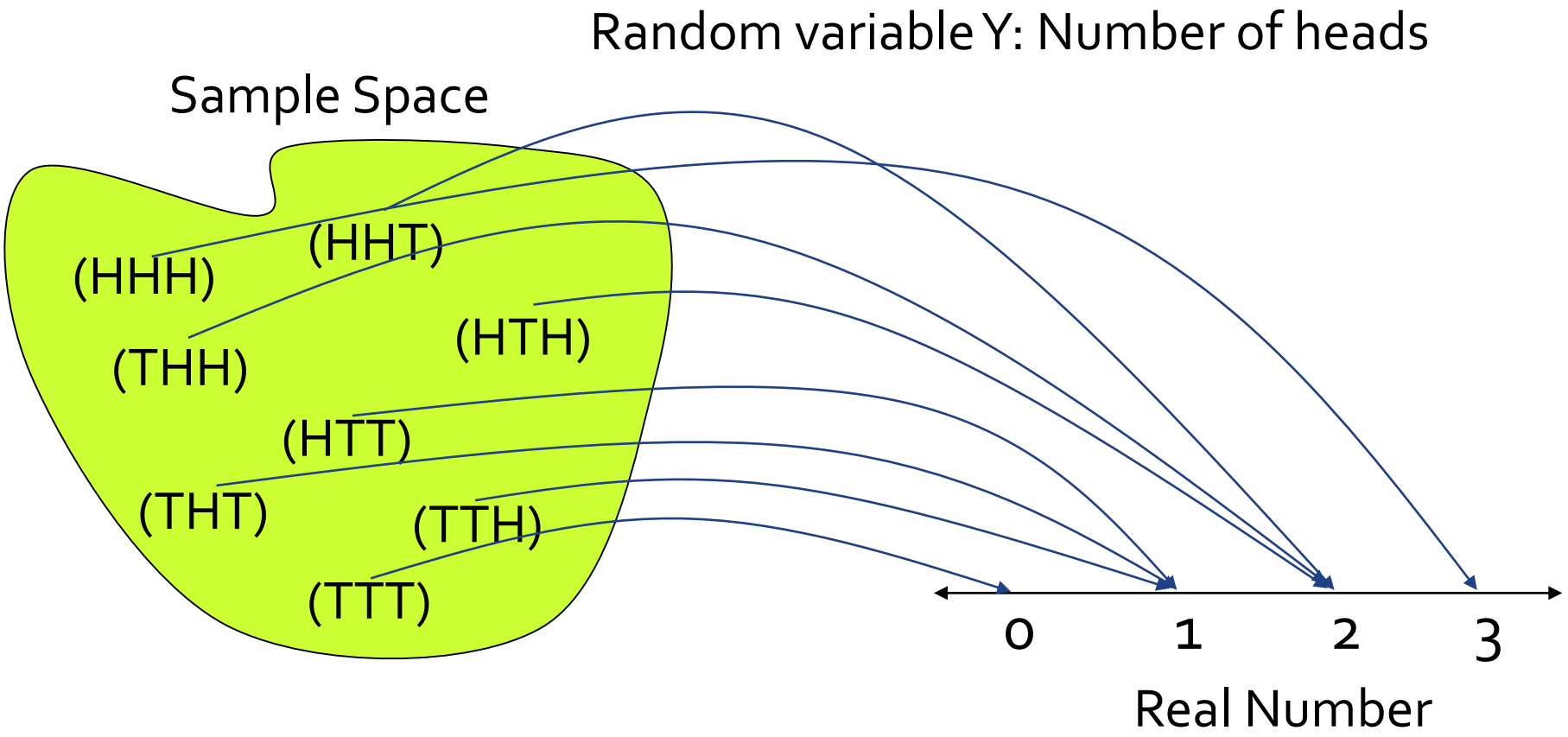
Random Variable - Example

- Flip 3 coins
- Sample space = { (HHH), (HHT), (HTH), (THH), (HTT), (THT), (TTH), (TTT) }
- Define *Y*=number of heads
- Random variable: *Y* (why??)
- Y={0, 1, 2, 3}

Y	outcomes
0	TTT
1	HTT, THT, TTH
2	HHT, HTH, THH
3	HHH

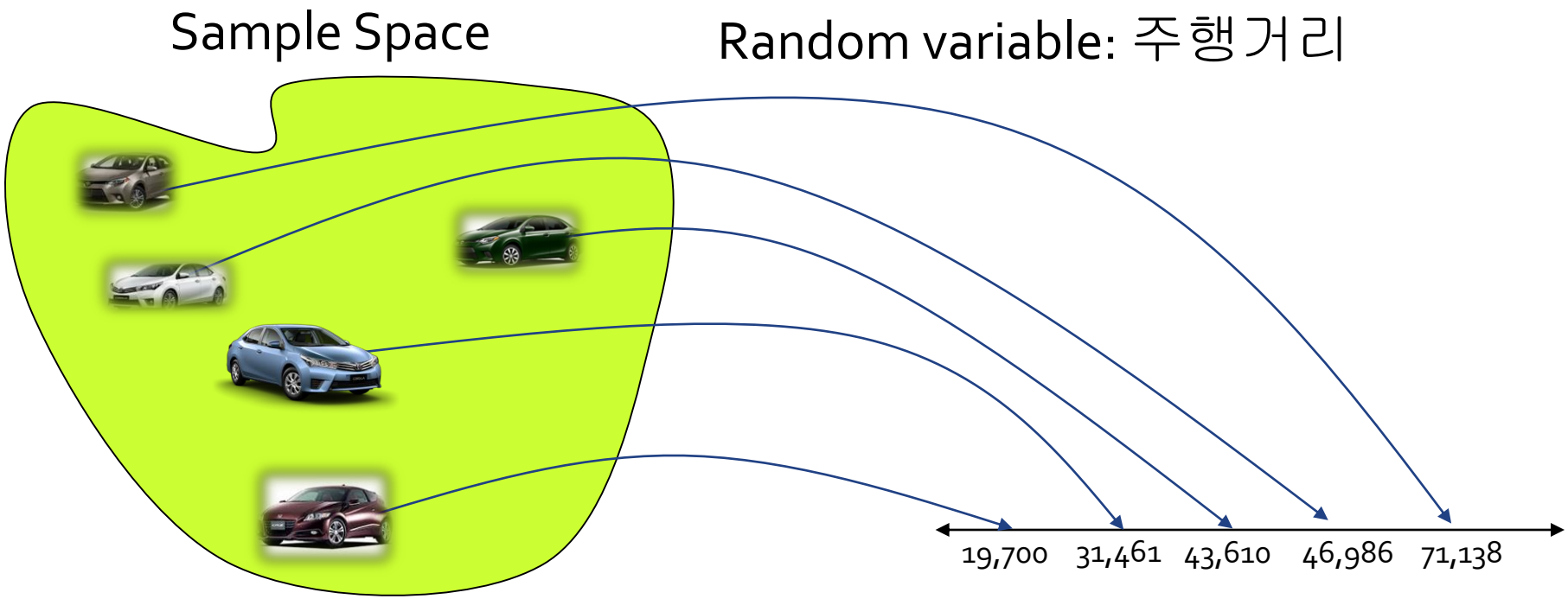
Random Variable - Example

- Flip 3 coins

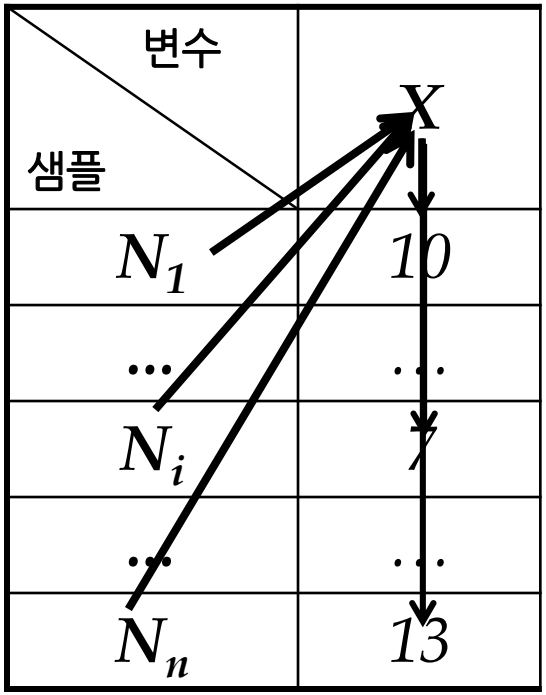


Random Variable - Example

No	모델	주행거리	마력	용량	가격
1	TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors	46,986	90	2,000	13,500
2	TOYOTA Corolla 1800 T SPORT VVT I 2/3-Doors	19,700	192	1,800	21,500
3	TOYOTA Corolla 1.9 D HATCHB TERRA 2/3-Doors	71,138	69	1,900	12,950
4	TOYOTA Corolla 1.8 VVTL-i T-Sport 3-Drs 2/3-Doors	31,461	192	1,800	20,950
5	TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT BNS 2/3-Doors	43,610	192	1,800	19,950



Random Variable - Example



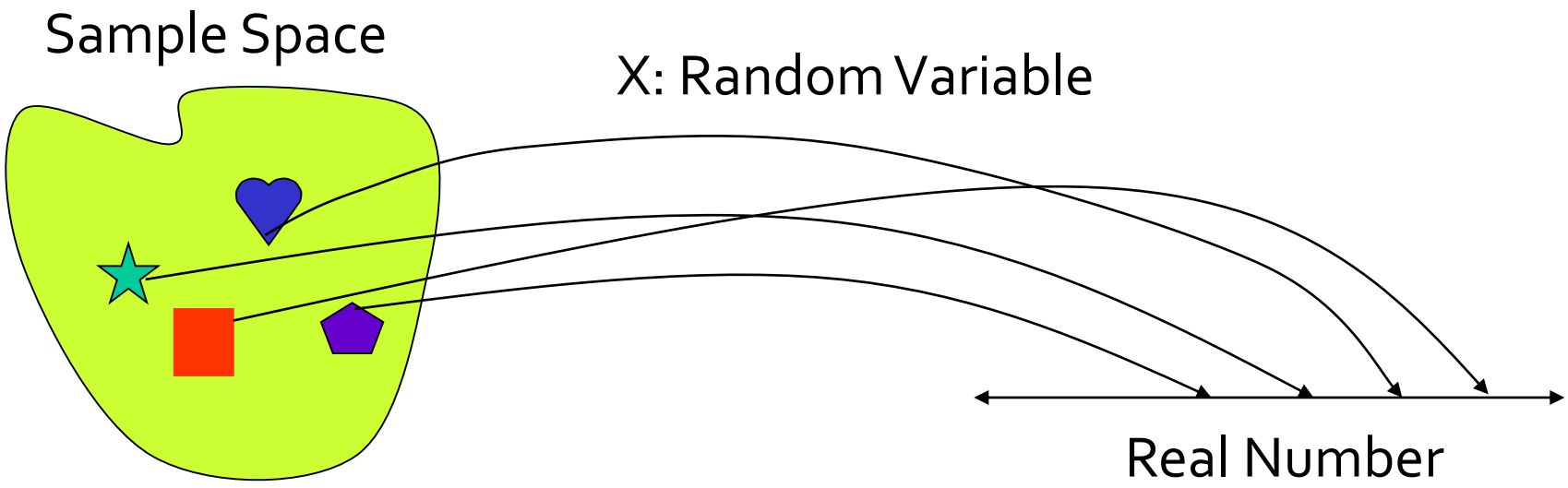
변수 / 샘플	X_1	...	X_p
N_1	x_{11}	...	x_{1p}
...	...	...	...
N_i	x_{i1}	...	x_{ip}
...	...	...	...
N_n	x_{n1}	...	x_{np}

Random Variable

- **Discrete random variables:** Random variable which can take only certain discrete values (0,1,2,….).
- *Example:*
  - The number of accidents per year in San Francisco in the US.
  - The number of eggs laid each month by a hen.

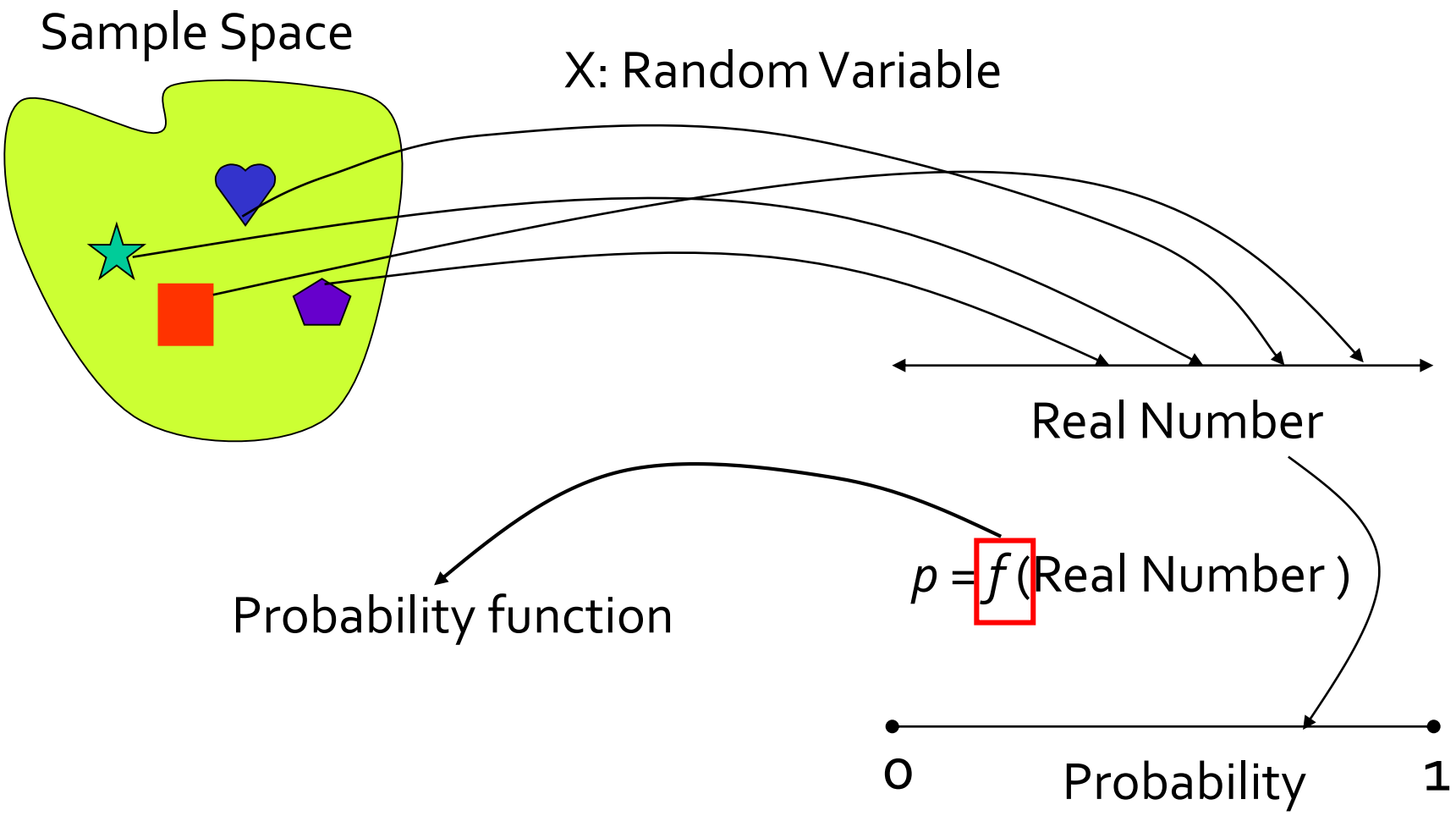
- **Continuous random variables:** Random variable which can take any values within some range (i.e., infinitely many values).
- *Example:*
  - The length of time to play a soccer game.
  - The annual income of Seoul residents.

Probability Function

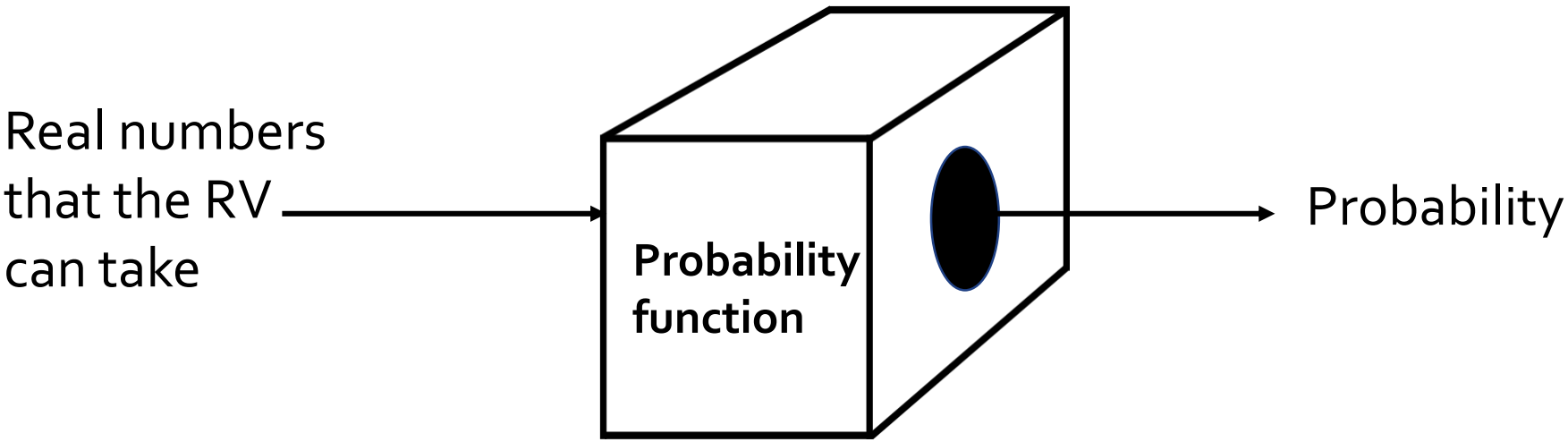


- A random variable can have a number of possible values (outputs)
- Each value of the random variable outputs has the corresponding probability
- Any relationship between random variable outputs and probabilities?

Probability Function

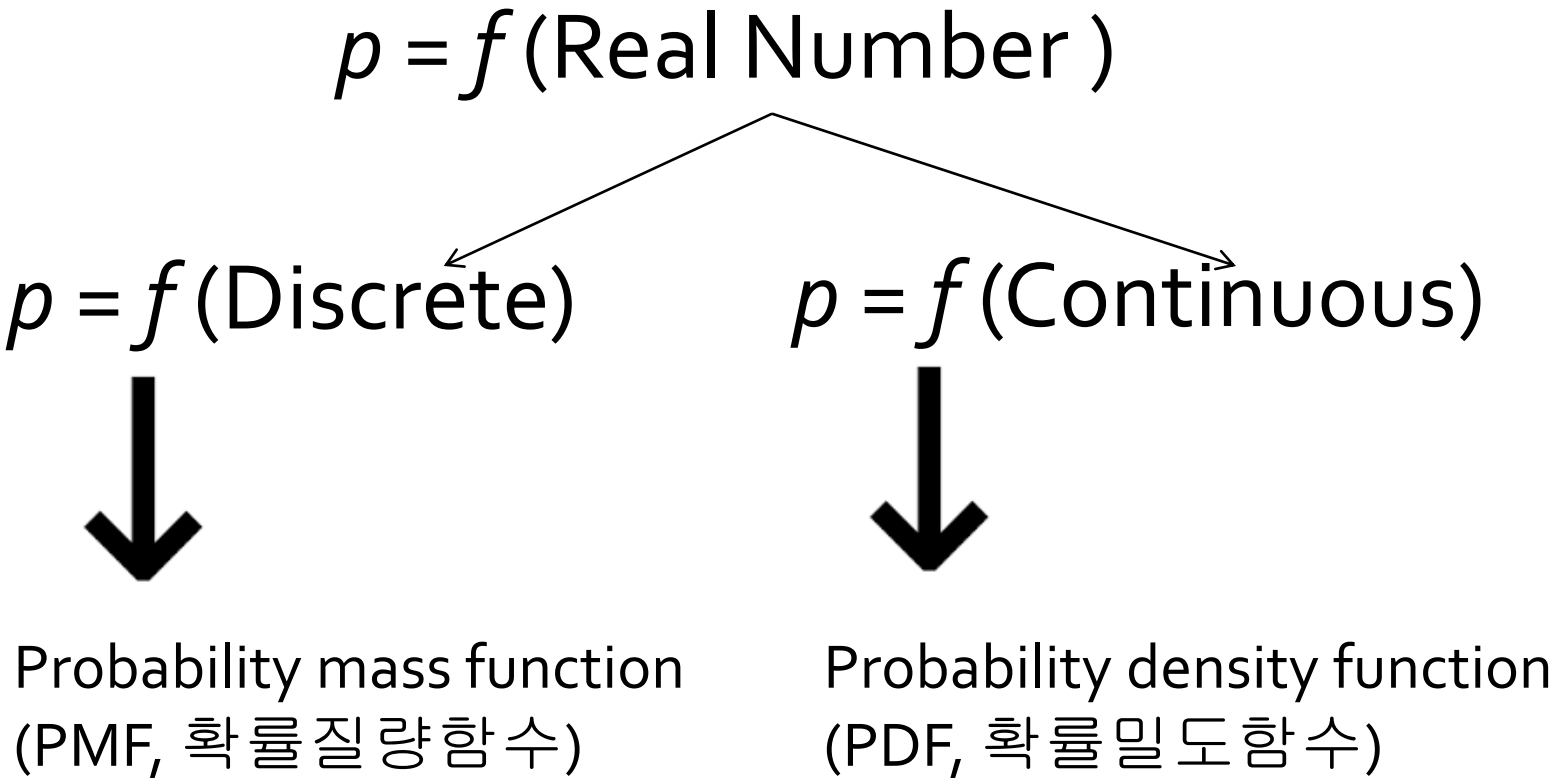


Probability Function

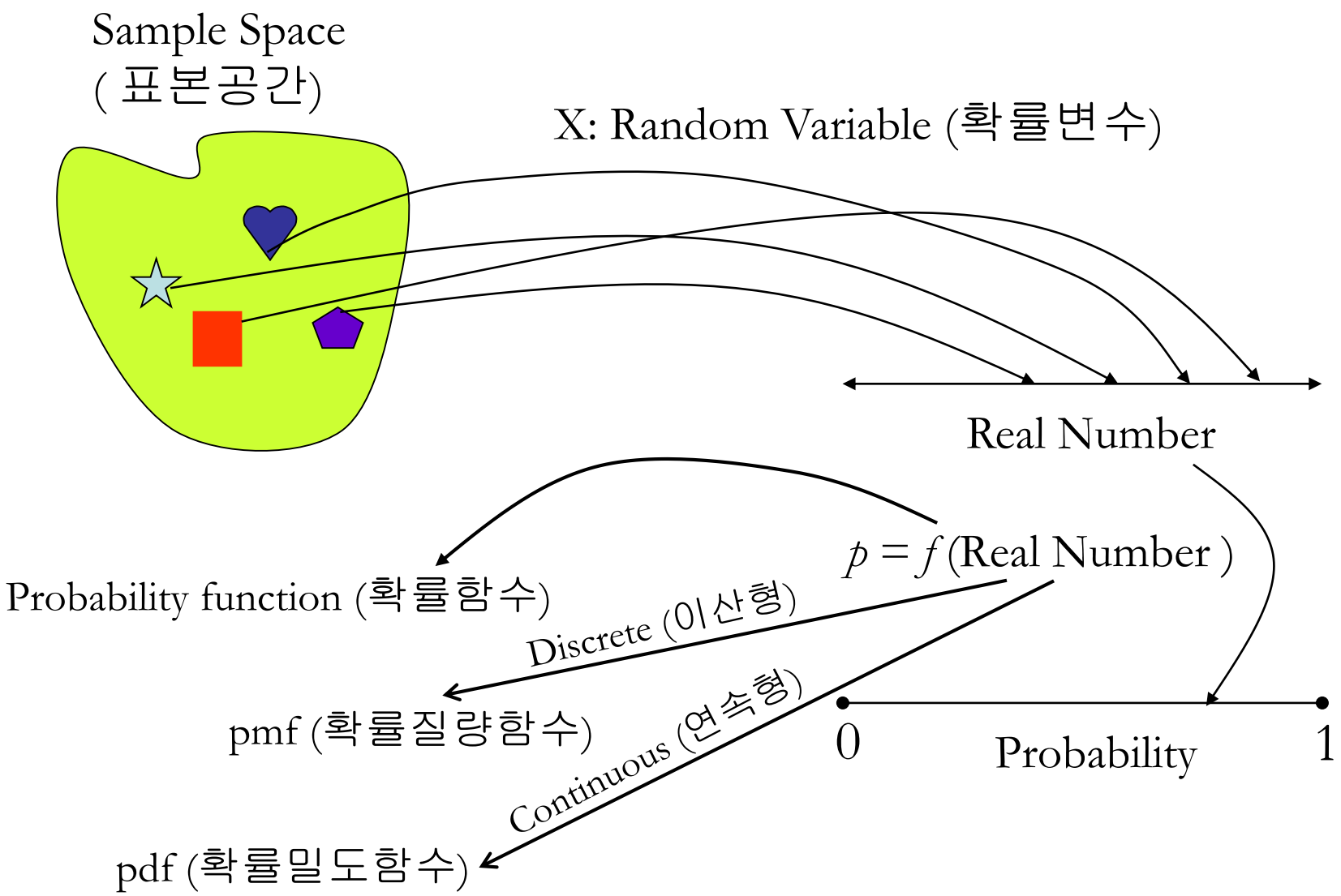


$$p = f(\text{Real Number})$$

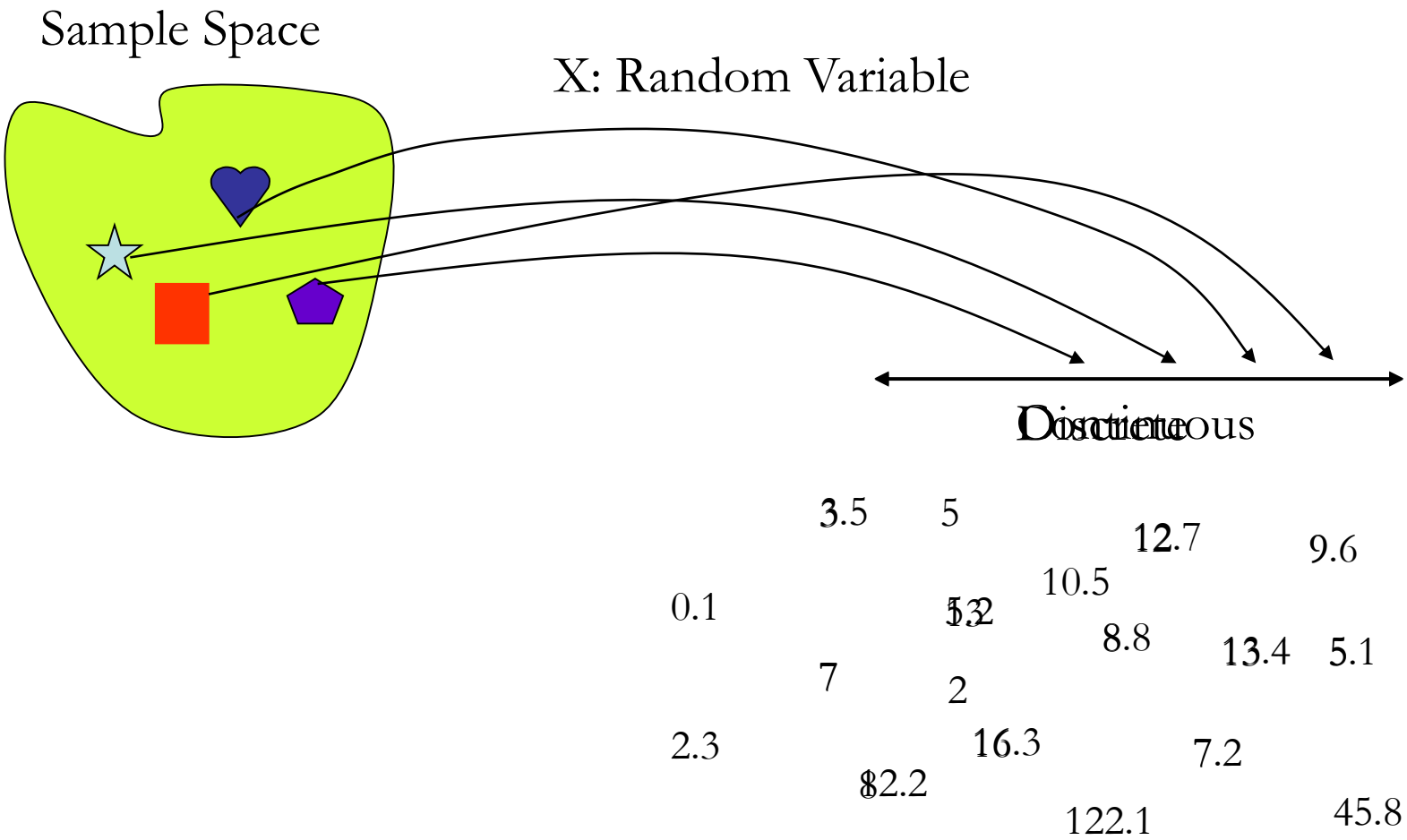
Probability Function



Summarization of Random Variable



Summarization of Random Variable



Expectation of a Discrete R.V.

- Expectation of a discrete random variable, X

$$E(X) = \sum_i x_i \cdot f_X(x_i)$$

Example

Q.) P(X=50)=0.3, P(X=200)=0.2, and P(X=350)=0.5
What is the expectation of R.V. of X ?

A.) E(X)=50·0.3 + 200·0.2 + 350·0.5 = 230 ($).

Expectation of a Continuous R.V.

- Expectation of a continuous random variable, X

$$E(X) = \int X \cdot f_X(X) dX$$

Example

Q.) Given a pdf, $f(X) = 1.5\text{-}6(X\text{-}50)^2$, $49.5 \le X \le 50.5$
What is the expectation of R.V. of X ?

A.) $$E(X) = \int_{49.5}^{50.5} X(1.5 - 6(X-50)^2) dX$$

Variance of a R.V.

- Definition: Measure of the spread of the distribution about its mean value.

$$
\begin{aligned}
V(X) &= E\left[\left\{X - E(X)\right\}^2\right] \\
&= E\left\{X^2 - 2XE(X) + E^2(X)\right\} \\
&= E(X^2) - 2E^2(X) + E^2(X) \\
&= E(X^2) - E^2(X).
\end{aligned}
$$

- Standard deviation

$$
\begin{aligned}
S.D.(X) &= \sqrt{Variance(X)}, \\
S.D.(X) &\geq 0.
\end{aligned}
$$

Covariance

- Suppose there are two R.V.s (say, X and Y)
- Covariance of X and Y: Measurement of the nature of the association between the two random variables.

$$Cov(X,Y) = E(XY) - E(X)E(Y).$$

- *Cov*(X,Y) is positively big → Positively strong relationship between the two.
- *Cov*(X,Y) is negatively big → Negatively strong relationship between the two.
- *Cov*(X,Y) = 0 → No relationship between the two.

Correlation

- Covariance is not scale free.
- Correlation coefficient: Scaled version of covariance.

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}},$$

$$-1 \leq Corr(X,Y) \leq 1.$$

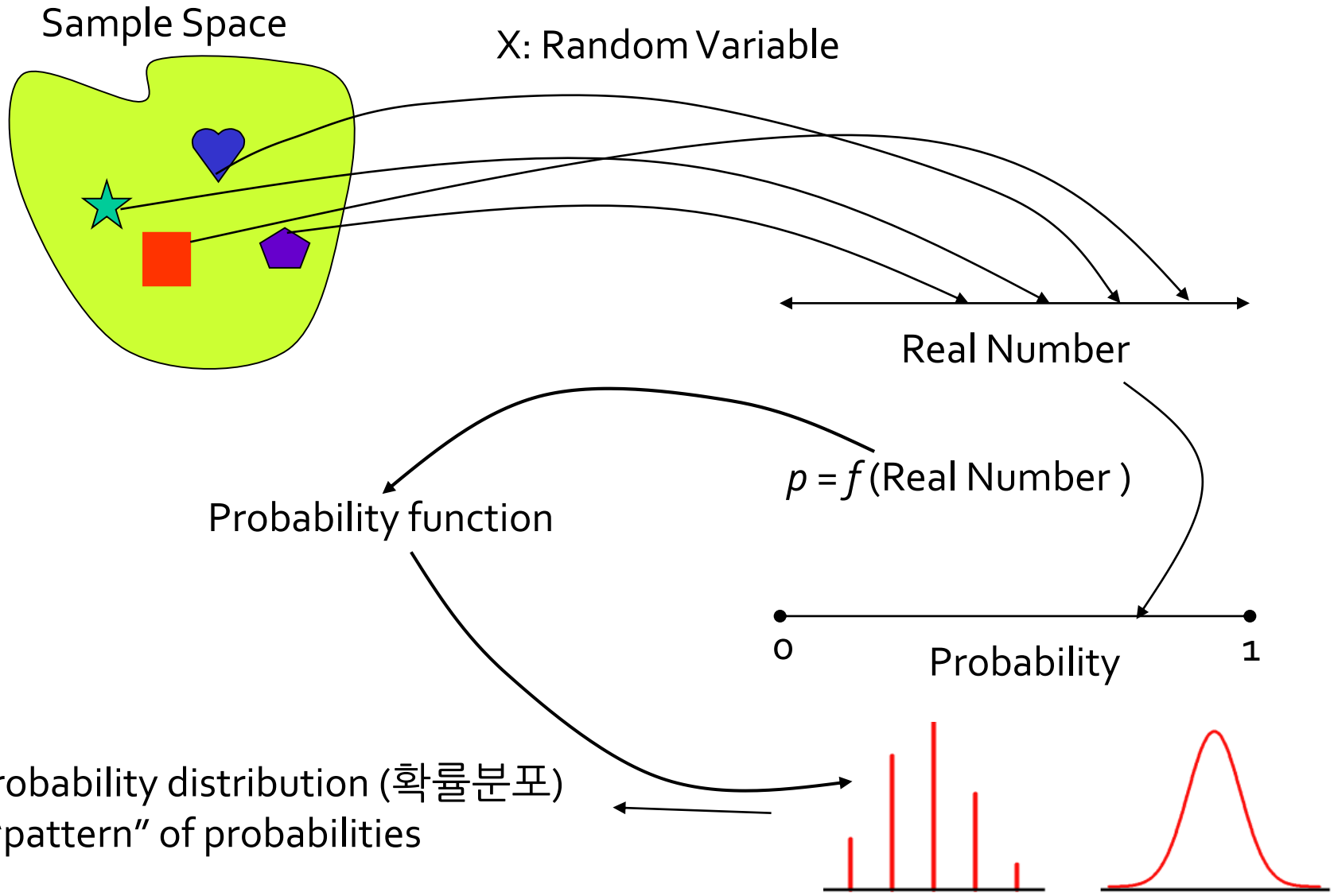
Mean and Variance of Linear Combinations of R.V.

For random variables X and Y with constant values a and b

$$E(aX+b)=aE(X)+E(b)=aE(X)+b.$$
$$E(g(X)\pm h(X))=E(g(x))\pm E(h(X)).$$

$$V(aX+b)=a^2V(X)+V(b)=a^2V(X).$$
$$V(aX+bY)=a^2V(X)+b^2V(Y)+2abCov(X,Y).$$

Sample Space → Random Variable → Probability Function → Probability Distribution



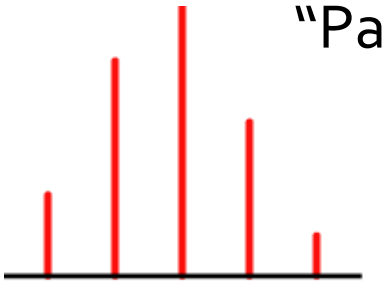
Probability distribution (확률분포)
: "pattern" of probabilities

Probability Distribution

$p = f(\text{Real Number})$

$p = f(\text{Discrete})$

Probability mass function



Discrete probability distribution
(이산형 확률분포)

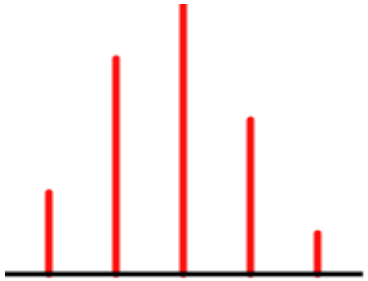
"Pattern of probabilities"

$p = f(\text{Continuous})$

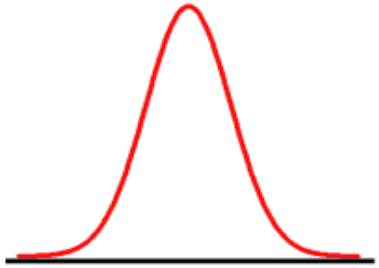
Probability density function

Continuous probability distribution
(연속형 확률분포)

Probability Distribution



- Bernoulli distribution
- Binomial distribution
- Poisson distribution
- Geometric distribution
- Negative Binomial distribution
- Hypergeometric distribution



- Uniform distribution
- Normal distribution
- Exponential distribution
- Gamma distribution
- Beta distribution

Bernoulli Distribution

- Consider a r.v. X with two possible outcomes
  - 0 = "failure" or 1 = "success"
- Define $p = P[\text{success}] = P[X = 1]$
  - The p.m.f. depends on the parameter p

$$f_X(x;p) = p^x(1-p)^{1-x} \quad \text{for } x = 0,1$$

- Expected value $E[X] = p$
- Variance $V(X) = p(1-p)$
- Example: Toss a coin
  - 0 = "tails" or 1 = "heads" and $p = 1/2$

Binomial Distribution

- One iteration of a Bernoulli experiment is called a Bernoulli trial
- Conduct n independent Bernoulli trials
- Let r.v. X be the # of successes in n trials
  - The p.m.f. depends on the parameters n and p

$$f(x;n,p)=\binom{n}{x}p^x(1-p)^{n-x} \quad \text{for } x=0,1,2,\ldots,n$$

- Expected value $E[X]=np$
- Variance $V(X)=np(1-p)$

Binomial Distribution

- Example: Count # of defectives
  - 0 = "nondefective" or 1 = "defective"
  - n = 10 manufactured parts, p = P[defective] = 0.1

$$P[X = 2] = b(2;10,0.1) = \binom{10}{2}(0.1)^2(0.9)^{10-2}$$

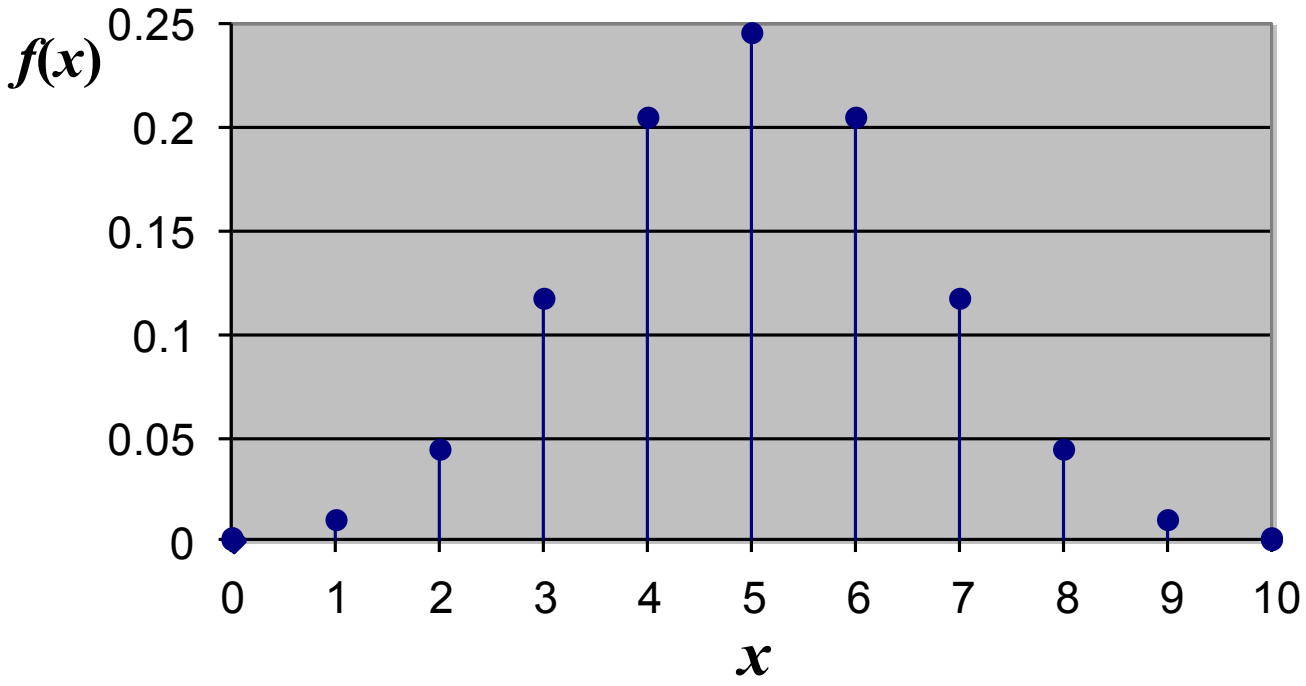
$$= 0.1937$$

$E[X]$ = (10)(0.1) = 1 defective

$V(X)$ = (10)(0.1)(0.9) = 0.9

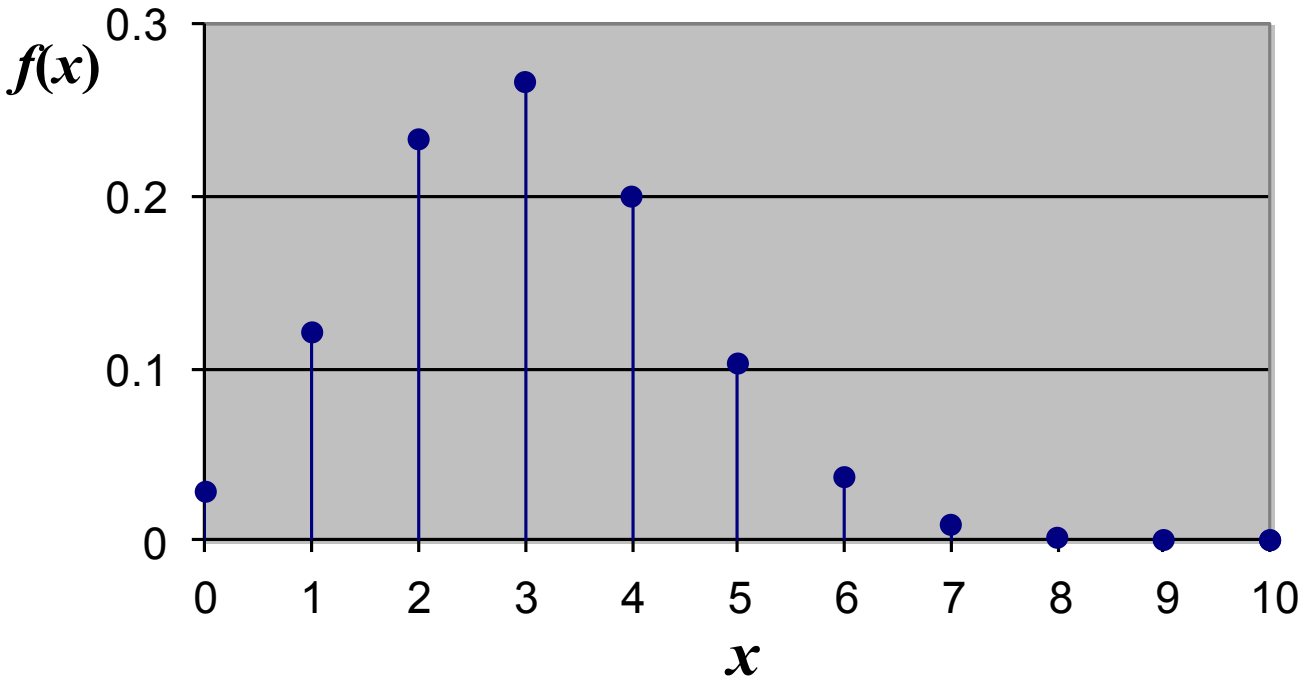
Binomial Distribution

- Example: $n = 10$, $p = 0.5$



Binomial Distribution

- Example: $n = 10$, $p = 0.3$



Geometric Distribution

- Conduct independent Bernoulli trials until the first "success" occurs ($p = P[\text{success}]$)
- Let r.v. X be the # of trials required
  - The p.m.f. depends on the parameter p

$$f(x;p) = p(1-p)^{x-1} \quad \text{for } x = 1,2,\ldots$$

- Expected value $E[X] = \frac{1}{p}$
- Variance $V(X) = \frac{1-p}{p^2}$

Geometric Distribution

- Example: Roll a die until we see a "6"
  - 0 = "not a 6" or 1 = "6"
  - $p = P[6] = 1/6$

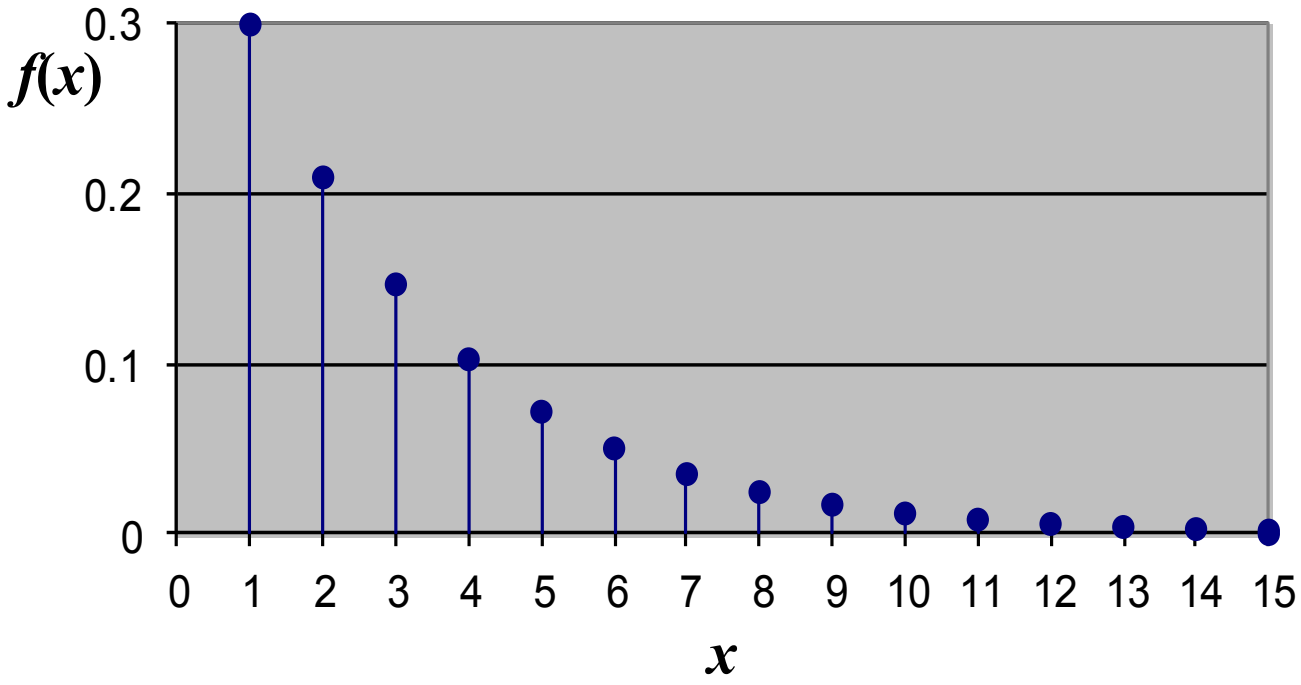
$$P[X=2] = f\left(2;\frac{1}{6}\right) = \left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{2-1} = \frac{5}{36}$$

$$E[X] = \frac{1}{1/6} = 6 \text{ trials}$$

$$V(X) = \frac{\frac{5}{6}}{\left(\frac{1}{6}\right)^2} = 30$$

Geometric Distribution

- Example: $p = 0.3$, $x = 1, 2, \ldots$



Poisson Distribution

- Count the occurrences of a specific event over a specific time period or a specific region.
  - # of customers entering a post office in one hour
  - # of misprints on a page (or group of pages)
  - # of car accidents in one month
- Define λ = rate of occurrence.
- Let r.v. X be the # of outcomes occurring over the time period (or region) t.
  - The p.m.f depends on the parameter $\mu = \lambda t$

$$p(x;\mu) = \frac{e^{-\mu}\mu^{x}}{x!} \quad \text{for } x = 0,1,2,\ldots$$

Poisson Distribution

- Expected value $E[X] = \mu = \lambda t$
- Variance $V(X) = \mu = \lambda t$

Poisson Distribution

- Example: Count # of radioactive particles over a 2.5-millisecond (ms) time period
  - λ = 4 particles/ms, t = 2.5 ms
  - $\mu = \lambda t = (4)(2.5) = 10$ particles

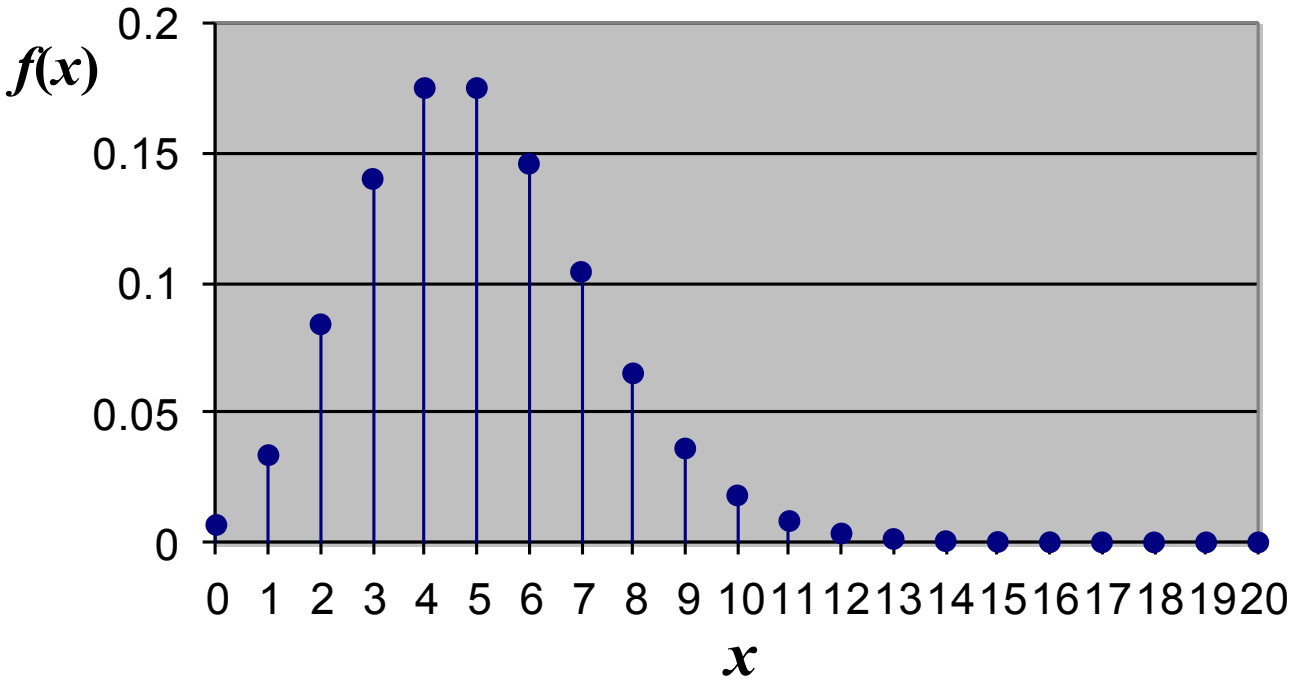
$$P[X = 15] = f(15;10) = \frac{e^{-10}(10)^{15}}{15!}$$

$$= 0.0348$$

$E[X] = V(X) = 10$ particles

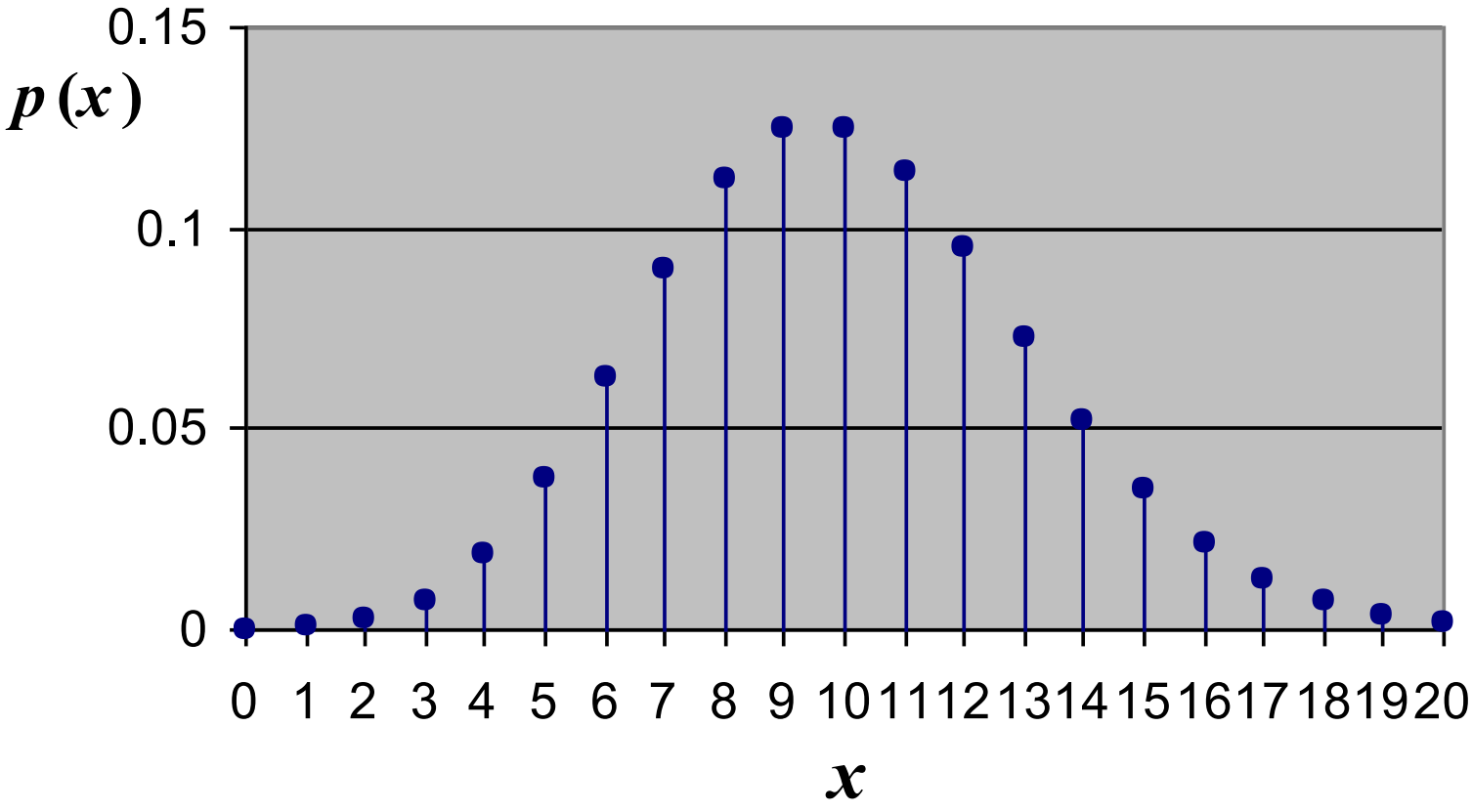
Poisson Distribution

- Example: $\mu = 5$, $x = 0, 1, 2, \ldots$

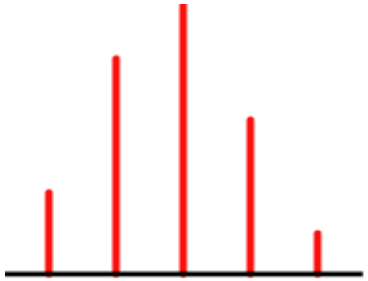


Poisson Distribution

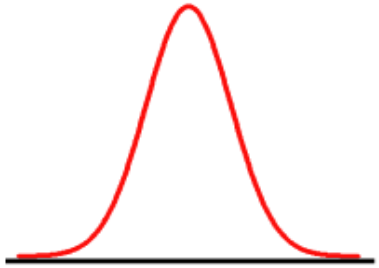
- Example: $\mu = 10$, $x = 0, 1, 2, \ldots$



Probability Distribution



- Bernoulli distribution
- Binomial distribution
- Poisson distribution
- Geometric distribution
- Negative Binomial distribution
- Hypergeometric distribution



- Uniform distribution
- Normal distribution
- Exponential distribution
- Gamma distribution
- Beta distribution

Normal Distribution

- A continuous r.v. X with a "bell curve" p.d.f.
  - One of the most important distributions in statistics
  - The p.d.f. depends on mean μ and variance σ^2

$$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for} \quad -\infty<x<\infty$$

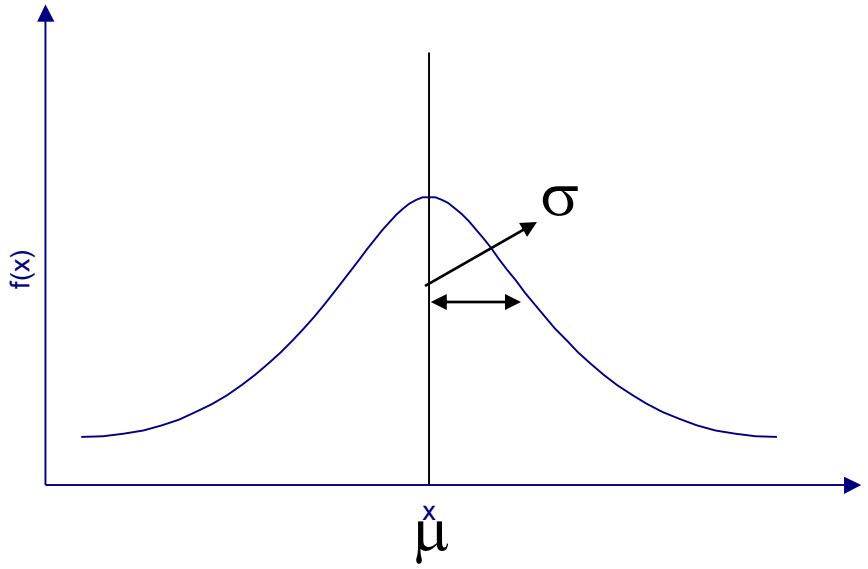
- Expected value $E[X]=\mu$

- Variance $V(X)=\sigma^2$

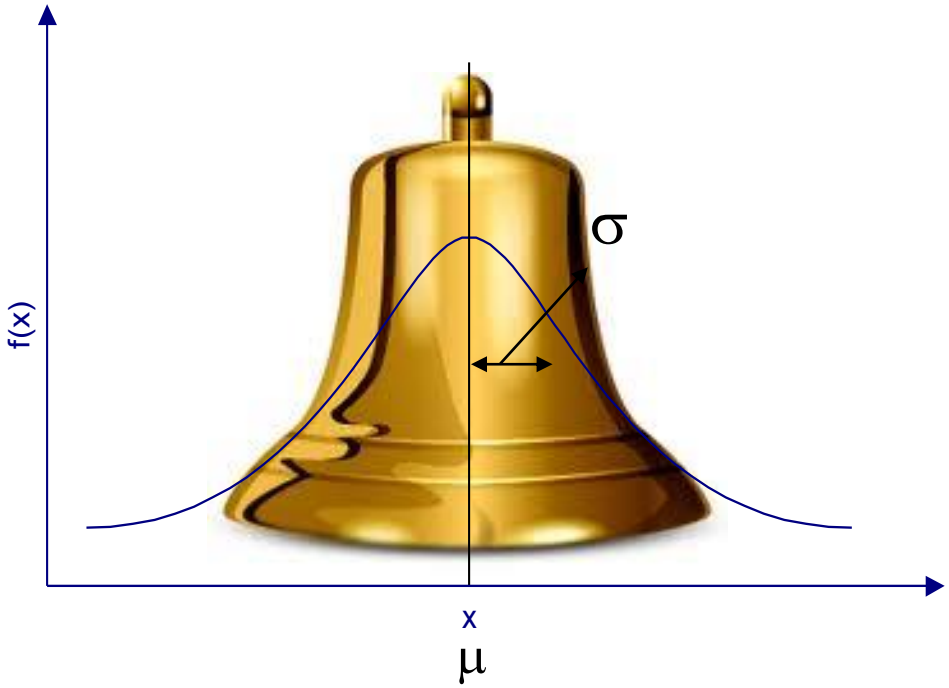
Normal Distribution

$$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for} \quad -\infty<x<\infty$$

$$E(X)=\mu,\ \ Var(X)=\sigma^2$$



Normal Distribution



- Symmetric about its mean (bell-shape curve)
- Probability is maximized when X= μ (unimodal)
- Whole area under the curve = 1 $P(\mu \leq X) =$?

Standard Normal Distribution

- Let X be distributed $N(\mu, \sigma^2)$
- Then $Z = \frac{X - \mu}{\sigma}$ is distributed $N(0, 1)$.
  - The p.d.f. for Z is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad \text{for} \quad -\infty < z < \infty$$

  - The c.d.f. for Z is tabulated in Table.

$$\Phi(z) = P[Z \le z] \quad \text{for} \quad -\infty < z < \infty$$

Standard Normal Table (*z* Table)

- **How to read Table.**

z	.00	.01	.02	.03
0.0	0.5000	0.5040	0.5080	0.5120
0.1	0.5398	0.5438	0.5478	0.5517
⋮				
1.2	0.8849	0.8869	0.8888	0.8907
1.3	0.9032	0.9049	0.9066	0.9082

$$P[Z \le 1.32] = \Phi(1.32) = 0.9066$$

$$P[Z \le -1.32] = 1 - \Phi(1.32) = 1 - 0.9066 = 0.0934$$

$$P[0.13 \le Z \le 1.31] = \Phi(1.31) - \Phi(0.13) = 0.9049 - 0.5517$$

Standard Normal Distribution

- **Let *X* be distributed $N(\mu, \sigma^2)$**

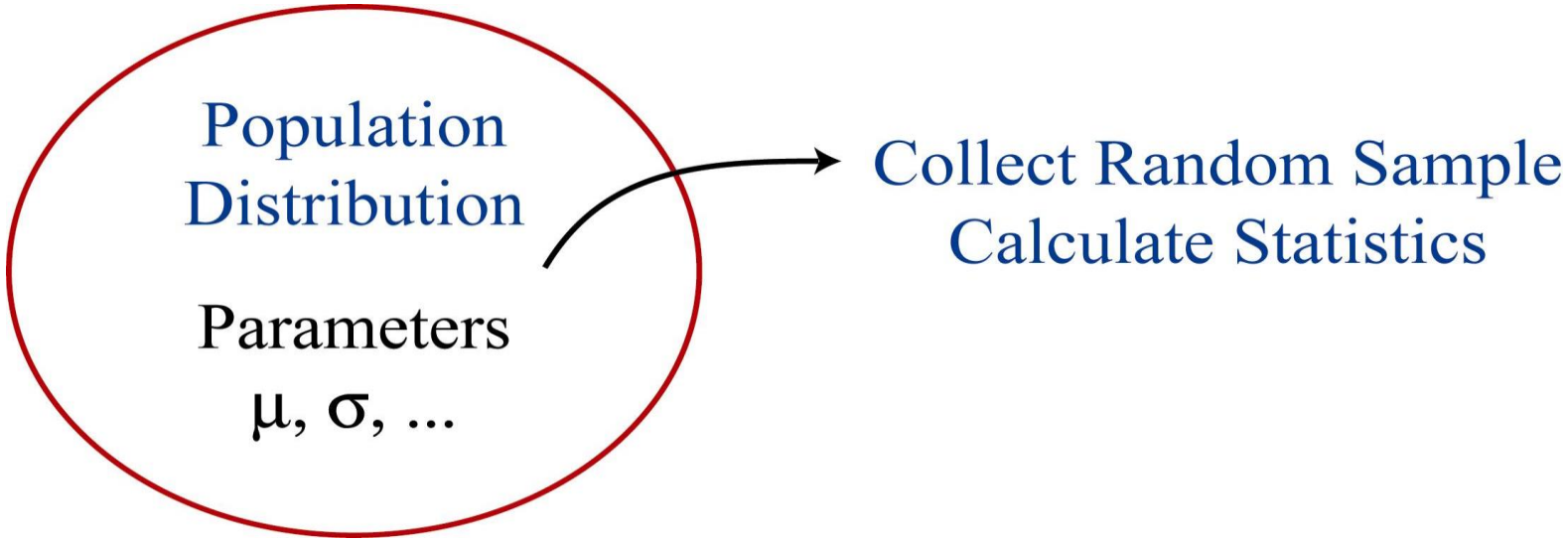
$$P[a \le X \le b] = P\left[\frac{a-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right]$$

$$= P\left[\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right]$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

- $z_1 = \frac{a-\mu}{\sigma}; \quad z_2 = \frac{b-\mu}{\sigma}$

Sampling



	Mean	Variance
Population	μ	σ^2
Sample	$\overline{X}$	S^2

Statistic (통계량)

- A statistic is a function of the samples

$$\text{Statistic} = f(x_1, x_2, \ldots, x_n)$$

$$\bar{X} = \frac{x_1 + x_2 + \ldots + x_n}{n} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

Sampling Distributions

- A sampling distribution is the distribution of a statistic.
- Two important statistics: $\overline{X}$, S^2
  - What is the distribution of $\overline{X}$?
  - What is the distribution of S^2?

Sampling Distributions - Mean

- Sample Mean (샘플 평균)

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right)$$

- Expected value of $\bar{X}$

$$E\left[\bar{X}\right] = E\left[\frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right)\right]$$

$$= \frac{1}{n}E\left[X_1\right] + \frac{1}{n}E\left[X_2\right] + \cdots + \frac{1}{n}E\left[X_n\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}(n\mu) = \mu$$

Sampling Distributions - Mean

- Variance of $\bar{X}$ (샘플 분산)
  - Note: $X_1, X_2, \ldots, X_n$ are independent r.v.'s

$$V(\bar{X}) = V\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right)$$

$$= \left(\frac{1}{n}\right)^2 V(X_1) + \left(\frac{1}{n}\right)^2 V(X_2) + \cdots + \left(\frac{1}{n}\right)^2 V(X_n)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

Sampling Distributions - Example

- If the population distribution is $N(\mu, \sigma^2)$
  - $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$
  - $\bar{X}$ is distributed $N(\mu, \sigma^2/n)$
  - $Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ is distributed $N(0, 1)$

- Example: $X_1, \ldots, X_{25}$ i.i.d. $N(\mu = 15, \sigma^2 = 100)$

$$P\left[\bar{X} \le 20\right] = P\left[\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \le \frac{20 - 15}{10/5}\right] = \Phi(2.5) = 0.9938$$

Sampling Distributions - Variance

- Sample Variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

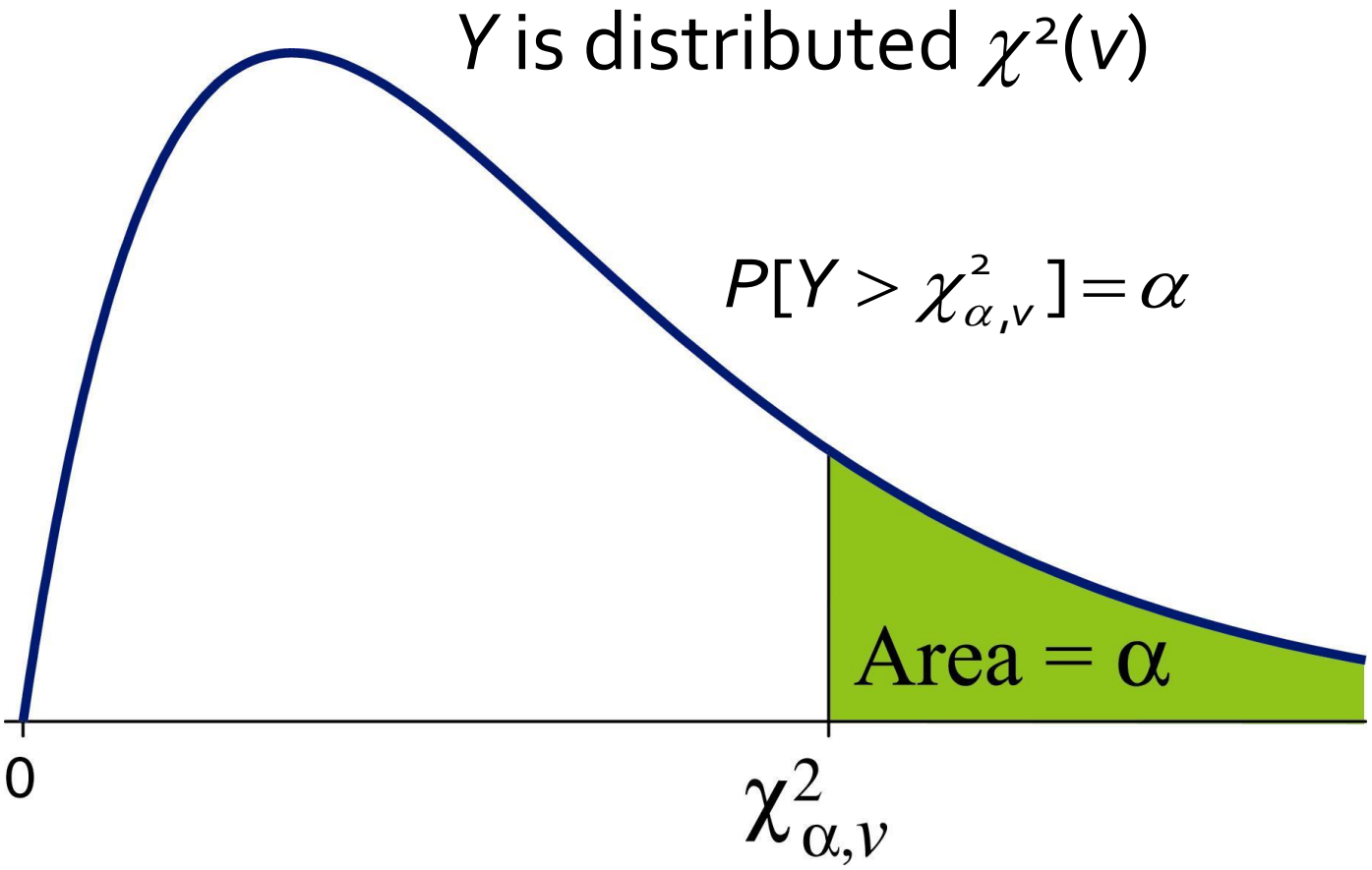
  - Note: $\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = (n-1)S^2$

- Assume $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$

  - Then $\frac{(n-1)S^2}{\sigma^2}$ is distributed $\chi^2(n-1)$

Sampling Distributions - Variance

- Upper tail percentiles of the $\chi^2(v)$ distribution



Sampling Distributions - Example

- Example: $X_1, \ldots, X_{25}$ i.i.d. $N(\mu = 15, \sigma^2 = 100)$
  - Then $Y = \dfrac{(n-1)S^2}{\sigma^2}$ is distributed $\chi^2(v = 24)$
  - $P[Y > \chi^2_{\alpha,24}] = \alpha$
- Find c such that $P[S^2 > c] = 0.95$

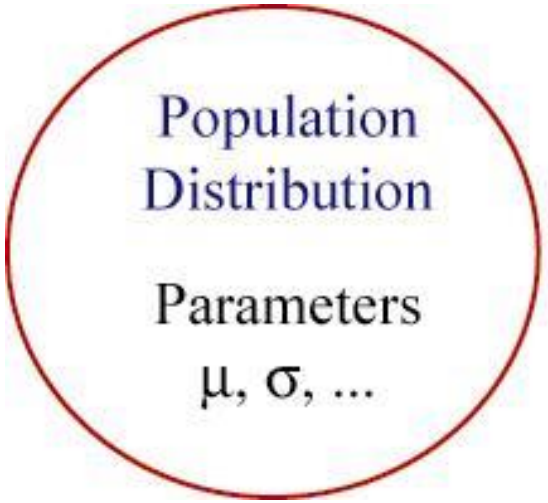
$$P\left[S^2 > c\right] = P\left[\frac{(n-1)S^2}{\sigma^2} > \frac{(25-1)c}{100}\right] = 0.95$$

$$\Rightarrow \frac{(24)c}{100} = \chi^2_{.95,24} = 13.848 \Rightarrow c = 57.7$$

Inferential Statistics (통계적 추론)

- Drawing conclusions on population based on sample(s)
- Estimation (추정)
- Hypothesis testing (가설검정)

Estimation



Setup:
Distribution form is known
Some parameters are unknown

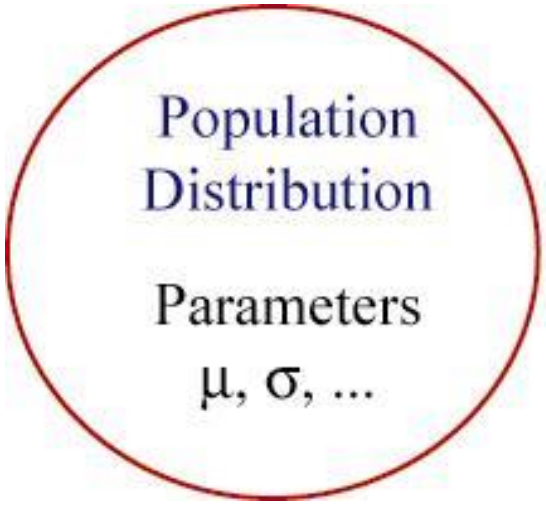
Goal:
Estimate unknown parameters

- An estimator is a special case of a statistic (추정량)
- An estimator is a function of the samples
- The purpose of estimator is to estimate an unknown parameter
- Two types of the estimator:
  (1) point estimator (점추정), (2) interval estimator (구간추정)

Point Estimator (점추정)

- An point estimator is designed to estimate an unknown parameter with a single value
  - θ = unknown parameter
  - $\hat{\theta}$ = point estimator (a function of the samples)
- Example: $\hat{\mu} = \bar{X}$ estimates μ, unknown parameter
- How to identify a good point estimator?
  - An estimator $\hat{\theta}$ is unbiased iff $E[\hat{\theta}] = \theta$
  - If an estimator $\hat{\theta}$ has the smallest variance, then it is the most efficient estimator of θ

Point Estimator



Normal distribution with unknown parameters, μ, σ^2

- $\overline{X}$ is a point estimator for the population mean μ
- S^2 is a point estimator for the population variance σ^2

Interval Estimation

- Estimate an unknown parameter θ **with an interval** that depends on the variance of $\hat{\theta}$
  - A confidence interval (C.I.) is constructed such that achieves a specified confidence level of $1-\alpha$

$$P[L(\hat{\theta}) \leq \theta \leq U(\hat{\theta})]$$

Interval Estimation – Confidence Interval

- The standard deviation (sd) of an estimator $\hat{\theta}$ is its standard deviation, $\text{sd}\{\hat{\theta}\} = \sqrt{\text{Var}(\hat{\theta})}$
  - If $\text{Var}(\hat{\theta})$ must be estimated, then $\text{sd}\{\hat{\theta}\}$ is the estimated standard error of $\hat{\theta}$

    Example: $\text{sd}\{\bar{X}\} = \sigma/\sqrt{n}$ or $s/\sqrt{n}$
- Note: Two-sided C.I.'s for means all have the form $\hat{\theta} \pm (C)\text{sd}\{\hat{\theta}\}$
  - C = critical value (e.g., $z_{\alpha/2}$)

CI Estimation of μ in Single Population

- Recall: If population distribution is $N(\mu, \sigma^2)$
  - $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$
  - $\hat{\mu} = \bar{X}$ is distributed $N(\mu, \sigma^2/n)$
  - $Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ is distributed $N(0, 1)$
- Construct 95% C.I. for μ when σ^2 is known:

$$P\left[-1.96 \le Z \le 1.96\right] = 0.95$$

$$\leftrightarrow P\left[-1.96 \le \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \le 1.96\right] = 0.95$$

CI Estimation of μ in Single Population

- Rearrange probability statement, so that μ is in the middle:

$$\leftrightarrow P\left[-1.96\frac{\sigma}{\sqrt{n}} \le \bar{X} - \mu \le 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

$$\leftrightarrow P\left[-\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le -\mu \le -\bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

$$\leftrightarrow P\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

CI for μ in Single Population

- Thus, the exact 95% C.I. for μ is:

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} , \bar{x} + 1.96\frac{\sigma}{\sqrt{n}} \right)$$

- Suppose the population variance is 9, and a sample of 16 results in a sample mean of 20

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}} = 20 \pm (1.96)\frac{3}{4} = (18.53, 21.47)$$

Estimation of μ in Single Population (unknown σ)

- Recall: If population distribution is $N(\mu, \sigma^2)$
  - $T = \dfrac{\bar{X} - \mu}{S / \sqrt{n}}$ is distributed $t(n-1)$
- Construct $100(1-\alpha)\%$ C.I. for μ when σ^2 is unknown:

$$P\left[-t_{\alpha/2,n-1} \le \frac{\bar{X} - \mu}{S/\sqrt{n}} \le t_{\alpha/2,n-1}\right] = 1-\alpha$$

$$\leftrightarrow P\left[\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right] = 1-\alpha$$

Estimation of μ in Single Population (unknown σ)

- Exact $100(1-\alpha)\%$ <u>two-sided</u> C.I. for μ is:

$$\left(\bar{x} - t_{\alpha/2,n-1} \frac{s}{\sqrt{n}} , \bar{x} + t_{\alpha/2,n-1} \frac{s}{\sqrt{n}} \right)$$

$$\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} , \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

Estimation of μ in Single Population (unknown σ)

- **Example:** 98% C.I. for μ, given $n = 16$ yields a sample mean of 20, sample variance of 9

$$\bar{x} \pm t_{.01,15} \frac{s}{\sqrt{n}} = 20 \pm (2.602)\frac{3}{4} = (18.05, 21.95)$$

Hypothesis Testing

- **가설 검정 = 가설(Hypothesis) + 검정(Testing)**
- **가설이란?**
  - 내가 주장하고 싶은 사실
- **가설**
  - 귀무가설(Null Hypothesis: H_0) "Equal(=) ; 같다"
    - 기존의 사실 (의미가 없다)
  - 대립가설(Alternative Hypothesis: H_1) "Not Equal(≠) ; 다르다"
    - 자료로부터 나온 증거에 의하여 새롭게 주장하고 싶은 사실 (의미가 있다)

보통 귀무가설을 기각하여 본인이 주장하고 싶은 사실이 의미가 있기를 바람.

Hypothesis Testing

Question: 렌즈의 뒤틀림이 글라스 두께에 영향을 미쳤나?

Step 1: 가설수립

H_0: 영향이 없었음 ***vs.*** H_1: 영향 미쳤음

Step 2: 샘플수집

Step 3: 검정통계량 T(X), based on H_0.

Step 4: p-value = $P(X>T(X))$, $X \sim D(\theta)$

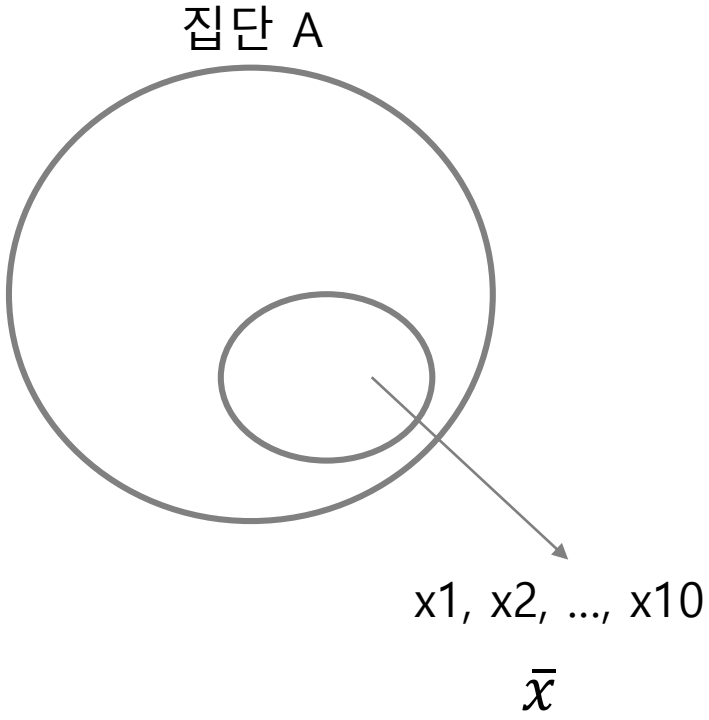
Step 5: 의사결정

If p-value < 0.01 (or 0.05), reject H_0.

If p-value > 0.1, accept H_0.

Hypothesis Testing

검정통계량: 귀무가설이 참이라는 가정 아래 얻은 통계량

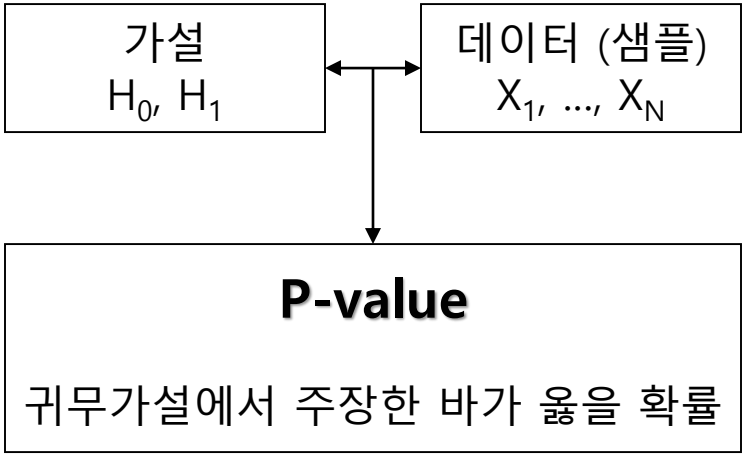


H_0: 집단 A의 평균이 10
H_1: 집단 A의 평균이 10이 아님

검정통계량: T = $\bar{x}$ - 10

큰 값의 T?
작은 값은 T?

P-value for Hypothesis Testing



- P-Value: 확률값, 0과 1사이로 표준화된 지표
- **귀무가설이 참이라는 가정 아래 얻은 통계량이 귀무가설을 얼마나 지지하는지를 나타낸 확률**
- 귀무가설을 채택할지 기각할지 기준으로 사용할 수 있는 값

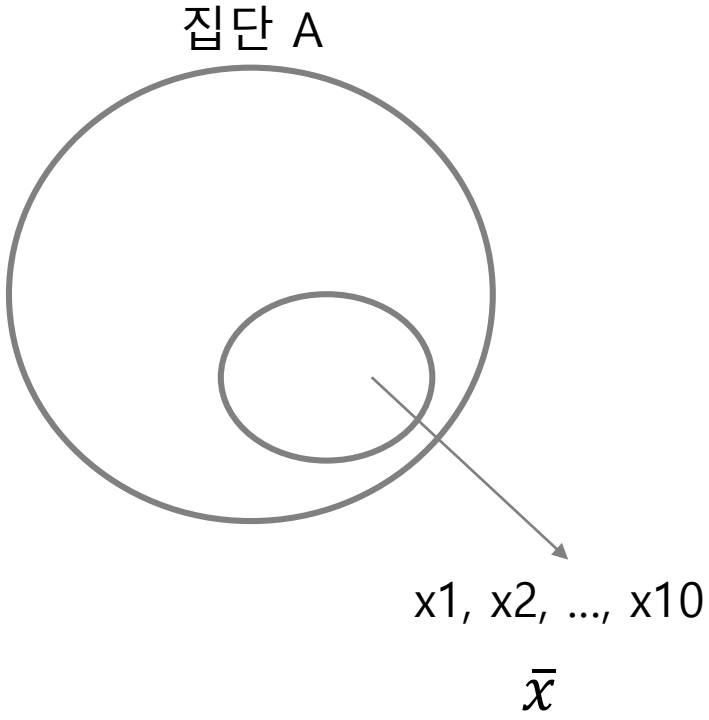
P-value for Hypothesis Testing

- 작은 P-value → 귀무가설이 참일 확률이 적어짐

- P-value < 0.05 (0.01): 귀무가설이 틀릴 확률이 매우 큼
- P-value > 0.1: 귀무가설이 틀리지 않을 확률이 매우 큼

P-value for Hypothesis Testing

검정통계량: 귀무가설이 참이라는 가정 아래 얻은 통계량



H_0: 집단 A의 평균이 10
H_1: 집단 A의 평균이 10이 아님

검정통계량: T = $\bar{x}$ - 10

P-value = P (Y > T), Y ~ 분포

P-value for Hypothesis Testing

- n=15, $\bar{X}$=10.6, s=1.61

$$H_0 : \mu = 10 \text{ vs. } H_1 : \mu \neq 10$$

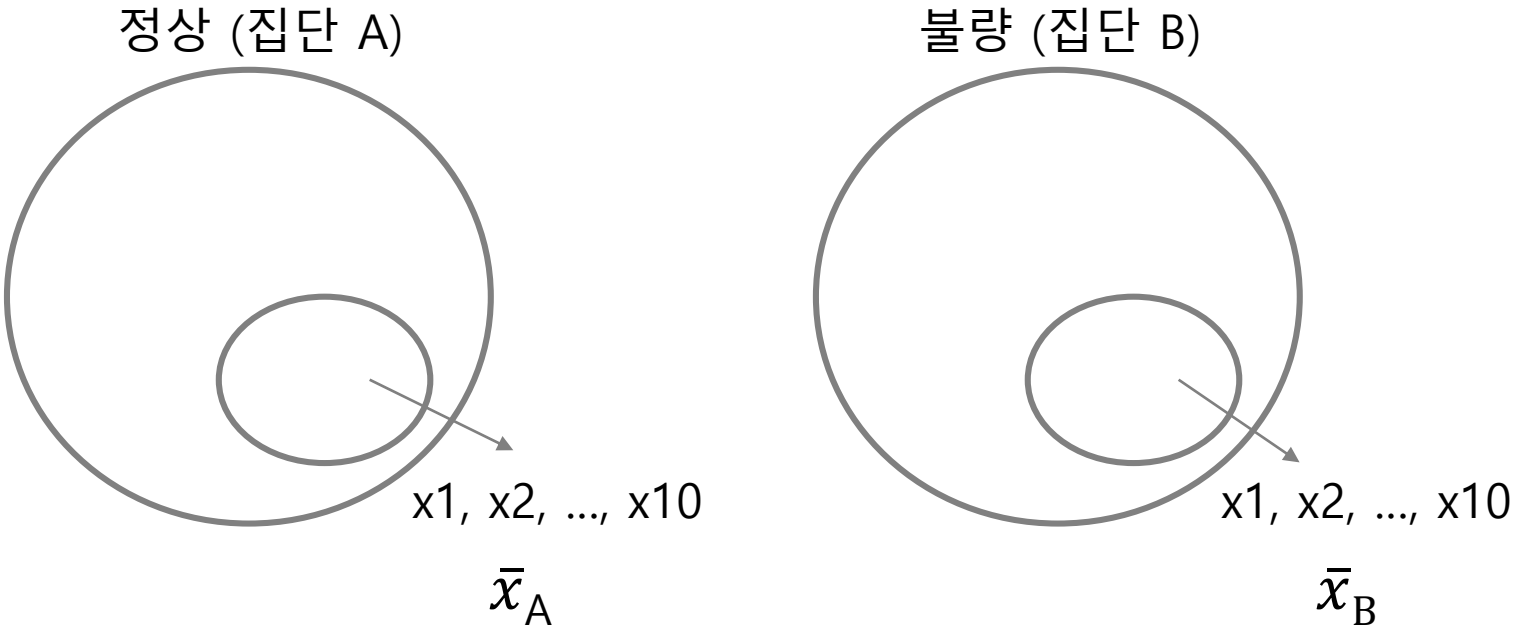
$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{10.6 - 10}{1.61/\sqrt{15}} = 1.44$$

$$\text{p} - \text{value} = 2 \times P(X \geq 1.44), \text{where } X \sim t(15-1)$$

$$= 2 \times 0.086$$

$$= 0.176$$

Two Samples: Hypothesis Testing



H_0: 정상집단과 불량집단의 평균이 같음
H_1: 정상집단과 불량집단의 평균이 다름

검정통계량: $T = (\bar{x}_A - \bar{x}_B) - 0$

P-value = P (Y > T), Y ~ 분포

Two Samples: Hypothesis Testing

정상

불량

x
75
85
90
88
96
78
82
84
65
70
68
74
80
84
70
75
64
60

변수 x는 정상과 불량상태에서 좌측과 같은 데이터를 생성하였다. 정상상태에서의 값이 불량상태에서의 값보다 크다고 할 수 있는가?

정상상태에서의 평균값을 μ_A, 불량상태에서의 평균값을 μ_B라고 하자.

$H_0: \mu_A = \mu_B$ vs $H_1: \mu_A > \mu_B$

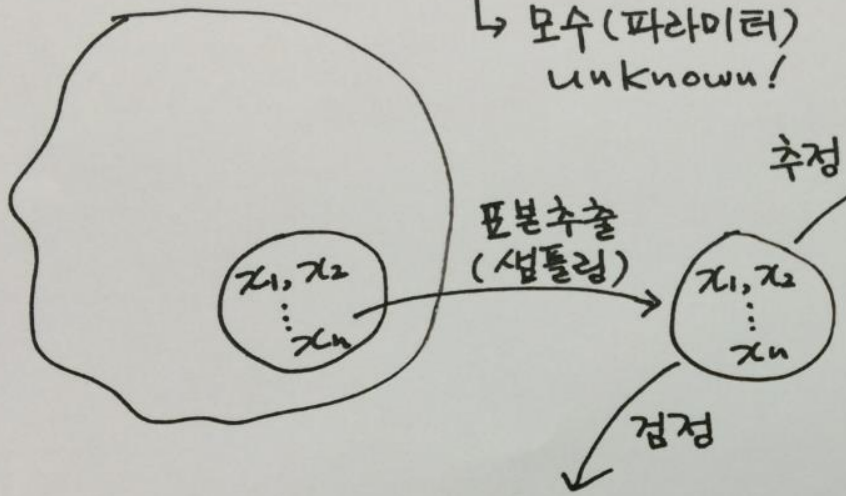
$$T^* = \frac{(\overline{X}_A - \overline{X}_B) - (\mu_A - \mu_B)}{\sqrt{S_A^2/n_A + S_A^2/n_B}} = \frac{(84.75 - 71) - 0}{\sqrt{44.78/8 + 54.67/10}} = 4.134$$

P-value = P(X>4.134) = 0.003, X~t(16=8+10-2)

귀무가설 기각. 정상상태에서의 값이 불량상태에서의 값보다 크다. 즉, 변수 x는 정상과 불량을 나누는데 중요한 변수이다.

모집단 ~ D(Θ)

↳ 모수(파라미터) unknown!



① $\hat{\Theta} = f(x_1, x_2, \cdots, x_n)$: 점추정

② $L(\hat{\Theta}) \le \Theta \le U(\hat{\Theta})$: 구간추정

$L(\hat{\Theta}) = \hat{\Theta} - C \cdot SD(\hat{\Theta})$

$U(\hat{\Theta}) = \hat{\Theta} + C \cdot SD(\hat{\Theta})$

$H_0 : \Theta = \Theta_0$ vs. $H_A : \Theta \neq \Theta_0$

검정통계량 (T) $= \dfrac{\hat{\Theta} - \Theta_0}{\text{Scaling}}$

① T의 분포 Known (Parametric)

$T \sim D(k)$

$T \gtrless D(\alpha, k)$ → Reject H_0 / → Accept H_0

② T의 분포 Unknown (Nonparametric)

보통 Ranking 이용.

* p-value $= P(X \ge T)$

용어정리

- 표본공간 (Sample Space): 실험으로 부터 얻은 결과값들의 집합
- 실험 (Experiment): 데이터를 생성하는 모든 과정
- 확률변수 (Random Variable): 표본공간의 요소를 실수로 대응시키는 함수 (법칙)
- 확률함수 (Probability Function): 확률값을 생성하는 함수
- 확률분포 (Probability Distribution): 확률함수로 부터 나온 확률 값들의 패턴
- 통계량 (Statistic(s)): 샘플의 함수
- 추정량 (Estimator): 샘플의 함수 → 확률분포의 모수를 추정하는 목적
- 추정 (Estimation); 알려지지 않은 확률분포의 모수를 추정하는 프로세스
- 검정 (Hypothesis Testing): 알려지지 않은 확률분포의 모수를 검정하는 프로세스

EOD

Hypothesis Tests for a Single Mean μ when σ^2 is known

Two-Sided Test Problems

1. Hypothesis: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$

2. Significance level: α

3. Test statistic: $$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

4. Decision making: If $|Z| \geq Z_{\alpha/2}$ Reject H_0

Hypothesis Tests for a Single Mean μ when σ^2 is known

One-Sided Test Problems

1. Hypothesis: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$

2. Significance level: α

3. Test statistic: $$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

4. Decision making: If $Z \geq Z_\alpha$ Reject H_0

Hypothesis Tests for a Single Mean μ when σ^2 is known

A random sample of 100 recorded deaths in the US during the past year showed an average life span of 71.8 years. Assuming a population standard deviation of 8.9 years, does this seem to indicate that the mean life span today is greater than 70 years? Use a 0.05 level of significance.

$$H_0 : \mu = 70 \quad \text{vs.} \quad H_1 : \mu > 70$$

$$Z = \frac{71.8 - 70}{8.9/\sqrt{100}} = 2.02$$

$$\text{If} \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = 2.02 \geq 0.17 = Z_{0.05}, \ \text{Reject} \ H_0$$

The mean life span today is greater than 70 years.

Hypothesis Tests for a Single Mean μ when σ^2 is unknown

Two-Sided Test Problems

1. Hypothesis: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$

2. Significance level: α

3. Test statistic: $t^* = \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$ $\qquad t^* \sim t_{(n-1)}$

4. Decision making: If $|t^*| \geq t_{\alpha/2, n-1}$ Reject H_0

Hypothesis Tests for a Single Mean μ when σ^2 is known

One-Sided Test Problems

1. Hypothesis: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$

2. Significance level: α

3. Test statistic: $t^* = \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$ $\qquad t^* \sim t_{(n-1)}$

4. Decision making: If $t^* \geq t_{\alpha, n-1}$ Reject H_0

Hypothesis Tests for a Single Mean μ when σ^2 is known

An engine oil supposed to have a mean viscosity of 85. A sample of *n*=25 viscosity measurements resulted in a sample mean of 88.3 and a sample standard deviation of 7.49. Conduct hypothesis test to check whether a mean viscosity is 85 or not with a significance level of 0.05.

$$H_0 : \mu = 85 \quad \text{vs.} \quad H_1 : \mu \neq 85$$

$$t^* = \frac{88.3 - 85}{7.49 / \sqrt{25}} = 2.203$$

$$\text{Since} \quad t^* = 2.203 \geq t_{0.025,24} = 2.064, \ \text{Reject} \ \ \mathrm{H}_0$$

The mean viscosity of an engine oil is not 85.