Probability & Distributions

Contents

Before Estimating & Testing

Demographic table

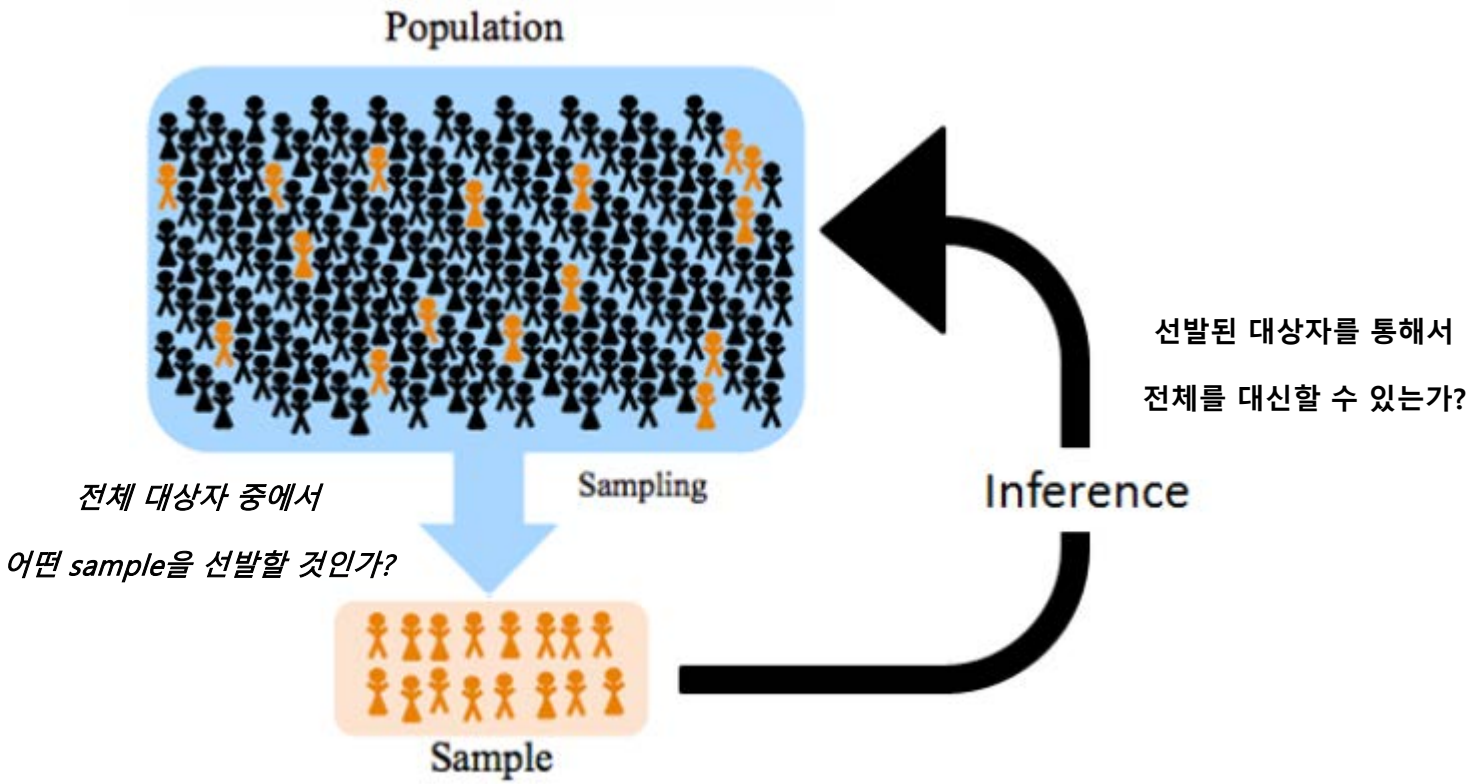
	Normal adrenal response (n = 34)	Deficient adrenal response (n = 11)	P
Age (years)*	63±12	57±15	0.25
Male/female	23/11	6/5	
Body-mass index (kg/m^2)	27.1	26.4	
Medication (n)			
Betablockers	18	4	
ACE-inhibitors	10	6	
Aspirin	15	8	
Oral anticoagulation	6	1	
Nitrates	8	3	
Calcium antagonists	6	5	
Systolic blood pressure (mm Hg)*	130±21	134±14	0.65
Diastolic blood pressure†	72 (65–80)	80 (76–83)	0.23
Potassium (mmol/l)*	3.9±0.4	3.8±0.6	0.54
Glucose (mmol/l)*	6.2±1.2	5.7±1.8	0.33
White-cell count (G/l)*	5.7±2.1	5.3±2.9	0.66
Eosinophils count (G/l)*	0.19±0.11 (3.3±1.9%)	0.23±0.20 (4.3±3.7%)	0.39
Basal cortisol concentration (nmol)†	387 (247–526)	178 (117–213)	0.001
Basal ACTH (pmol/l)*	9.2±4.4	6.5±5.3	0.19
Left ventricular ejection fraction (%)*	48±17	42±15	0.54

*Mean±S.D.; difference tested by Student's t-test; †median (interquartile range); difference tested by Mann-Whitney test.
G/l, the absolute number of eosinophils (=10^9/l), corresponding to the relative number (percentage) of 3.3±1.9% and 4.3±3.7%, respectively.

1. Basic concepts in probability

Basic concepts in probability

Beginning of statistics

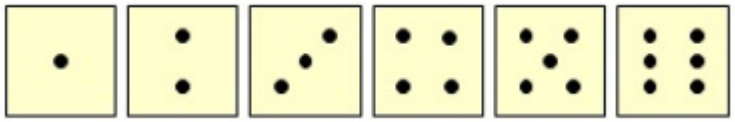


https://newonlinecourses.science.psu.edu/stat200/book/export/html/21

Basic concepts in probability

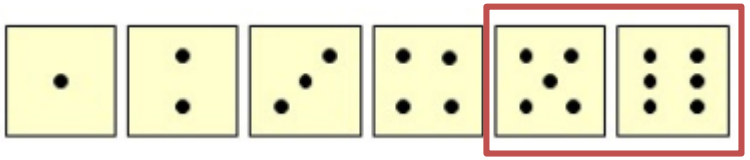
Definition of sample space and event

- 표본공간(sample space)
  - ✓ 확률실험을 실시할 때 나타날 수 있는 모든 결과들의 조합



$$S = \{1, 2, 3, 4, 5, 6\}$$

- 사상 또는 사건(event)
  - ✓ 표본공간을 구성하고 있는 원소들 중에서 관심 대상 원소들이 모여진 부분집합을 의미

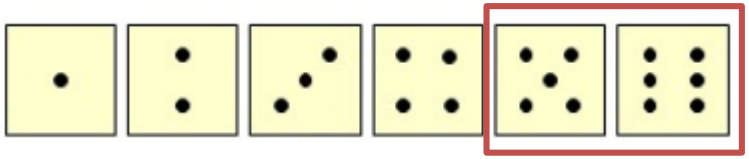


$$A = \{5, 6\}$$

Basic concepts in probability

Definition of probability

- 확률(probability)
  - ✓ 어떤 사건이 발생할 수 있는 가능성을 숫자로 표현. 표본공간의 모든 원소가 동일한 발생확률을 갖는다면 확률은 전체 원소의 개수 대비 사건이 발생된 비율을 의미



$$P = \frac{\#\ of\ \{5,6\}}{\#\ of\ \{1,2,3,4,5,6\}} = \frac{1}{3}$$

게임 내 확률

당첨 확률

$$_{45}C_6 = \frac{45!}{39! \times 6!}$$

$$= 8{,}145{,}060$$

Basic concepts in probability

Types of probability

실험적(experimental) 확률

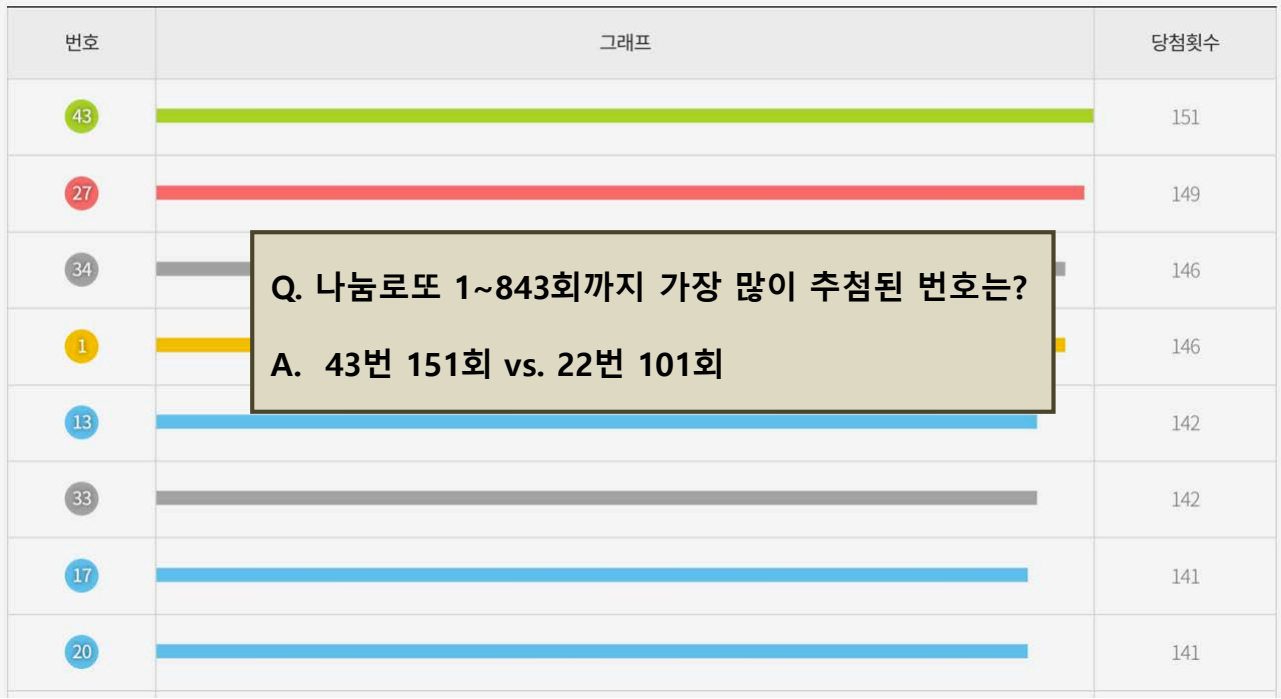
- 동전 던지기 10회 시행 시 앞면이 6번 나왔다.
- 실험적 확률은

$$P(head) = \frac{6}{10} \; and \; P(tail) = \frac{4}{10}$$

이론적(theoretical) 확률

- 동전 던지기 시행의 이론적 확률은

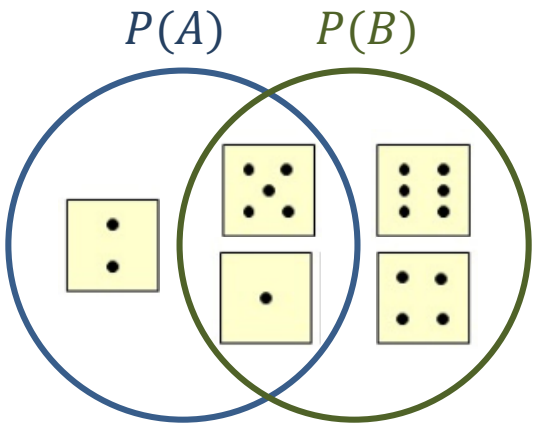
$$P(head) = P(tail) = \frac{5}{10}$$



Basic concepts in probability

Types of probability

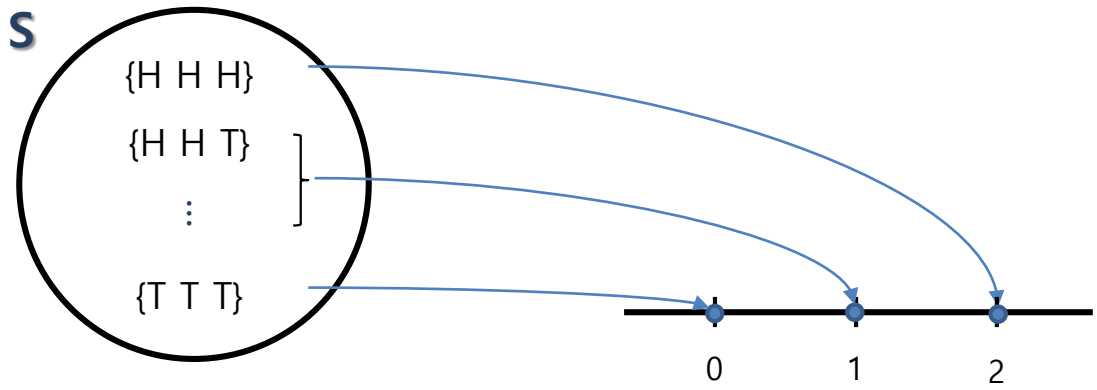
- 조건부 확률(conditional probability)
  - ✓ 어떤 사건 B가 발생했다는 조건 하에서 또 다른 사건 A가 발생할 확률로 정의한다.
  - ✓ $P(A|B) = \frac{P(A \cap B)}{P(B)}$, 단 $B \neq \phi$.



Basic concepts in probability

Random variable

- 확률변수(random variable)
  - ✓ 표본공간의 각 원소를 실수 값으로 바꾸는 함수로 확률분포를 가진다.



Basic concepts in probability

Random variable

- 확률변수(random variable)
  - ✓ 이산확률변수(discrete random variable)는 확률변수 X가 유한 개 또는 셀 수 있는 무한 개의 값을 가질 때 정의한다. 예를 들면 성별, Stage, 질병발생 유무 등
  - ✓ 연속확률변수(continuous random variable)는 표본공간을 직선 상의 임의의 구간으로 나타낼 수 있을 때 정의한다. 예를 들면 나이, lab 측정 수치 등
  - ✓ 연속확률변수 중 하나인 나이 변수를 40대 이하(=1), 5-60대(=2), 70대 이상(=3) 등으로 이산확률변수로 정의할 수 있다.

2. Basic concepts in distribution

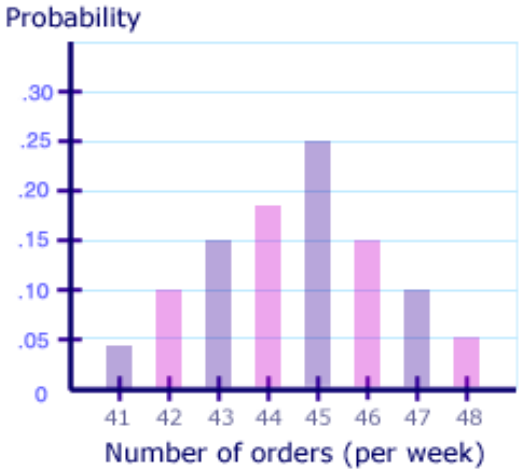
Basic concepts in distribution

Discrete probability distribution

- 이산확률분포(discrete probability distribution)
  - ✓ 이산적인(셀 수 있는) 값을 갖는 확률변수의 확률분포로서, 확률분포함수 f(x)는 확률변수 X가 'x'의 값을 가지는 확률인 P(X=x)를 의미하며 확률질량함수(probability mass function)라고도 함
  - ✓ 다음의 조건을 만족해야 함

    1) $\sum_x f(x) = 1$

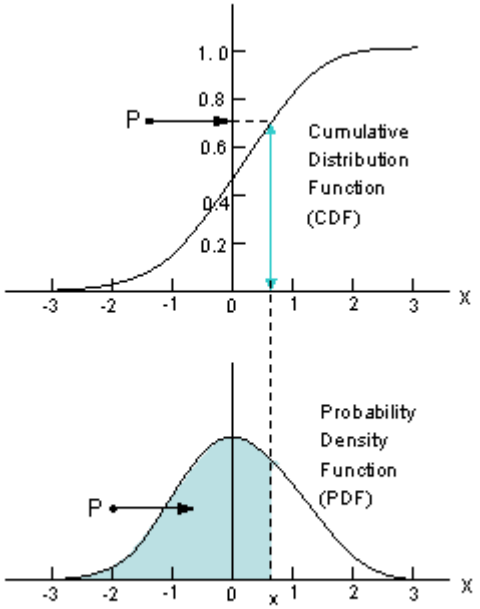
    2) $0 \leq f(x) \leq 1$
  - ✓ 확률 값은 언제나 0이상 1이하의 값을 가지며, 모든 가능한 결과에 대한 확률의 합은 1이어야 함



Basic concepts in distribution

Continuous probability distribution

- 연속확률분포(continuous probability distribution)
  - ✓ 연속적인(셀 수 없는) 값을 갖는 확률변수의 확률분포로서, 확률분포함수 f(x)는 확률 $P(a < X < b) = \int_a^b f(x)dx$를 구하기 위한 확률밀도함수(probability density function)를 의미함
  - ✓ 다음의 조건을 만족해야 함
    1) $\int_{-\infty}^{\infty} f(x) = 1,$
    2) $f(x) \geq 0$
  - ✓ 모든 x에 대해서 언제나 0 이상의 값을 가지며, 아래 면적의 합은 1이다.
  - ✓ 누적분포함수(cumulative distribution function): 확률변수 X가 특정한 값 x 이하일 확률로 $F(x) = P(X \leq x)$로 정의

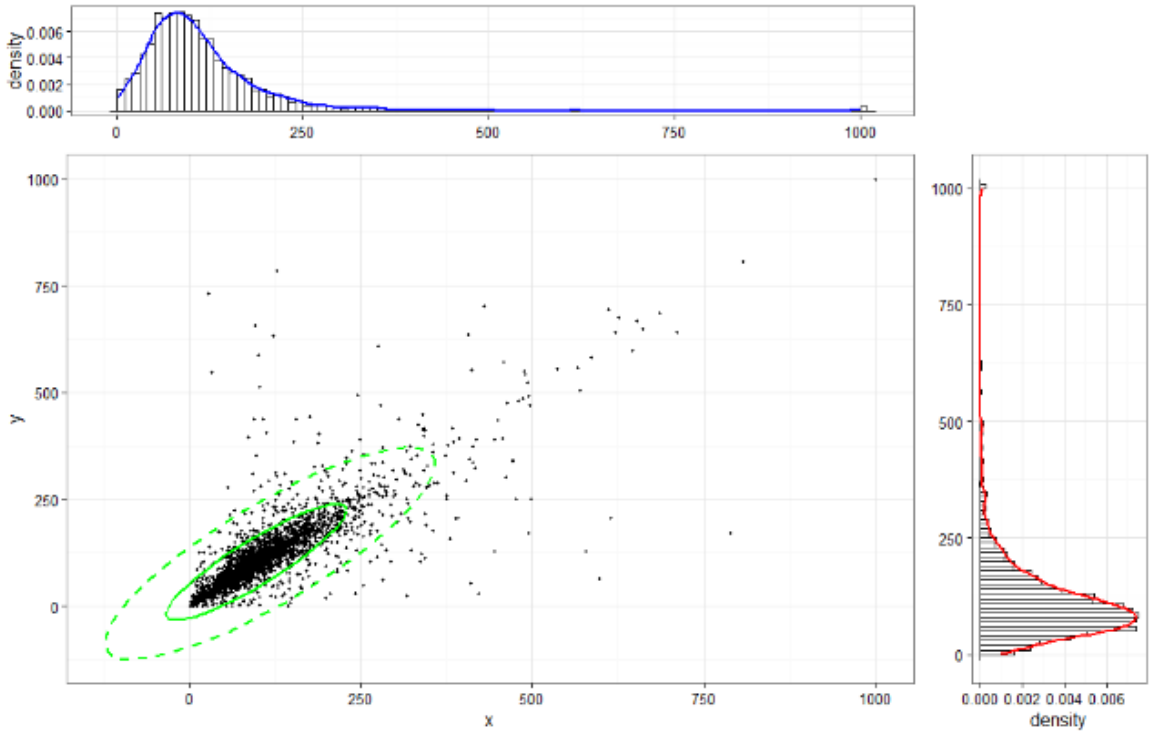


Relations Between Two Different Typical Representations of a Population

Basic concepts in distribution

Joint probability distribution

- 결합확률분포(joint probability distribution)
  - ✓ 2개 이상의 확률변수에 대한 확률분포를 의미하며, 확률변수 X와 Y의 결합확률분포는 f(x,y)로 나타내며 이산확률 변수의 경우에는 $f(x,y) = P(X = x, Y = y)$가 된다.



Basic concepts in distribution

Joint probability distribution

- 주변확률분포(marginal probability distribution)
  - ✓ 주변확률분포는 결합확률분포로부터 각각의 확률변수 X, Y에 대한 분포로 표현할 때 주변확률분포 라고 정의한다.
    - 이산형: $f_X(x) = \sum_y f(x,y)\,, f_Y(y) = \sum_x f(x,y)$
    - 연속형: $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy\,, f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$

- 통계적 독립(statistically independent)
  - ✓ 확률변수 X와 Y가 통계적으로 독립이기 위한 필요충분조건은 결합확률함수가 $f(x,y) = g(x)h(y)$을 만족할 때 확률 변수 X와 Y는 통계적으로 독립이다. 여기서, $g(x)$는 x만의 함수이고 $h(y)$는 y만의 함수이다.

Basic concepts in distribution

Examples

		질병발생여부		
		발생O	발생X	
복용 여부	Yes	12	16	28
	No	8	4	12
		20	20	40

Joint probability

Marginal probability

$$P(A) = P(\text{복용}Yes) = \frac{28}{40} = 0.7$$

$$P(A \cap B) = \frac{12}{40} = 0.3$$

$$P(B) = P(\text{질병발생}O) = \frac{20}{40} = 0.5$$

Conditional probability

$$P(B|A) = P(\text{질병발생}O \mid \text{복용}Yes)$$

$$= \frac{P(A \cap B)}{P(A)} = \frac{12/40}{28/40} = 0.43$$

→ 복용을 한 대상자들의 질병발생 확률이 0.43으로 더 낮다!

Statistically independent

$$P(A \cap B) = 0.3 \neq 0.35 = P(A) \times P(B)$$

→ 변수 A와 B는 독립이 아니다!

Basic concepts in distribution

Simpson's paradox

- A와 B 중 어떤 treatment가 효과가 좋은가?

Treatment A

		Successful		
		Yes	No	
Stone size	Small	81	6	87
	Large	192	71	263
		273	77	350

Treatment B

		Successful		
		Yes	No	
Stone size	Small	234	36	270
	Large	55	25	80
		289	61	350

$$P(A\text{치료 성공}) = \frac{273}{350} = 0.78 \quad < \quad P(B\text{치료 성공}) = \frac{289}{350} = 0.826$$

$$P(A\text{치료 성공}|Small) = \frac{81}{87} = 0.931 \quad > \quad P(B\text{치료 성공}|Small) = \frac{234}{270} = 0.867$$

$$P(A\text{치료 성공}|Large) = \frac{192}{263} = 0.73 \quad > \quad P(B\text{치료 성공}|Large) = \frac{55}{80} = 0.688$$

→ 어떤 treatment가 더 효과적이라고 결론 내릴 수 있는가?

Basic concepts in distribution

Simpson's paradox

- A와 B 중 어떤 treatment가 효과가 좋은가?

Treatment A

		Successful		
		Yes	No	
Stone size	Small	81	6	87
	Large	192	71	263
		273	77	350

Treatment B

		Successful		
		Yes	No	
Stone size	Small	234	36	270
	Large	55	25	80
		289	61	350

$$P(A\text{치료 성공}) = P(A\text{치료 성공}|Small)P(Small) + P(A\text{치료 성공}|Large)P(Large)$$
$$= 0.931 \times 0.249 + 0.73 \times 0.751$$
$$= 0.78$$

$$P(B\text{치료 성공}) = P(B\text{치료 성공}|Small)P(Small) + P(B\text{치료 성공}|Large)P(Large)$$
$$= 0.867 \times 0.771 + 0.688 \times 0.229$$
$$= 0.826$$

→ 변수의 범주가 서로 다른 비율의 양상을 보인다면 역설적인 현상이 발생됨!

Basic concepts in distribution

Bayes theorem

베이즈 정리(Bayes theorem)

표본공간 S를 공사상이 아닌 사상 $B_1, B_2, \cdots, B_k$들로 분할하면, 공사상이 아닌 $(P(A) \neq 0)$ 임의의 사상 A에 대하여 아래의 식이 성립한다.

$$P(B_r|A) = \frac{P(B_r \cap r)}{\sum_{i=1}^{k} P(B_i \cap r)} = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)}$$

Basic concepts in distribution

Distributions

연속형 확률변수의 분포

이산형 확률변수의 분포

평균 비교

비율 비교

X	X Measures	f(x)	Values of X	E(x)	V(x)
Continuous uniform	Outcomes with equal density (continuous)	$\frac{1}{b-a}$	$a \le x \le b$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$
Exponential	Time between events; time until an event	$\lambda e^{-\lambda x}$	$x \ge 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal	Values with a bell-shaped distribution (continuous)	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$-\infty < x < \infty$	μ	σ
Standard normal (Z)	Standard scores	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$	$Z = \frac{x-\mu}{\sigma}$	0	1
Binomial approximation	Number of successes in large number of trials	Approx. normal if $np \ge 5$ and $n(1-p) \ge 5$ by CLT	$Z = \frac{x-np}{\sqrt{np(1-p)}}$	np	$np(1-p)$
Poisson approximation	Number of occurrences in a fixed time period (large average)	Approx. normal if $\lambda > 30$	$z = \frac{x-\lambda}{\sqrt{\lambda}}$	λ	λ
$\bar{X}$	Average of $x_1, x_2, \ldots, x_n$	Exactly normal if x is normal. Approx. normal if $n \ge 30$ by CLT	$Z = \frac{\bar{x}-\mu_x}{\sigma_x/\sqrt{n}}$	μ_x	$\frac{\sigma_x^2}{n}$
$\hat{p}$	Proportion or percentage of successes in binomial with $np \ge 5$, $n(1-p) \ge 5$	Approx. normal if $np \ge 5$ and $n(1-p) \ge 5$ by CLT	$Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$	p	$\frac{p(1-p)}{n}$

Basic concepts in distribution

Distributions

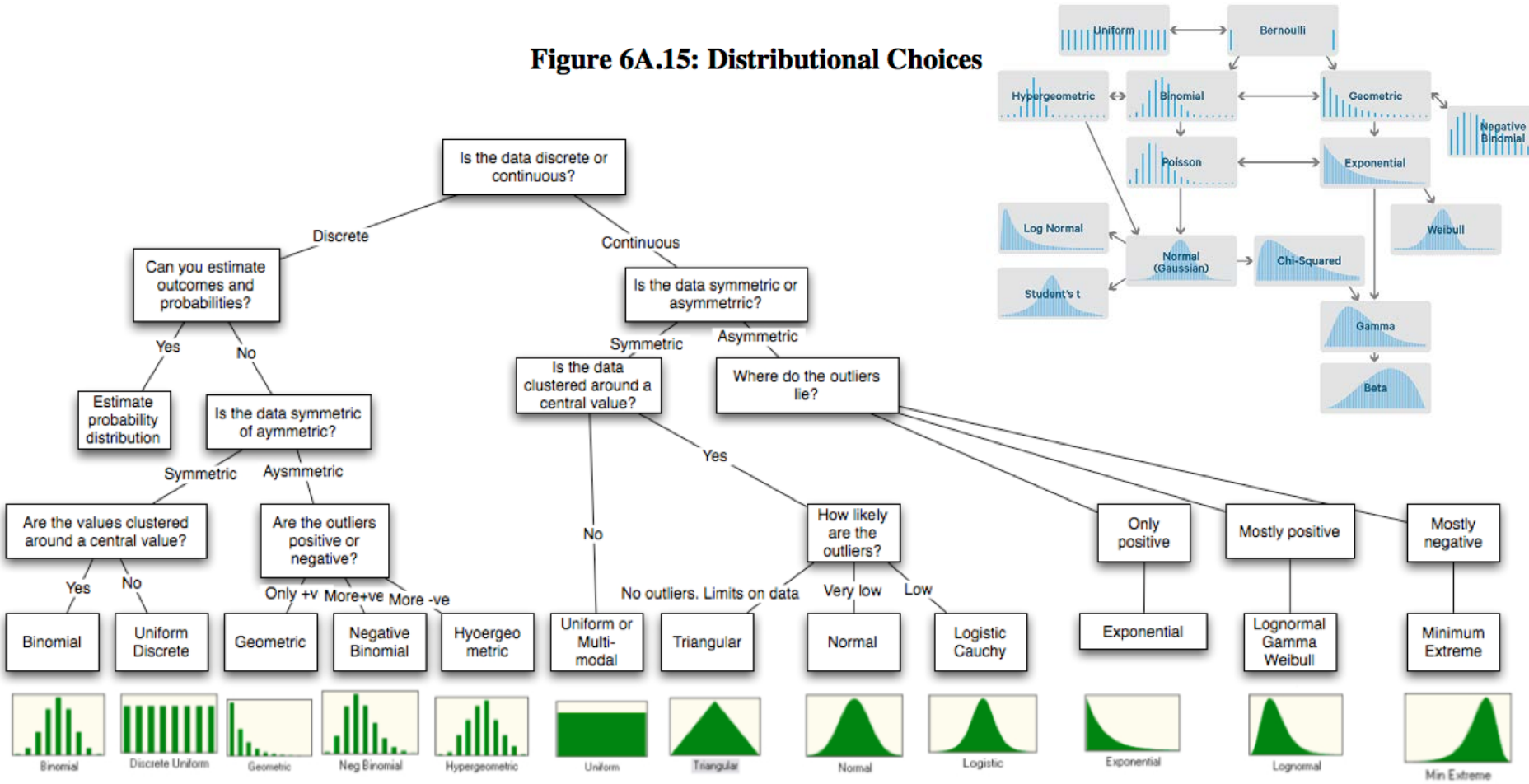


Figure 6A.15 from Aswath Damondaran

3. Normal distribution and related distributions

- Normal distribution
- t-distribution
- Chi-square distribution
- F-distribution

Normal distribution & related distributions

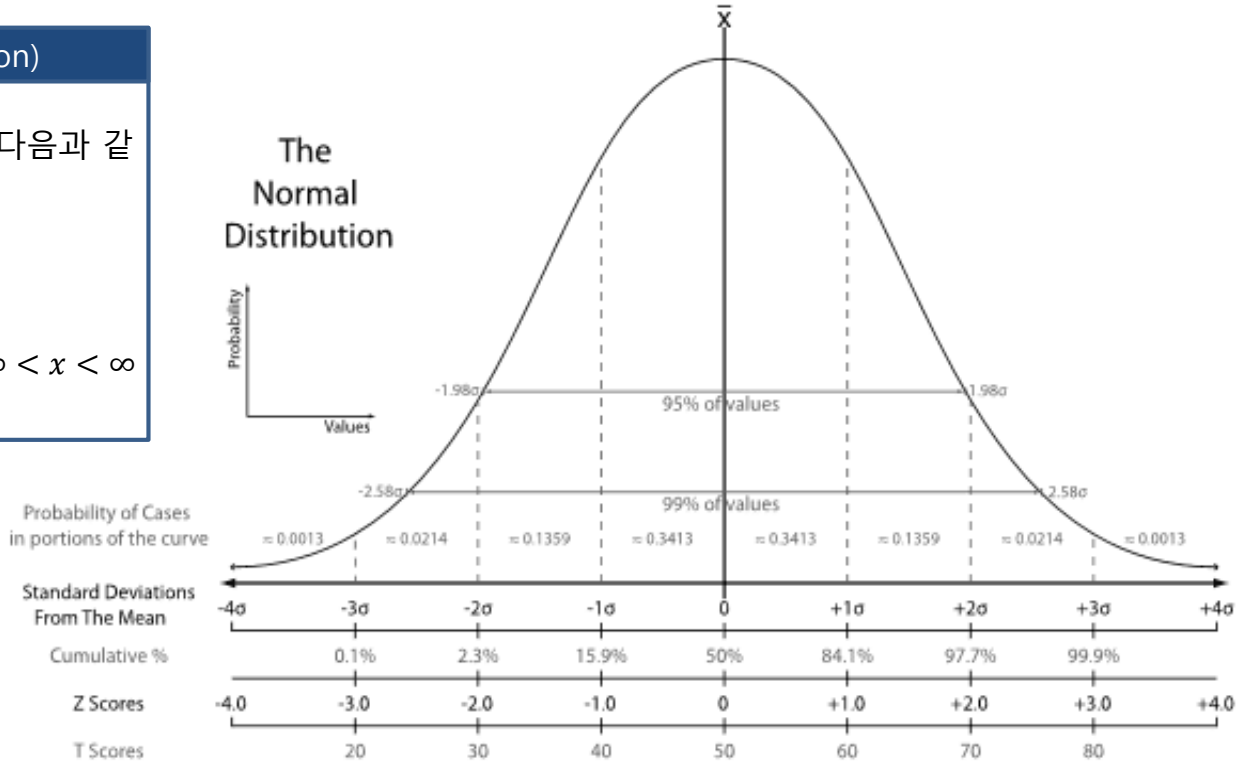
Normal distribution

- 정규분포(normal distribution)
  - ✓ 통계분석에서 가장 많이 사용되며, 가우스(Gauss) 분포라고 불림
  - ✓ 기댓값을 중심으로 대칭이며 산포는 표준편차에 의해 엎어놓은 종 모양의 형태가 결정됨
  - ✓ 기댓값이 0이고 표준편차는 1인 정규분포를 표준(standard)정규분포라고 정의

정규분포(normal distribution)

정규분포는 확률변수 X에 대해서 다음과 같은 확률분포함수를 따른다.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], -\infty < x < \infty$$



Investopedia

Normal distribution & related distributions

Normal distribution

- 정규분포(normal distribution)
  - ✓ 많은 통계적 방법론에서 정규분포를 가정
  - ✓ 실제로는 모집단의 경우에만 국한하여 사용하며, 표본 수 산출 시 활용

정규분포의 선형변환

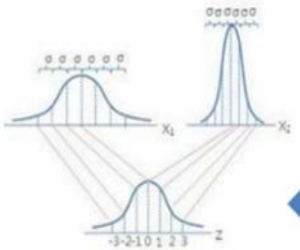
$X \sim N(\mu, \sigma^2)$이고 $Y = a + bX$라 하면

$Y \sim N(a + b\mu, b^2\sigma^2)$도 정규분포를 따른다.

정규분포의 표준화(standardization)

$X \sim N(\mu, \sigma^2)$이고 $Z = (X - \mu)/\sigma$라 하면

$Z \sim N(0, 1)$인 정규분포를 따르고, 이를 표준정규분포라고 한다.



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

이항분포의 정규근사

$X \sim B(n, p)$이면 $Z \equiv \frac{X - np}{\sqrt{np(1-P)}} \sim N(0,1)$ 인

표준정규분포를 따른다.

정규분포의 가법성

$X \sim N(\mu_1, {\sigma_1}^2)$이고 $Y \sim N(\mu_2, {\sigma_2}^2)$ 이고, X와

Y가 독립이면 $X + Y \sim N(\mu_1 + \mu_2, {\sigma_1}^2 + {\sigma_2}^2)$

인 정규분포를 따른다.

Normal distribution & related distributions

Normal distribution

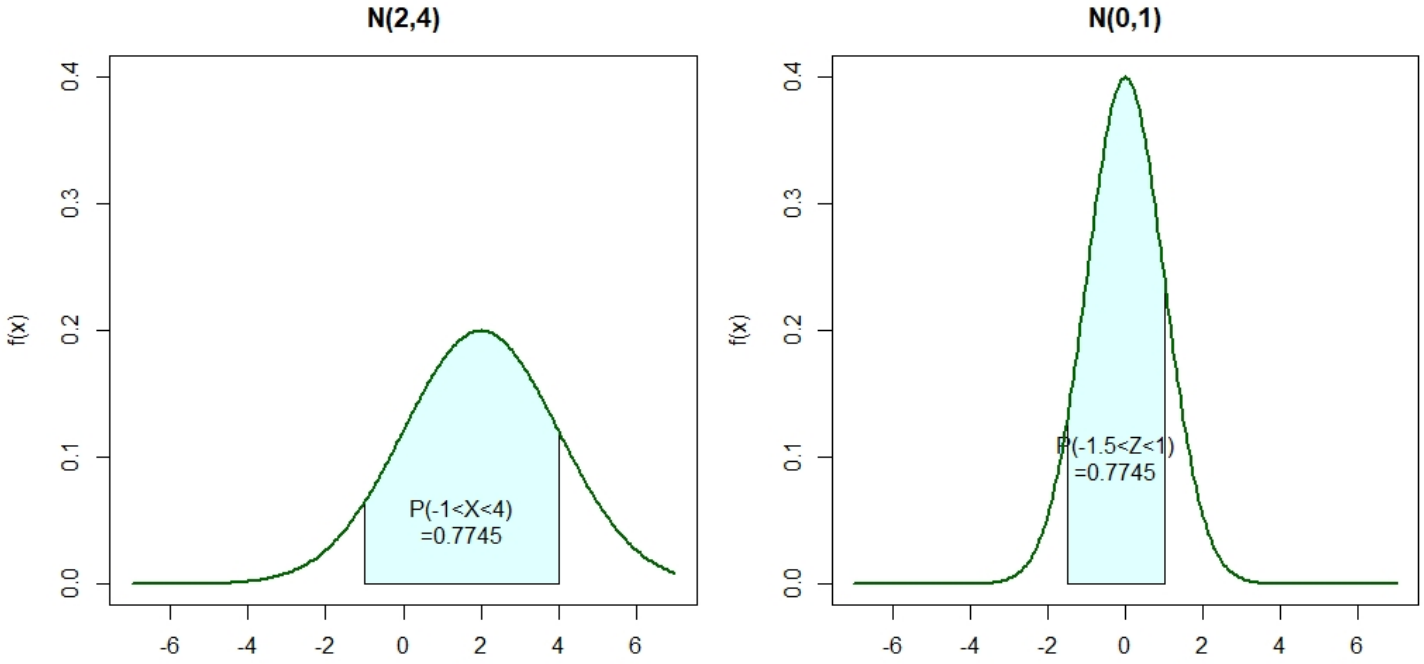
- 확률변수 X가 $N(2, 2^2)$를 따를 때, $P(-1 < X < 4)$를 구하시오. – *R 실습!*

(풀이)

$$P(-1 < X < 4) = P\left(\frac{-1-2}{2} < \frac{X-2}{2} < \frac{4-2}{2}\right) = P(-1.5 < X < 1)$$

정규분포표로부터 $P(Z < 1) \doteq 0.8413, P(Z < 1.5) \doteq 0.9332$ 이므로,

$\Rightarrow P(-1.5 < X < 1) \doteq 0.8413 - (1 - 0.9332) = 0.7745$



Normal distribution & related distributions

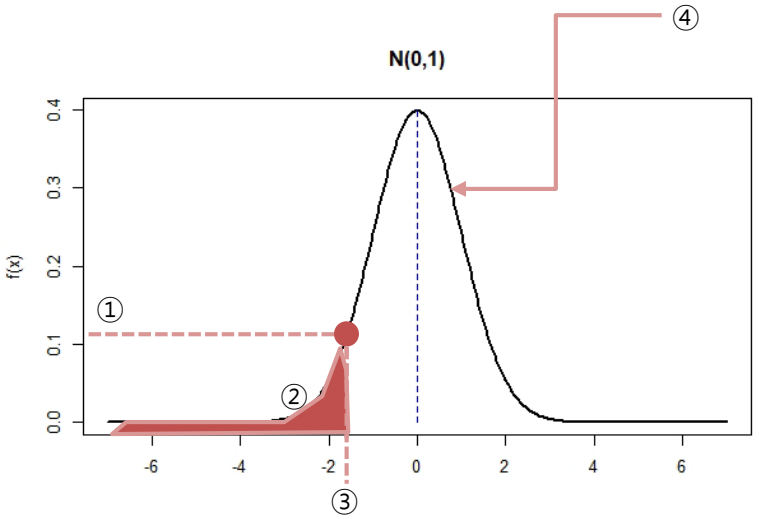
Normal distribution

확률밀도함수 f(x)

> dnorm(x, mean, sd) # mean과 sd를 생략하면 표준정규분포로 계산함(mean=0, sd=1)

누적확률분포 F(x)

> pnorm(x, mean, sd, lower.tail=TRUE)



분위수

> qnorm(p, mean, sd, lower.tail=TRUE)

정규 확률변수 (n= 난수의 개수)

> rnorm(n, mean, sd)

동일한 결과를 얻기 위해서 set.seed(난수#)를 설정 후 위의 함수들을 실행함

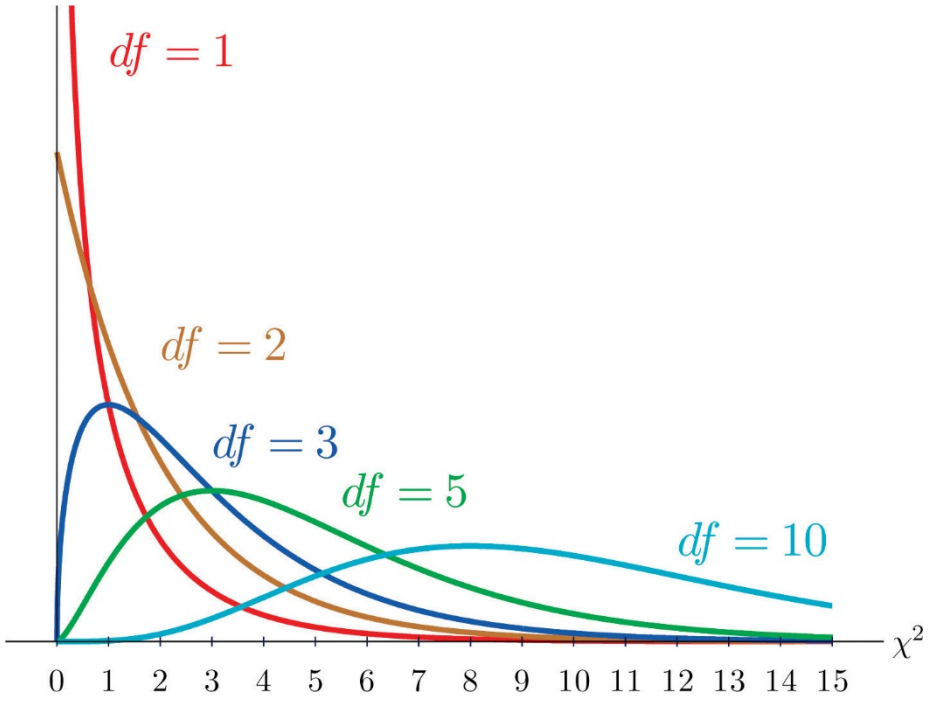
Normal distribution & related distributions

Chi-square distribution

- 카이제곱분포(χ^2; chi-square distribution)

카이제곱분포(chi-square distribution)

표준정규모집단으로부터 추출된 확률표본을 $Z_1, Z_2, \cdots, Z_n$이라 하면, $\sum_{i=1}^{n} {Z_i}^2$는 자유도가 n인 카이제곱분포를 따른다.



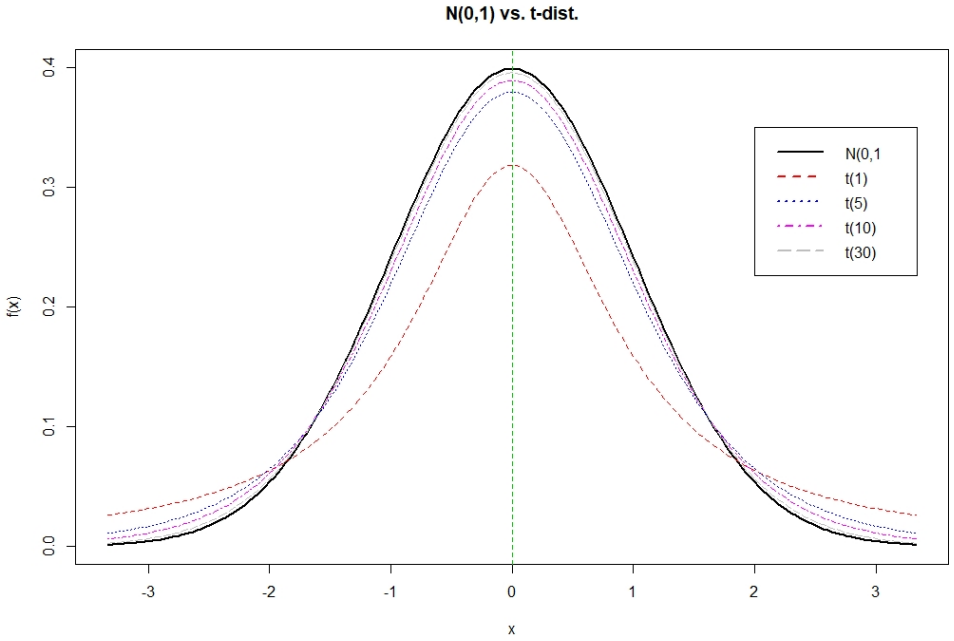
Normal distribution & related distributions

t-distribution

- t-분포
  - ✓ Student에 의해 고안되어 Student t-distribution 이라고도 부름
  - ✓ 표준정규분포와 유사한 형태를 지니고, 자유도가 커질수록 표준정규분포에 가까워진다.

t-분포(t-distribution)

표준정규분포를 따른 확률변수를 Z라 정의하고 Z와는 독립적으로 자유도 ν인 카이제곱(χ^2) 분포를 따르는 확률변수를 Y라 하면, 확률변수 $T \equiv \frac{Z}{\sqrt{Y/\nu}}$는 자유도 ν인 t-분포를 따른다.



Normal distribution & related distributions

Chi−square & *t*-distribution

Table 1. **Baseline Demographic and Clinical Characteristics of the Patients.***

Characteristic	CPAP (N=307)	Intubation (N=303)	P Value†
Gestational age (wk)	26.91±1.0	26.87±1.0	0.63
Gestational age of 25 or 26 wk (%)	33	35	0.59
Birth weight (g)	964±212	952±217	0.48
Use of antenatal corticosteroids (%)	94	94	0.76
Cesarean section (%)	66	69	0.51
Mother in labor (%)	65	66	0.82
Rupture of membranes (days before birth)			
Median	0	0	0.65
Interquartile range	0–2	0–1	
Male sex (%)	49	56	0.05
Multiple births (%)	35	32	0.57
Resuscitation device used (%)‡			0.13
None	19.5	13.9	
Self-inflating bag	14.7	16.5	
Self-inflating bag plus CPAP	13.0	14.2	
Flow-inflating bag	5.2	6.6	
Neopuff or bubble CPAP	47.2	46.5	
Apgar score at 5 minutes			
Median	9	8	0.001
Interquartile range	8–9	8–9	

* Plus–minus values are means ±SD. CPAP denotes continuous positive airway pressure.
† P values were calculated by the t-test, the chi-square test, or the Mann–Whitney test.
‡ In this category, eight infants (1.3%) were excluded because the resuscitation method was classified as "other."

Normal distribution & related distributions

Chi−square & *t*-distribution

	Normal adrenal response (n = 34)	Deficient adrenal response (n = 11)	P
Age (years)*	63±12	57±15	0.25
Male/female	23/11	6/5	
Body-mass index (kg/m^2)	27.1	26.4	
Medication (n)			
Betablockers	18	4	
ACE-inhibitors	10	6	
Aspirin	15	8	
Oral anticoagulation	6	1	
Nitrates	8	3	
Calcium antagonists	6	5	
Systolic blood pressure (mm Hg)*	130±21	134±14	0.65
Diastolic blood pressure†	72 (65–80)	80 (76–83)	0.23
Potassium (mmol/l)*	3.9±0.4	3.8±0.6	0.54
Glucose (mmol/l)*	6.2±1.2	5.7±1.8	0.33
White-cell count (G/l)*	5.7±2.1	5.3±2.9	0.66
Eosinophils count (G/l)*	0.19±0.11 (3.3±1.9%)	0.23±0.20 (4.3±3.7%)	0.39
Basal cortisol concentration (nmol)†	387 (247–526)	178 (117–213)	0.001
Basal ACTH (pmol/l)*	9.2±4.4	6.5±5.3	0.19
Left ventricular ejection fraction (%)*	48±17	42±15	0.54

*Mean±S.D.; difference tested by Student's *t*-test; †median (interquartile range); difference tested by Mann-Whitney test.
G/l, the absolute number of eosinophils (=10^9/l), corresponding to the relative number (percentage) of 3.3±1.9% and 4.3±3.7%, respectively.

Normal distribution & related distributions

F-distribution

- F-분포
  - ✓ ANOVA 분석 시 활용되며, 모형 평가의 척도로도 활용된다.

F-분포(F-distribution)

카이제곱분포를 따르며 독립적인 두 개의 확률변수를 각각의 자유도로 나누어 비율을 취하면 F-분포를 따른다. 즉, 자유도 ν_1 과 ν_2인 카이제곱분포를 따르며 독립적인 확률변수를 각각 U와 V라 하면 다음과 같이 카이제곱을 따른다.

$$F \equiv \frac{\frac{U}{\nu_1}}{\frac{V}{\nu_2}} \sim F(\nu_1, \nu_2)$$

Normal distribution & related distributions

중심극한정리

- 중심극한정리
  - ✓ 실제 자료에서 모집단의 분포가 정규분포를 완전히 따르는 경우는 드물다.
  - ✓ 통계적 추론을 위해서 정규분포를 가정해야 하는 상황들이 빈번하다.
  - ✓ 중심극한정리란 모집단으로부터 충분히 큰 수의 확률표본을 추출한다면 표본평균의 분포는 모집단의 본포와는 상관없이 근사적으로 정규분포를 따른다는 것이다.
  - ✓ 즉, 표본의 개수가 충분히 크다면 모집단이 정규분포를 따르지 않는다 하더라도 표본평균을 통한 추론 시 표준정규분포를 사용할 수 있다.

중심극한정리(central limit theorem)

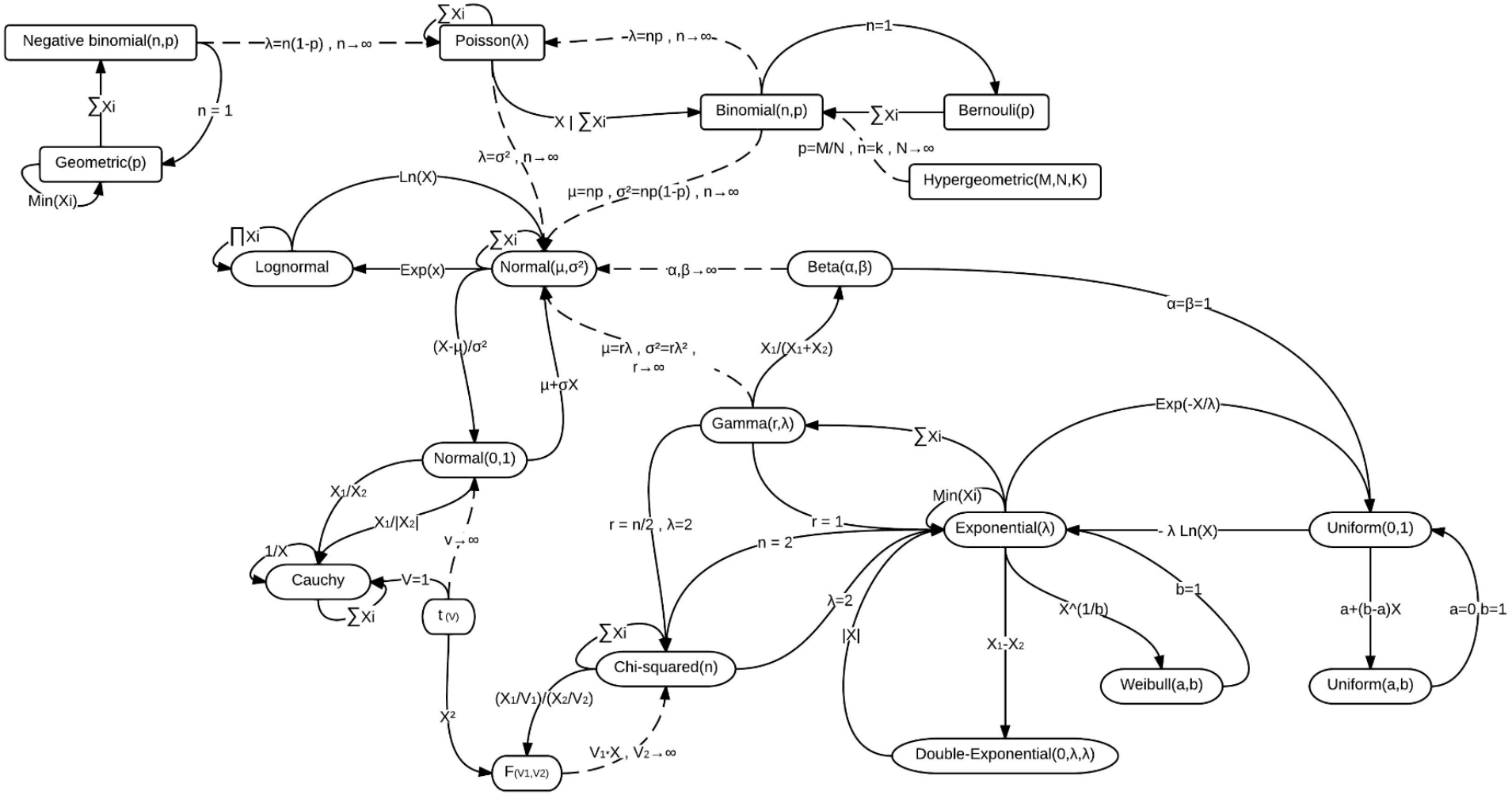
기댓값이 μ이고 분산이 σ^2인 정규모집단의 확률표본을 $X_1, X_2, \cdots, X_n$이라 할 때, 표본의 크기 n이 충분히 크다면 확률표본의 합, 혹은 표본평균의 분포는 근사적으로 다음과 같이 정규분포를 따른다.

$$S_n = X_1 + X_2 + \cdots + X_n \sim N(n\mu, n\sigma^2)$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$$

Normal distribution & related distributions

Relationship between distributions



Wikimedia Commons

Thank you!