Data Science Basic

2018 Fall (October-January)

(I) A rudimentary knowledge of multivariate analysis (Obata)

- 10/5 (5) Probability distributions
- 10/19 (4) Statistical inference
- 10/19 (5) Normal linear models

(II) Introduction to computability theory (Uramoto)

- 10/26 (4) Computability and its hierarchy
- 11/9 (4) Formal languages and automata
- 11/16 (4) Higher hierarchies

(III) Graph theory (Irie)

- 11/30 (4) Basics of graph theory
- 12/7 (4) Graph analysis
- 12/7 (5) Graph search algorithms

(IV) Geographical information system and complex networks (Fujiwara)

- 12/14 (4) Geographical information system and geographical information science
- 12/21 (4) Geospatial data analysis
- 1/11 (4) Complex networks and geographical networks

(V) Data classification and visualization (Nishi)

- 1/18 (4) Principal component analysis
- 1/25 (4) Factor analysis
- 2/1 (4) Clustering

A Rudimentary Knowledge of Multivariate Analysis

Nobuaki Obata

www.math.is.tohoku.ac.jp/~obata

October 5, 19, 2018

Contents of Lectures

Lecture 1. Probability Distributions

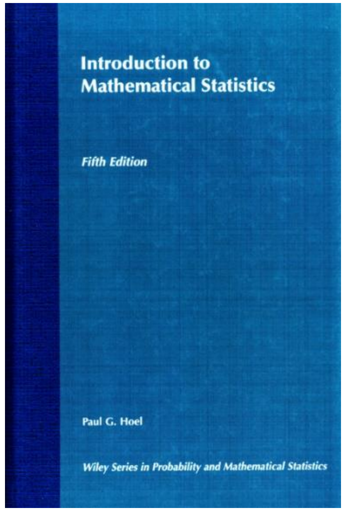
[Hoel] Chaps 2-3, 6 [Dobson] Chap 1

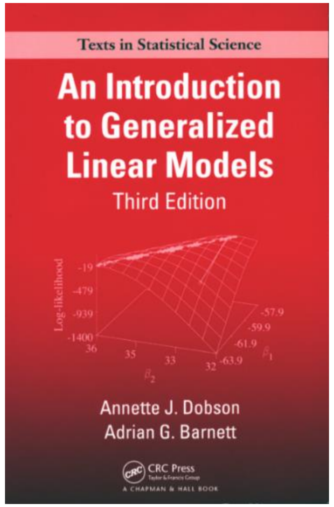
Lecture 2. Statistical Inference

[Hoel] Chaps 4-5, 8 [Dobson] Chaps 4-5

Lecture 3. Normal Linear Models

[Hoel] Chaps 6-7 [Dobson] Chap 6





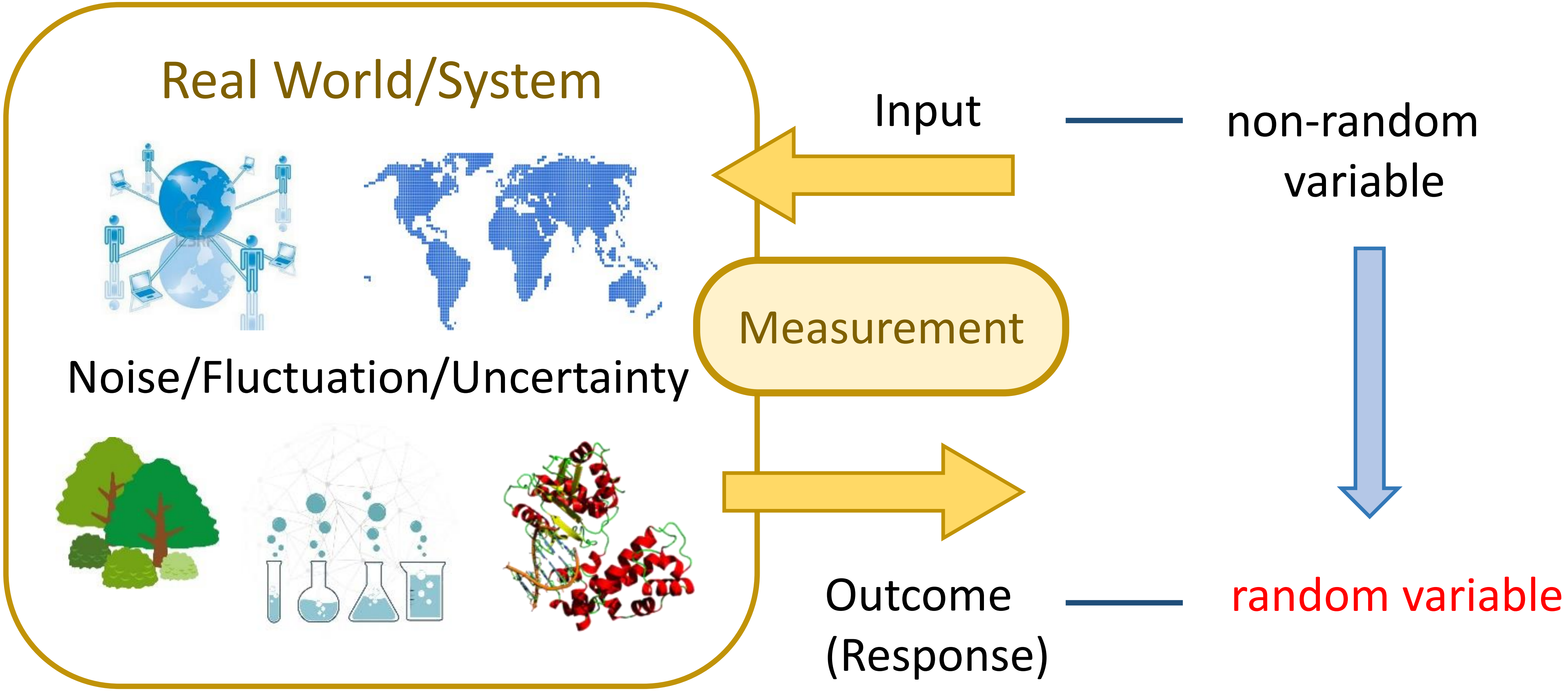
[1] P. G. Hoel: Introduction to Mathematical Statistics, 5th Ed. Wiley, 1984. [Japanese translation available for 4th Edition]

[2] A. J. Dobson and A. G. Barnett: An Introduction to Generalized Linear Models, 3rd Ed. CRC Press, 2008. [Japanese translation available for 2nd Edition]

Lecture 1

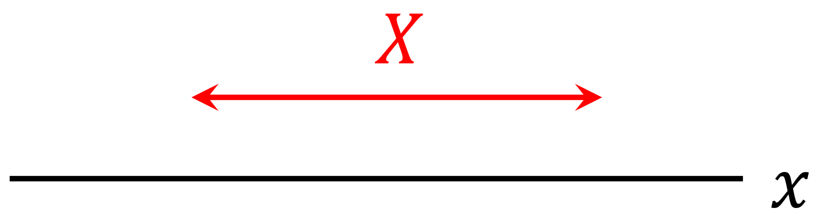
Probability Distributions

1. Statistical analysis

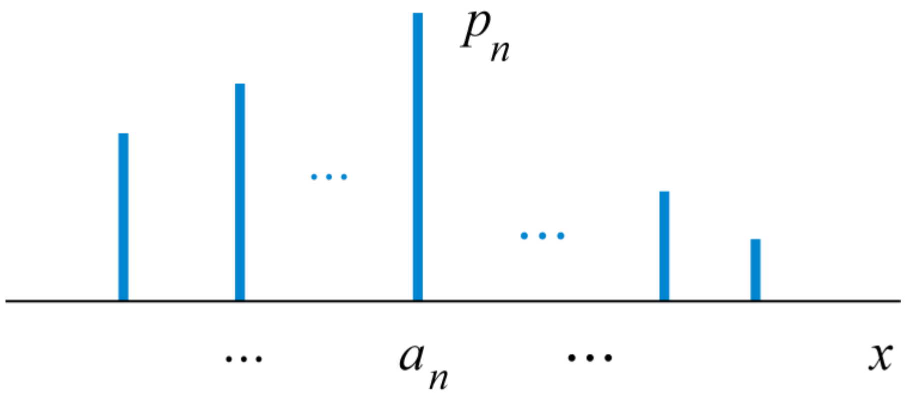


2. Random variables

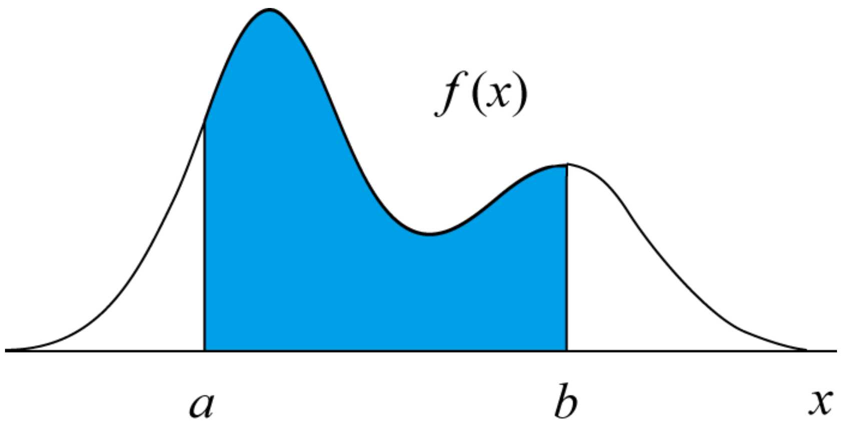
A *random variable X* varies over a domain in the real line with certain tendency (probability) of occurrence of its values.



- discrete random variable



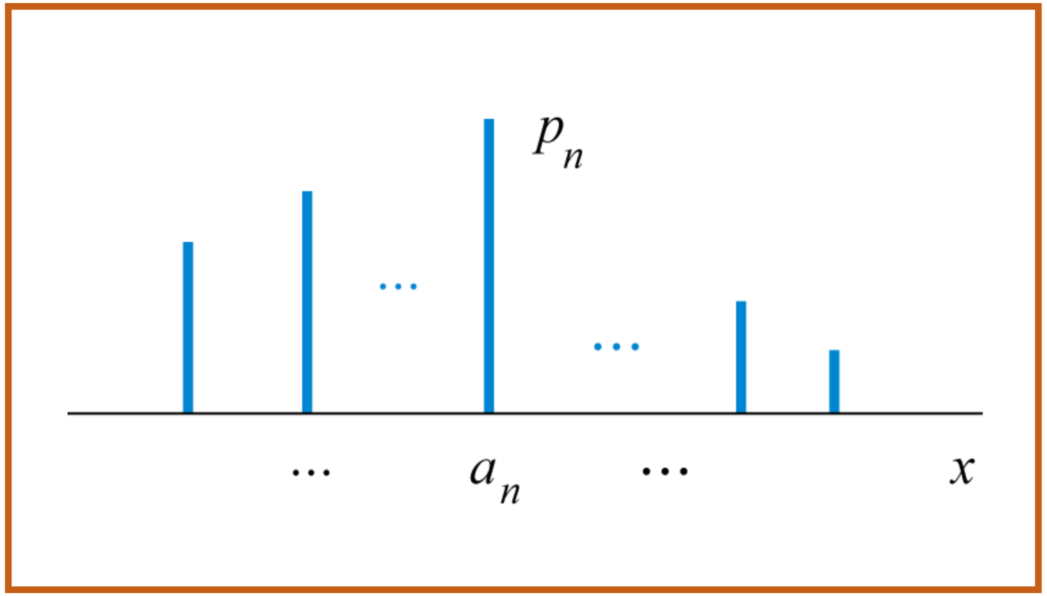
- continuous random variable



3. Probability distributions I: Discrete case

- range of values $\{a_1, a_2, \ldots, a_n, \ldots\}$
- distribution by a sum of point masses

$$P(X = a_n) = p_n, \quad p_n \geq 0, \quad \sum_n p_n = 1$$



- mean value

$$m = m_X = \mathbf{E}[X] = \sum_n a_n \, p_n$$

- variance

$$\sigma^2 = \sigma_X^2 = \mathbf{V}[X] = \mathbf{E}[(X-m)^2] = \sum_n (a_n - m)^2 \, p_n$$

$$= \mathbf{E}[X^2] - m^2 = \sum_n {a_n}^2 \, p_n - m^2$$

3. Probability distributions II: Continuous case

- range of values $I \subset \mathbf{R} = (-\infty, +\infty)$
- distribution by a density function

$$f_X(x) \geq 0, \qquad \int_{-\infty}^{+\infty} f_X(x)dx = 1$$

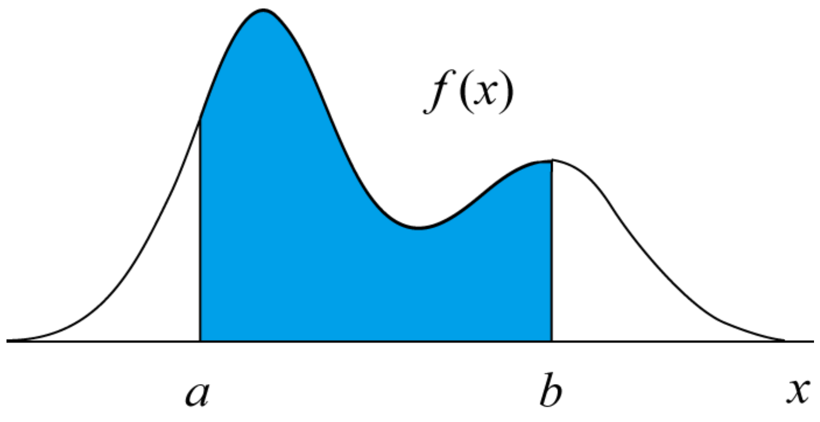
- mean value

$$m = m_X = \mathbf{E}[X] = \int_{-\infty}^{+\infty} x f_X(x)dx$$

- variance

$$\sigma^2 = \sigma_X^2 = \mathbf{V}[X] = \mathbf{E}[(X-m)^2] = \int_{-\infty}^{+\infty} (x-m)^2 f_X(x)dx$$

$$= \mathbf{E}[X^2] - m^2 = \int_{-\infty}^{+\infty} x^2 f_X(x)dx - m^2$$



$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

$$f_X(x) = \frac{d}{dx} P(X \leq x)$$

3. Probability distributions: A list

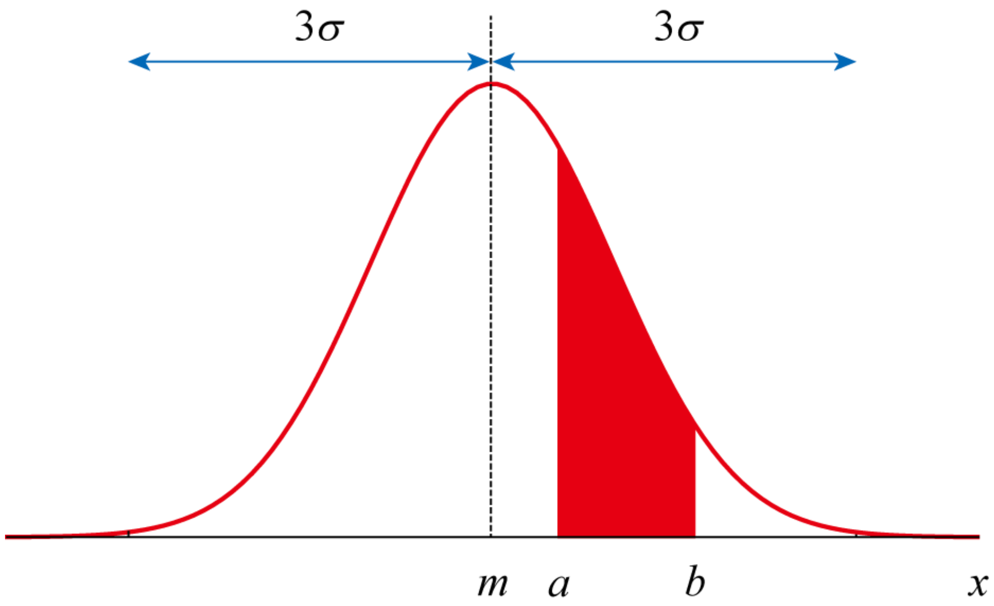
Discrete distributions	**mean**	**variance**
binomial distribution $B(n,p)$	np	$np(1-p)$
Bernoulli distribution $B(1,p)$	p	$p(1-p)$
geometric distribution with parameter p	$1/p$	$1/p^2$
Poisson distribution with parameter λ Po(λ)	λ	λ

Continuous distributions	**mean**	**variance**
uniform distribution on $[a,b]$	$(a+b)/2$	$(b-a)^2/12$
exponential distribution with parameter λ	$1/\lambda$	$1/\lambda^2$
normal (or Gaussian) distribution $N(m,\sigma^2)$	m	σ^2
chi-square distribution χ_n^2	n	$2n$
t-distribution t_n	0	$n/(n-2)$
F-distribution $F(m,n) = F_n^m$	$n/(n-2)$	$2n^2(m+n-2)$ $/m(n-2)^2(n-4)$

4. Normal distribution $N(m, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}$$

- mean value $$m = \int_{-\infty}^{+\infty} x f(x) dx$$
- variance $$\sigma^2 = \int_{-\infty}^{+\infty} (x-m)^2 f(x) dx$$



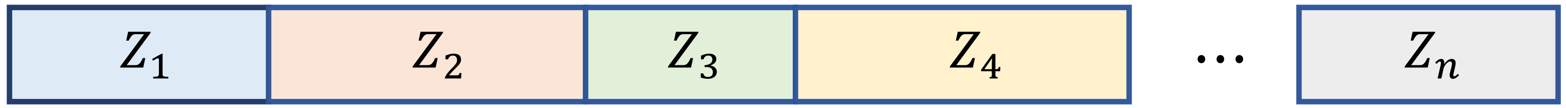
5. Central Limit Theorem (CLT)

$Z_1, Z_2, \cdots, Z_n, \cdots$: independent identically distributed (iid) random variables

mean value $= m$, variance $= \sigma^2$

Consider the sum:

$$S_n = Z_1 + Z_2 + \cdots + Z_n = \sum_{k=1}^{n} Z_k$$



mean value

$$\mathbf{E}[S_n] = \sum_{k=1}^{n} \mathbf{E}[Z_k] = mn$$

CLT $S_n - mn \sim N(0, n\sigma^2)$ holds as $n \to \infty$

- Accumulation of small fluctuation gives rise to a normal distribution.

6. An example of data

Two sets of numerical data are shown in the following table.

- dried weight of plants grown under two conditions

Is there any significant difference?

Details Later

A	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.98	4.90	5.75
	5.36	3.48	4.69	4.44	4.89	4.71	5.48	4.32	5.15	6.34
B	4.17	3.05	5.18	4.01	6.11	4.10	5.17	3.57	5.33	5.59
	4.66	5.58	3.66	4.50	3.90	4.61	5.62	4.53	6.05	5.14

[Dobson] Exercise 2.1

A	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.98	4.90	5.75
	5.36	3.48	4.69	4.44	4.89	4.71	5.48	4.32	5.15	6.34
B	4.17	3.05	5.18	4.01	6.11	4.10	5.17	3.57	5.33	5.59
	4.66	5.58	3.66	4.50	3.90	4.61	5.62	4.53	6.05	5.14

X_k : k th output

x_k : k th data

just one of the possible values of X_k

Input

A biological system

- noise
- fluctuation

Outcome

X

By empirical knowledge X is often assumed to obey a normal distribution [see also CLT].

$$X \sim N(m, \sigma^2)$$

7. Sample distributions I: Sample mean

$X_1, X_2, X_3, \cdots, X_N$: random samples
(iid random variable)

sample mean

$$\bar{X} = \frac{1}{N}\sum_{k=1}^{N} X_k$$

Theorem If $X_k \sim N(m, \sigma^2)$, then we have

$$\bar{X} \sim N\left(m, \frac{\sigma^2}{N}\right)$$

PROOF: (1) If X and Y are independent, and

$$X \sim N(m_X, \sigma_X^2)$$

$$Y \sim N(m_Y, \sigma_Y^2)$$

then

$$X + Y \sim N(m_X + m_Y, \sigma_X^2 + \sigma_Y^2)$$

(2) If $X \sim N(m_X, \sigma_X^2)$ and a is a real constant, then

$$aX \sim N(am_X, a^2\sigma_X^2)$$

Check the details!

7. Sample distributions II: Unbiased variance

$X_1, X_2, X_3, \cdots, X_N$: random samples

unbiased variance

$$U^2 = \frac{1}{N-1}\sum_{k=1}^{N}(X_k - \bar{X})^2$$

Theorem If $X_1, X_2, X_3, \cdots, X_N$ are iid random variable with mean m and variance σ^2, we have

$$\mathbf{E}[\bar{X}] = m, \qquad \mathbf{E}[U^2] = \sigma^2$$

Check the details!

Theorem If $X_k \sim N(m, \sigma^2)$, then we have

$$\frac{N-1}{\sigma^2}U^2 \sim \chi^2_{N-1}$$

Here χ^2_{N-1} is the *chi-square distribution* of $N-1$ degrees of freedom.

- *Sum of squares* appears in many contexts.

$$\sum_{k=1}^{N}(X_k - \bar{X})^2$$

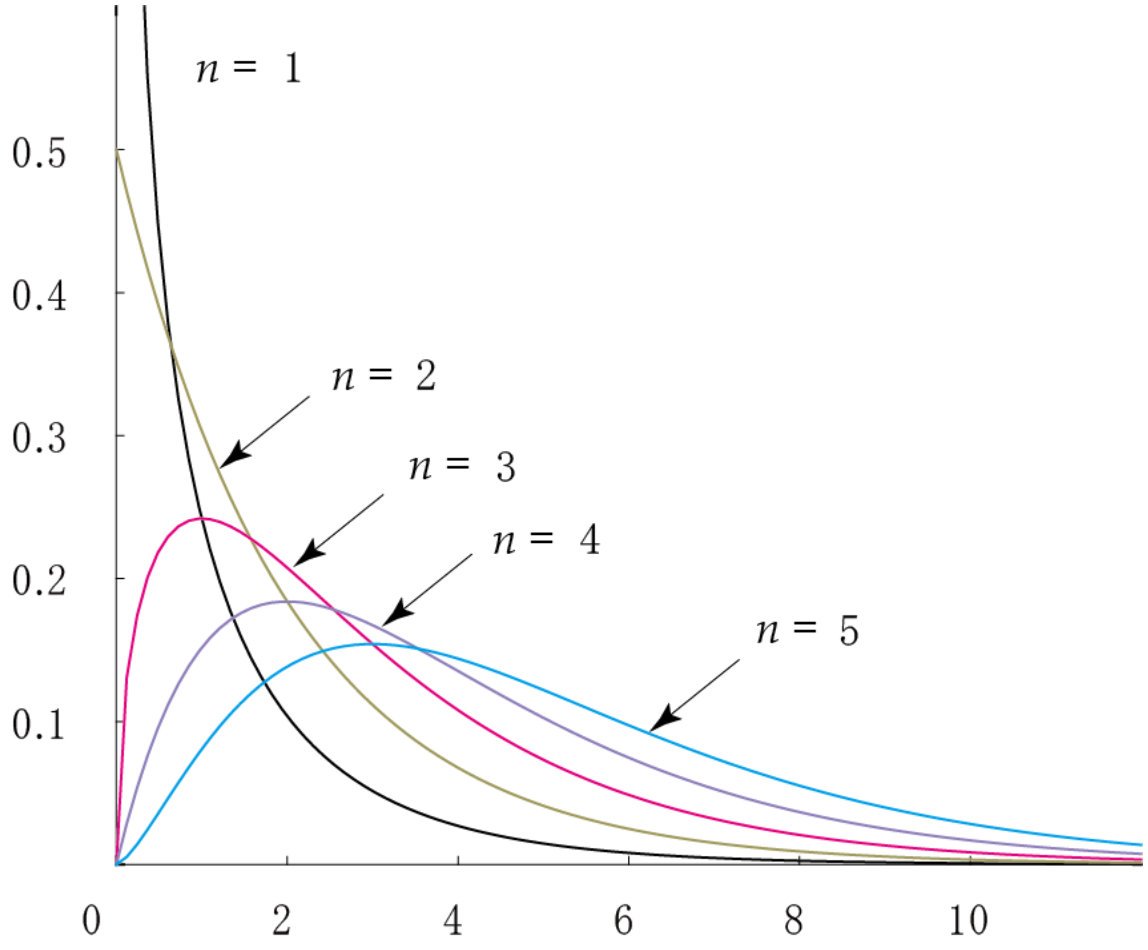
7. Sample distributions III: χ_n^2-distribution

$$f_n(x) = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \quad (x \geq 0)$$

- mean value $m = n$
- variance $\sigma^2 = 2n$
- This is defined to be the distribution of

$$\sum_{k=1}^{n} Z_k^2$$

where Z_k are iid and $\sim N(0,1)$



7. Sample distributions IV: t-distribution

$X_1, X_2, X_3, \cdots, X_N$: random samples

sample mean

$$\bar{X} = \frac{1}{N}\sum_{k=1}^{N} X_k$$

unbiased variance

$$U^2 = \frac{1}{N-1}\sum_{k=1}^{N}(X_k - \bar{X})^2$$

Theorem If $X_k \sim N(m, \sigma^2)$, then we have

$$\frac{\bar{X} - m}{U/\sqrt{N}} \sim t_{N-1}$$

where t_{N-1} is the t-distribution of $N-1$ degrees of freedom.

Cf:

$$\bar{X} \sim N\left(m, \frac{\sigma^2}{N}\right) \Longleftrightarrow \frac{\bar{X} - m}{\sigma/\sqrt{N}} \sim N(0,1)$$

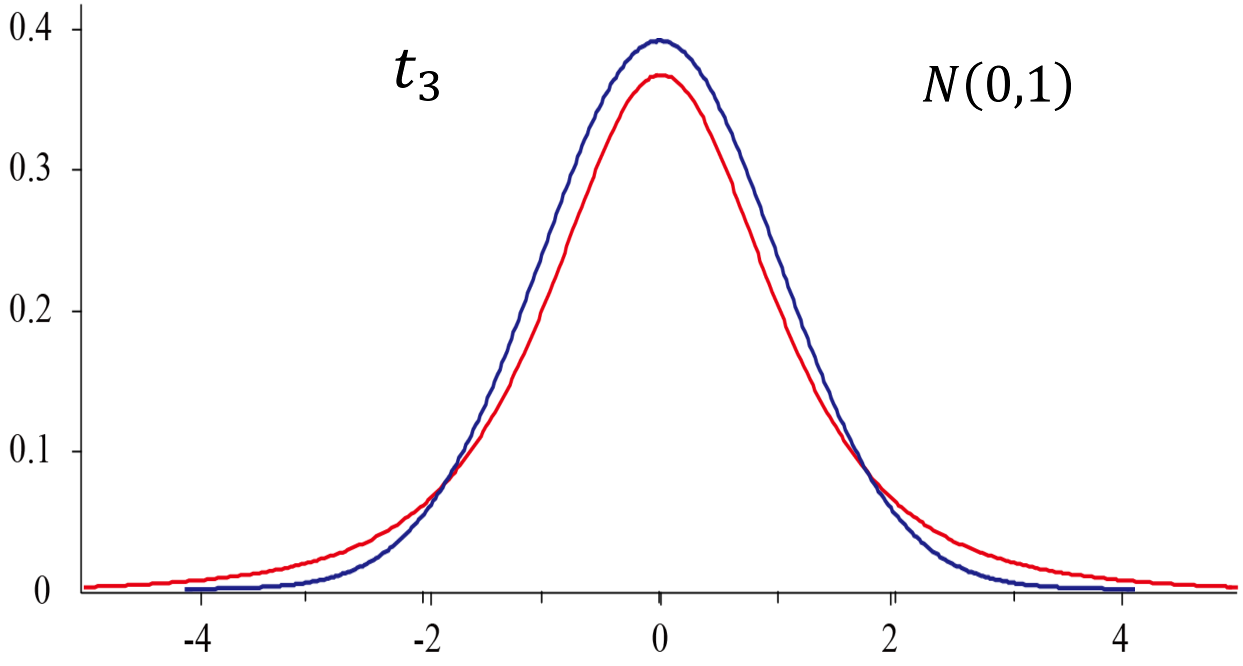
Check the details!

7. Sample distributions IV: t_n-distribution

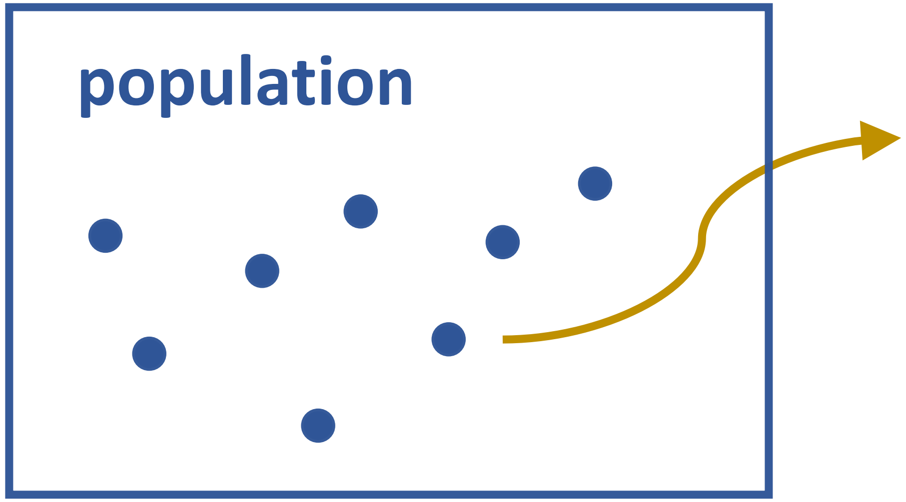
t_n-distribution of n degrees of freedom

$$f_n(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

- Fat tail - compare with $N(0,1)$
- $t_n \rightarrow N(0,1)$ as $n \rightarrow \infty$
- $t_n \approx N(0,1)$ for $n \geq 30$ in practice



8. Random vectors



- sampling
- measuring d quantities for each sample

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = [x_1 \quad x_2 \quad \cdots \quad x_d]^T$$

one sample $\leftrightarrow$ one d dimensional vector

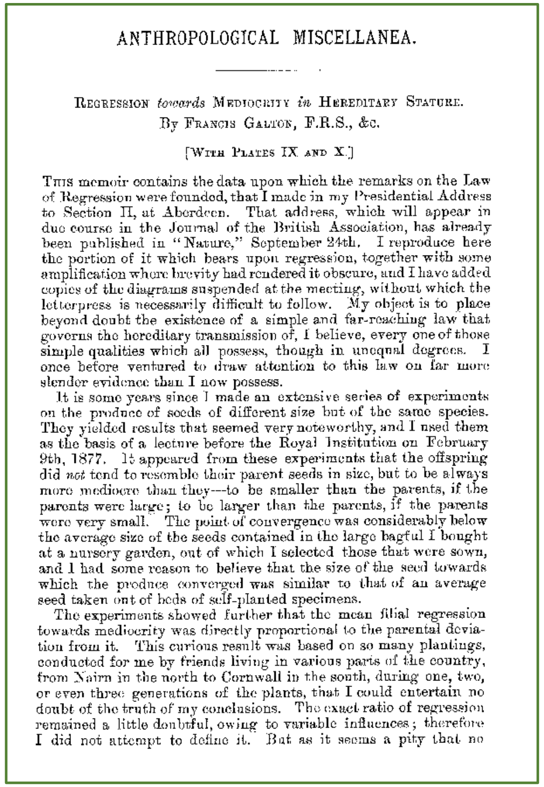
➢ This sampling is modelled by d dimensional random vector

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{bmatrix} = [X_1 \quad X_2 \quad \cdots \quad X_d]^T$$

9. An example of 2-dim data

		Mid-Heights of Parents (x)										
Heights of Adult Children (y)		below	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	above	sum
	above							5	3	2	4		14
	73.2						3	4	3	2	2	3	17
	72.2			1		4	4	11	4	9	7	1	41
	71.2			2		11	18	20	7	4	2		64
	70.2			5	4	19	21	25	14	10	1		99
	69.2	1	2	7	13	38	48	33	18	5	2		167
	68.2	1		7	14	28	34	20	12	3	1		120
	67.2	2	5	11	17	38	31	27	3	4			138
	66.2	2	5	11	17	36	25	17	1	3			117
	65.2	1	1	7	2	15	16	4	1	1			48
	64.2	4	4	5	5	14	11	16					59
	63.2	2	4	9	3	5	7	1	1				32
	62.2		1		3	3							7
	below	1	1	1			1		1				5
	sum	14	23	66	78	211	219	183	68	43	19	4	928

F. Galton:
Regression towards mediocrity in hereditary stature,
Anthropological Miscellanea (1886)



ANTHROPOLOGICAL MISCELLANEA.

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.
By FRANCIS GALTON, F.R.S., &c.

[WITH PLATES IX AND X.]

THIS memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. I reproduce here the portion of it which bears upon regression, together with some amplification where brevity had rendered it obscure, and I have added copies of the diagrams suspended at the meeting, without which the letterpress is necessarily difficult to follow. My object is to place beyond doubt the existence of a simple and far-reaching law that governs the hereditary transmission of, I believe, every one of those simple qualities which all possess, though in unequal degrees. I once before ventured to draw attention to this law on far more slender evidence than I now possess.

It is some years since I made an extensive series of experiments on the produce of seeds of different size but of the same species. They yielded results that seemed very noteworthy, and I used them as the basis of a lecture before the Royal Institution on February 9th, 1877. It appeared from these experiments that the offspring did *not* tend to resemble their parent seeds in size, but to be always more mediocre than they—to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small. The point of convergence was considerably below the average size of the seeds contained in the large bagful I bought at a nursery garden, out of which I selected those that were sown, and I had some reason to believe that the size of the seed towards which the produce converged was similar to that of an average seed taken out of beds of self-planted specimens.

The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it. This curious result was based on so many plantings, conducted for me by friends living in various parts of the country, from Nairn in the north to Cornwall in the south, during one, two, or even three generations of the plants, that I could entertain no doubt of the truth of my conclusions. The exact ratio of regression remained a little doubtful, owing to variable influences; therefore I did not attempt to define it. But as it seems a pity that no

10. Joint probability distributions

For a d dimensional random vector $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{bmatrix} = [X_1 \quad X_2 \quad \cdots \quad X_d]^T$

the *joint distribution* is the most fundamental.

(1) For discrete random variables: $P(X_1 = a_1, X_2 = a_2, \cdots, X_d = a_d)$

(2) For continuous random variables we use the joint density function:

$$P(X_1 \le a_1, X_2 \le a_2, \cdots, X_d \le a_d) = \int_{-\infty}^{a_1} \int_{-\infty}^{a_2} \cdots \int_{-\infty}^{a_d} f(x_1, x_2, \cdots, x_d)\, dx_1 dx_2 \cdots dx_d$$

10. Joint probability distributions: An example

Rolling two dices

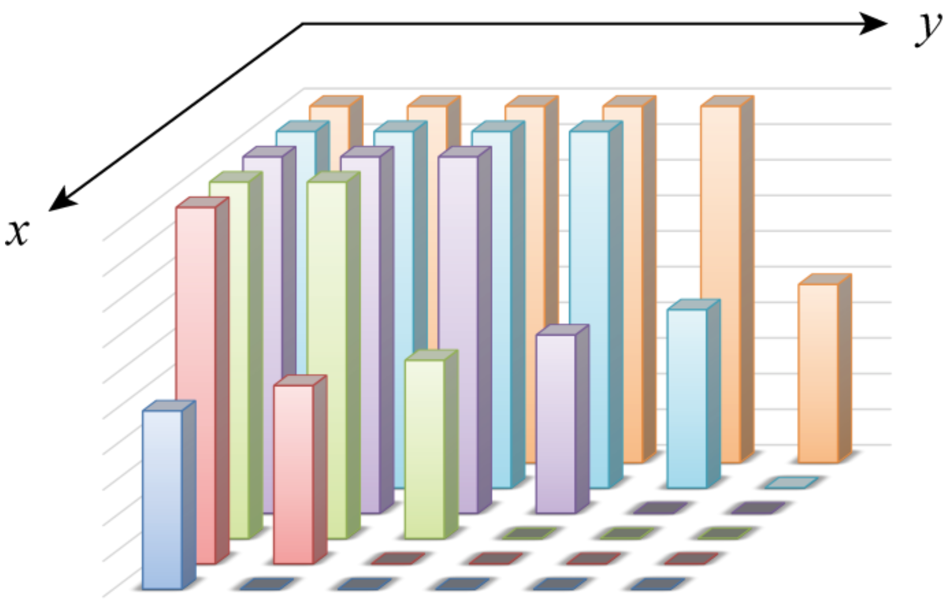
Measurement

$$\mathbf{X} = \begin{bmatrix} X \\ Y \end{bmatrix},\quad \begin{aligned} X &= \text{maximal slot} \\ Y &= \text{minimal slot} \end{aligned}$$

2-dimensional random vector

joint distribution

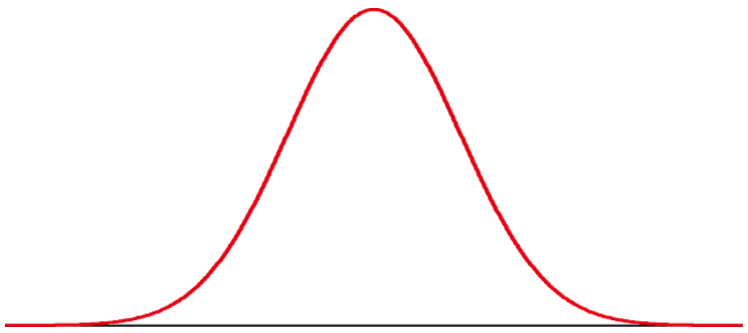
X \ Y	1	2	3	4	5	6
1	1/36	0	0	0	0	0
2	2/36	1/36	0	0	0	0
3	2/36	2/36	1/36	0	0	0
4	2/36	2/36	2/36	1/36	0	0
5	2/36	2/36	2/36	2/36	1/36	0
6	2/36	2/36	2/36	2/36	2/36	1/36



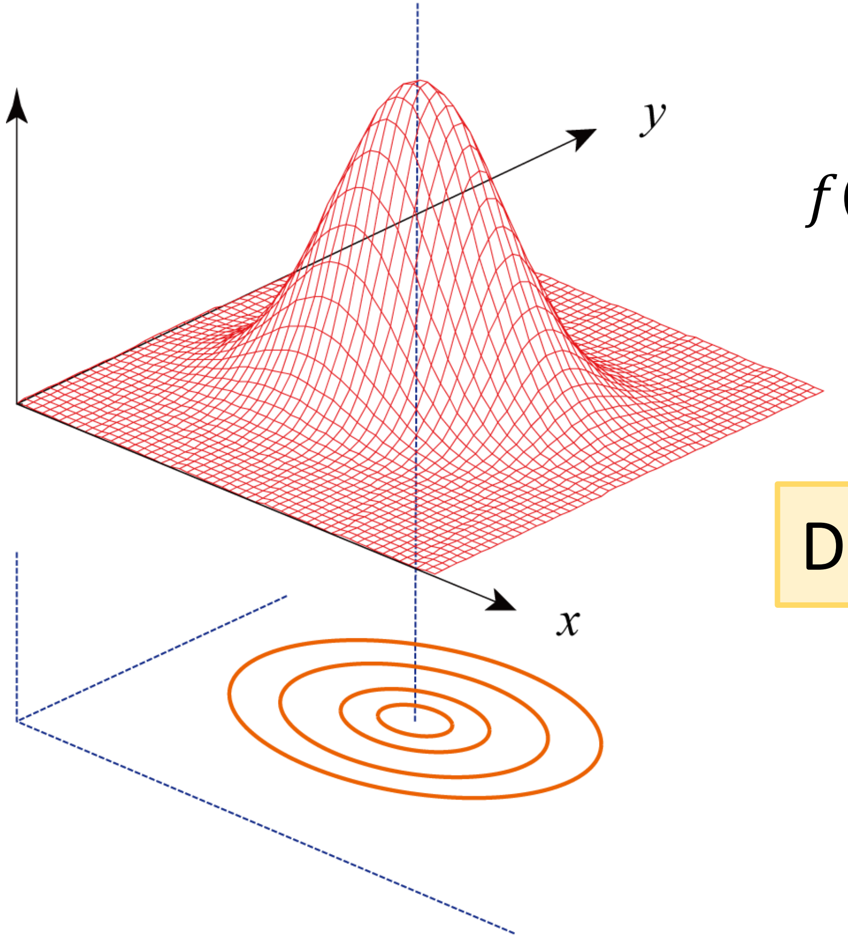
10. Joint probability distributions: $N(\boldsymbol{m}, \Sigma)$

- 1-dimensional case $N(m, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}$$



- 2-dimensional case $N(\boldsymbol{m}, \Sigma)$



$$f(x, y) = \cdots$$

Details Later

11. Marginal distributions

Joint distribution of 2-dimensional discrete random vector $\mathbf{X} = \begin{bmatrix} X \\ Y \end{bmatrix}$

$$P(X = a_i, Y = b_j) = p_{ij}$$

$X \backslash Y$	b_1	$\cdots$	b_j	$\cdots$	b_n
a_1	p_{11}	$\cdots$	p_{1j}	$\cdots$	p_{1n}
$\vdots$					
a_i	p_{i1}	$\cdots$	p_{ij}	$\cdots$	p_{in}
$\vdots$					
a_m	p_{m1}	$\cdots$	p_{mj}	$\cdots$	p_{mn}

11. Marginal distributions

Joint distribution of 2-dimensional <u>discrete</u> random vector $\mathbf{X} = \begin{bmatrix} X \\ Y \end{bmatrix}$

$$P\left(X = a_i, Y = b_j\right) = p_{ij}$$

$X \backslash Y$	b_1	$\cdots$	b_j	$\cdots$	b_n	**sum**
a_1	p_{11}	$\cdots$	p_{1j}	$\cdots$	p_{1n}	$p_{1\cdot}$
$\vdots$						
a_i	p_{i1}	$\cdots$	p_{ij}	$\cdots$	p_{in}	$p_{i\cdot}$
$\vdots$						
a_m	p_{m1}	$\cdots$	p_{mj}	$\cdots$	p_{mn}	$p_{m\cdot}$

$$P(X = a_i) = \sum_{j=1}^{n} P\left(X = a_i, Y = b_j\right)$$

Marginal distribution

11. Marginal distributions

Joint distribution of 2-dimensional <u>discrete</u> random vector $\mathbf{X} = \begin{bmatrix} X \\ Y \end{bmatrix}$

$$P(X = a_i, Y = b_j) = p_{ij}$$

$X \backslash Y$	b_1	$\cdots$	b_j	$\cdots$	b_n	**sum**
a_1	p_{11}	$\cdots$	p_{1j}	$\cdots$	p_{1n}	$p_{1\cdot}$
$\vdots$						
a_i	p_{i1}	$\cdots$	p_{ij}	$\cdots$	p_{in}	$p_{i\cdot}$
$\vdots$						
a_m	p_{m1}	$\cdots$	p_{mj}	$\cdots$	p_{mn}	$p_{m\cdot}$
sum	$p_{\cdot 1}$		$p_{\cdot j}$		$p_{\cdot n}$	1

$$P(X = a_i) = \sum_{j=1}^{n} P(X = a_i, Y = b_j)$$

Marginal distribution

$$P(Y = b_j) = \sum_{i=1}^{m} P(X = a_i, Y = b_j)$$

12. Conditional distributions

Discrete case

Marginal distribution of Y

$$P(Y = b_j) = \sum_{i=1}^{m} P(X = a_i, Y = b_j)$$

Conditional distribution

$$P(X = a_i \mid Y = b_j) = \frac{P(X = a_i, Y = b_j)}{P(Y = b_j)}$$

Conditional expectation

$$\mathbf{E}[X|Y = b_j] = \sum_i a_i \, P(X = a_i \mid Y = b_j)$$

$X \backslash Y$	b_1	$\cdots$	b_j	$\cdots$	b_n	sum
a_1			p_{1j}			
$\vdots$			$\vdots$			
a_i			p_{ij}			
$\vdots$			$\vdots$			
a_m			p_{mj}			
sum			$P(Y = b_j)$			1

Marginal distribution of Y

Exercise 1 (5min)

Suppose that the joint distribution of (X, Y) is given by the following table:

$X \backslash Y$	1	2	3	4
1	1/16	1/16	0	0
2	1/16	2/16	0	1/16
3	2/16	2/16	0	1/16
4	1/16	1/16	2/16	1/16

(1) Find the marginal distributions.

(2) Calculate $\mathbf{E}[X]$ and $\mathbf{V}[X]$.

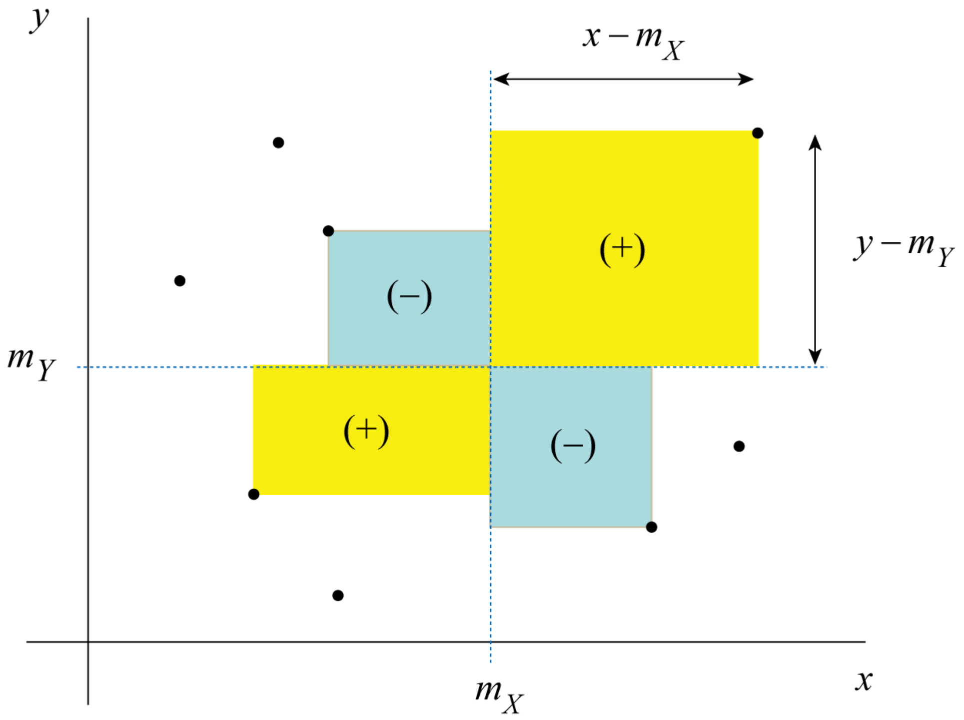
(3) Find $P(X = 2|Y = 1)$.

(4) Find $P(Y = 2|X = 3)$.

(5) Find $\mathbf{E}[Y|X = 3]$.

14. Covariance and correlation coefficient

$$\mathbf{Cov}(X,Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$$
$$= \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$$
$$\rho(X,Y) = \frac{\mathbf{Cov}(X,Y)}{\sqrt{V[X]}\sqrt{V[Y]}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$



Normalization of a random variable:

$$\tilde{X} = \frac{X - \mathbf{E}[X]}{\sqrt{V[X]}} = \frac{X - m}{\sigma}$$

$$E[\tilde{X}] = 0, \qquad V[\tilde{X}] = 1$$

Check the details!

FORMULA:

$$\rho(X,Y) = E[\tilde{X}\tilde{Y}] = \mathbf{Cov}(\tilde{X},\tilde{Y}) = \rho(\tilde{X},\tilde{Y})$$

Exercise 2 (10min)

Suppose that the joint distribution of (X, Y) is given by the following table:

$X \backslash Y$	1	2	3	4
1	1/16	1/16	0	0
2	1/16	2/16	0	1/16
3	2/16	2/16	0	1/16
4	1/16	1/16	2/16	1/16

(1) Calculate $\mathbf{Cov}(X, Y)$.

(2) Calculate $\rho(X, Y)$.

15. Independent random variables

A set of random variables $X_1, X_2, \cdots, X_n$ is called *independent* if

$$P(X_1 \le a_1, X_2 \le a_2, \cdots, X_n \le a_n) = \prod_{k=1}^{n} P(X_k \le a_k) \qquad (*)$$

If $X_1, X_2, \cdots, X_n$ are discrete random variables, (*) is equivalent to

$$P(X_1 = a_1, X_2 = a_2, \cdots, X_n = a_n) = \prod_{k=1}^{n} P(X_k = a_k)$$

If $X_1, X_2, \cdots, X_n$ are continuous random variables, (*) is equivalent to

$$f_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n) = \prod_{k=1}^{n} f_{X_k}(x_k)$$

FORMULA: For any random variables X and Y, we have

(1) $\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y]$

(2) $\mathbf{V}[aX + bY] = a^2\mathbf{V}[X] + b^2\mathbf{V}[Y] + 2ab\mathbf{Cov}(X, Y)$

Check the details!

PROOF (1) We deal with discrete random variables. For continuous case we need only to replace the joint probability by joint density function.

$$\mathbf{E}[aX + bY] = \sum_z zP(aX + bY = z) = \sum_z z \sum_{ax+by=z} P(X = x, Y = y) = \sum_{x,y} (ax + by)P(X = x, Y = y)$$

$$= a\sum_{x,y} xP(X = x, Y = y) + b\sum_{x,y} yP(X = x, Y = y) = a\mathbf{E}[X] + b\mathbf{E}[Y]$$

(2) By definition,

$$\begin{aligned}\mathbf{V}[aX + bY] &= \mathbf{E}[(aX + bY)^2] - \mathbf{E}[aX + bY]^2 \\ &= a^2\ \mathbf{E}[X^2] + 2ab\mathbf{E}[XY] + b^2\mathbf{E}[Y^2] - a^2\mathbf{E}[X]^2 - 2ab\mathbf{E}[X]\mathbf{E}[Y] - b^2\mathbf{E}[Y]^2 \\ &= a^2\mathbf{V}[X] + b^2\mathbf{V}[Y] + 2ab\mathbf{Cov}(X, Y)\end{aligned}$$

16. Uncorrelated random variables

Theorem If X and Y are independent, they are uncorrelated, i.e., $\mathbf{Cov}(X,Y) = 0$.
Remark The converse assertion is not true in general.

Lemma If X and Y are independent, then $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$.

PROOF We consider the discrete case. The continuous case is similar.

$$\mathbf{E}[XY] = \sum_z zP(XY = z) = \sum_z z \sum_{xy=z} P(X = x, Y = y) = \sum_{x,y} xy\, P(X = x, Y = y)$$

$$= \sum_x xP(X = x) \sum_y yP(Y = y) = \mathbf{E}[X]\mathbf{E}[Y]$$

Check the details!

PROOF of Theorem
Only need to apply Lemma to the definition $\mathbf{Cov}(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$.

Submission of reports for evaluation

- ✓ During my lectures you will be given 10 Problems.
- ✓ Choose 3 problems at your own taste and write up a short report.
- ✓ Submission deadline: November 5 (Mon), 2018.
- ✓ Way of submission: (a) directly hand to Prof Obata

  or (b) send in PDF by e-mail to obata@tohoku.ac.jp

  or (c) bring to the secretary on 6F GSIS and ask her politely.

Problem 1

Suppose that the joint distribution of (X, Y) is given by the following table:

X \ Y	1	2	3	4	5	6
1	1/36	0	0	0	0	0
2	2/36	1/36	0	0	0	0
3	2/36	2/36	1/36	0	0	0
4	2/36	2/36	2/36	1/36	0	0
5	2/36	2/36	2/36	2/36	1/36	0
6	2/36	2/36	2/36	2/36	2/36	1/36

(1) Find the marginal distributions.

(2) Calculate $\mathbf{E}[X]$ and $\mathbf{V}[X]$.

(3) Calculate $\mathbf{Cov}(X, Y)$ and ρ(X, Y).

(4) Find $P(X = 4|Y = 2)$.

(5) Find $\mathbf{E}[X|Y = 2]$.

(6) [challenge] Since $\mathbf{E}[X|Y{=}k]$ $(k = 1,2,\cdots,6)$ may be considered as a function of Y, it is a random variable, denoted by $\mathbf{E}[X|Y]$ and called the conditional expectation. Examine that $\mathbf{E}[\mathbf{E}[X|Y]]{=}\mathbf{E}[X]$.

Problem 2

Four cards are drawn from a deck (of 52 cards). Let X be the number of aces and Y the number of kings that show.

(1) Show the joint distribution of (X, Y), and marginal distributions of X and Y.

(2) Find the mean values $\mathbf{E}[X]$ and $\mathbf{E}[Y]$.

(3) Find the variances $\mathbf{V}[X]$ and $\mathbf{V}[Y]$.

(4) Find the covariance $\mathbf{Cov}(X, Y)$ and correlation coefficient $\rho(X, Y)$.

(5) Find $\mathbf{E}[X|Y = 1]$

- Answer by the ratios of integers, do not use decimal expressions.

Problem 3

Let X, Y be random variables taking just two values, say,

$$P(X = a) = p, \qquad P(X = b) = 1 - p$$
$$P(Y = c) = q, \qquad P(Y = d) = 1 - q$$
$$0 < p < 1, \qquad 0 < q < 1.$$

Show that X and Y are independent if and only if $\mathbf{Cov}(X, Y) = 0$.

Note: 'only if' part is straightforward (see also the general theorem in Section 16). The point here is to show 'if' part.

Problem 4

(1) Show the histogram of each group.

(2) Calculate the mean value and unbiased variance of each group.

(3) Judge by hypothesis testing whether these groups are random samples from the normal population $N(4.82, 0.04) = N(4.82, 0.2^2)$?

➢ If you are not familiar with the hypothesis testing, study it on this occasion!

A	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.98	4.90	5.75
	5.36	3.48	4.69	4.44	4.89	4.71	5.48	4.32	5.15	6.34
B	4.17	3.05	5.18	4.01	6.11	4.10	5.17	3.57	5.33	5.59
	4.66	5.58	3.66	4.50	3.90	4.61	5.62	4.53	6.05	5.14