Statistics and data analysis I

Week 13
“Hypothesis test”

2019.9th July.
Hirotada Honda

1. Summary on Interval estimation

Interval estimation on population mean

- Two cases:

  ① In case the population variance, σ^2 , is known;

  ② In case the population variance is unknown.

Interval estimation on population mean

- 95%-CI when the population variance is known to be σ² :

$$\bar{X} - 1.96 \times \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}$$

Upper2.5-percentile of z-dist.

- 95%-CI when the population variance is unknown :

$$\bar{X} - t_{n-1}\left(\frac{0.05}{2}\right) \times \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1}\left(\frac{0.05}{2}\right) \times \frac{S}{\sqrt{n}}$$

S:Unbiased SD.

Upper2.5-percentile of t-dist.

Two differences; percentile and SD.

Interval estimation with python

Python の人・・・ Population variance is known→norm.interval

Population variance is unknown→t.interval. You need df (=degree of freedom) also!

CI by python （Population SD is <u>known</u>）

```
import numpy as np
import scipy.stats as st

x=np.array([120])

#Sample size.
n=x.size

#Sample mean.
x_mean=x.mean()

#Population SD.
x_sd=6

st.norm.interval(alpha=0.95,loc=x_mean,scale=x_sd/np.sqrt(n))
```

```
(108.24021609275968, 131.75978390724032)
```

Population SD

CI by python （Population SD is unknown）

```
import numpy as np
import scipy.stats as st

x=np.array([7.86, 7.89, 7.84, 7.90, 7.82])

#Sample size.
n=x.size

#Sample mean.
x_mean=x.mean()

#Unknown SD.
x_sd=np.std(x,ddof=1)

st.t.interval(alpha=0.95,df=n-1,loc=x_mean,scale=x_sd/np.sqrt(n))
```

```
(7.820445974652658, 7.903554025347343)
```

Unbiased SD

Don’t forget *df*!

【Ref.】 Sample SD and unbiased SD

- In case the population variance σ² is unknown, you need to estimate it from the sample.
- The estimator for the population variance σ² is, usually,the unbiased variance defiend as below ：

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

2． Hypothesis test

Hypothesis test

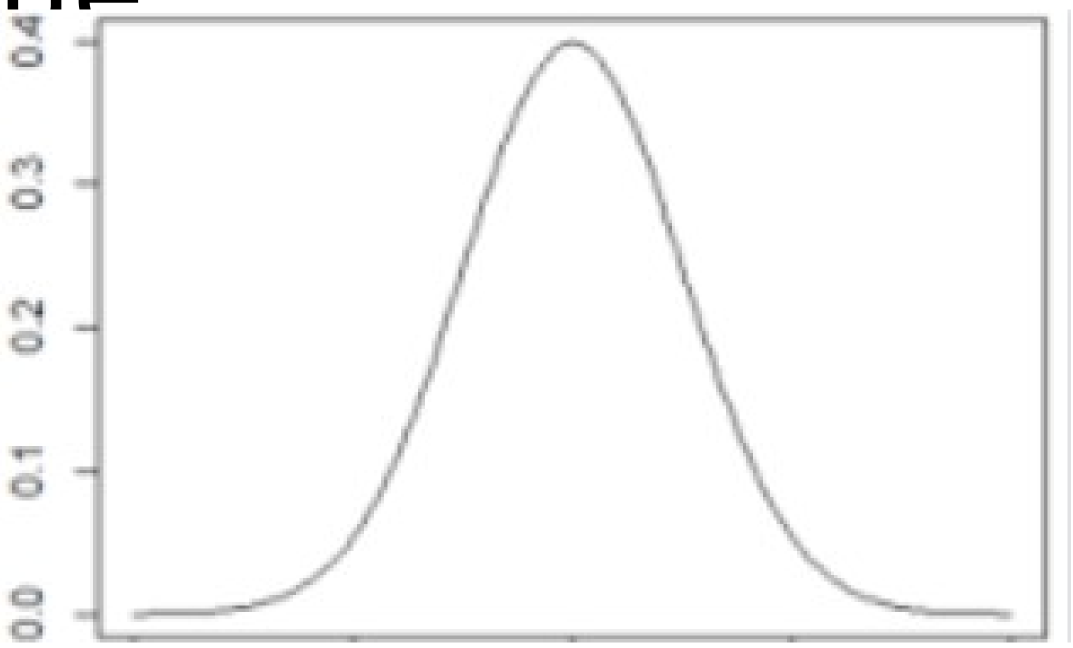
- "Verifying a hypothesis though a statistical method"

Use cases

- Ex①） A certain lady says concerning the milk tea: “By drinking a cup of milk tea, I can detect which of milk tea is inserted first.”
- We tested this statement, and she really answered correctly 5 times in a row. Can we say she says “truth”?
- Ex②)In a certain company, they trained the employees in different ways. After half a year passed from the start of the course, examine whether the training programs work well.

Quiz for population mean

- Now we generate r.v.s that follow the normal dist. with R.
- Based on the observed values, estimate the value of its expected value



Quiz for population mean

- Mr. A and Mr. B now make hypotheses (=null hypothesis).
- Answer true or false.

```
> rnorm([redacted])
 [1] -0.7434768  0.1437100 -1.2882375 -1.1411836 -0.2146270 -0.6403474
 [7] -0.3248116  1.5498379  0.5170777 -0.4697282
```

```
> rnorm(          )
 [1] -0.7434768  0.1437100 -1.2882375 -1.1411836 -0.2146270 -0.6403474
 [7] -0.3248116  1.5498379  0.5170777 -0.4697282
```

- Mr.A：「Expected value is 0.」
- T/F?

```
> rnorm(          )
 [1] -0.7434768  0.1437100 -1.2882375 -1.1411836 -0.2146270 -0.6403474
 [7] -0.3248116  1.5498379  0.5170777 -0.4697282
```

- Mr.B： 「Expected value is 10.」

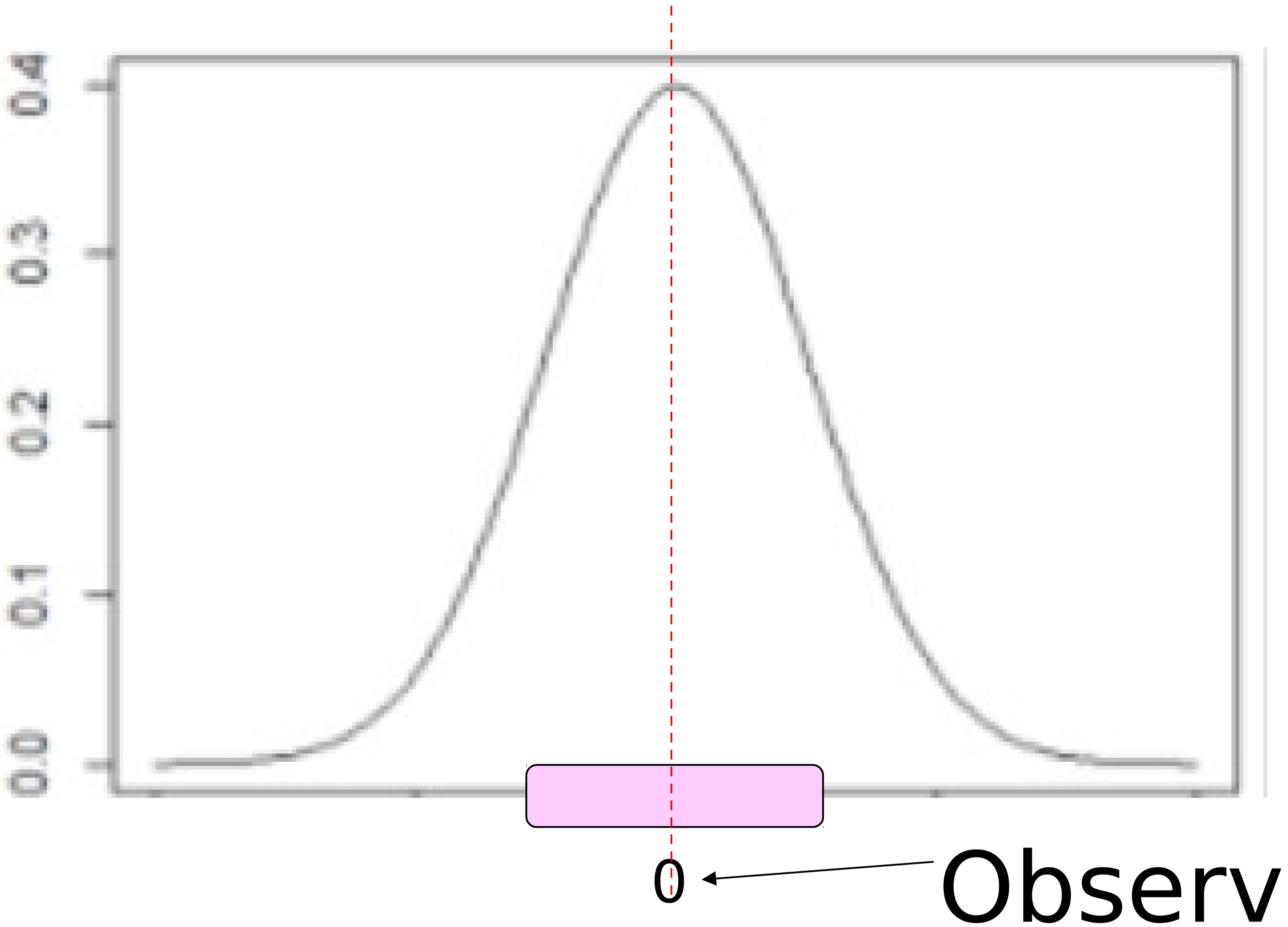
- T/F?

```
> rnorm(        )
 [1] -0.7434768  0.1437100 -1.2882375 -1.1411836 -0.2146270 -0.6403474
 [7] -0.3248116  1.5498379  0.5170777 -0.4697282
```

- Mr.A seems (likely to be) true.
- What Mr.B said is <u>probably</u> not true.

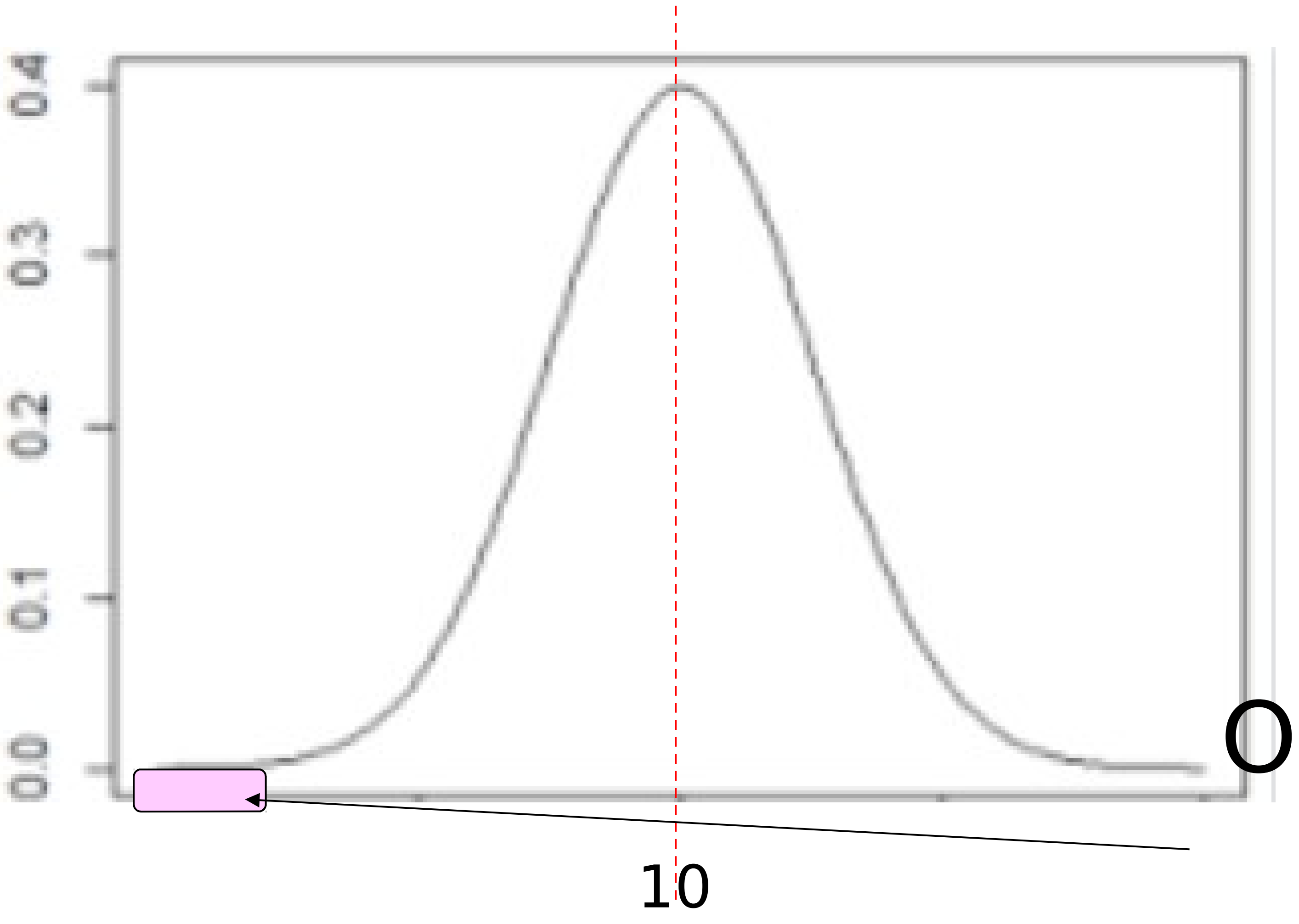
- Why did you think so?

- Normality is assumed.
- If Mr. A is correct…



Observed data are around here.

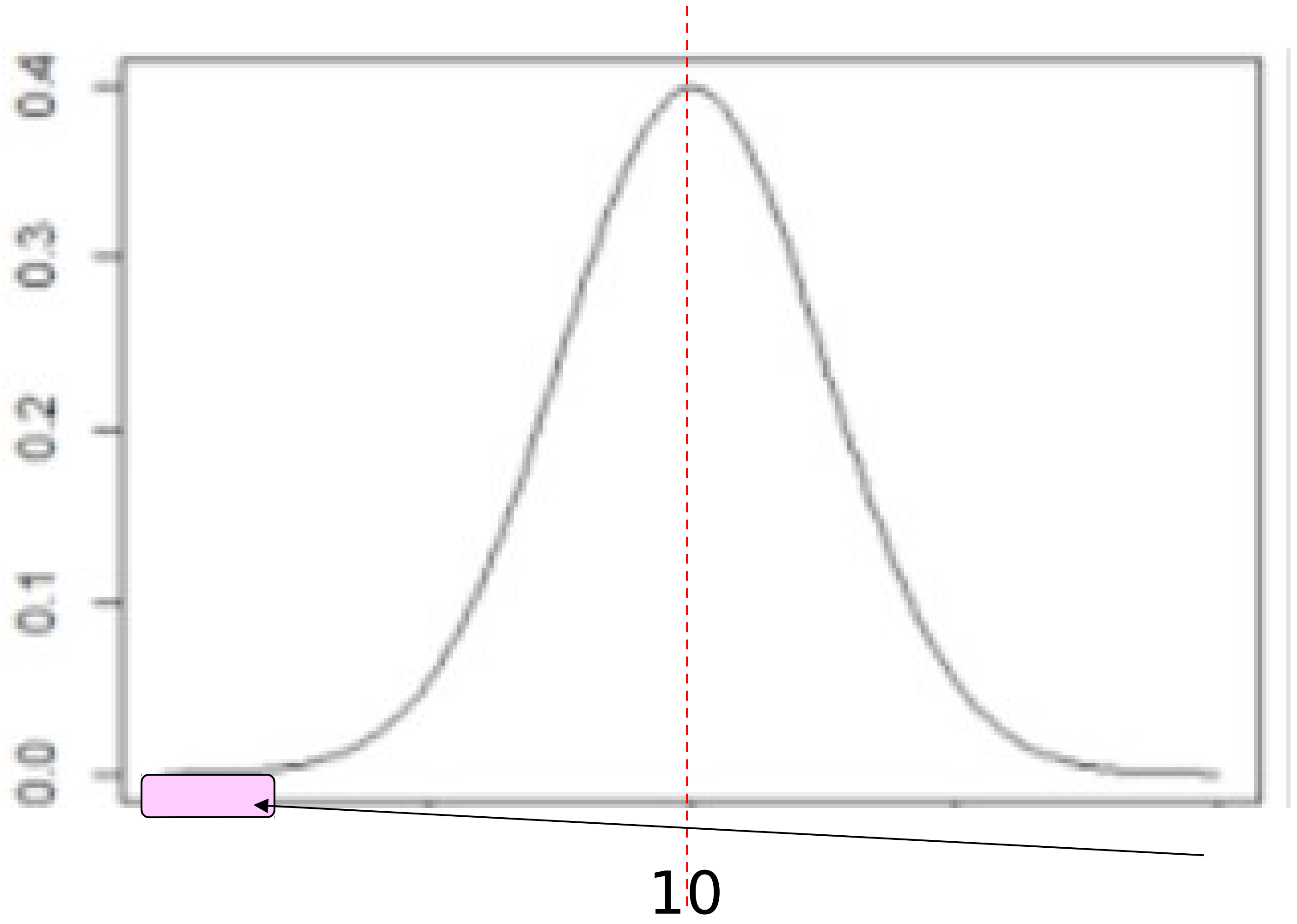
- If Mr. B is correct…



Observed data are around here.
→Seems not likely.

- However, it might happen with a very low probability.
- So we cannot assert that ‘what Mr.B said is absolutely false’.

⇒ The possibility of ‘incorrect conclusion’ should be taken into account.



- However, it might happen with a very low probability.
- So we cannot assert that ‘what Mr.B said is absolutely false’.

⇒　The possibility of ‘incorrect conclusion’ should be taken into account.

-’Might be incorrect with 5%-probability, but Mr.B’s hypothesis is probably not true.’

- However, it might happen with a very low probability.
- So we cannot assert that ‘what Mr.B said is absolutely false’.

⇒ The possibility of ‘incorrect conclusion’ should be taken into account.

-’Might be incorrect with 5%-probability, but Mr.B’s hypothesis is probably not true.’

- However, it might happen with a very low probability.
- So we cannot assert that 'what Mr.B said is absolutely false'.

⇒ The possibility of 'incorrect conclusion' should be taken into account.

-'Might be incorrect with 5%-probability, but Mr.B's hypothesis is probably not true.'

Significance level.

If this conclusion is acutually incorrect, then we say that it is type-I error.

(asserted a correct hypothesis as incorrect)

- The spirit of hypothesis test is ‘The benefit of the doubt ’
- In the sense that..

- We assert ‘the (null) hypothesis is not true’ if and only if we can say so almost surely (i.e., 95%).

- On the other hand, Mr.A’s hypothesis is actually true?

- It is true that ‘it does not seem incorrect.’
- But then, you certainly guarantee that ‘the expected value is actually 0.00’?

- Might be 0.0001…. (Such a case is sufficiently possible.)

- On the other hand, Mr.A’s hypothesis is actually true?

- It is true that ‘it does not seem incorrect.’
- But then, you certainly guarantee that ‘the expected value is actually 0.00’?
- Might be 0.0001…. (Such a case is sufficiently possible.)

We do not say ‘the null hypothesis is true’. We say
‘We cannot reject the (null) hypothesis with significance level of 5%.’(seems not incorrect)

- 「Seems not incorrect」
- Very passive representation. On the other hand, we like to detect an incorrect hypothesis surely.

The prob. with which an incorrect hypothesis is detected as ‘not true’

： Called power.

Contrary, the issue that you cannot detect an incorrect hypothesis is

Called type-II error.

The power means that prob. that you don’t make a type-II error.

2-2. Flow of hypothesis test

Flow of hypothesis test

- i) Set the population （☚ Similar to confidence interval ）
- ii) Set the null hypothesis H_0.
- iii) Extract samples $x_1, x_2, \ldots, x_N$ from the population.
- iv) Find a statistics $T(x_1, x_2, \ldots, x_N)$ from the sample above.
- v) Calculate the probability density of the statistics $T(X_1, X_2, \ldots, X_N)$ for r.v.s $X_1, X_2, \ldots, X_N$.

Flow of hypothesis test

- vi) For a certain significance level α, find a region R, where

$$P(T(X_1,X_2,\ldots,X_N)\in R) = \alpha$$

Usually, α=0.01 or 0.05

holds(This region R is called as the critical region)

- vii) If $T(x_1,x_2,\ldots,x_N)\in R$, reject the null hypothesis H_0 / otherwise, H_0 cannot be rejected.

To the definition of terms

2-3. definition of terms

Hypothesis

- Null hypothesis H_0 usually takes the following form (in case of the test of mean)

  " the population mean is μ_0"

- A hypothesis different from H_0 is called as alternative hypothesis H_1
  - In the example above, there are 3 possible H_1:

    " The population mean μ is not μ_0", i.e., $\mu \neq \mu_0$

    "The population mean μ is larger than μ_0", i.e., $\mu > \mu_0$

    "The population mean μ is smaller than μ_0", i.e., $\mu < \mu_0$
  - Depends on the purpose of the survey / experiment.

Null hypothesis and alternative hypothesis

- You must write them correctly.

- Null hypothesis is:
- hypothesis that assumes ‘not changed’ in many cases.
- Pointwise statement like ‘μ=1.0’. -> hard to verify!

- We can’t support its actual correctness. Called ’null hypothesis’ in that sense.

Ex.

- Answer the null and alternative hypotheses:

A certain product is specified as its mean and SD of weight are 12 [kg] and 1[kg2], resp.
Now, as a result of a sample survey, they observed:
11, 12,15,14,17,20,18,14,18,11,17,14,16,13,15,19.
Now, check if you can say the strength of this product is improved or not. Do the hypothesis tes
with the significance level of 5%.

Ex

- H0:’not changed’ (the mean is 12, is also allowed.)

Ex

- H1:’the strength has increased’ (the mean is larger than 12, is also allowed.)

Significance level

- Significance level： probability that you may incorrectly reject a correct null hypothesis. Should be determined in advance! Usually, α=0.05 or 0.01. (0.05 throughout this course)
  - On the basis of the correctness of H_0, determine whether a rare situation happens or not, from the observed samples.
  - If you judged that a rare situation happens, then the null hypothesis H_0 is incorrect. The probability that the test statistic takes the observed or more extreme value under H_0 is called as p-value.

P-value

- Smaller p-value = The observed result is likely to be extreme. = H0 seems incorrect.

Significance level

- The smaller the p-value is, the smaller the probability is that the test statistic takes the observed value.
- Under the significance level of 5%, if the p-value is 4.5%, then, the null hypothesis is rejected.

Significance level, type I and II errors

- The significance level is the probability that “you may incorrectly reject a correct null hypothesis”.
- This type of error is called as the “type I error”.
- In other words, the significance level α denotes the probability that you make a type I error.

Power

- The probability that you correctly reject a incorrect null hypothesis H_0.

- The error that you cannot reject a incorrect null hypothesis is called as the “type II error”.

- In other words, the power means the probability that you do not make the type II error.

Results and errors

		Actual situation	
		H0	H1
Results of test	H0	Correct （Probability： 1-α）	Type II error （Probability： β）
	H1	Type I error （Probability： α）	Correct （Probability： 1－β=power）

Test statistic

- In hypothesis test, we transfer the measured value (height, weight and so on) into the value for the test.
- This value is called as the test statistic.
- In hypothesis test, you should carefully watch whether the observed value of test statistics belongs to the rejection region or not.

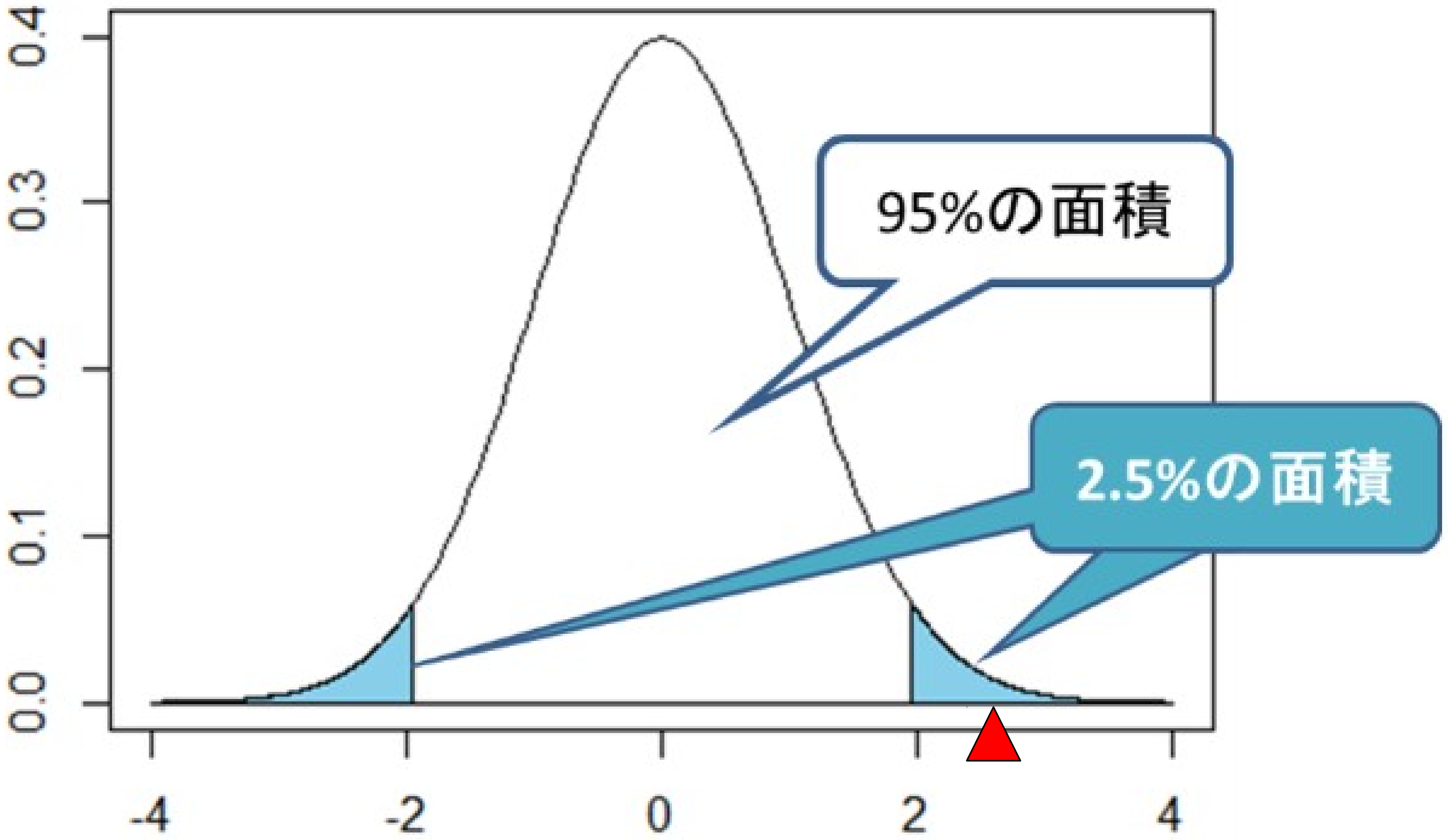
Two-sided test / one-sided test

- Depends on the alternative hypothesis.

- In the former example,
- H_1: 「$\mu \neq \mu_0$」 is the two-sided test.
- H_1 : 「$\mu > \mu_0$」 「$\mu < \mu_0$」 are one-sided tests.

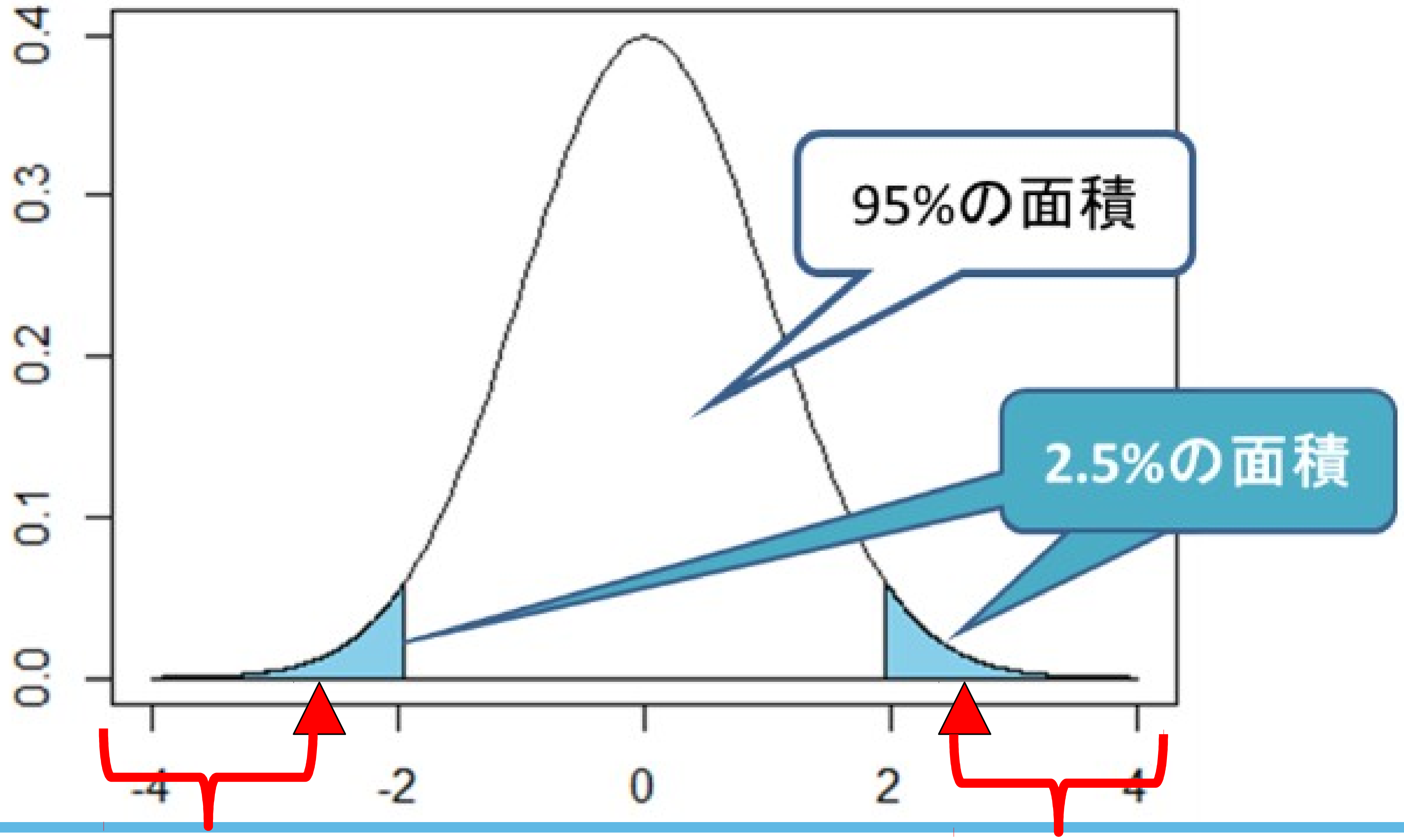
Two-sided p-value and one-sided p-value

- In the two-sided test, we should think of the prob. of the extreme situations ‘*in both sides*’.



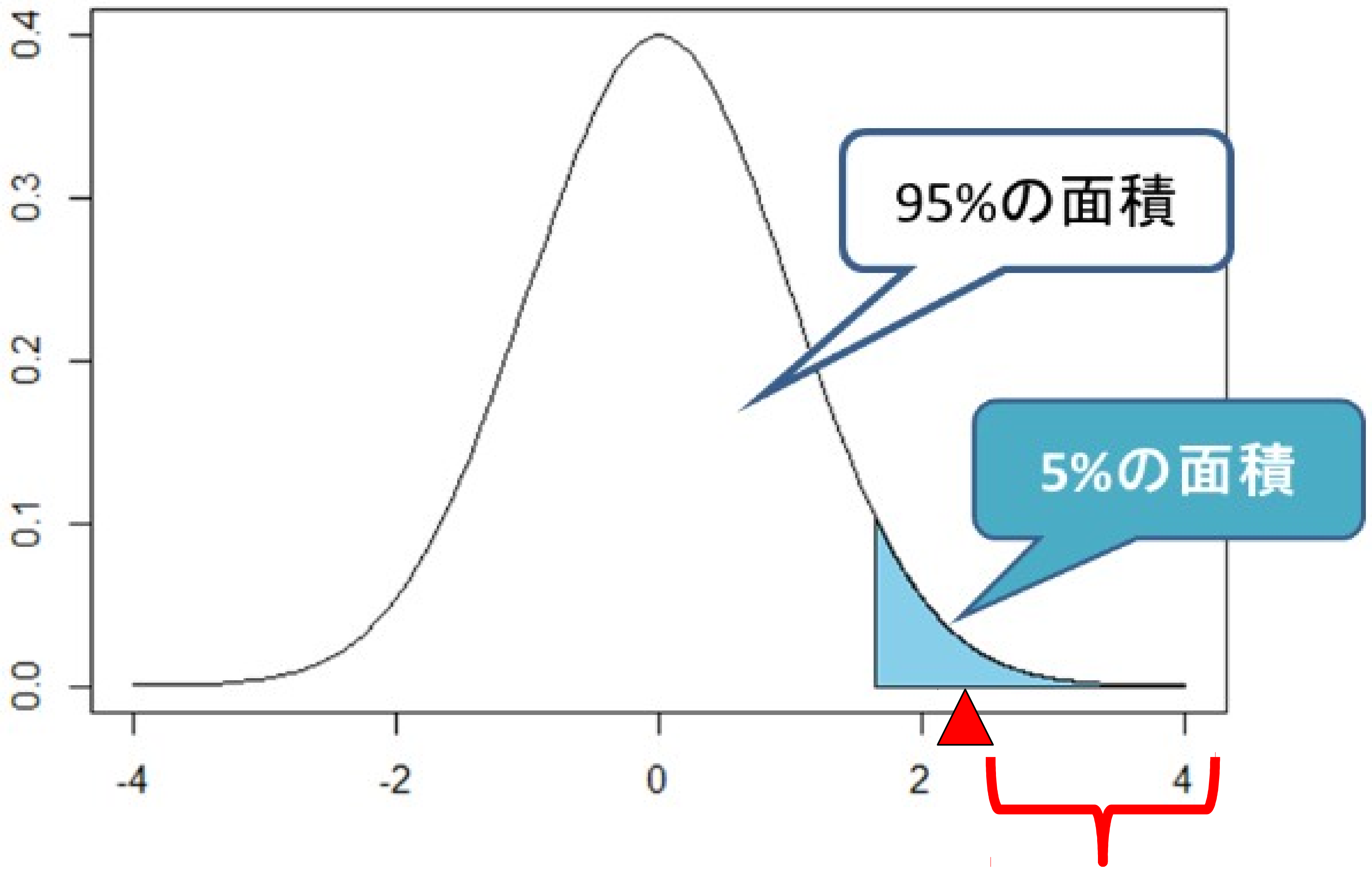
Two-sided p-value and one-sided p-value

- In the two-sided test, we should think of the prob. of the extreme situations ‘*in both sides*’.



Two-sided p-value and one-sided p-value

- In the one-sided test, the extreme situation only in either side is considered.



Two-sided test / one-sided test: examples

- Survey of element B in drug A. How much B is included in A?
- Let us sample 25 tablets of drug A, and then measure the weight of B per tablet. The sample mean is $\bar{x} = 98$ [mg]、 unbiased variation is $S^2 = 1$
- Null hypothesis H_0:”100mg of B on average is included in A.” ⇒ 3 possible alternative hypotheses;

  i) The content of B in A is not 100mg per tablet.

  ii) The content of B in A is larger than 100mg per tablet.

  iii) The content of B in A is smaller than100mg per tablet.

Two-sided test / one-sided test: examples

- i) The content of B in A is not 100mg per tablet.

  ⇒ Check that the content of B is 100mg or not.
  (Two-sided)

ii) The content of B in A is larger than 100mg per tablet.

⇒ Check whether the conte is larger than100mg or not.
(One-sided. We don’t c her it’s smaller or not) .

One-sided

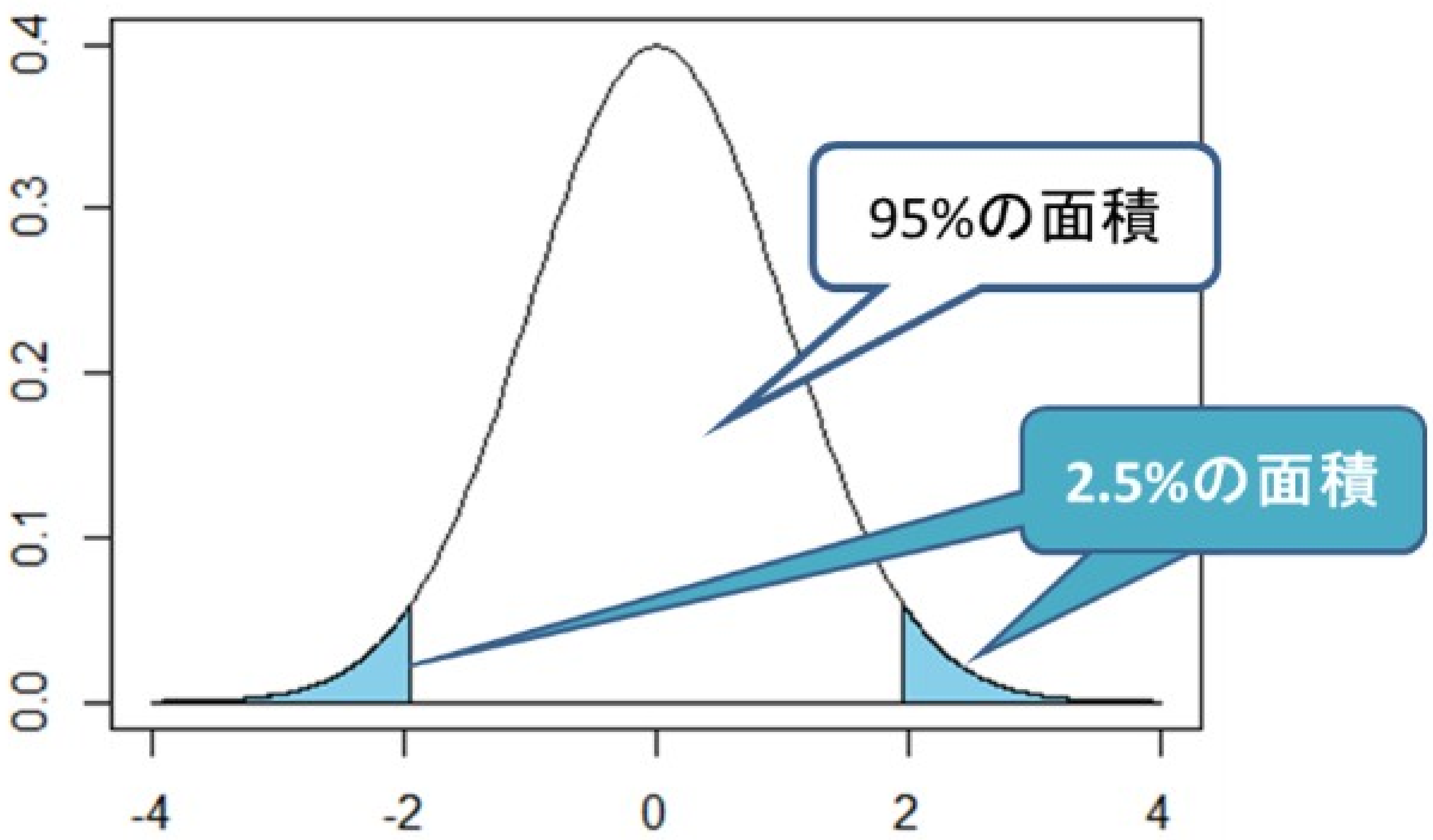
iii) The content of B in A is smaller t mg per tablet.

⇒ Check whether the content of B is smaller than100mg or not
(One-sided. We don’t care whether it’s larger or not) .

Two-sided test / one-sided test

- In case of significance level of 5%:

Two-sided （Alternative hypothesis： The content of B is not 100mg）

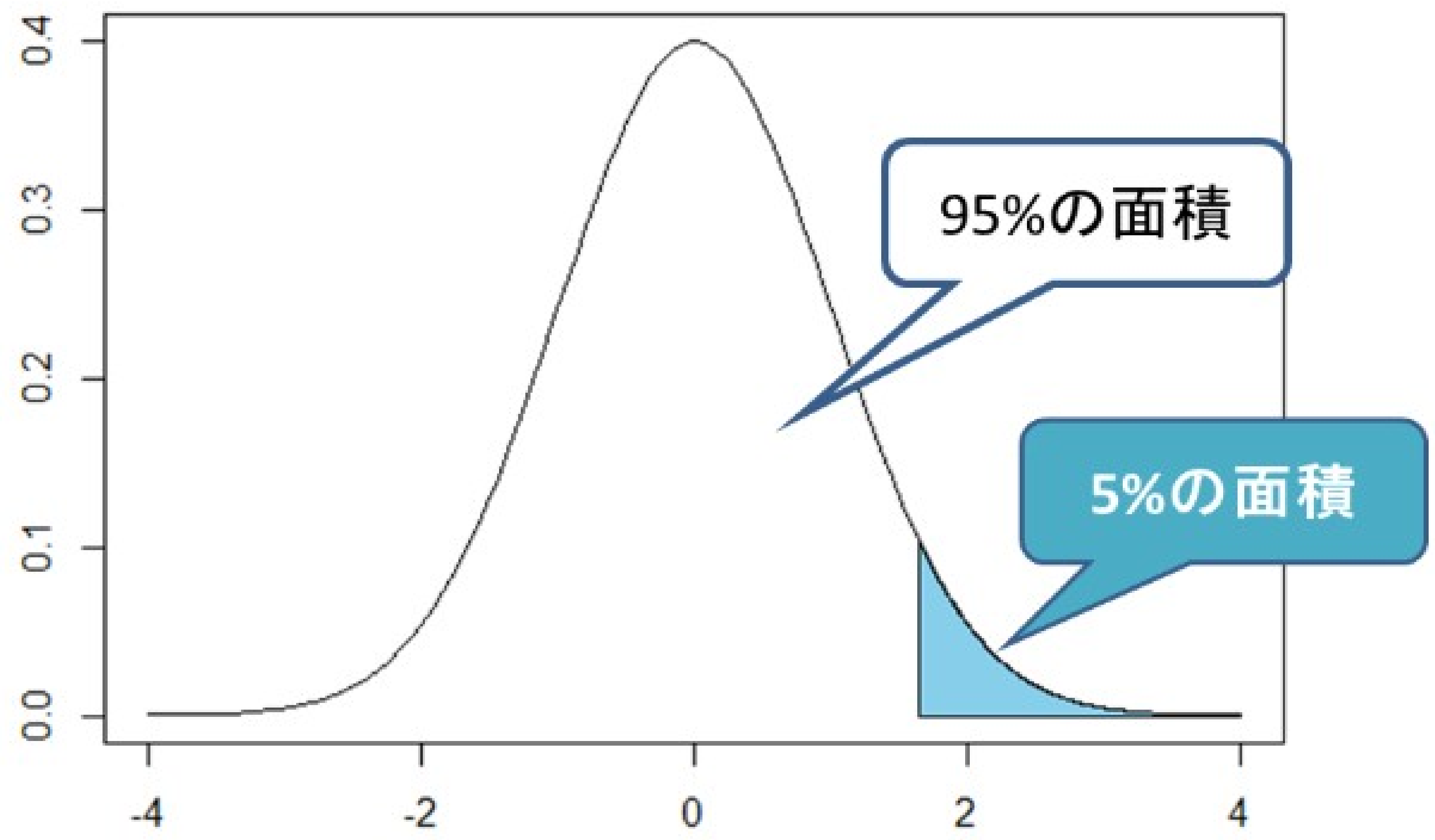


The rejection region is outside of the percentiles (both tails).
Reject H0 in case the content of B is extremely large or small.

Two-sided test / one-sided test

- In case of significance level of 5%:

One-sided （ Alternative hypothesis： The content of B is larger than 100mg ）

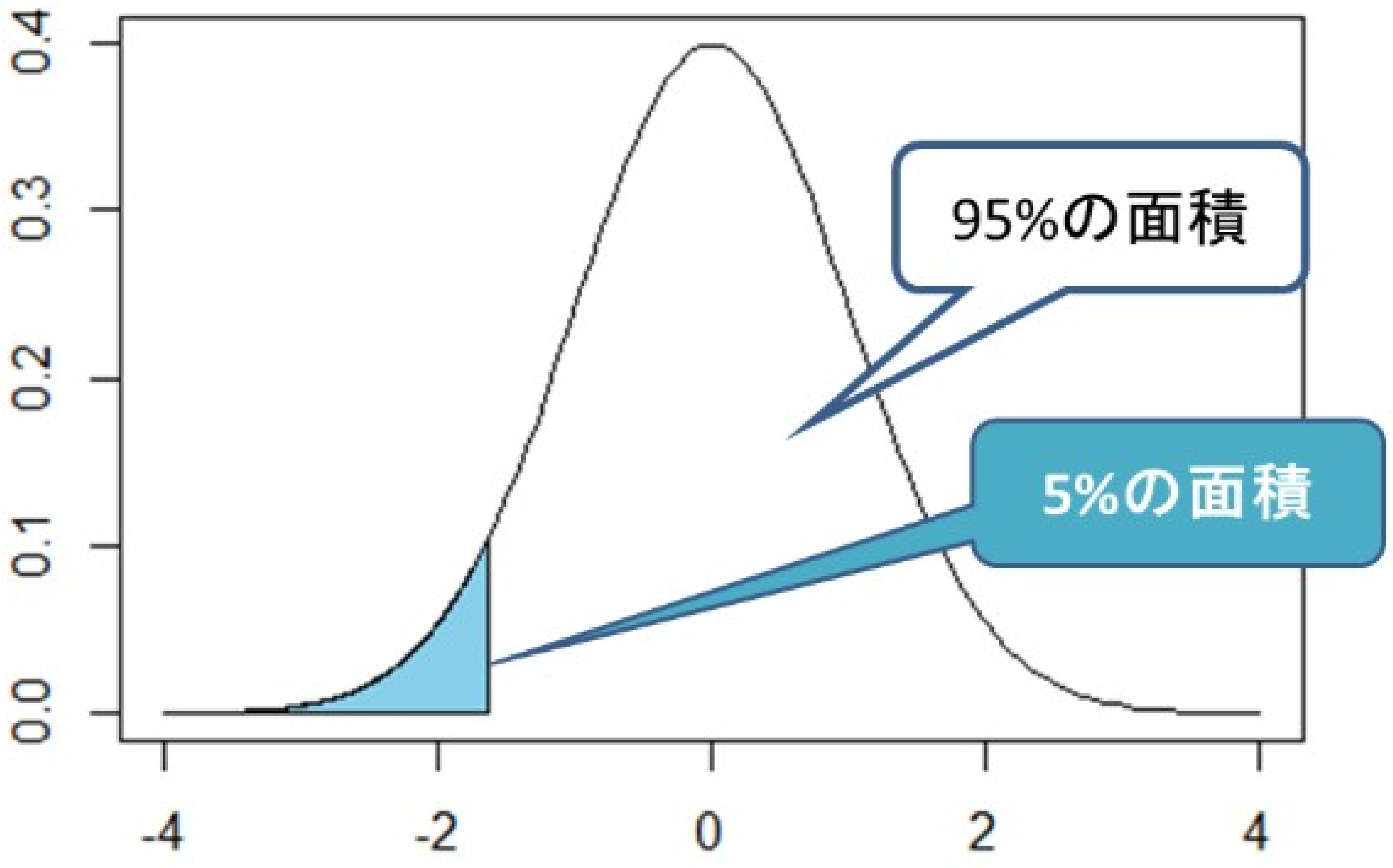


The rejection region is the right part of the tail. Reject H0 in case the content of B is extremely larger than 100mg.

Two-sided test / one-sided test

- In case of significance level of 5%:

One-sided （ Alternative hypothesis： The content of B is smaller than 100mg ）



The rejection region is the left part of the tail. Reject H0 in case the content of B is extremely smaller than 100mg.

Which you choose?

You should determine which of two- / one- sided test is used in advance.

Trying another way of test after obtaining a result is a serious mistake!

Examples.

Example of binomial test

- After you tossed a coin 20 times, a specific side appeared 15 times. Can you say that both sides appear equivalently?

cl1

Intuitively, it’s distorted.
→ Let us do the hypothesis test.

Example of binomial test

- Let us denote the probability that a specific side appears as *p*, and verify the null hypothesis H_0: "p=1/2".

On the basis of the hypothesis above, the probability that a $P(\{X \le 5\} \bigcup \{X \ge 15\} | p = 1/2) = 2 \sum_{i=15}^{20} {}_{20}C_i p^i (1-p)^{20-i} = 0.042$

→ H_0 is rejected under the significance level of 5%.

(i.e., the coin is distorted.)

You can't reject H_0 under the significance level of 1%

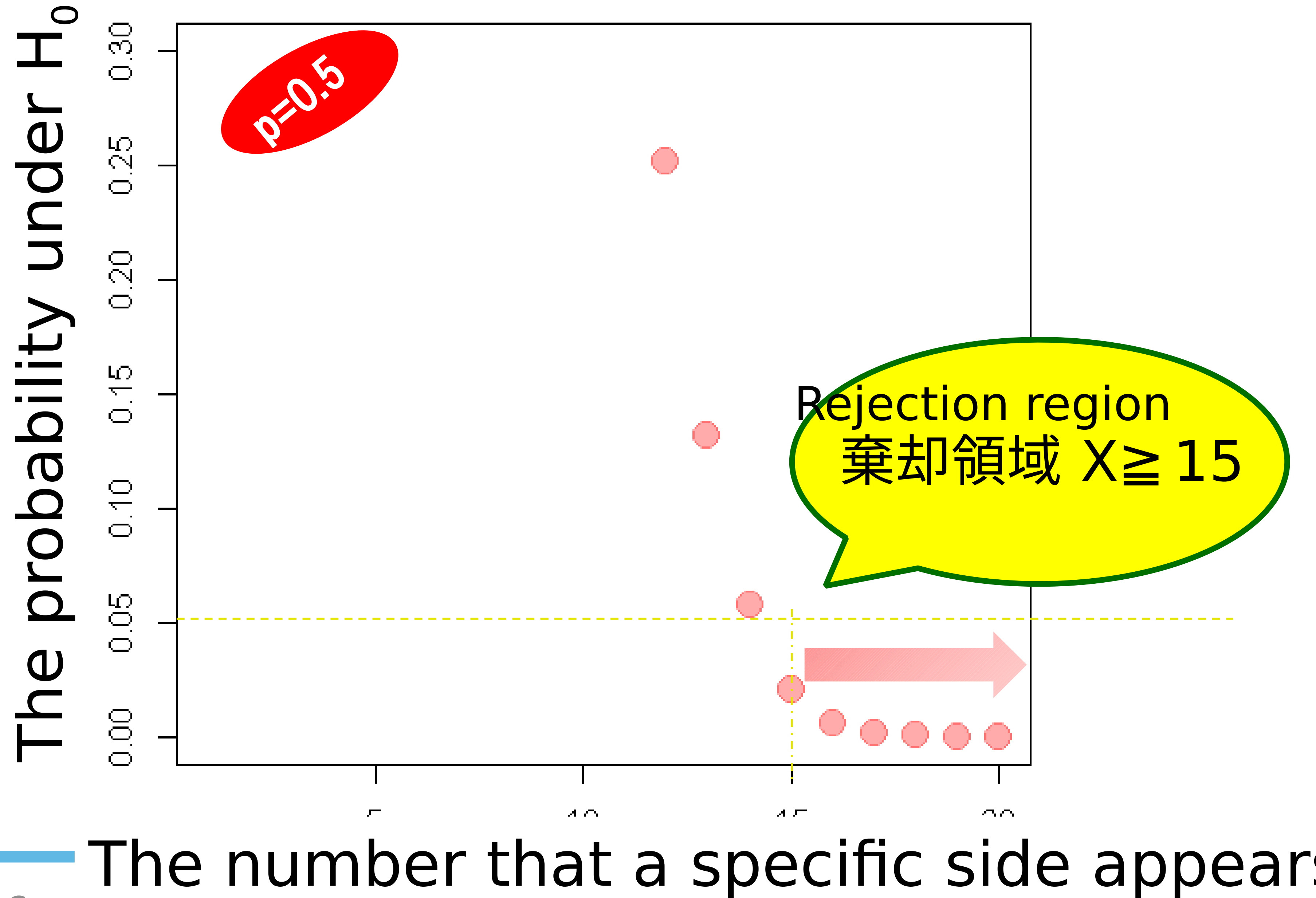
With N large, we can approximate the p.d.f by N(np,np(1-p)).

One- / two- sided test

- In this example, the null hypothesis will be rejected under the too many appearance of either side of the coin.

Rejection region

- Rjection region means the range of a r.v. X, where the null hypothesis is rejected. In the former example, “a specific side appears more than 15 times ” is the rejection region.



Statistical power

- New issues:
- By the way, if the coin is actually distorted, can we correctly detect that through the trial of 20 times coin tossing?
- Let us assume, for instance, p=0.6.
- Either side of the coin should appear 15 times or more so that the incorrect null hypothesis “p=0.5” is correctly rejected.
- → Find the probability of such situations under the situation “p=0.6”.

Statistical power

- → The probability of rejecting H_0 under “p=0.6”.

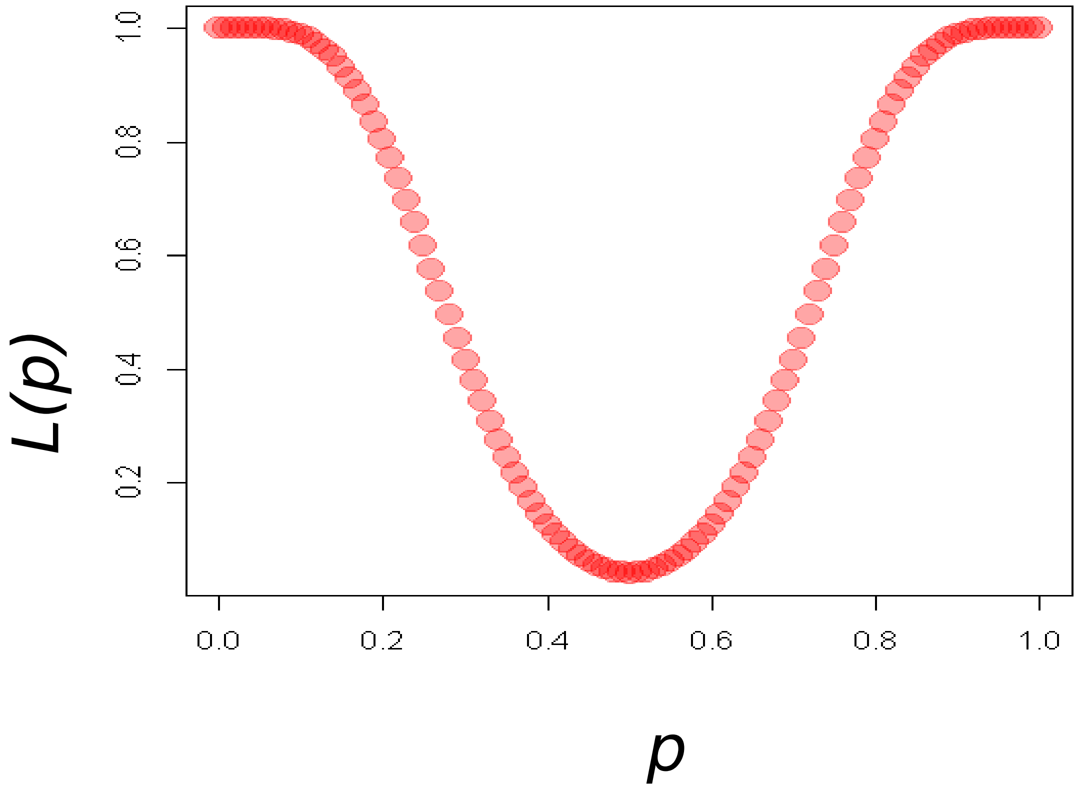
$$P(\{X \leq 5\} \bigcup \{X \geq 15\} | p = 0.6) =$$
$$= \sum_{i=15} {}_{20}C_i \{p^i (1-p)^{20-i} + p^{20-i}(1-p)^i\}$$
$$= 0.125$$

- →We can reject H_0 with the probability of only 12.5%..

Power function

- As p changes, power also changes.
- → Power function *L(p)*

$$L(p) = \sum_{i=15}^{20} {}_{20}C_i \{p^i(1-p)^{20-i} + p^{20-i}(1-p)^i\}$$



Requirements for power

- Requirement:
- “We should detect the distortion of the coin with the probability of 90% if p=0.6.”

- How can we attain this?

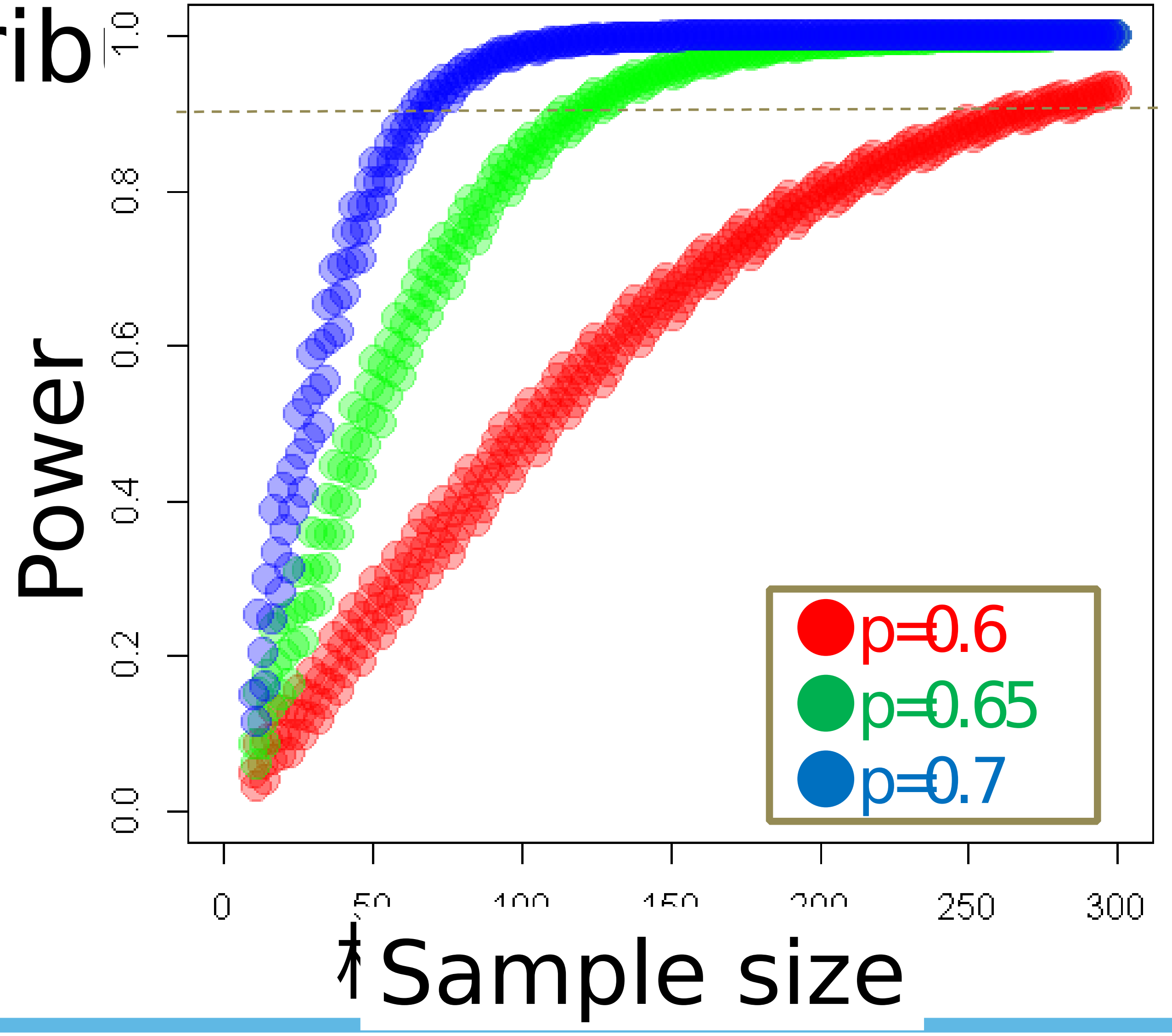
- → Increase the sample size!

Sample size and power

- For each p, the corresponding power is plotted.
- If p=0.6, we can attain the requirement with the sample size of 260 or more.

  → With such a large sample size, we can approximately employ the normal distrib

- As p increases, we can attain the requirement with fewer sample size.



3. Distributions of test statistic

Distributions of test statistic

- We will often use:
- T-test, χ^2-test， and F-test.
- These names come from t-dist.， χ^2-dist.， and F-dist.

3-1. Normal distribution

On test statistic

- A function $T=f(X_1, X_2, \ldots, X_n)$ of r.v.s $X_1, X_2, \ldots, X_n$ is called as test statistic.
- 【Typical example】
- Let r.v.s $X_1, X_2, \ldots, X_n$ be independent, and follow the distributions $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, ..., $N(\mu_n, \sigma_n^2)$.
- Then, $T=a_1X_1+a_2X_2+\ldots+a_nX_n$

follows

$$N(a_1\mu_1+a_2\mu_2+\ldots+a_n\mu_n,\ a_1\sigma_1^2+a_2\sigma_2^2+\ldots+a_n\sigma_n^2).$$

Sum of normal dist.

- As a special case of the former slide,
- 【Special case】
- Let X_1, X_2, …,X_n be independent and follow the same dist.： N（μ， σ²） . Then,

$$T=(X_1+X_2+\ldots+X_n)/n$$

follows

$$N(\mu, \sigma^2/n)$$

3-2. t-dist.

T-distribution

- 【Ex】
- Let r.v.s X_1, X_2, ...,X_n be independent with each other, and subject to N（μ， σ²）. Then, the quantity

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

is subject to t-distribution of (n-1) degree of freedom. Here,

$$s = \sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n-1}}$$

t-distribution

- In other words…
- Let Z ~ N(0, 1) and W be subject to χ^2 -distribution of n degree of freedom. We also assume that they are independent of each other. Then, the following quantity is subject to t-distribution of n degree of freedom.

$$t = \frac{Z}{\sqrt{\frac{W}{n}}}$$

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{S^2}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{\sigma^2}}{n-1}}} = \frac{Z}{\sqrt{\frac{W}{n-1}}}$$

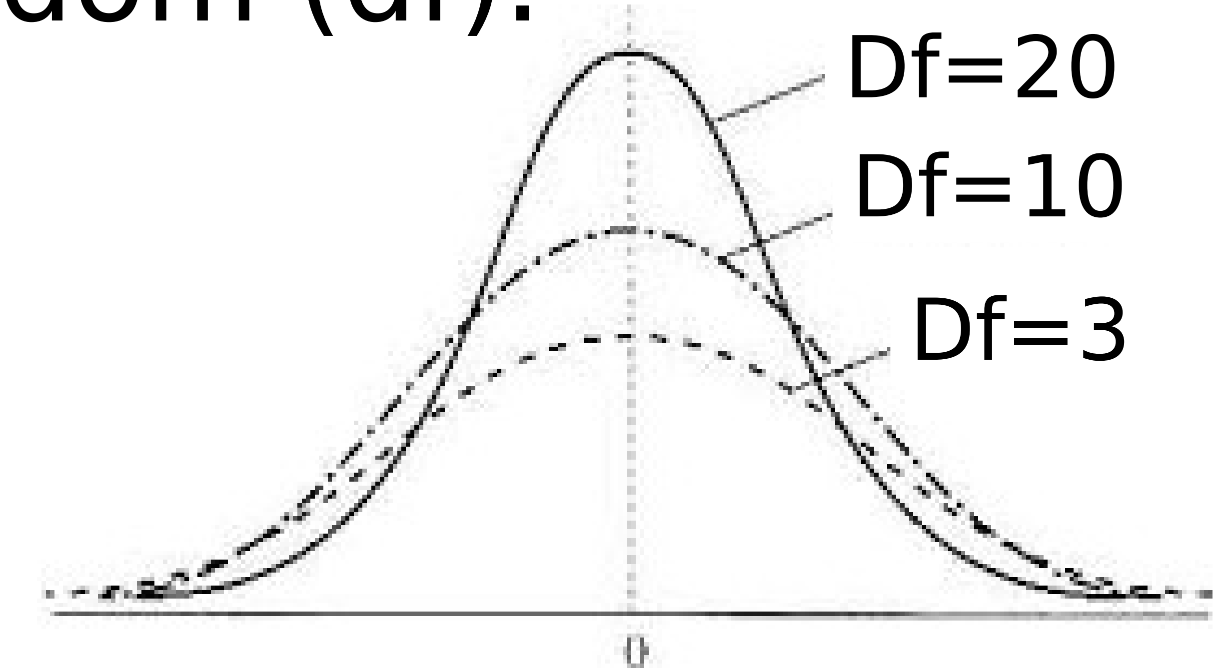
t-distribution

- T-distribution has degree of freedom.
- Used for the interval estimation / hypothesis testing of population mean.
- 【Probability density】
- The probability density of t-distribution of n degree of freedom is

$$f(x;n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}\Gamma\left(\frac{n}{2}\right)}$$

t-distribution

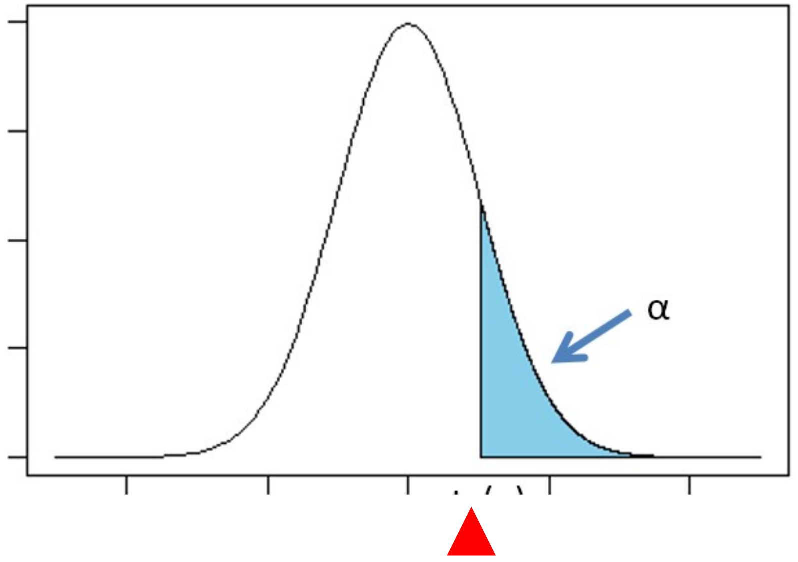
- Probability density of t-distributions of various degree of freedom (df).



  - Symmetric with respect to x=0 (as z-dits.) !
  - Asymptotically tends to z-dist. as df=n→∞.
  - If df=n is n large (n≧30, for instance), can be regarded as z-dist.

Percentile of t-distribution

- We denote t-distribution of n degree of freedom as t_n hereafter.
- It's upper α-percentile is denoted as $t_n(\alpha)$.
  - Ex :
  - Upper 5-percentile of t-distribution of df=5.



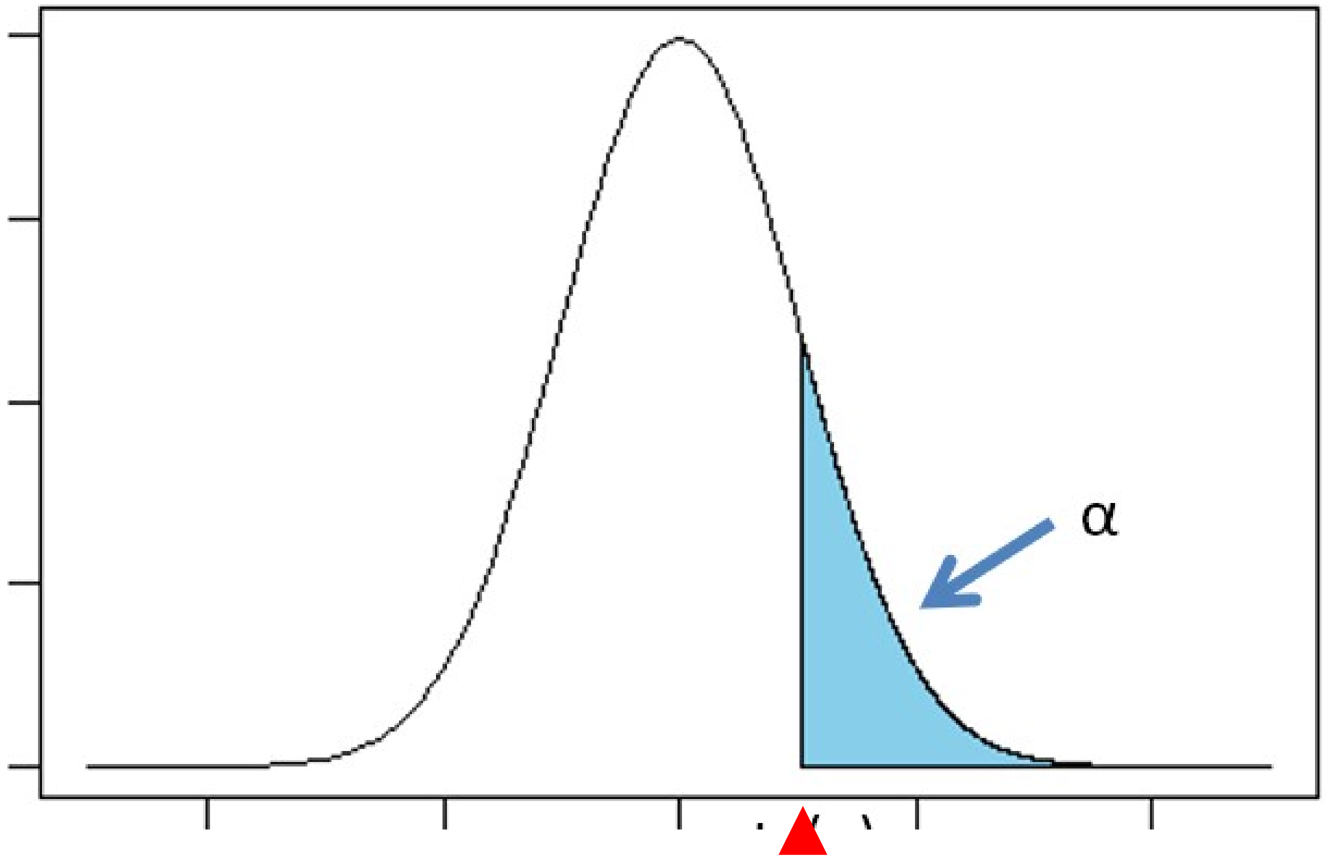
$t_5(0.05)=2.015$

How to find percentile?

- Tables （z-table / t-table）
- Python

Percentile of t-distribution

- For instance, the upper 2.5-percentile of t-distribution of df=5 is about 2.57.



$t_5(0.025)=2.57$

- Using python;
  - E.g.） Find the upper 2.5-percentile of t-dist. With df=9.

```
from scipy.stats import t
t.ppf(0.975,9)
```

```
2.2621571627409915
```

T-table

- In making 95% confidence interval, upper 2.5-percentile is needed.
- Therefore, you should look into the pink column of “2.5%” in “one-side(片側)”

	有意確率								
	0.10	0.05	0.01	0.001	両側	0.10	0.05	0.01	0.001
df	0.05	0.025	0.005	0.0005	片側	0.05	0.025	0.005	0.0005
1	6.3138	12.706	63.657	636.62	18	1.7341	2.1009	2.8784	3.922
2	2.9200	4.3027	9.9248	31.598	19	1.7291	2.0930	2.8609	3.883
3	2.3534	3.1825	5.8409	12.941	20	1.7247	2.0860	2.8453	3.850
4	2.1318	2.7764	4.6041	8.610	21	1.7207	2.0796	2.8314	3.819
5	2.0150	2.5706	4.0321	6.859	22	1.7171	2.0739	2.8188	3.792
6	1.9432	2.4469	3.7074	5.959	23	1.7139	2.0687	2.8073	3.767
7	1.8946	2.3646	3.4995	5.405	24	1.7109	2.0639	2.7969	3.745
8	1.8595	2.3060	3.3554	5.041	25	1.7081	2.0595	2.7874	3.725
9	1.8331	2.2622	3.2498	4.781	26	1.7056	2.0555	2.7787	3.707
10	1.8125	2.2281	3.1693	4.587	27	1.7033	2.0518	2.7707	3.690
11	1.7959	2.2010	3.1058	4.437	28	1.7011	2.0484	2.7633	3.674
12	1.7823	2.1788	3.0545	4.318	29	1.6991	2.0452	2.7564	3.659
13	1.7709	2.1604	3.0123	4.221	30	1.6973	2.0423	2.7500	3.646