CS 215- Data Interpretation and Analysis
(Post Midsem)

Pushpak Bhattacharyya
Computer Science and Engineering Department
IIT Bombay

Lecture-5

Hypothesis Testing cntd

19oct23

Recap

Terminology for Test of hypothesis

Terminology

Null and Alternative Hypothesis

- H_0: Null Hypothesis → the hypothesis we want to reject

- H_A or H_1: Alternative Hypothesis → opposite of H_0

- We use the sample statistics, trying to reject H_0

H_0 and H_A for manufacturing-part problem

Dimensions

5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5

Variance: 4 (somehow known)

Sample mean: 9

Company “claims” av. dimension as 15

H_0: Dim>=15, H_A: Dim < 15 (one sided)

Type I and Type II error

- Type I: incorrectly reject H_0, when it should have been accepted.

- Type II: incorrectly accept H_0 when it should have been rejected.

More on H_0

- The data would be unlikely to occur if the null hypothesis were true.

- In logical form:

$$N_H \ \vdash \sim D$$

- Where N_H is the proposition "null hypothesis true" and D is the proposition "Data occurs"

Digression: Hypothesis Testing in Logic

Using Predicate Calculus

Himalayan Club example

- Introduction through an example *(Zohar Manna, 1974):*
  - Problem: A, B and C belong to the Himalayan club. Every member in the club is either a mountain climber or a skier or both. A likes whatever B dislikes and dislikes whatever B likes. A likes rain and snow. No mountain climber likes rain. Every skier likes snow. *Is there a member who is a mountain climber and not a skier?*
- Given knowledge has:
  - Facts
  - Rules

Example contd.

- Let *mc* denote mountain climber and *sk* denotes skier. Knowledge representation in the given problem is as follows:
  1. *member(A)*
  2. *member(B)*
  3. *member(C)*
  4. $\forall x[member(x) \rightarrow (mc(x) \vee sk(x))]$
  5. $\forall x[mc(x) \rightarrow \sim like(x,rain)]$
  6. $\forall x[sk(x) \rightarrow like(x, snow)]$
  7. $\forall x[like(B, x) \rightarrow \sim like(A, x)]$
  8. $\forall x[\sim like(B, x) \rightarrow like(A, x)]$
  9. *like(A, rain)*
  10. *like(A, snow)*
  11. *Question:* $\exists x[member(x) \wedge mc(x) \wedge \sim sk(x)]$
- We have to infer the 11th expression from the given 10.
- Done through Resolution Refutation.

Club example: Inferencing

1. $member(A)$
2. $member(B)$
3. $member(C)$
4. $\forall x[member(x) \rightarrow (mc(x) \vee sk(x))]$
   - Can be written as $[member(x) \rightarrow (mc(x) \vee sk(x))]$
   - $\sim member(x) \vee mc(x) \vee sk(x)$
5. $\forall x[sk(x) \rightarrow lk(x, snow)]$
   - $\sim sk(x) \vee lk(x, snow)$
6. $\forall x[mc(x) \rightarrow \sim lk(x, rain)]$
   - $\sim mc(x) \vee \sim lk(x, rain)$
7. $\forall x[like(A, x) \rightarrow \sim lk(B, x)]$
   - $\sim like(A, x) \vee \sim lk(B, x)$

8. $\forall x[\sim lk(A,x) \to lk(B,x)]$

– $lk(A,x) \vee lk(B,x)$

9. $lk(A, rain)$

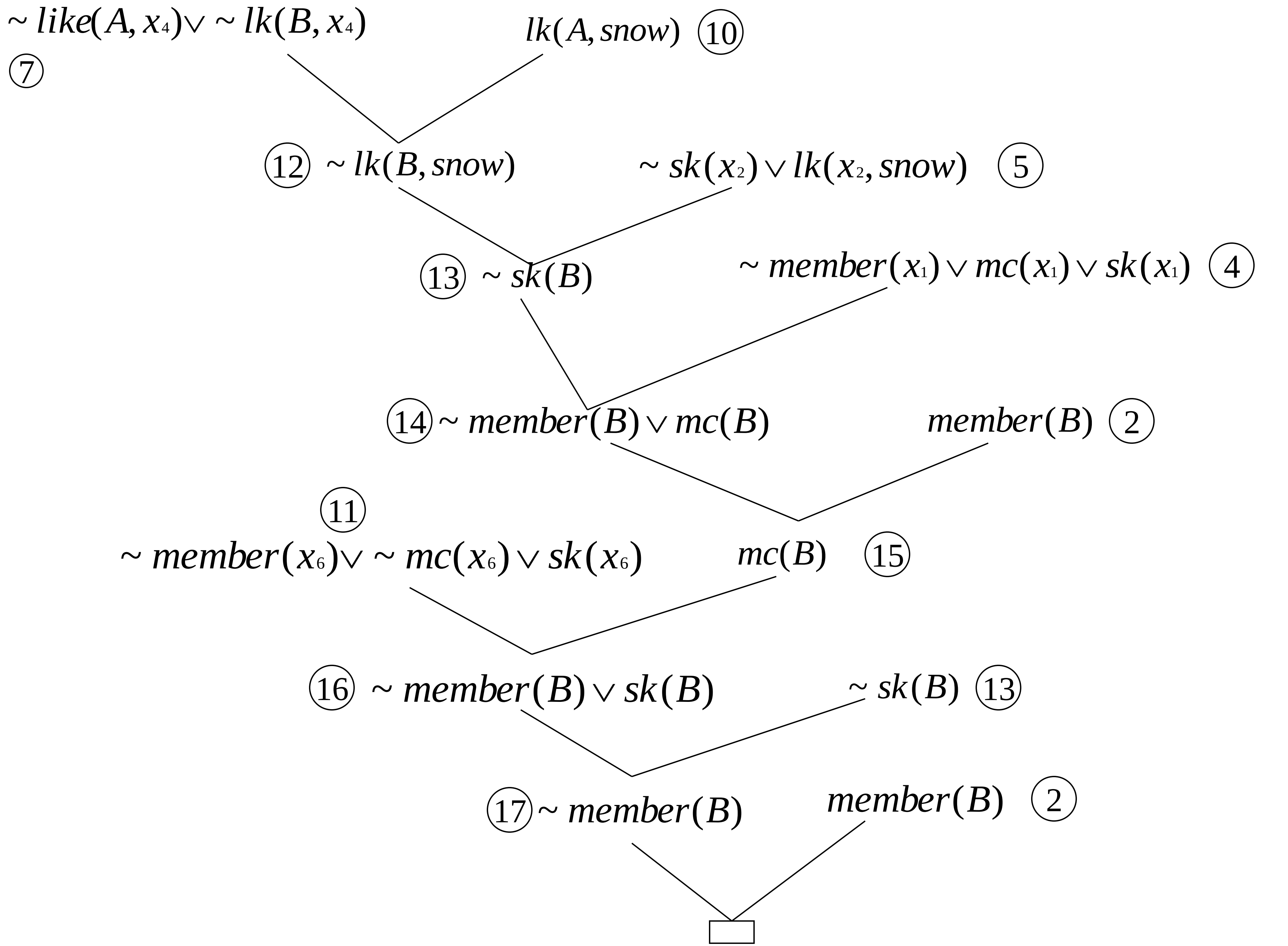
10. $lk(A, snow)$

11. $\exists x[member(x) \wedge mc(x) \wedge \sim sk(x)]$

– Negate– $\forall x[\sim member(x) \vee \sim mc(x) \vee sk(x)]$

- Now standardize the variables apart which results in the following

1. $member(A)$
2. $member(B)$
3. $member(C)$
4. $\sim member(x_1) \vee mc(x_1) \vee sk(x_1)$
5. $\sim sk(x_2) \vee lk(x_2, snow)$
6. $\sim mc(x_3) \vee \sim lk(x_3, rain)$
7. $\sim like(A, x_4) \vee \sim lk(B, x_4)$
8. $lk(A, x_5) \vee lk(B, x_5)$
9. $lk(A, rain)$
10. $lk(A, snow)$
11. $\sim member(x_6) \vee \sim mc(x_6) \vee sk(x_6)$


$\sim like(A, x_4) \vee \sim lk(B, x_4)$ 7
$lk(A, snow)$ 10
12 $\sim lk(B, snow)$
$\sim sk(x_2) \vee lk(x_2, snow)$ 5
13 $\sim sk(B)$
$\sim member(x_1) \vee mc(x_1) \vee sk(x_1)$ 4
14 $\sim member(B) \vee mc(B)$
$member(B)$ 2
11
$\sim member(x_6) \vee \sim mc(x_6) \vee sk(x_6)$
$mc(B)$ 15
16 $\sim member(B) \vee sk(B)$
$\sim sk(B)$ 13
17 $\sim member(B)$
$member(B)$ 2

Null Hypothesis: H_0

- H_0: The club does NOT have any member that is a mountain climber (MC) and not a skier (SK)

- Key question: Under H_0, is the observation valid?

- In other words: is the hypothesis consistent with the data?

Methodology

- If Hypothesis not consistent with data, hypothesis must be rejected
- Data cannot be rejected
- Data is GOLD!

Data for Himalayan Club Example

(1) A, B and C belong to the Himalayan club.

(2) Every member in the club is either a mountain climber or a skier or both.

(3) A likes whatever B dislikes and

(4) dislikes whatever B likes.

(5) A likes rain and snow.

(6) No mountain climber likes rain.

(7) Every skier likes snow

Null Hypothesis for Himalayan Club Example

- H_0: *There is NOT a single member who is a mountain climber and not a skier*
- H_0 inconsistent with data

- So must be rejected
- Methodology: Logical Inferencing-Resolution-Refutation

More on H_0

- Maximum Likelihood in action
- Move the ball to the “court” of observations
- Formulate H_0 in such a way that high probability of H_0 makes the data probability low

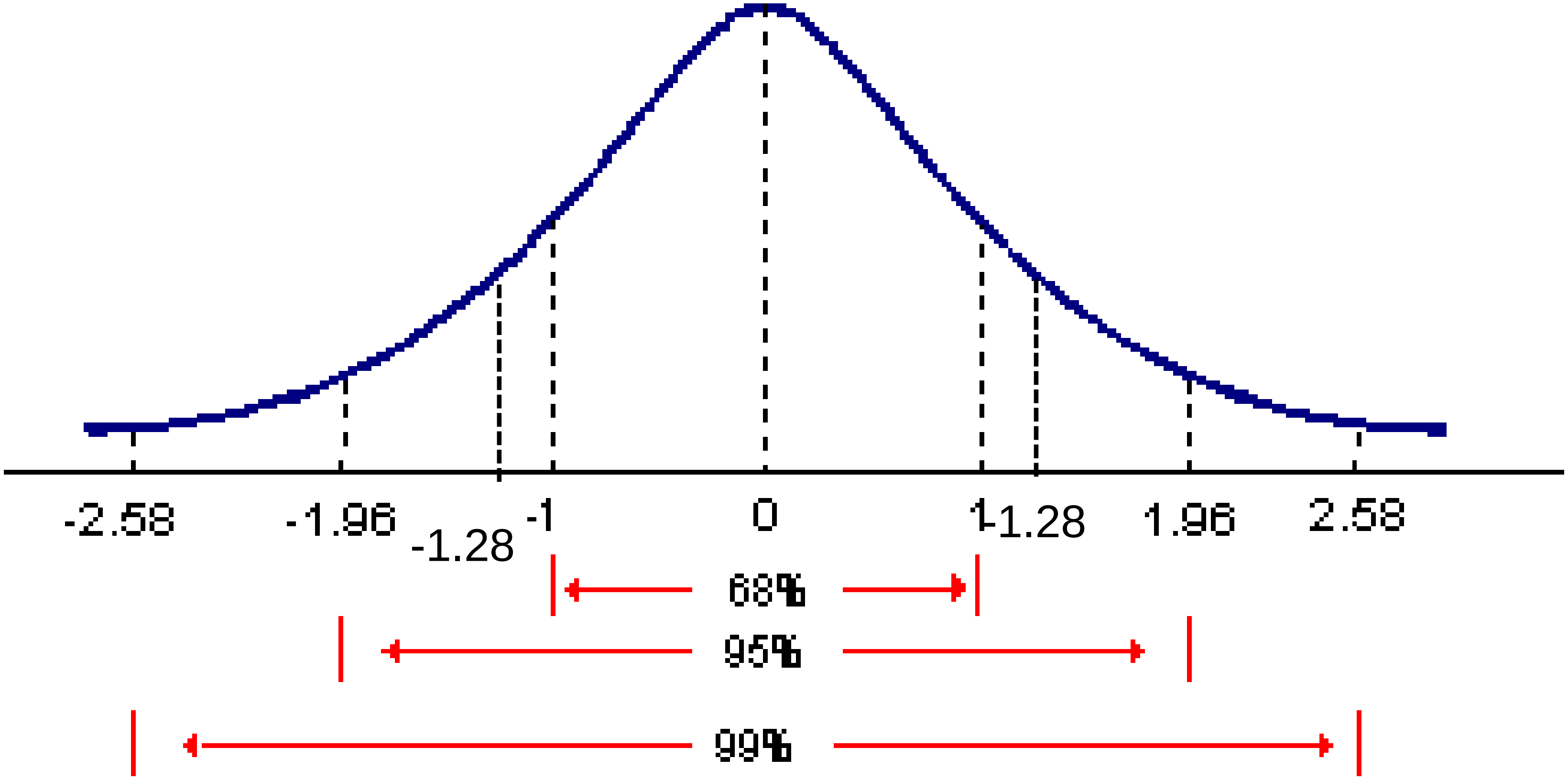
End Recap

Useful concepts for hypothesis testing

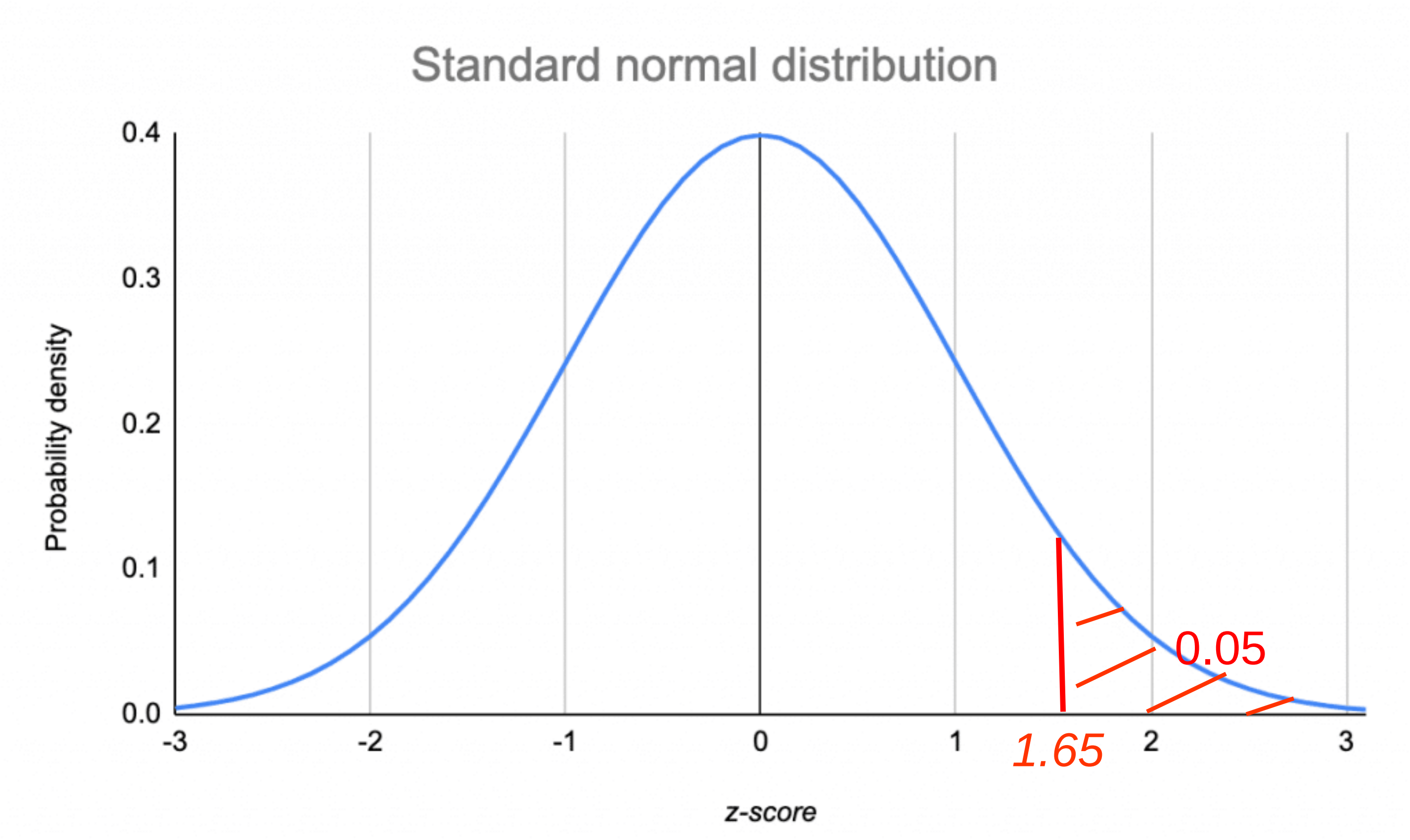
A useful table

test-type (col)			
vs.	**Two-Tail**	**1 sided to +inf**	**1 sided from -inf**
Confidence Interval (significance level)			
90% (0.10)	**(- and +) 1.65**	**-1.28 to +inf**	**-inf to +1.28**
95% (0.05)	**(- and +) 1.96**	**-1.65 to +inf**	**-inf to +1.65**
99% (0.01)	**(- and +) 2.58**	**-2.33 to +inf**	**-inf to 2.33**

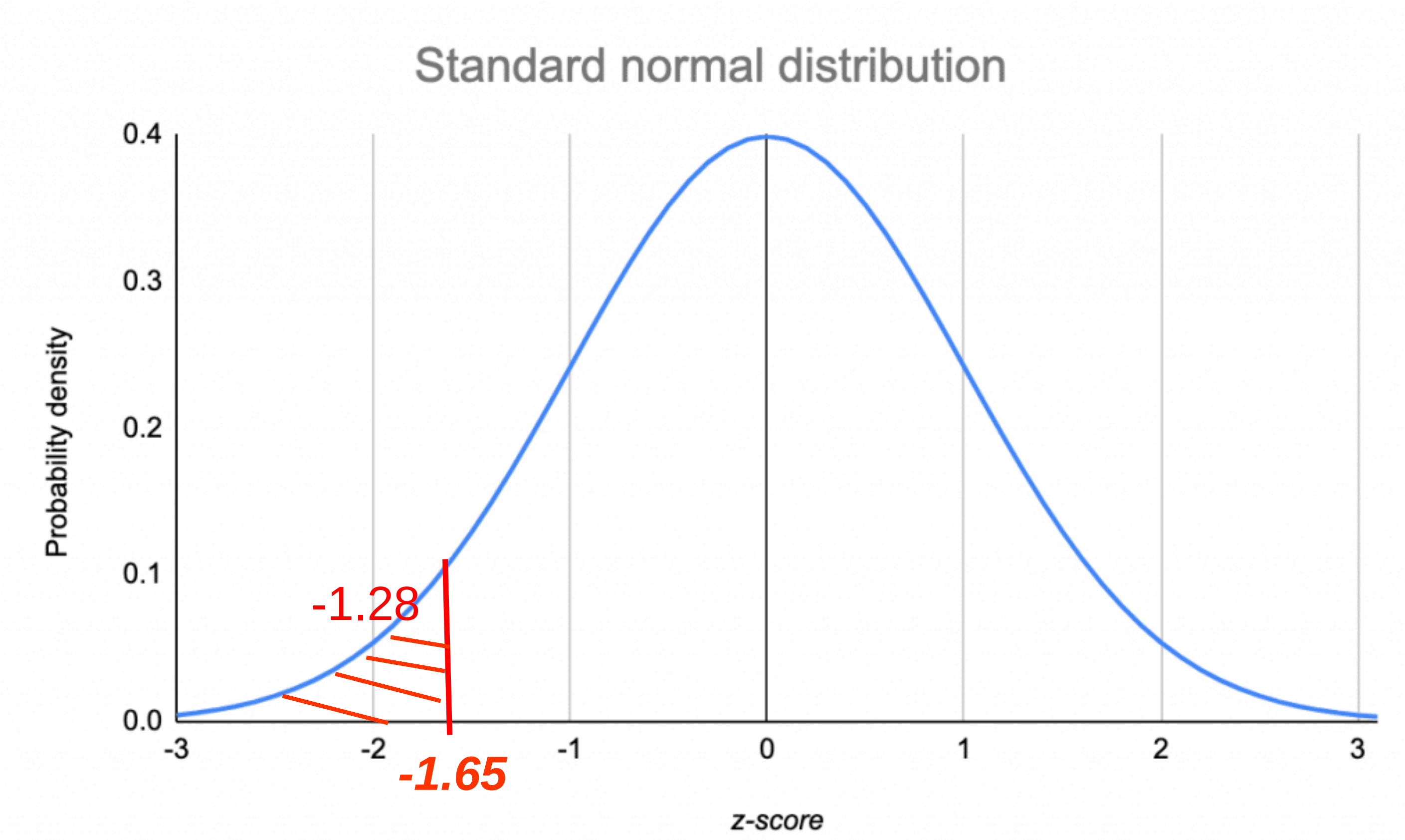
2 sided 95% confidence interval



1-sided confidence interval (upper/right)



1-sided confidence interval (lower/left)



Illustration

Problem Statement: bottling of fluid

- A factory has a machine that- the factory claims- dispenses 80mL of fluid in a bottle. This needs to be tested. A sample of 40 bottles is taken. The average amount of fluid is 78mL with standard deviation of 2.5. Verify the factory's claim.

https://www.youtube.com/watch?v=zJ8e_wAWUzE

Essential elements (1/n)

- Examine the problem carefully, read the problem statement again and again, discuss the issue threadbare
- ***1. Formulate H_0***
- 2. Formulate H_A
- 3. *Decide confidence interval* (usually 90, 95 or 99%)→ this automatically fixes the significance level (0.10, 0.05 or 0.01)

This sets the rule of the game, cannot change going forward !!

Essential elements (2/n)

- From H_0 and H_A, decide 2-sided or 1 sided test
  - 2 sided for ‘=‘ or ‘≠’
  - 1 sided for ‘>=’ or ‘=<‘

Essential elements (3/n)

- Decide Z-test/T-test/F-test/ChiSquare test
- If Z-test, Z_c (critical value)=

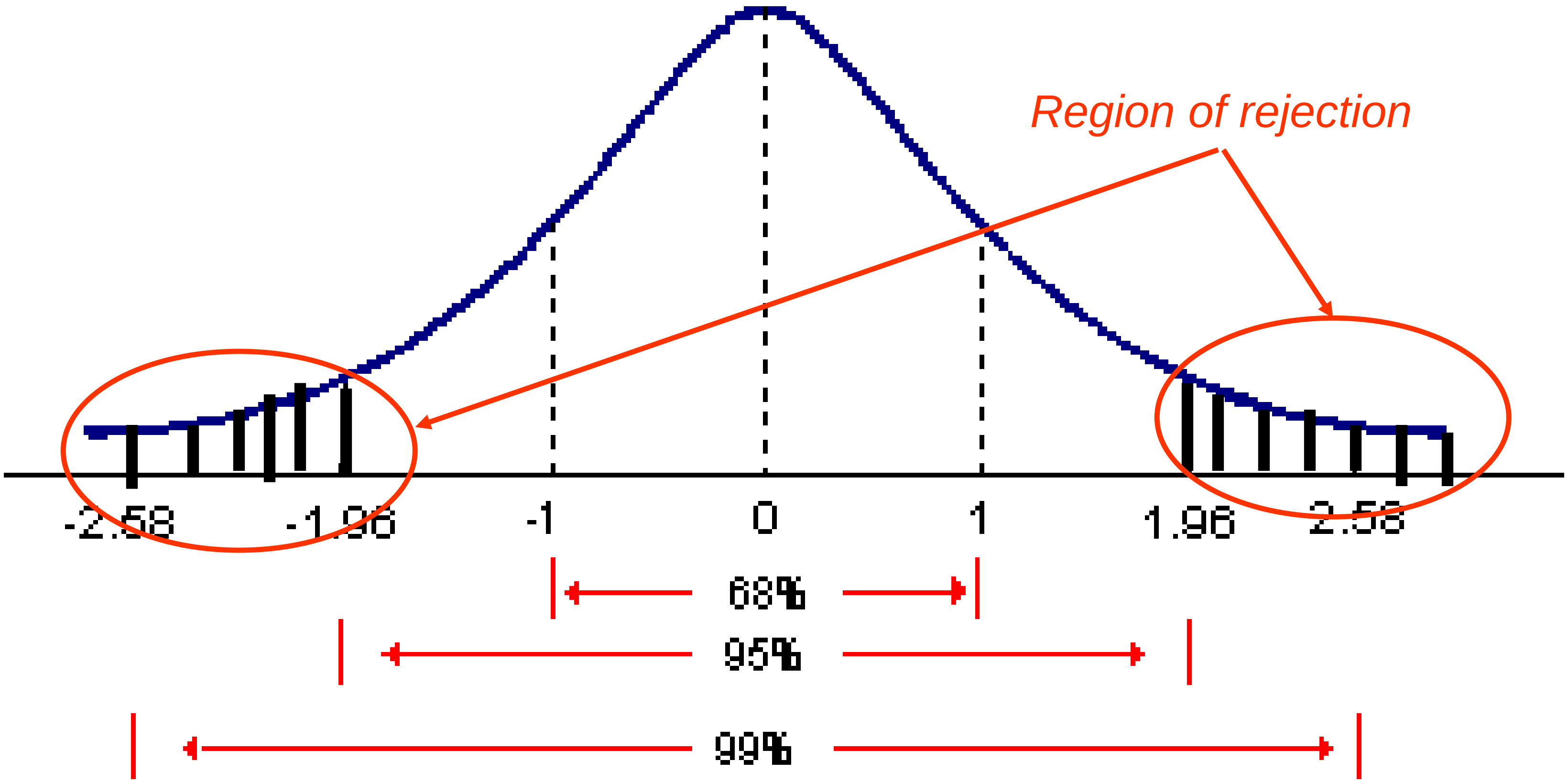
  +- 1.65 for 90% confidence interval,

  -inf to +1.28, -1.28 to +inf

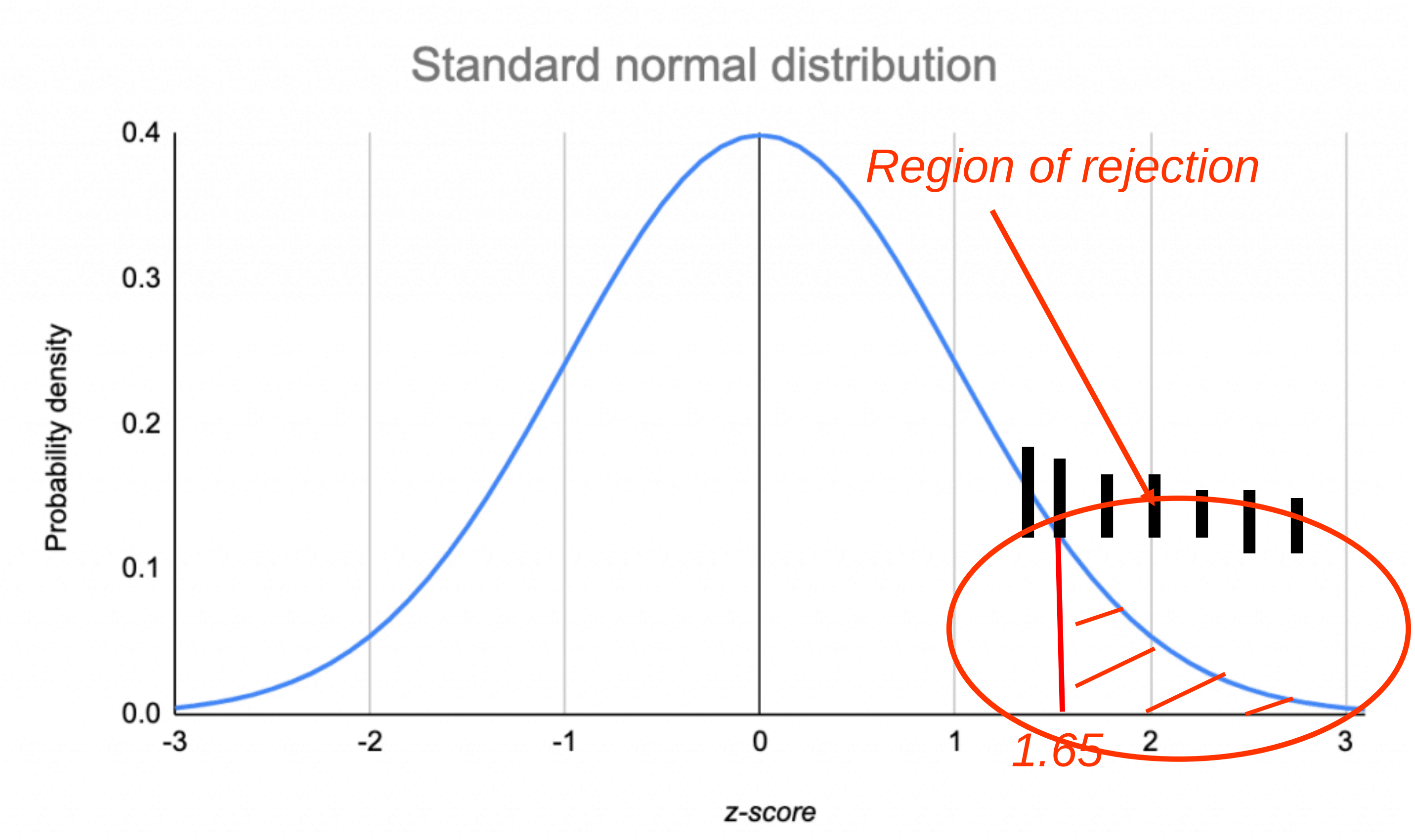
  +- 1.96 for 95% confidence interval, -inf to +1.65, -1.65 to +inf

  +-2.58 for 99% confidence interval, -inf to +2.33, -2.33 to +inf

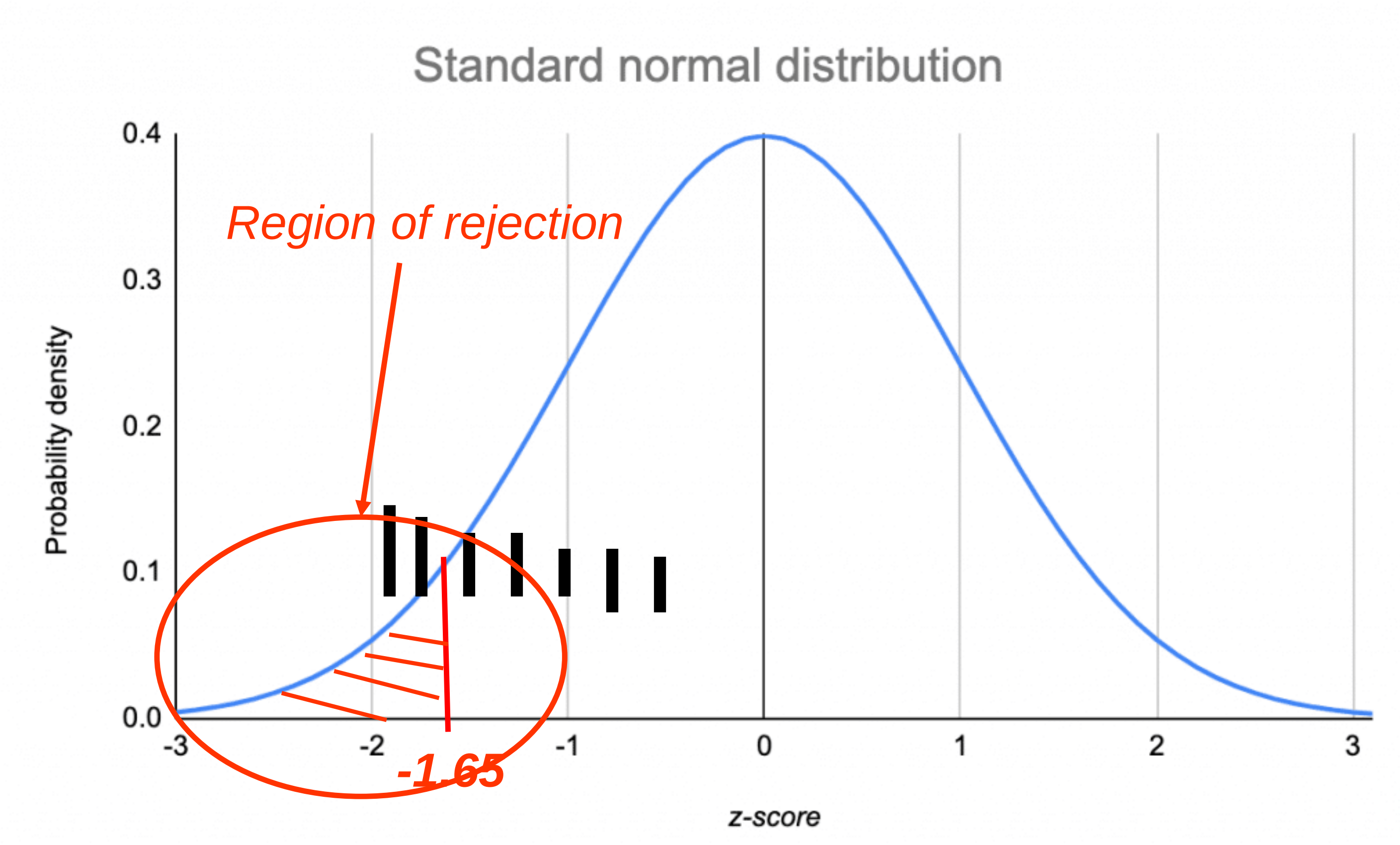
2 sided 95% confidence interval



1-sided confidence interval



1-sided confidence interval



Essential elements (4/n)

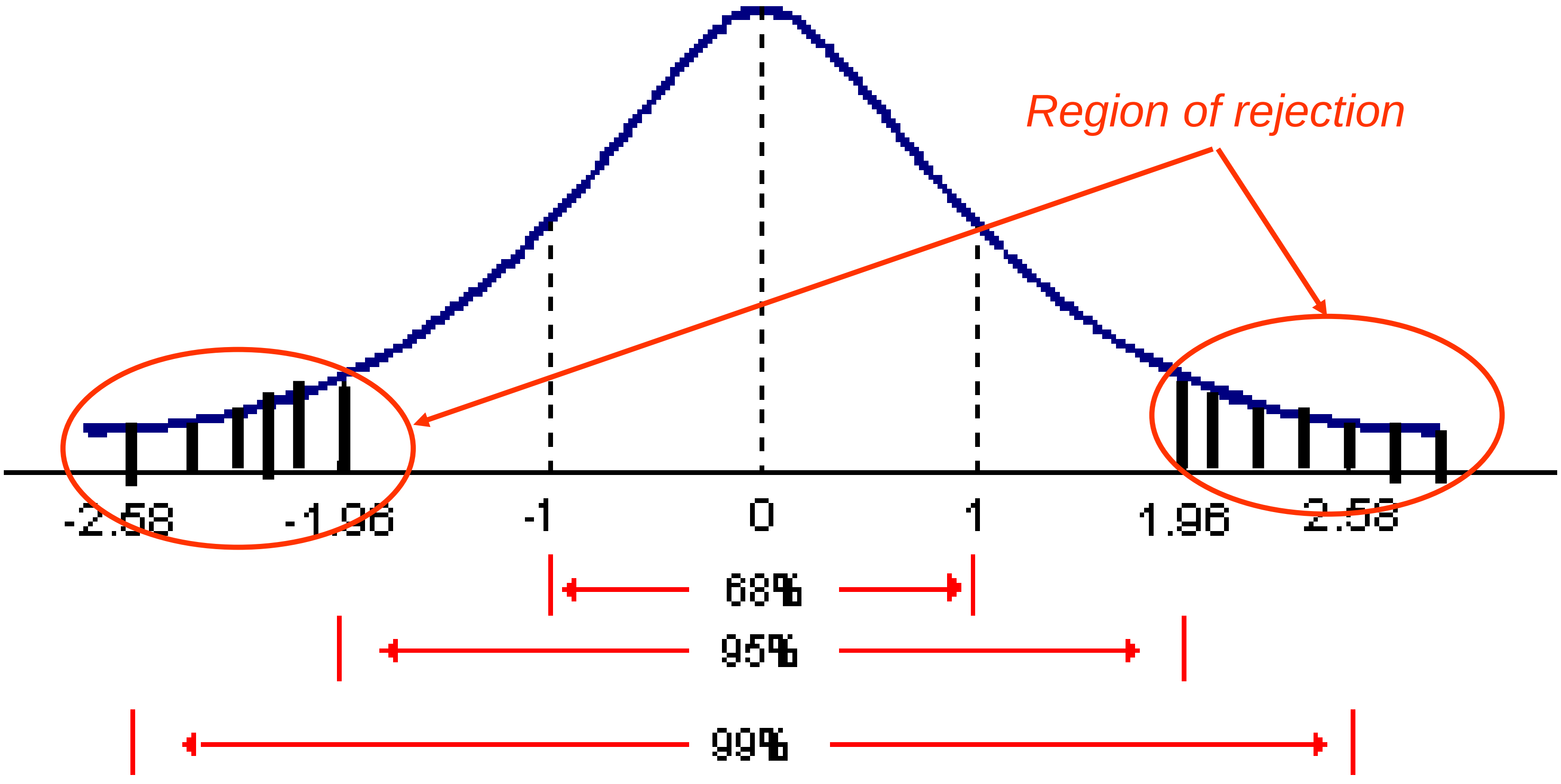
- **<u>Under H0</u>**, determine test statistic from the data called OBSERVED

- For Z-test, Z_o (observed)
- If Zo is inside the “REJECT” region, reject H_0
- Else cannot reject H_0

Back to Illustration: bottling of fluid

- Claimed population mean, $\mu=80$
- $n=40$, sample mean, $\mu_{obs}=78$, sample standard deviation, $\sigma_{obs}=2.5$
- H_0: $\mu=80$
- H_A: 3 options
  - $\mu \neq 80$ (2-sided test)
  - $\mu >= 80$ (1-sided)
  - $\mu <= 80$ (1-sided)

2 sided 95% confidence interval



2-tailed analysis

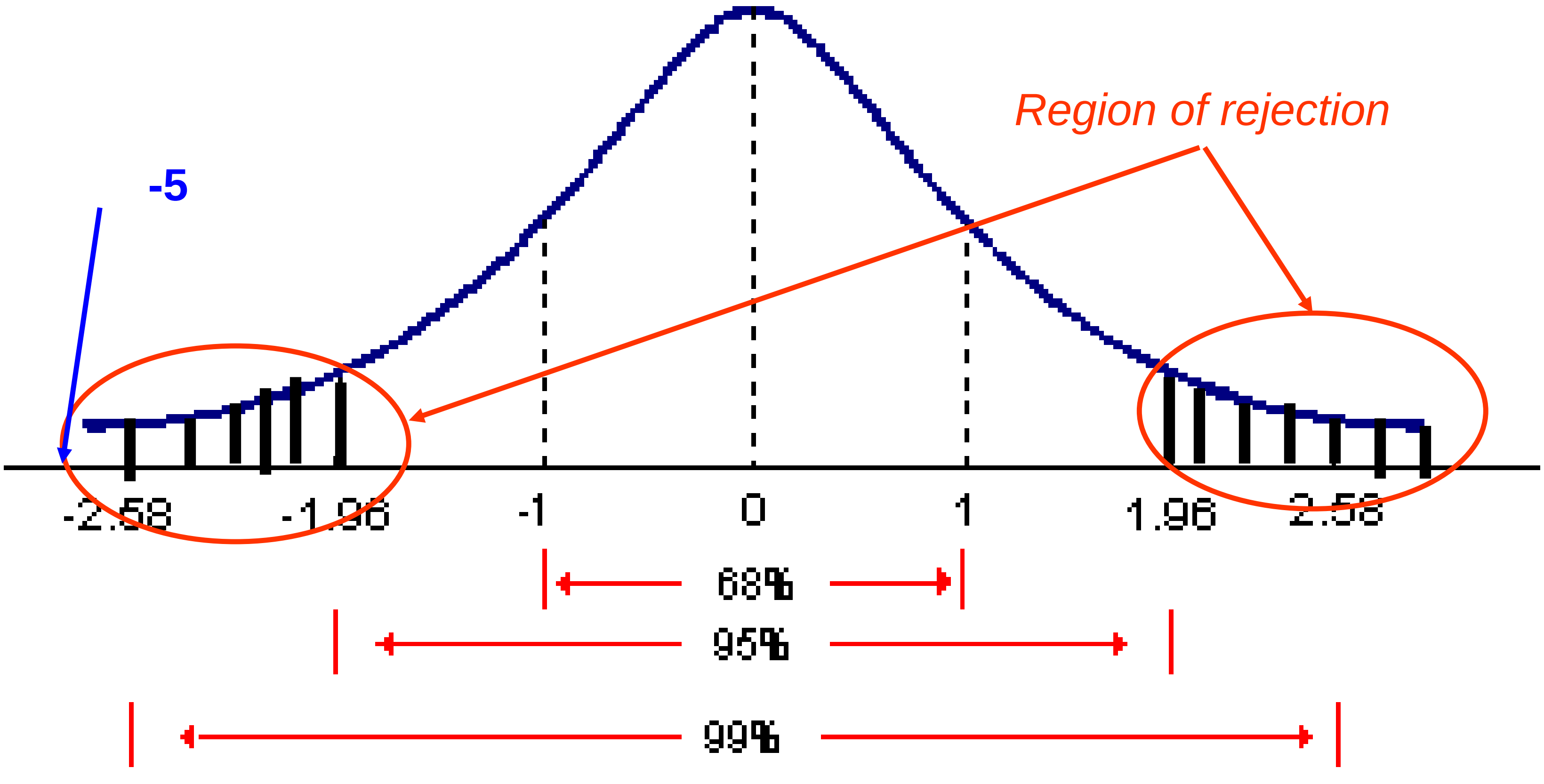
- Z_c= +- 1.96
- Z_{obs}=

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}=\frac{78-80}{\frac{2.5}{\sqrt{40}}}$$

$$=-5\,(approx.)$$

- Falls in rejection region

2 sided 95% confidence interval



Z-test based observation (2-tailed)

- $-5 < -1.96$
- We reject the null hypothesis
- The claim that the machine fills bottles with 80mL fluid is rejected based on the evidence

99% confidence interval

- -5 still in rejection region
- $-5.0 < -2.58$
- So for 99% confidence interval also the hypothesis is rejected

90% confidence interval

- -5 still in rejection region

- $-5.0 < -1.28$

- So for 90% confidence interval also the hypothesis is rejected

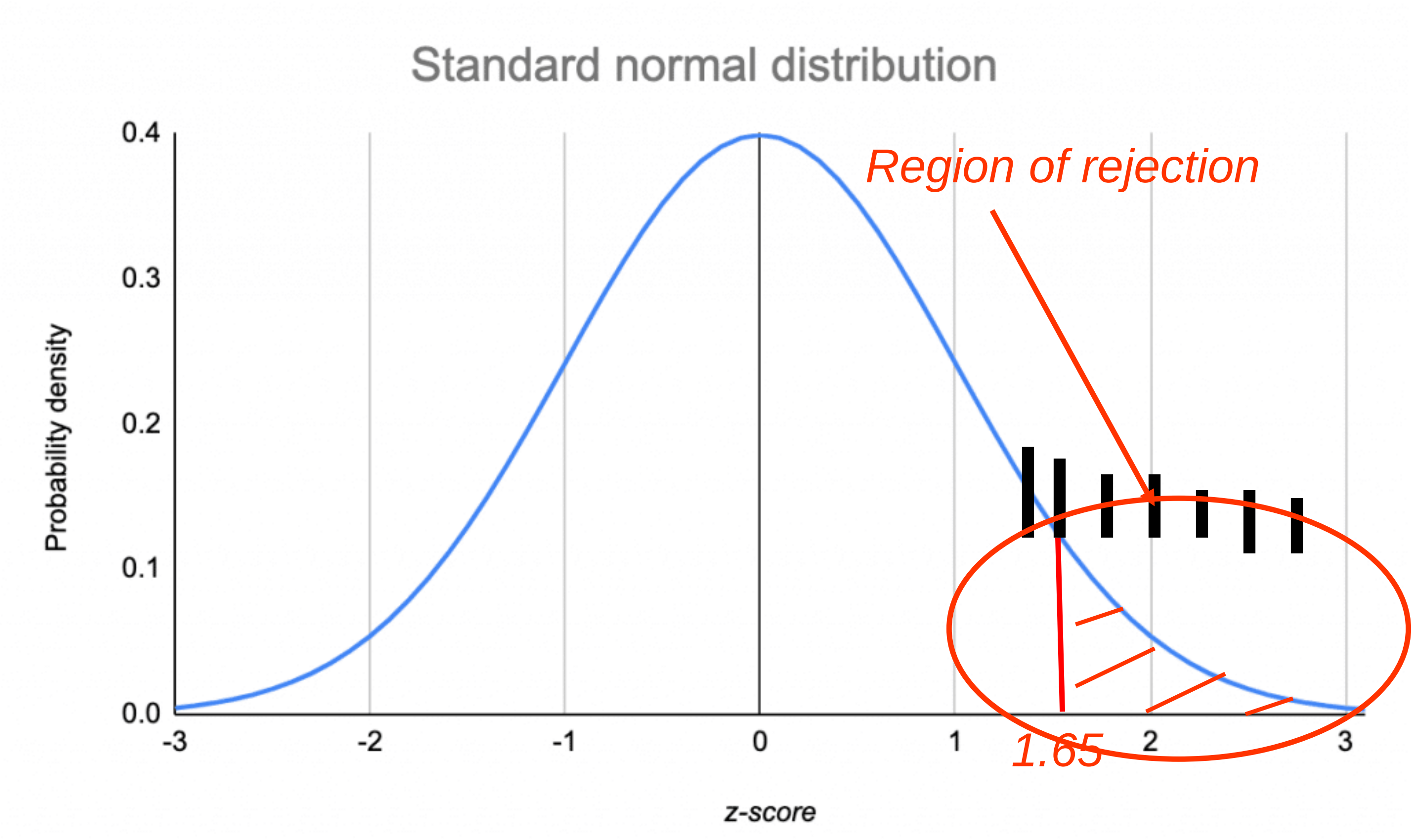
What about '>' test?

- H_A: The factory fills bottles with more than 80mL fluid
- Z_c= +1.65 or -1.65
- *what can we conclude?*

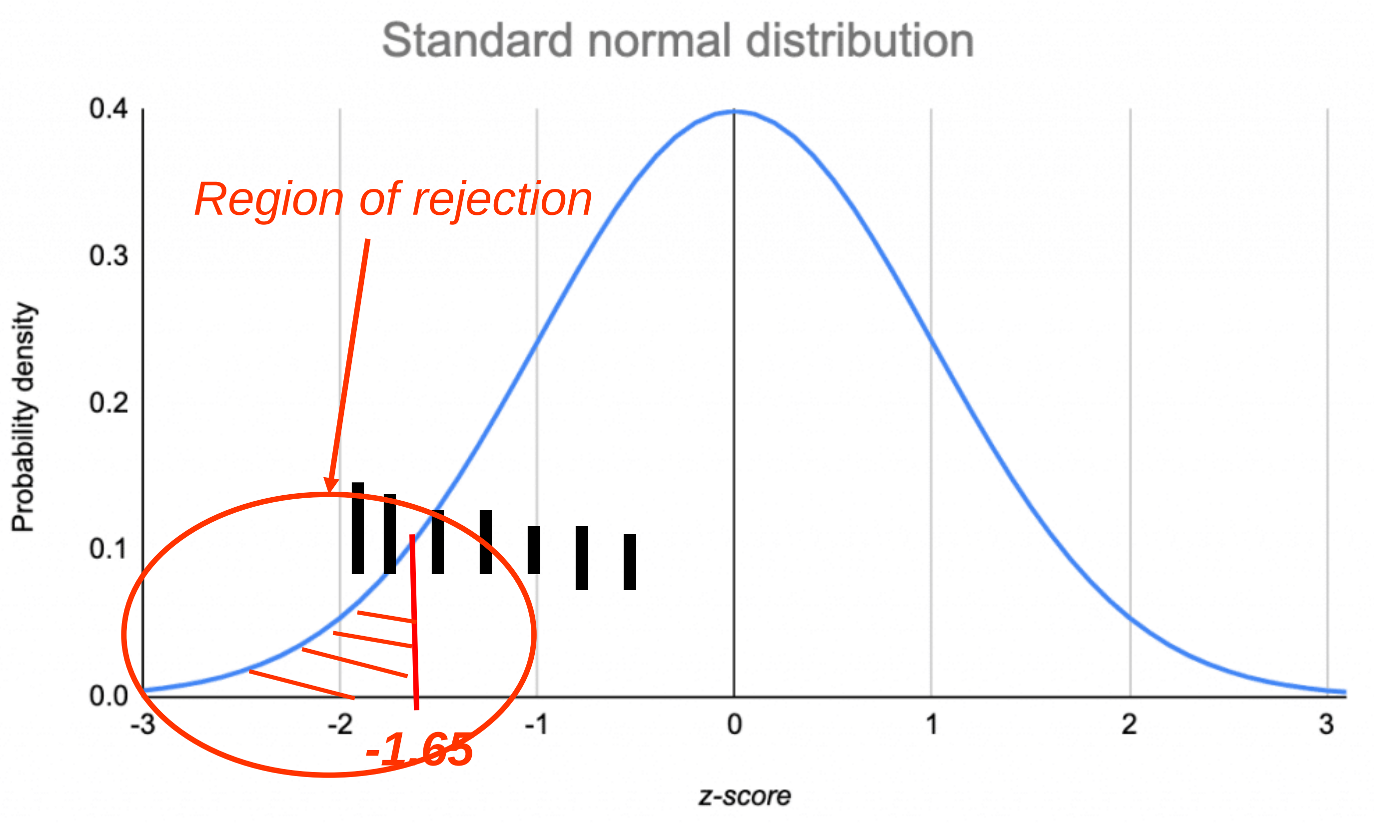
What about ‘<’ test?

- H_A: The factory fills bottles with less than 80mL fluid
- Will have to perform lower half Z-tests
- Z_c= +1.65 or -1.65
- *what can we conclude?*

1-sided confidence interval (upper/right)



1-sided confidence interval (lower/left)



Nicotine Problem

Problem statement *(Sheldon M. Ross, PSES, 2004)*

All cigarettes presently on the market have an average nicotine content of at least 1.6mg per cigarette. A firm that produces cigarettes claims that it has discovered a new way to cure tobacco leaves that will result in the average nicotine content of a cigarette being less than 1.6 mg. To test this claim, a sample of 20 of the firms cigarettes were analysed. If it is known that the standard deviation of a cigarette's nicotine content is 0.8 mg., what conclusions can be drawn at the 5% level of significance if the average nicotine content of the 20 cigarettes is 1.54?

Solution to the Nicotine problem

- First decide H_0
- Requirement: the probability of rejecting H_0 when it is true will never exceed α
- We should test
- H_0: $\mu >= 1.6$ versus H_1: $\mu < 1.6$

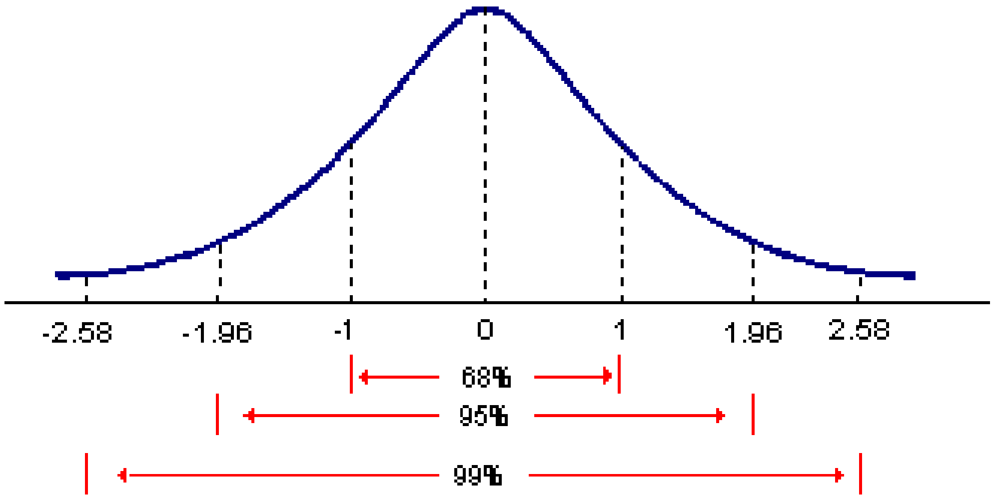
Nicotine problem μ=1.6

Value of test statistic is

$$\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma}=\frac{\sqrt{20}(1.54-1.6)}{0.8}$$

$$=-0.3$$

Z=-0.3 >-1.96



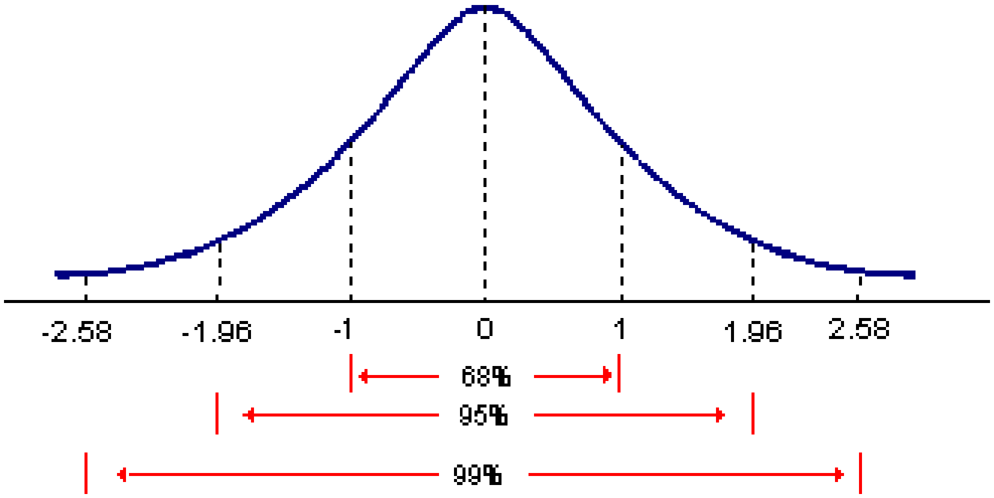
Nicotine problem μ=1.7

Value of test statistic is

$$\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma}=\frac{\sqrt{20}(1.54-1.7)}{0.8}$$

$$=-0.8$$

Z=-0.8 >-1.96



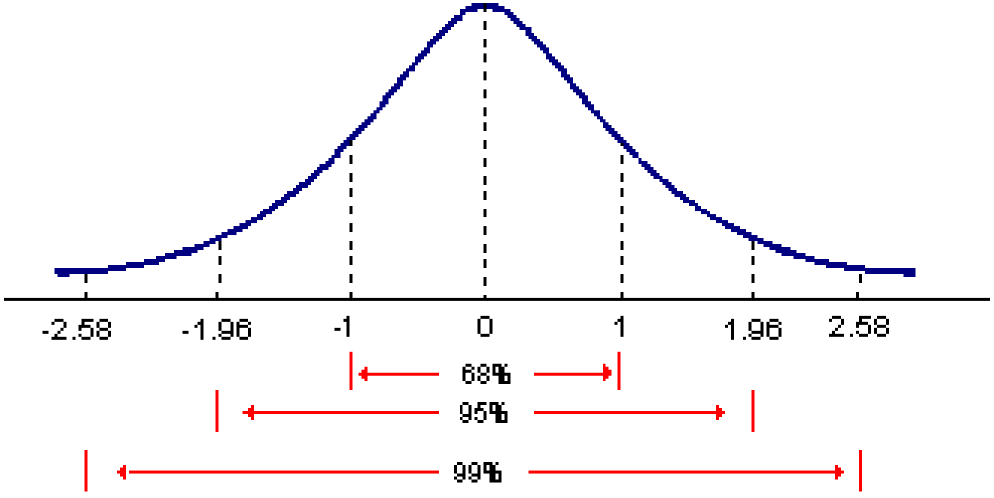
Nicotine problem μ=1.8

Value of test statistic is

$$\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma}=\frac{\sqrt{20}(1.54-1.8)}{0.8}$$

$$=-1.3$$

Z=-1.3 >-1.96



Conclusion from the nicotine problem

- 0.4 > 0.05; 0.2>0.05; 0.1>0.05
- Foregoing data do not enable us to reject at the 0.05 percent level of significance the hypothesis that “mean nicotine content exceeds 1.6”
- In other words, the evidence though supporting the cigarette producer’s claim is not strong enough to prove the claim

Coin toss problem

Problem Statement and Solution

Q: Find the probability of getting between 40 and 60 heads both inclusive in 100 tosses of a fair coin

A: According binomial distribution, the required probability is

$$^{100}C_{40}(1/2)^{40}(1/2)^{60} + {}^{100}C_{41}(1/2)^{41}(1/2)^{59} + \ldots {}^{100}C_{60}(1/2)^{60}(1/2)^{40}$$

Cumbersome to compute

Normal Approximation to Binomial

Mean=μ=np=100.(1/2)=50

Standard deviation=

σ=sqrt(npq)=sqrt[100.(1/2).(1/2)]=5

Since both np and nq are greater than 5, normal approx. to the binomial can be used to evaluate the sum.

On a continuous scale, 40 and 60 heads inclusive is same as between 39.5 to 60.5 heads

Z values for 39.5 and 60.5

$(39.5-50)/5=-2.10$

$(60.5-50)/5=+2.10$

The area under the normal curve between

-2.10 to +2.10= 0.96

A Hypothesis wrt coin toss

H_0: The coin is fair when the number of heads is between 40 and 60, both inclusive in a sample of tosses of 100

What is the probability of Type I error?

Ans: 1-0.96= 4%

Important Terminology

The probability of Type I error is called the

LEVEL OF SIGNIFICANCE