CS 215- Data Interpretation and Analysis
(Post Midsem)

Pushpak Bhattacharyya
Computer Science and Engineering Department
IIT Bombay

Lecture-6

Hypothesis Testing

23oct23

Recap

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

Type I and Type II error

- Type I: incorrectly reject H_0, when it should have been accepted. H_0 is actually true in the population

- Type II: incorrectly accept H_0 when it should have been rejected. H_0 is actually false in the population.

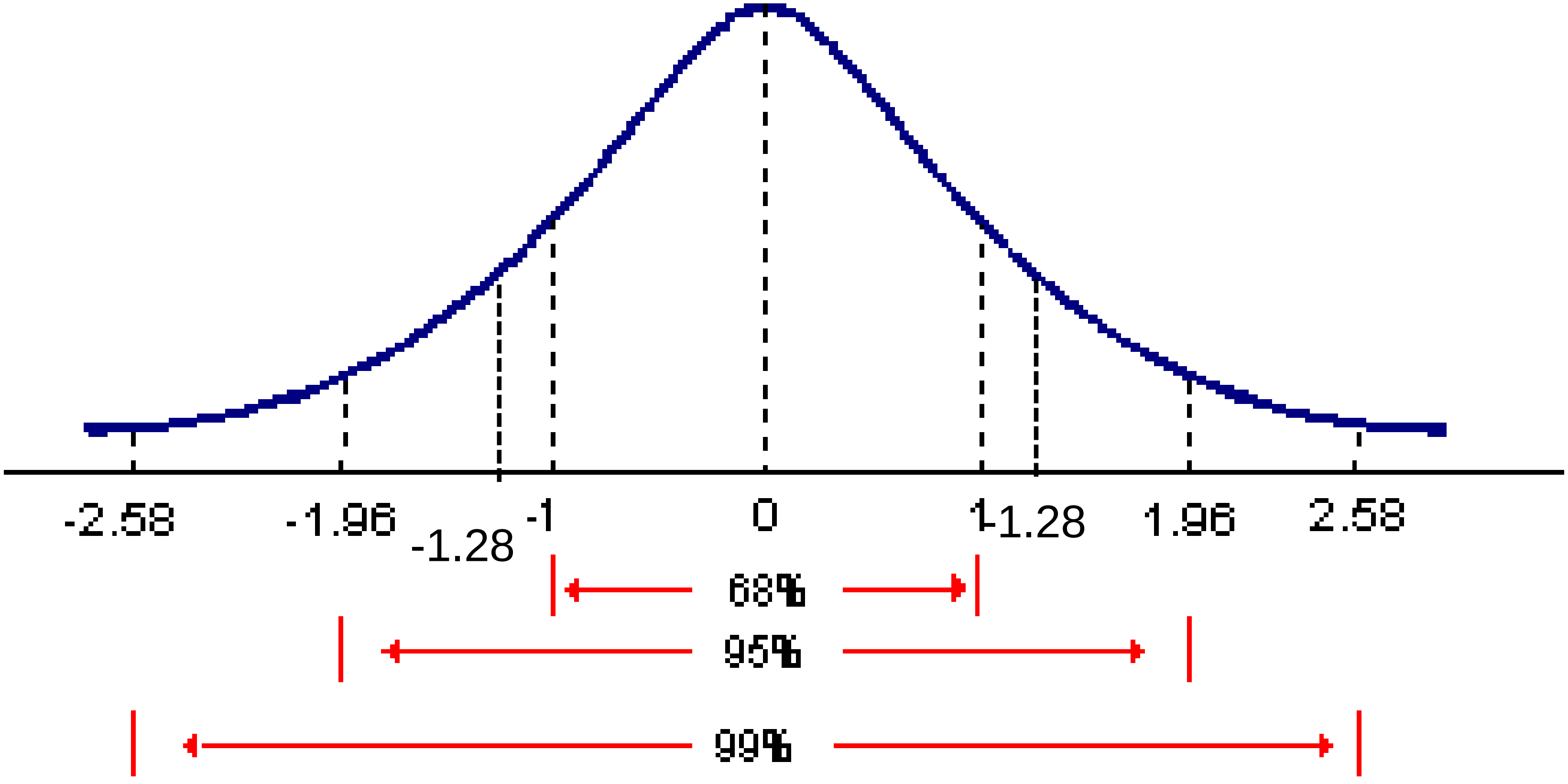
More on H_0

- **Data**: (a) All men are mortal, (b) Shakespeare is a man
- **H_0**: Shakespeare is not mortal
- **Contradiction**: Shakespeare is not a man
- **Conclusion**: Reject H_0

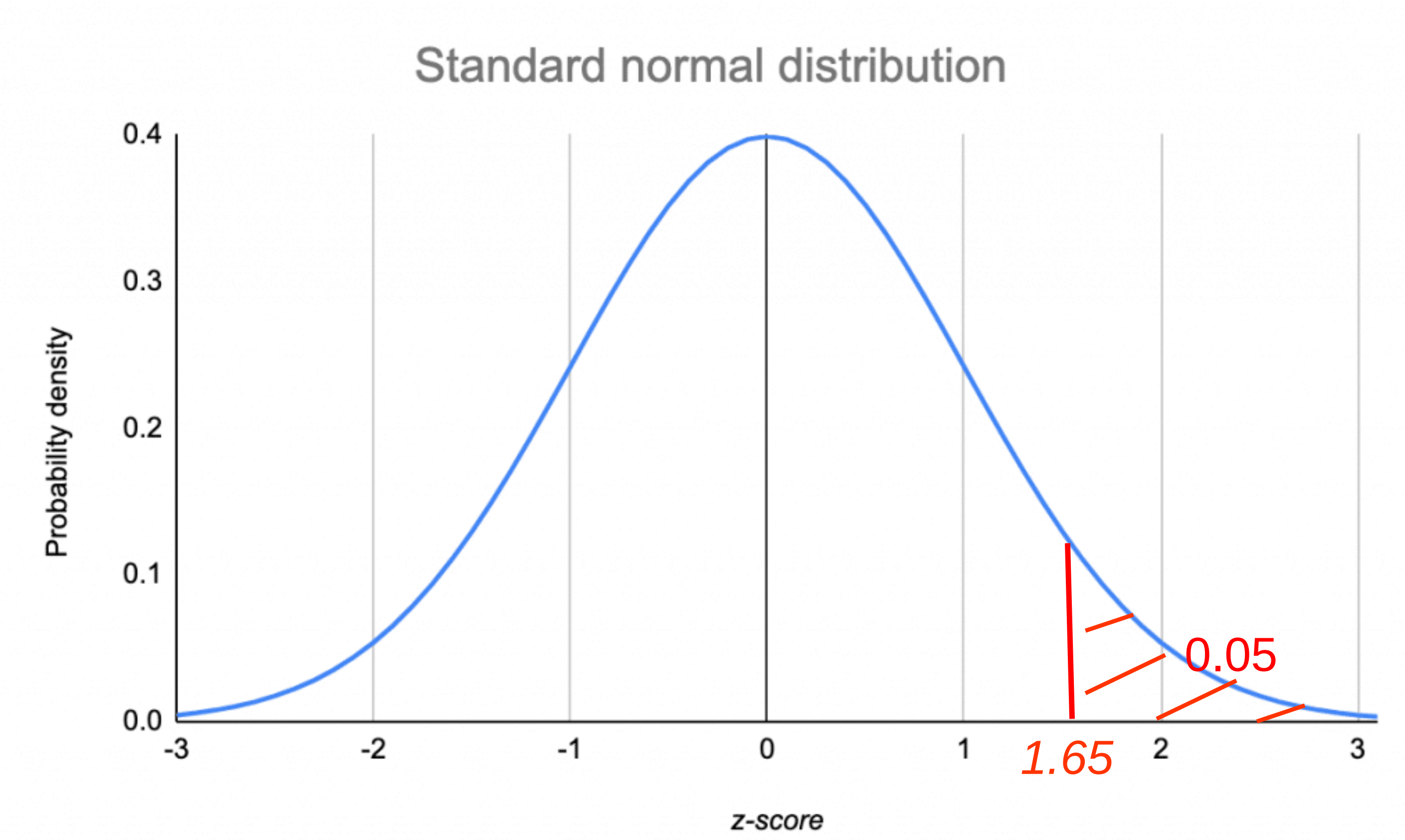
A useful table

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

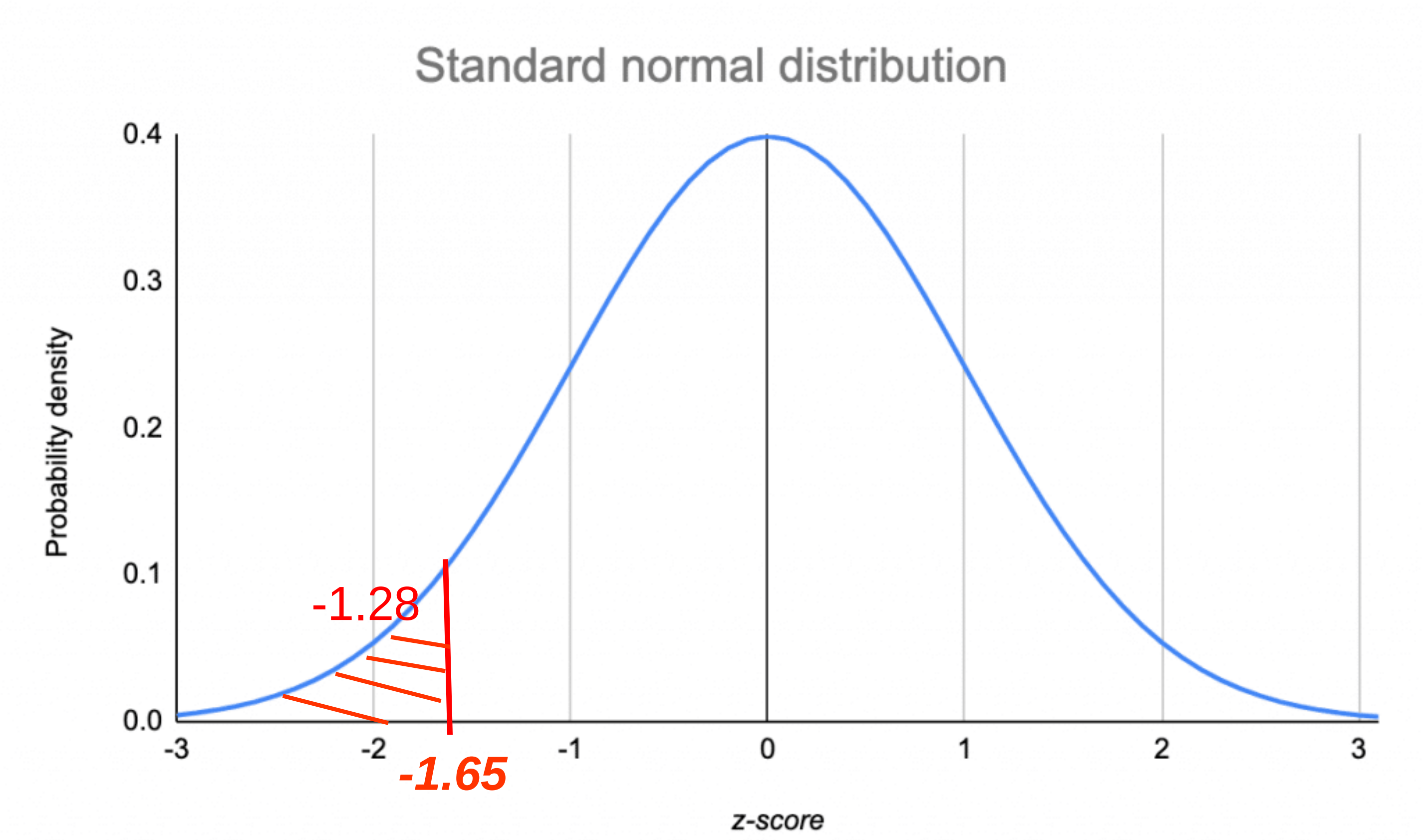
2 sided 95% confidence interval



95% 1-sided confidence interval (from –inf)



95% 1-sided confidence interval (to +inf)



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

Solution

- 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:
  - $\mu \neq 80$ (2-sided test)

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

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, $Z_c = +- 2.58$

- $-5.0 < -2.58$
- So for 99% confidence interval also the hypothesis is rejected

90% confidence interval, Z_c=+-1.28

- $-5.0 < -1.28$
- So for 90% confidence interval also the hypothesis is rejected

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 $+2.10 = 0.96$

End Recap

Hypothesis Testing wrt coin toss

H_0: The coin is fair, $p_H=1/2$

Confidence level: 95%

Data: 53 heads in 100 tosses

Foundation-1

The no. *K* of tosses in *N* trials follows the binomial distribution

Let p_H be the prob. of head in one toss

Then np_H is the mean number of heads

sqrt[$np_H(1-p_H)$] is the standard deviation, if we take many samples of n tosses

Foundation-2

If $np_H > 5$ as also $n(1-p_H)$, then K can be approximated by a normal distribution with $\mu = np_H$ as the mean and $\sigma = sqrt[np_H(1-p_H)]$ as the standard deviation

Solution to the coin toss hypothesis testing (1/2)

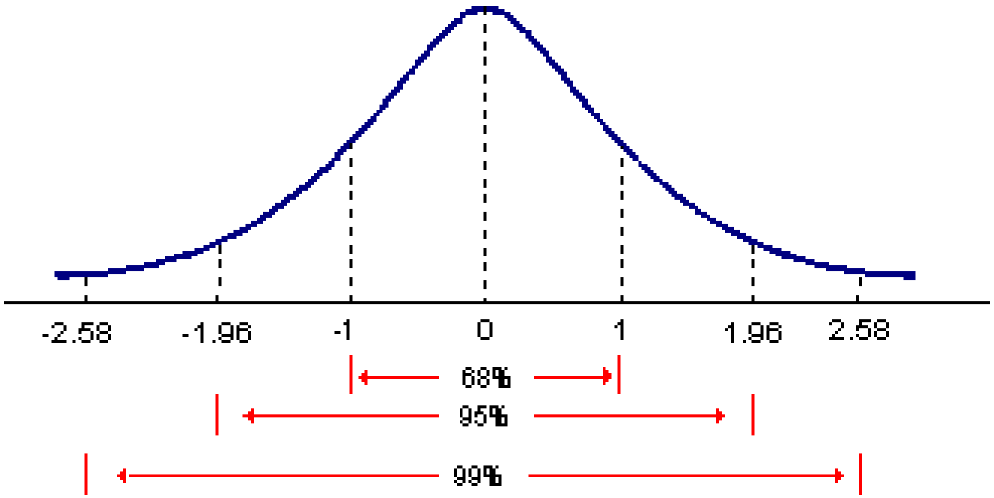
$K=53$, $\mu=100.(1/2)=50$,
$\sigma=\text{sqrt}[100.(1/2).(1/2)]=5$

Since both np_H and $n(1-p_H)$ are >=5, normal approx. to binomial can be done

$Z_o=(53-50)/5=0.6$

Solution to the coin toss hypothesis testing (2/2)

$Z_o = 0.6 < 1.96$, falls within 95% confidence interval (CI)



Cannot reject H_0 that the coin is fair

Important Terminology

The probability of Type I error is also called the

LEVEL OF SIGNIFICANCE (α)

Type-I and Type-II errors: **Always wrt Null Hypothesis H_0**

as per data / actual	ACCEPT	REJECT
TRUE	*No Error*	*Type- I error*
FALSE	*Type-II Error*	*No Error*

A useful table

test-type (col)			
vs.	Two-Tail	Left-Tail	Right-Tail
Confidence Interval (significance level)			
90% (0.10)	(- and +) 1.65	-1.28	1.28
95% (0.05)	(- and +) 1.96	-1.65	1.65
99% (0.01)	(- and +) 2.58	-2.33	2.33

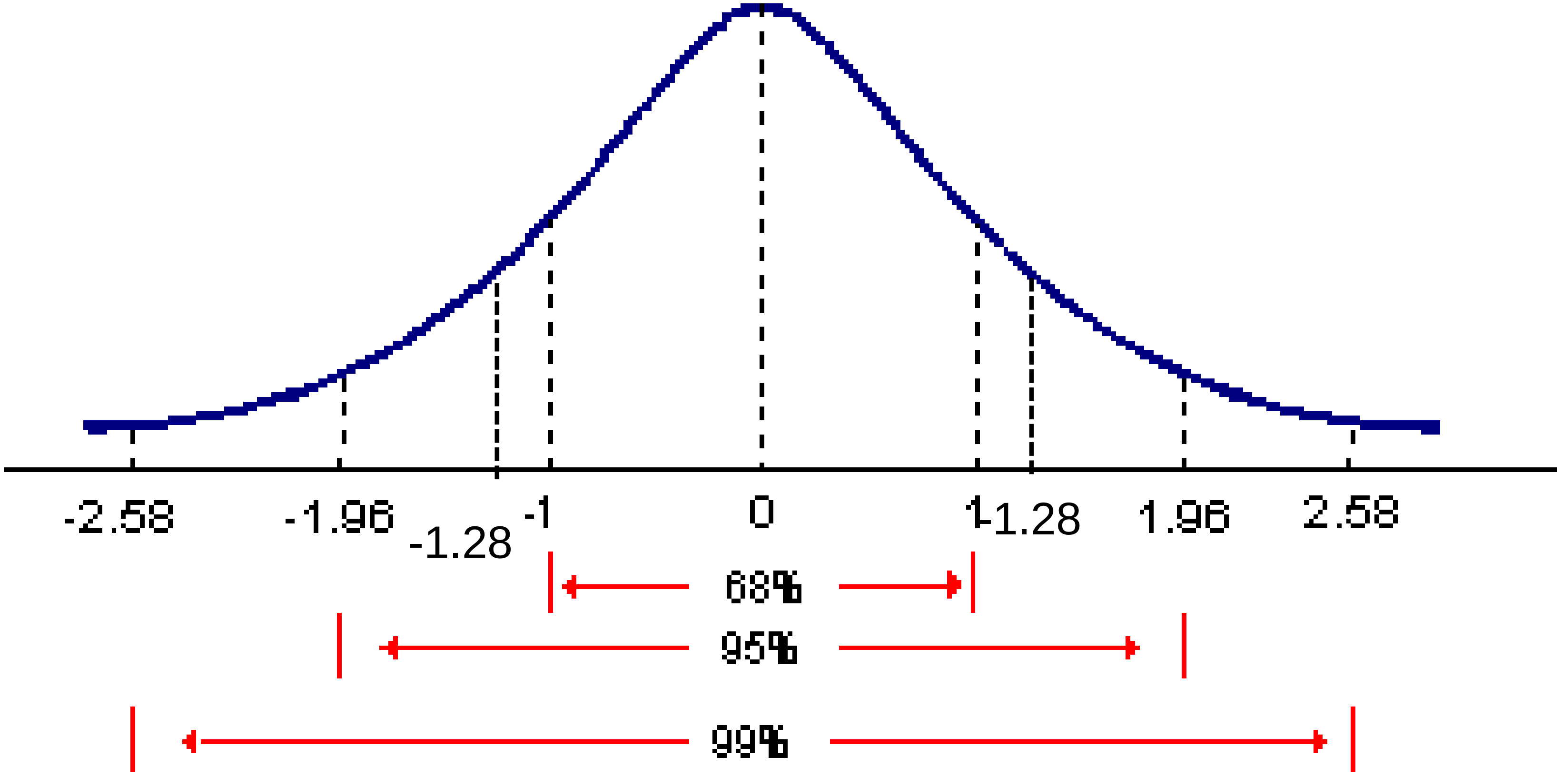
Significance of level of significance α

- 90% confidence interval, $\alpha=0.10$
  - ➔Prepared to tolerate 10% Type-I error
  - ➔Probability of wrong rejection of H_0 is 10%
- 95% confidence interval, $\alpha=0.05$
  - ➔Prepared to tolerate 5% Type-I error
  - ➔Probability of wrong rejection of H_0 is 5%
- 99% confidence interval, $\alpha=0.01$
  - ➔Prepared to tolerate 1% Type-I error
  - ➔Probability of wrong rejection of H_0 is 1%

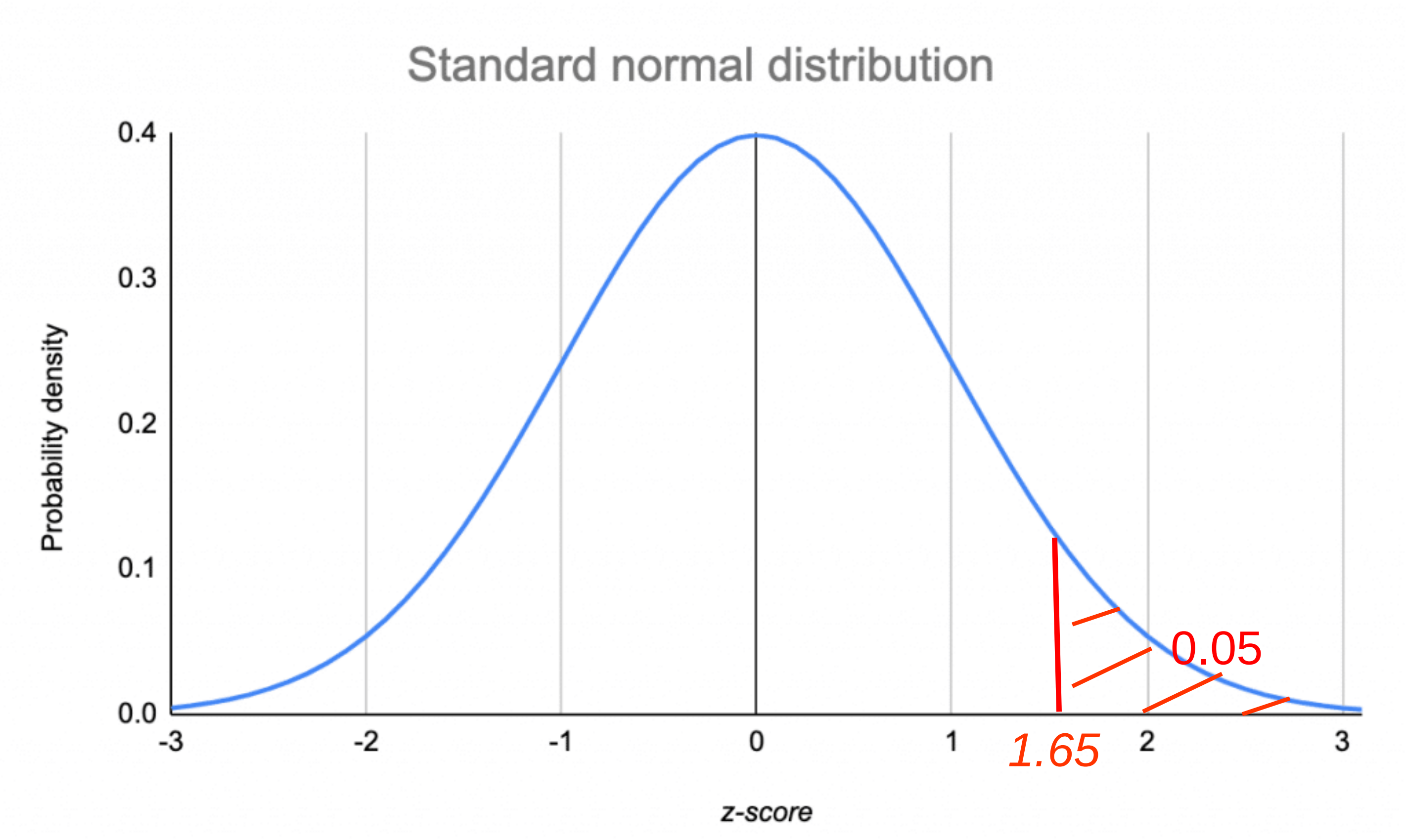
Significance of level of significance

As α decreases, the test becomes more and more stringent (strict, demanding)

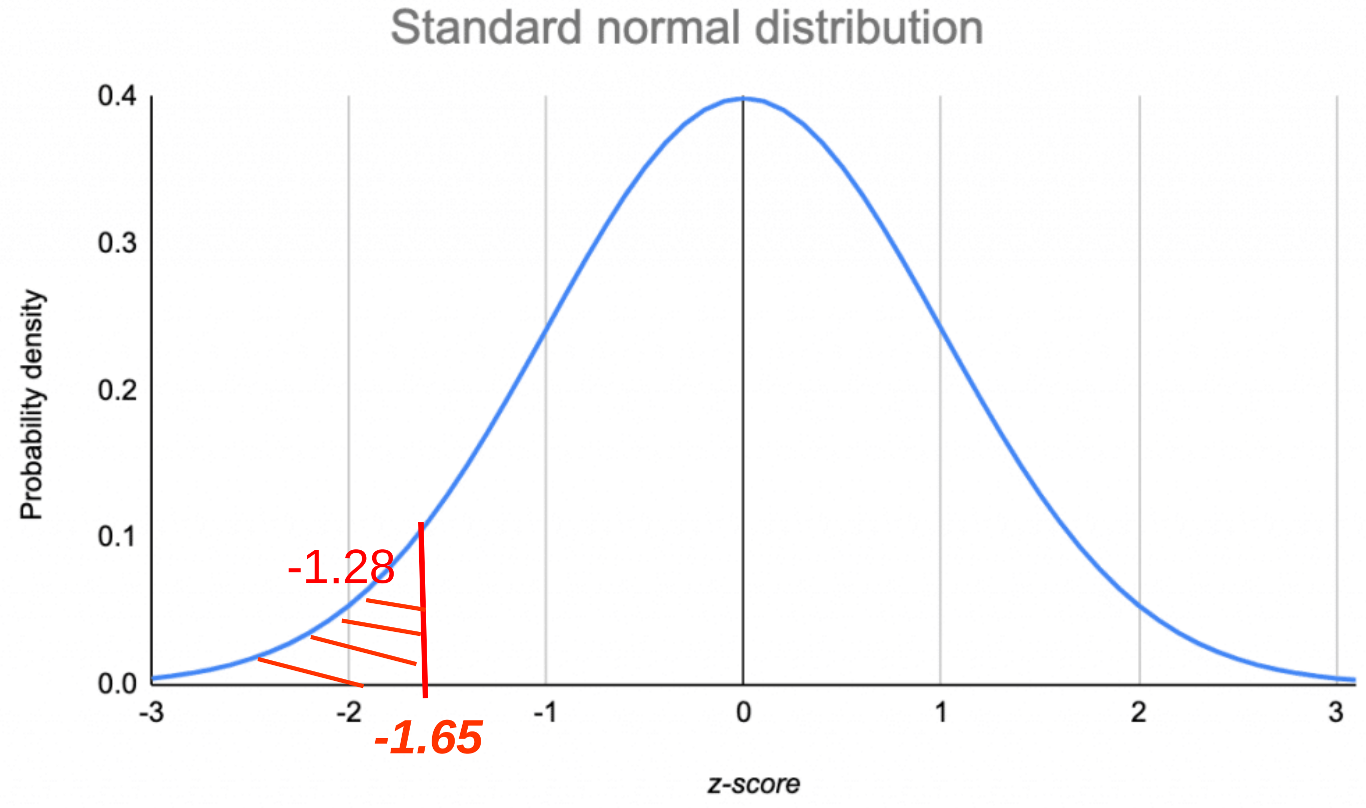
2 sided 95% confidence interval



1-sided confidence interval (rejection region: upper/right)



1-sided confidence interval (rejection region: lower/left)



An NLP problem

A NEI tagging situation

An NEI tagger is given a sequence of 50 words and asked to identify if the words are *name* or *no_name*. If the tagger tags 32 words correctly, determine if the results are significant at (a) 0.10, (b) 0.05 and (c) 0.01 level of significance.

Input-Output

Input- *In the recent address to Congress, the US President Biden had two ladies Pelosi and Harris sitting behind him.*

Output- *In_0 the_0 recent_0 address_0 to_0 Congress_1, the_0 US_1 President_0 Biden_1 had_0 two_0 ladies_0 Pelosi_1 and_0 Harris_1 sitting_0 behind_0 him_0* ._0

What is the question here?

- On the face of it, the tagger accuracy as evidenced from the data is 32/50=64%.
- Compare with random guessing
- To construct a random guesser we do not need any resources or techniques
- Just toss a coin: Head→ *name*, Tail→ *no_name*

Question- is the tagger doing random guessing?

Solution: start with H_0

- *p= probability of the tagger labelling correctly*
- H_0: Null Hypothesis (that we want to demolish)- *p=50%*
  - *Equivalent to random guessing*
- Confidence intervals are given, as 90%, 95% and 99%
- Now the process can start

Given H_0, fix H_A

- H_A: $p>50\%$
- We choose one-tailed test (the greater-than test), since we want the tagger to perform with high score

Examine consequence of H_0

- $\mu = np = 50.\ (0.5) = 25$

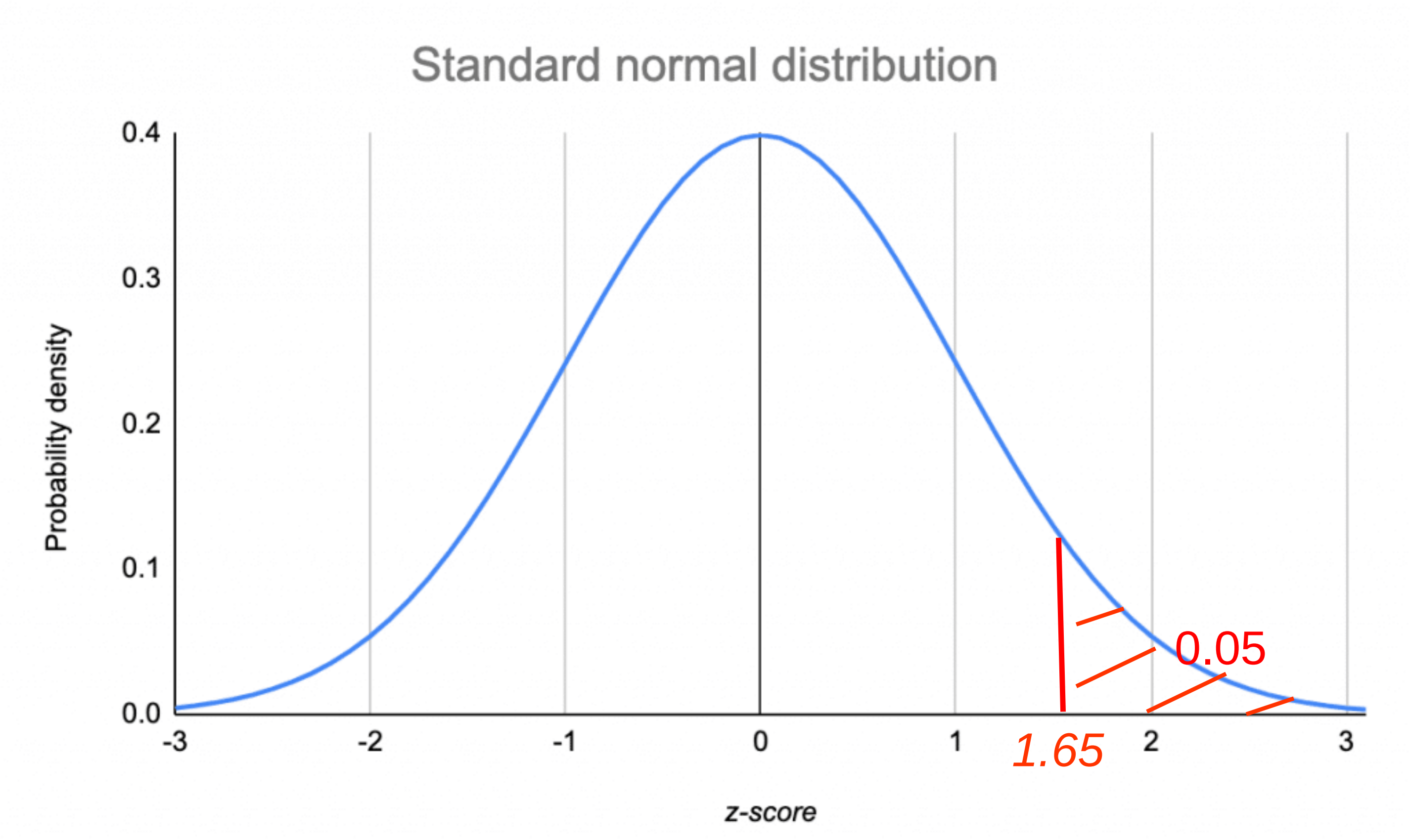
- This is the population mean too as per MLE principle

- $\sigma = (npq)^{1/2} = (50.0.5.0.5)^{1/2} = (12.5)^{1/2} = 3.54$

Relook at confidence intervals

test-type (col)			
vs.	Two-Tail	Left-Tail	Right-Tail
Confidence Interval (significance level)			
90% (0.10)	(- and +) 1.65	-1.28	1.28
95% (0.05)	(- and +) 1.96	-1.65	1.65
99% (0.01)	(- and +) 2.58	-2.33	2.33

1-sided confidence interval (upper/right)



Compute Z_o, under H_0 and examine confidence intervals

- $Z_o = (32-25)/3.54 = \underline{1.98}$

- $Z_o > Z_c$ (=1.28), for one sided 90% confidence interval
- $Z_o > Z_c$ (=1.65), for 95% one sided confidence interval
- $Z_o < Z_c$ (=2.33), for one sided 99% confidence interval

Implications: 90% Confidence Interval (1/2)

- Z_o (=1.98) > Z_c (=1.28), for 90% confidence interval
  - Enters the rejection region
- H_0 can be rejected
- Reject H_0 ($p=0.5$) and accept H_A
- In other words, the data in conjunction with Confidence Interval suggests that the NEI tagger is doing better than random guessing

Implications: 90% Confidence Interval (2/2)

- If we are prepared to accept 10% Type-I error, H_0 can be rejected

- If we are prepared to tolerate that in 10 out of 100 cases we will go wrong, we can reject H_0

Interpretation in the NEI situation (1/2)

- Suppose the NEI tagger is actually doing random guessing
- And, we reject H_0 (p=50%)
- That is we say, “no the NEI tagger is not doing random guessing, but internally employing an algorithm that is better than random guessing”

Interpretation in the NEI situation (2/2)

- As per the confidence interval, I am prepared to tolerate 10% Type-I error
- **As per data**, the probability of going wrong with the stand “no, the NEI tagger is better than a random guesser” is at most 10%
- Because Z_o has entered into the critical region

What about 95% CI

- Z_o (=1.98) > Z_c (=1.65), for 90% confidence interval
  - Enters the rejection region
- H_0 can be rejected
- Reject H_0 and accept HA
- In other words, the data in conjunction with the CI suggests that the NEI tagger is doing better than random guessing

Implications: 99% CI

- Z_o (=1.98) > Z_c (=2.33), for 99% confidence interval
  - Does not enters the rejection region
- Now H_0 cannot be rejected
- On the basis of the data and CI, we cannot rule out the possibility that the NEI tagger is just doing random guessing

Which implication is correct?

- All implications are correct!
- Depends on our tolerance level, how much Type-I error we are prepared to accept
- Only 1 error case out of 100 → very stringent
- 5 errors case out of 100 → less stringent
- 10 errors case out of 100 → much less stringent, more permissive

Anti-Allergen Medicine problem

Problem statement

The manufacturer of a patent medicine claimed that it was 90% effective in relieving the allergy. In a sample of 200 people who had the allergy, the medicine provided relief for 160 people. Determine if the manufacturer’s claim is legitimate.

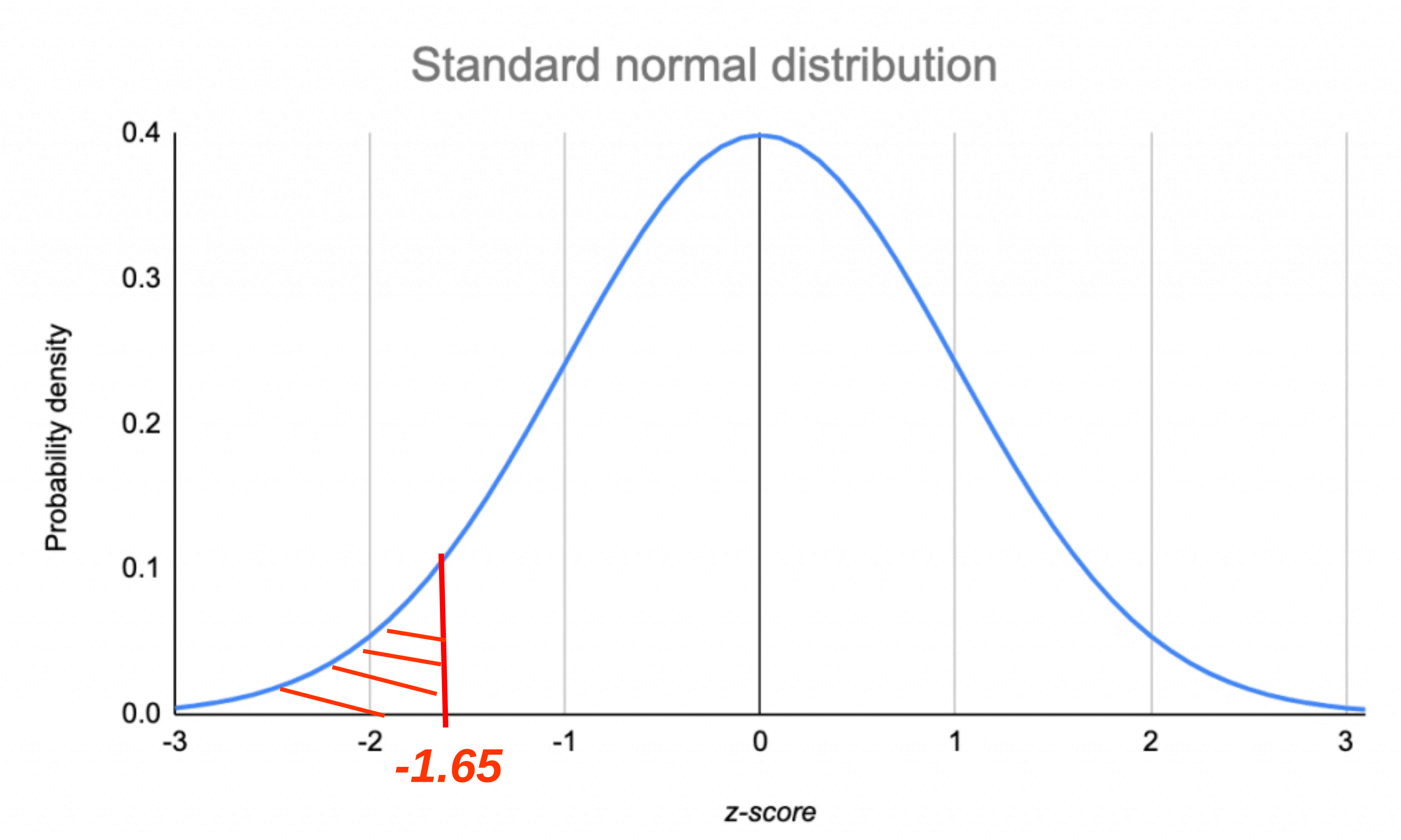
Solution to the anti-allergen medicine problem

- H_0: p=0.9 versus H_1: p<0.9
- If H_0 is true, $\mu = np = 200.(0.9) = 180$
- $\sigma = (npq)^{1/2} = (200.0.9.0.1)^{1/2} = 4.23$
- Now, 160 gives rise to $Z_o = 160 - 180/4.23 = -4.73$

Recall: 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

1-sided confidence interval (Rejection: lower/left)



$Zo=-4.73$

- $Zo<-1.65$ (falls in rejection region)
- So, can reject H_0 at 95% CI
- Thus the claim that **the medicine is 90% effective in relieving allergy** is not tenable in the face of the data, viz., 160 out of 200 people who took the medicine got relief.

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?

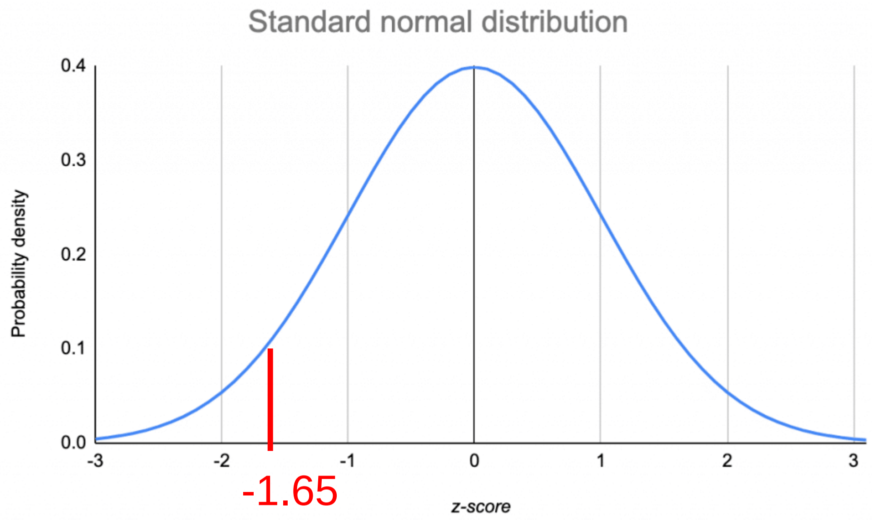
Solution to the Nicotine problem

H_0: μ>=1.6 versus H_1: μ<1.6

With μ=1.6,

$$Z_o = \frac{\sqrt{n}(\overline{X} - \mu\ )}{\sigma} = \frac{\sqrt{20}(1.54 - 1.6)}{0.8} = -0.33$$

$Z_o > Z_c$, so cannot reject H_0



Conclusion from the nicotine problem

- H_0 cannot be rejected
- Data suggests nicotine content is >=1.6 mg
- Company's claim that the new method of cigarette making ensures < 1.6mg nicotine content is not consistent with the data