

# **CS 215- Data Interpretation and Analysis (Post Midsem)**

Pushpak Bhattacharyya

Computer Science and Engineering Department

IIT Bombay

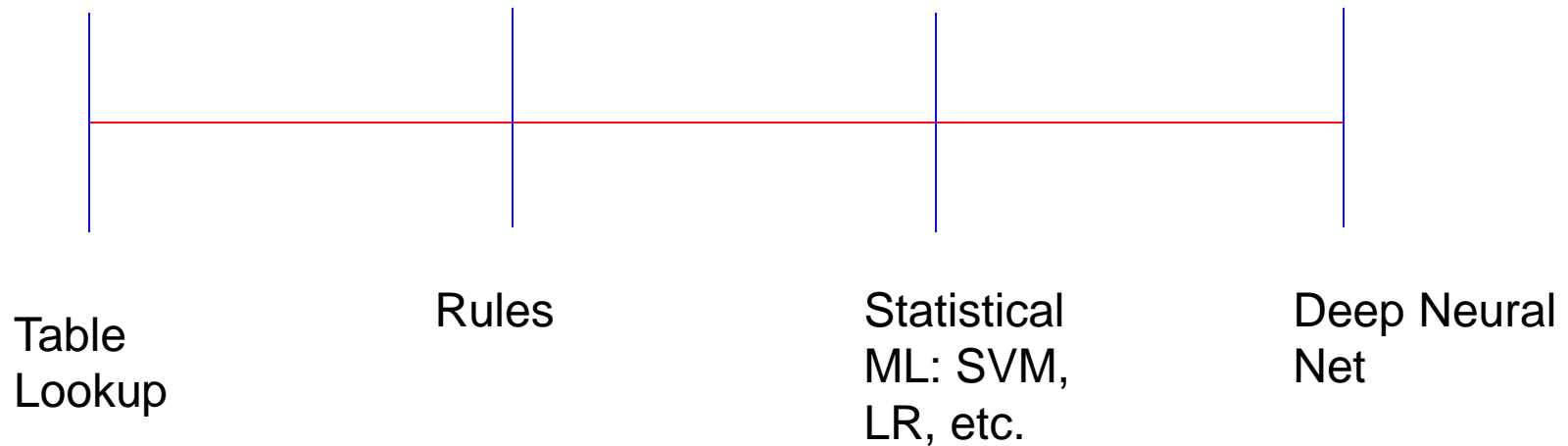
Lecture-2

Why probability cntd., reading Z-score table, interval estimate

9oct23

Recap

# A Perspective on Machine Learning



# A Practical Problem

- A bridge is being built. The weight it can tolerate has a normal distribution with  $\mu=400$  and  $\sigma=40$ . A car that goes on the bridge has weight distribution (again normal) given by  $\mu=3$  and  $\sigma=0.3$ . We want the probability that the bridge is damaged to be less than  $0.1$ . How many cars can we allow to go on the bridge?

# When does the bridge break?

$$W_{total} > W_{tolerance}$$

# Bring in probability

- #cars = N
- Each car's weight is normal with  $\mu=3$  and  $\sigma=0.3$

## Apply CLT

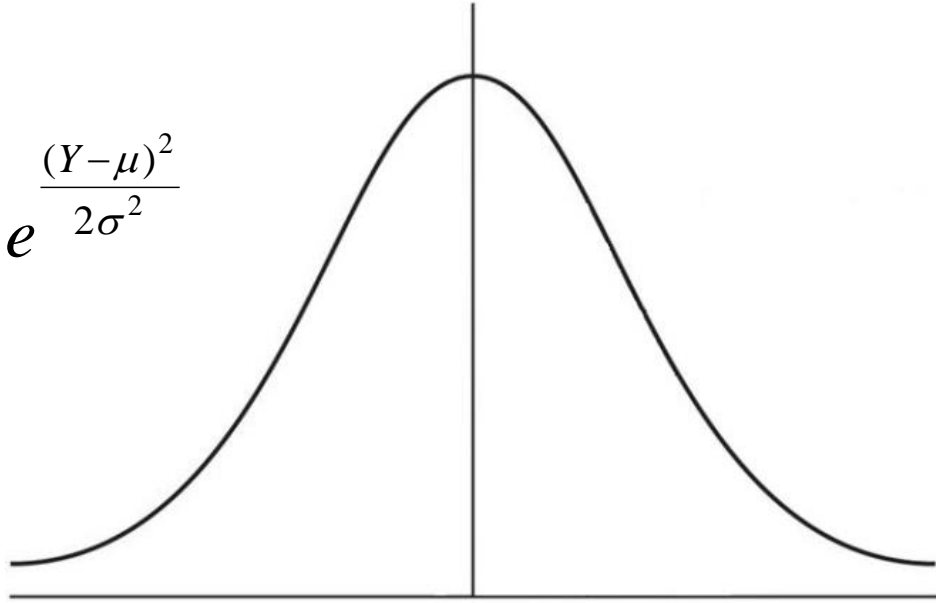
- By central limit theorem, the sum of Gaussian Random Variables is Gaussian with mean and variance being sums of individual means and variances



***total weight of  $N$  cars is normal  
with  $\mu=3N$  and  $\sigma^2=0.09N$***

$W_{\text{tolerance}}$  looks like this...

$$F(Y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Y-\mu)^2}{2\sigma^2}}$$



$\mu=400$  and  $\sigma=40$



# We want

- What no. of cars will cause the probability to exceed 0.1?

$$\textit{Probability}(W_{total} - W_{tolerance}) > 0.1$$

LHS is a function of N

$W_{total}$  is a function of N by CLT

# Bring $N$ , the number of cars in picture

- Central Limit Theorem applied again
- $W_{total} - W_{tolerance}$  is a random variable
- Follows Normal Distribution
- *Mean* =  $3N - 400$
- *Variance* =  $0.09N + 1600$

# Convert to Standard Normal Form

$$z \equiv \frac{(W_{total} - W_{tolerance}) - (3N - 400)}{\sqrt{0.09N + 1600}}$$

We want this event...

$$(W_{total} - W_{tolerance}) > 0$$

$$\Rightarrow \frac{(W_{total} - W_{tolerance}) - (3N - 400)}{\sqrt{0.09N + 1600}} > \frac{-(3N - 400)}{\sqrt{0.09N + 1600}}$$

$$\Rightarrow z > \frac{-(3N - 400)}{\sqrt{0.09N + 1600}}$$

When will this Probability exceed  
0.1

$$P\left(z > \frac{-(3N - 400)}{\sqrt{0.09N + 1600}}\right) > 0.1$$

**Solving this gives  $N \leq 117$**

**How?**

## Use Standard Normal Form Table

$$P(z < V) = \int_{-\infty}^V \frac{1}{\sqrt{2\pi}} \exp(-y^2 / 2) dy$$

$$\text{Now } P(z > V) = 1 - P(z < V)$$

$$\text{Since we want } P(z > V) > 0.1$$

$$\Rightarrow 1 - P(z < V) > 0.1$$

$$\Rightarrow P(z < V) \leq 0.9$$

$V=1.28$ , consulting the table

$V=1.28$

### Standard Normal Probabilities

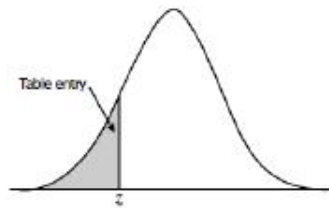


Table entry for  $z$  is the area under the standard normal curve to the left of  $z$ .

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

### Standard Normal Probabilities

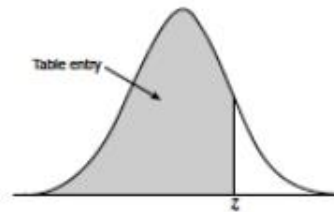


Table entry for  $z$  is the area under the standard normal curve to the left of  $z$ .

[illegible]

Get N from...

$$1.28 = \frac{-(3N - 400)}{\sqrt{1600 + 0.09N}}$$

$$N = \sim 117$$



# Conclusion

*If we allow more than 117 cars on the bridge, then in 10 out 1000 such cases the BRIDGE WILL BREAK!!*

End recap

Another problem

# Problem Statement

- We have to estimate the percentage of sand grains in a pile of sand resulting from the fragmentation of a mineral compound which fall in a particular range.



# What is Mineral Sand Used for

- Mineral Sand is an essential component of consumer products such as paint, plastics and paper
- Almost all titanium **minerals** are **used** as feedstock to produce titanium dioxide pigment **used** in products such as paints, paper and plastics
- ***For the purpose of paints for example, sand particles should be in a range of sizes***

# Problem Statement Continued

A sample of 10 grains taken from a large pile have respective sizes as

– 2.2, 3.4, 1.6, 0.8, 2.7, 3.3, 1.6, 2.8, 2.5, and 1.9

Estimate the % of sand grains in the entire pile whose size is between 2 and 3



# What do we know about the properties of mineral sand?

- Kolmogorov's Law of Fragmentation
- “The size of particles resulting from the fragmentation of a mineral compound has a **LOGNORMAL DISTRIBUTION**”

# What is lognormal distribution?

- A random variable  $X$  has a lognormal distribution if  $\log X$  has a normal distribution

$$Y = \log X$$

$F(Y)$  = Probability Density  
Function of  $Y$

$$F(Y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Y-\mu)^2}{2\sigma^2}}$$



# Naïve Approach to solving the sand particle size problem

- Sample size= 10
- Respective sizes:
  - 2.2, 3.4, 1.6, 0.8, 2.7, 3.3, 1.6, 2.8, 2.5, and 1.9
- Number of particles in the range 2-3= 4
- $\% = 4/10 = 40\%$
- Is this correct?

# Is this correct?

- May or may not be
- If it is correct, then so by chance
- Not principled
- When is the conclusion drawn from a sample correct?
- When the conclusion holds for the population too
- Does 40% in the range 2-3 hold for the pile too?

# May or may not be: but not PRINCIPLED

- Important knowledge not used
- KNOWLEDGE OF DISTRIBUTION
- Knowledge that the sizes of sand particles is lognormally distributed

# IMPORTANT!!

- If we know the distribution, we must use it

***HOW?***

Use of knowledge of distribution

# Original goal: Sand Particle Size

- Percentage of sand particles in the range 2-3
- Percentage=Probability
- Frequentist view of probability
- Goal:  $P(2 \leq X \leq 3)$

# To use the distribution

- $P(\log 2 \leq \log X \leq \log 3)$
- Then use the MACHINERY OF PROBABILITY

# Machinery of Probability

- Sample size= 10
- Respective sizes:
  - 2.2, 3.4, 1.6, 0.8, 2.7, 3.3, 1.6, 2.8, 2.5, and 1.9
- Log values:
  - 0.79, 1.22, 0.47, -0.22, 0.99, 1.19, 0.47, 1.03, 0.91, and 0.64



# Calculations (somehow known- $\mu=0.75$ , $\sigma=0.44$ )

$$P(\log 2 \leq \log X \leq \log 3)$$

$$= P\left(\frac{\log 2 - 0.75}{0.44} \leq \frac{\log X - 0.75}{0.44} \leq \frac{\log 3 - 0.75}{0.44}\right)$$

$$= P(-0.13 \leq Z \leq 0.80)$$

$$= \Phi(0.8) - \Phi(-0.13)$$

$$= \Phi(0.8) - [1 - \Phi(0.13)]$$

$$= \Phi(0.8) + \Phi(0.13) - 1$$

$$= 0.79 + 0.55 - 1$$

$$= 0.32$$

So the answer is...

Not 40%

But 32% !!!

More principled

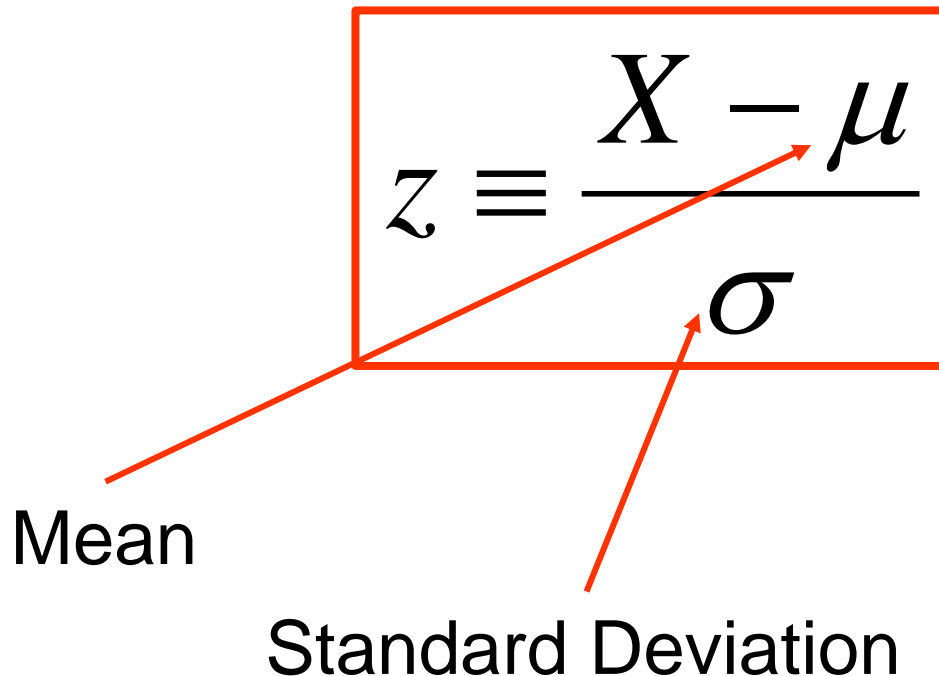
Can answer “why”

# Which estimate is correct?

- 32% or 40%
- 32% of the sand pile population in size range 2 to 3
- Topic of TEST OF HYPOTHEIS

Important Concept: Z-score

Any normally distributed random variable  $X$  can be converted to standard normal form



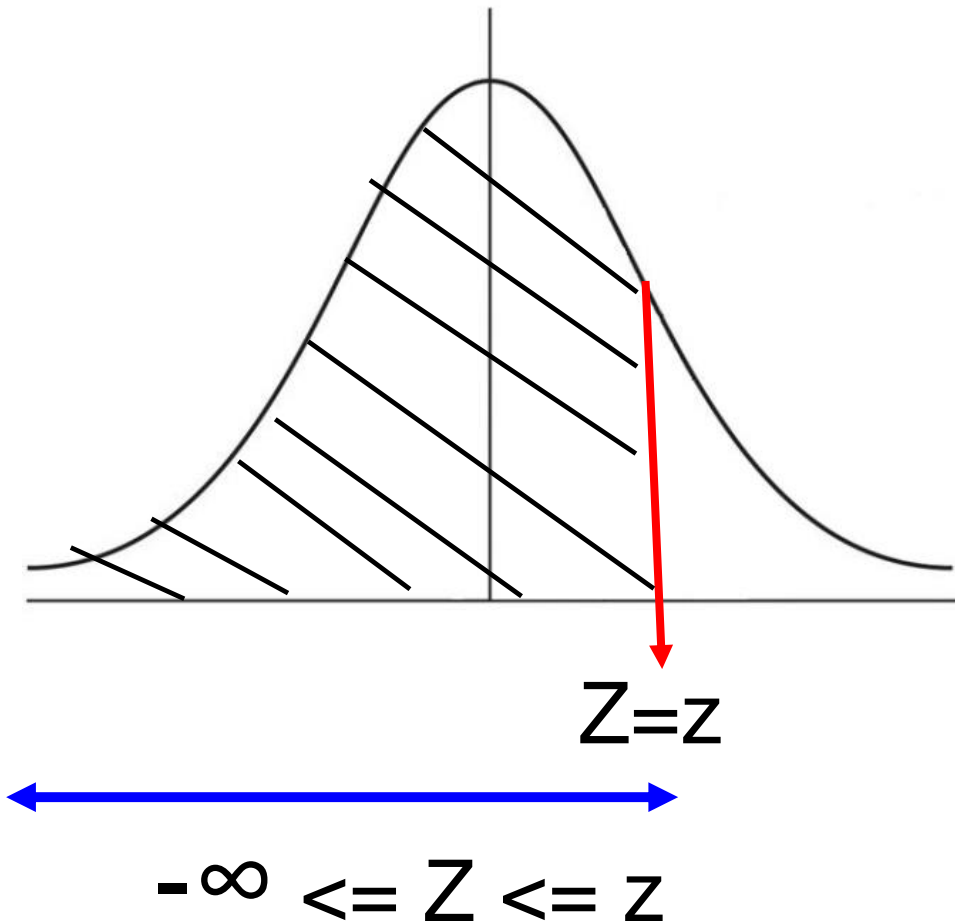
The diagram shows the formula for converting a normally distributed random variable  $X$  to standard normal form,  $z \equiv \frac{X - \mu}{\sigma}$ . The formula is enclosed in a red rectangular box. Two red arrows point from labels below to the formula: one from the word "Mean" to the  $\mu$  in the numerator, and another from the words "Standard Deviation" to the  $\sigma$  in the denominator.

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

Mean

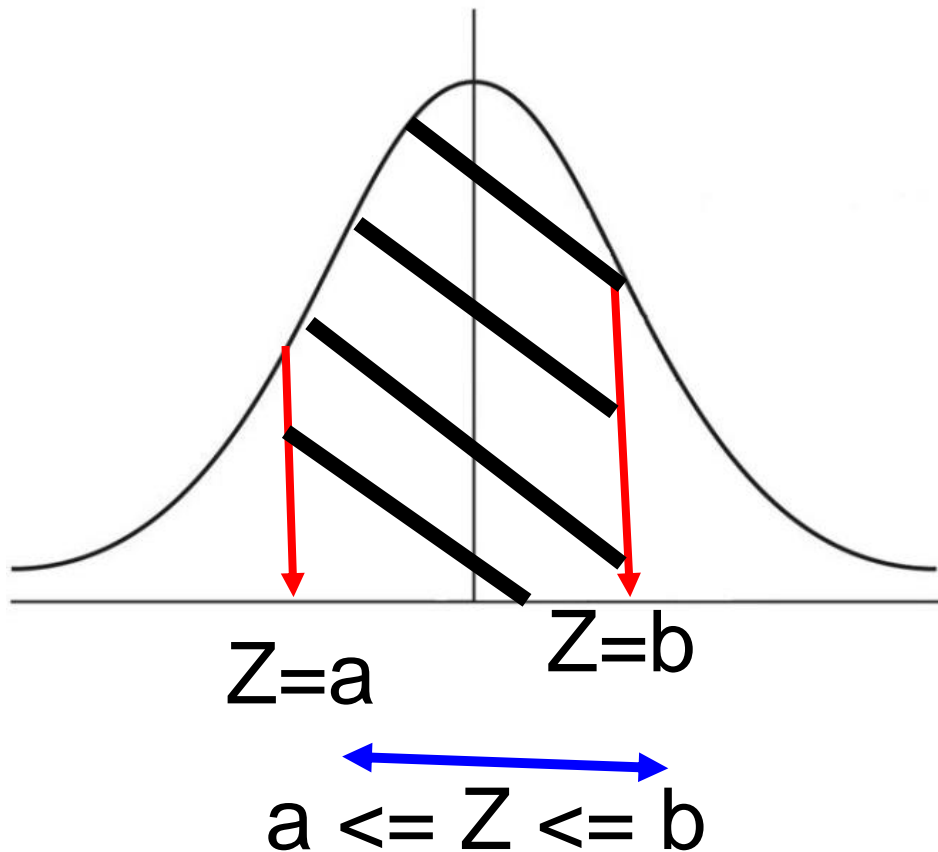
Standard Deviation

# Areas under normal curve $\rightarrow$ z-score



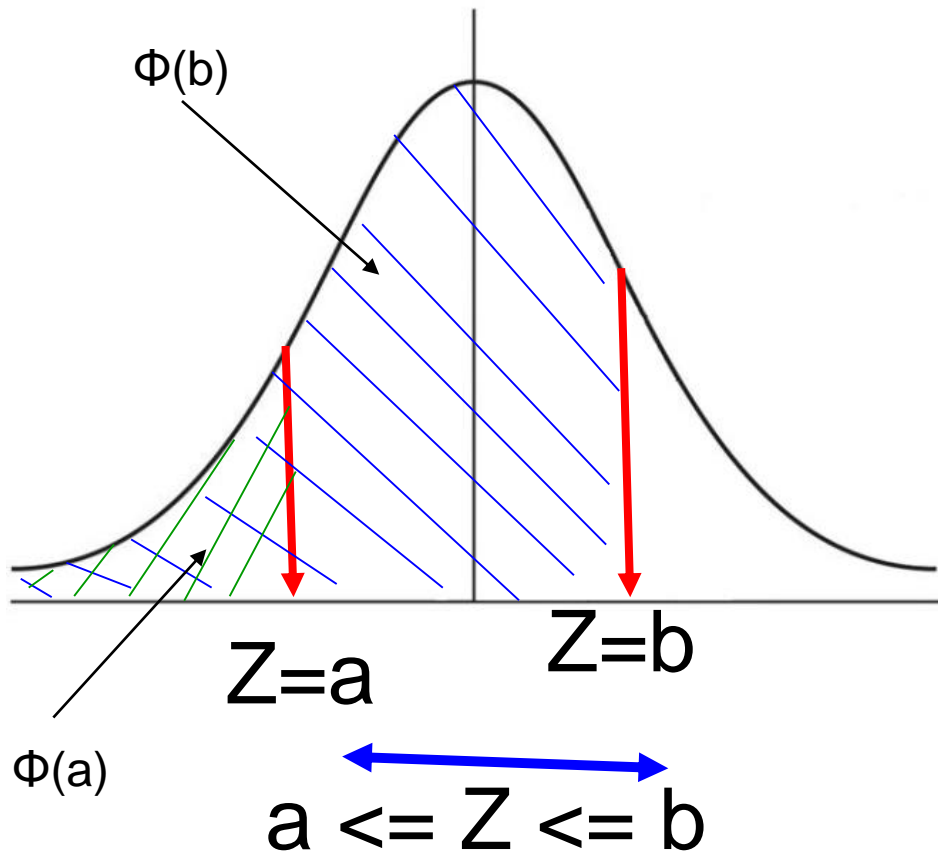
- Z-score = area under the curve from  $-\infty$  to  $z$
- Is the probability of  $Z$  being in interval  $[-\infty, z]$

# Interval Probability



- Area under normal the curve from  $a$  to  $b$  is the probability of  $Z$  being in interval  $[a, b]$

# $\Phi(b) - \Phi(a)$ : difference of cumulative probability



- Area under the curve from  $a$  to  $b$  is the probability of  $Z$  being in interval

$$P(a \leq Z \leq b)$$



# PDF and CDF

- Probability density function (PDF)
- Cumulative Distribution Function
- PDF denoted by  $P(X)$
- CDF denoted by  $\Phi(X)$

*Integration-Differentiation Relationship*

# Integration of Normal PDF

## cumbersome: Use Standard Normal Form Table

$$P(Z < V) = \int_{-\infty}^V \frac{1}{\sqrt{2\pi}} \exp(-y^2 / 2) dy$$

- Difficult to find an algebraic closed form expression
- Values found numerically
- Tabulated in standard normal form tables
- Z-scores

# Z-score table

### Standard Normal Probabilities

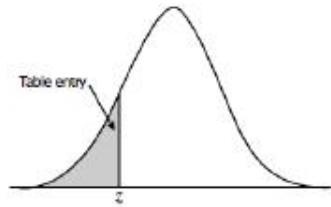


Table entry for  $z$  is the area under the standard normal curve to the left of  $z$ .

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

### Standard Normal Probabilities

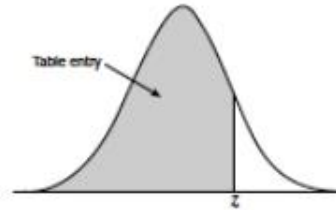


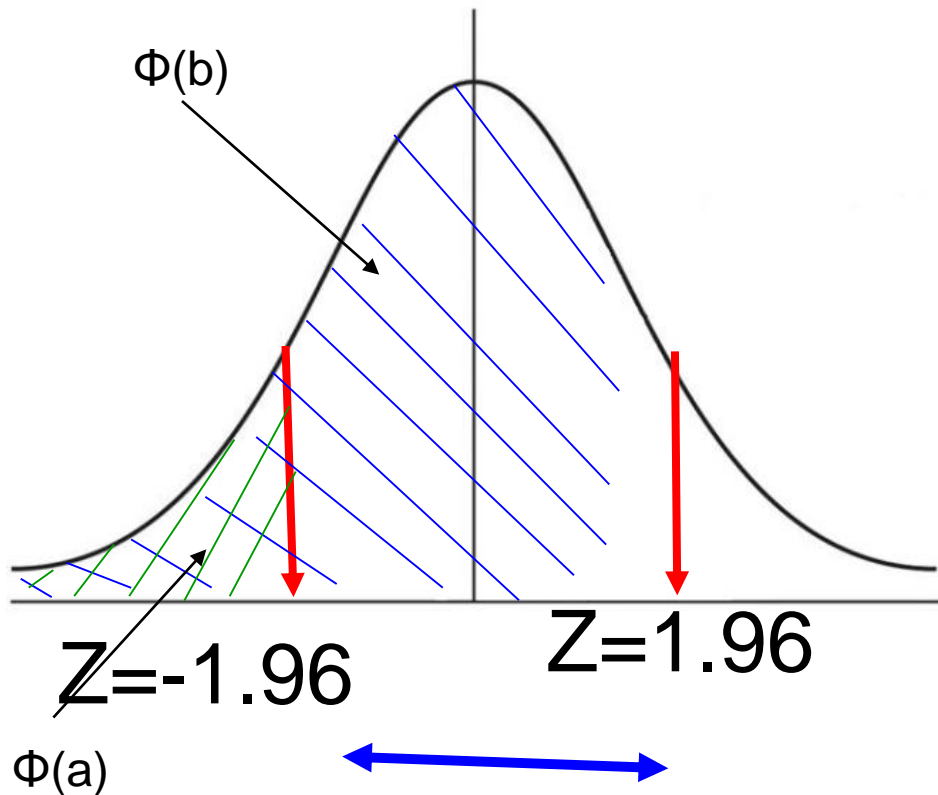
Table entry for  $z$  is the area under the standard normal curve to the left of  $z$ .

[illegible]

# How to use Z-score table

- Read off values, adding row heading and column heading
- Lets verify that 95% of the area under the normal curve lies within  $\pm 1.96\sigma$ ,  
**plus and minus 1.96 times standard deviation**
- The so called 95% confidence interval
- Often used in “Hypothesis Testing”

$$\Phi(1.96) - \Phi(-1.96) = 0.95$$



- Area under the curve from  $-1.96$  to  $+1.96$  is the probability of  $Z$  being in interval

$$-1.96 \leq Z \leq +1.96$$

$$P(-1.96 \leq Z \leq +1.96)$$

# The importance of 95% confidence interval is supreme !

- Often used backward
- We know that  $Z = \pm 1.96$  in  $Z = z$ , that gives me 95% confidence interval

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

- But
- This gives me values of  $X$ , mean, standard deviation if 2 out of 3 quantities are known



# Z-score table

### Standard Normal Probabilities

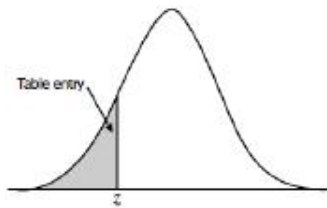


Table entry for  $z$  is the area under the standard normal curve to the left of  $z$ .

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

### Standard Normal Probabilities

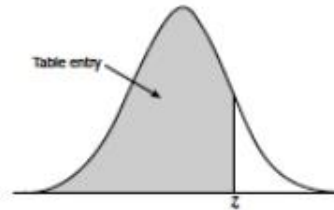
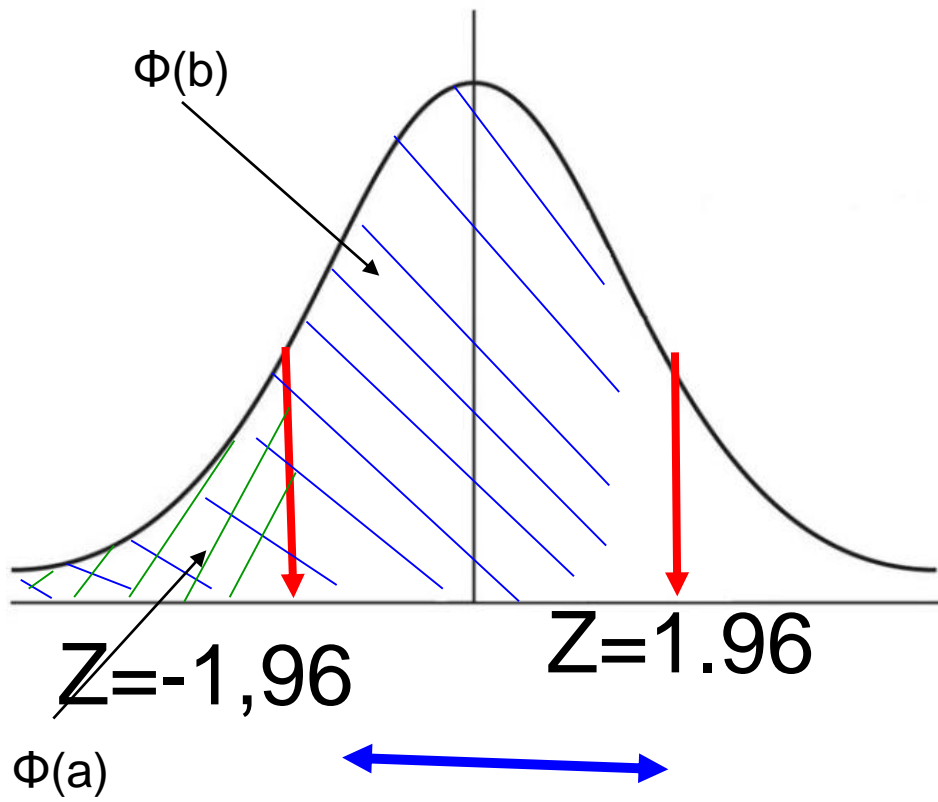


Table entry for  $z$  is the area under the standard normal curve to the left of  $z$ .

[illegible]

# The 95% confidence interval



$$-1.96 \leq Z \leq +1.96$$

- $\Phi(1.96)=0.9750$

- By symmetry

$$\Phi(-1.96)= 1-\Phi(1.96)$$

$$\begin{aligned} \Rightarrow \Phi(1.96)-\Phi(-1.96) &= \\ 2.\Phi(1.96)-1 &= 2 \times 0.975- \\ 1.0 &= 1.95-1.0=0.95 \end{aligned}$$



Showing the calculation again

$$\Phi(1.96)=0.9750$$

By symmetry

$$\Phi(-1.96)= 1-\Phi(1.96)$$

$$\Rightarrow \Phi(1.96)-\Phi(-1.96)$$

$$= 2.\Phi(1.96)-1$$

$$= 2 \times 0.975-1.0 =1.95-1.0=0.95$$

# Interval Estimate

# Sample Mean and Population Mean

- $X_1, X_2, X_3, \dots, X_n$  is a sample from a normal distribution having unknown mean  $\mu$  and known variance  $\sigma^2$ .
- Maximum likelihood point estimator of  $\mu$  is

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

$$\bar{X}$$

- We know that  $\bar{X}$  is normally distributed with mean  $\mu$  and known standard deviation  $\sigma/\sqrt{n}$
- So the following is standard normal distribution:

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

# 95% confidence interval

$$P\left[-1.96 < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1.96\right] = 0.95$$

$$\Rightarrow P\left[-1.96 \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1.96 \frac{\sigma}{\sqrt{n}}\right] = 0.95$$

$$\Rightarrow P\left[\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right] = 0.95$$