upGrad

#LifeKoKaroLift

498

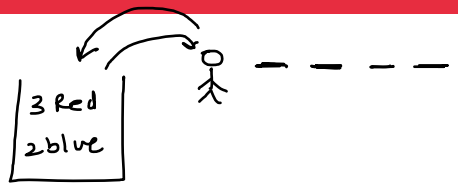
- Random variables
- Expected values
- Prob.
- CLT
- Hypothesis testing
- EDA
- Python
- SQL

Data Science Program


upGrad
Course : Data Science
Lecture On : Course 1 & 2
Practice Session
Instructor : Sumit Shukla

Expected Value

upGrad



Take the ball out of the bag note the colour and put the ball back to the bag } x 4

X : The count of Red balls a player may get.

→ { 4, 3, 2, 1, 0 }

Random variable

That maps the outcome of a random process to a number

Random process

→ The toss of two coin

→ The Roll of a die

→ Pick a ball from a bag

HH
HT] -
TH
TT -

What you are interested in?

X : The count of heads in a toss of two fair coins

{ 2, 1, 0 }

Y : The count of tails in a toss of two fair coins

{ 2, 1, 0 }

If the person gets
two heads → win +20 Rs
else for any other combination
the person will loose → -10 Rs

Z : The amount a player may win

{ +20, -10 }

Expected value : The avg value of the random variable over a large number of experiments

$$E(X) = \sum_{i=1}^{n} x_i \, P(X = x_i)$$

	P(X=xi)
+20	X 0.25] → -2.5
-10	X 0.75

on a long run, each player on an avg is going to loose 2.5 Rs in this game.

Expected Value

upGrad

What to do, If the game is not actually played? How we get the probabilities?

$$P(x) = {}_nC_x\, p^x q^{n-x} = \frac{n!}{(n-x)!\,x!}\, p^x q^{n-x}$$

- n = number of trials
- p = probability of success
- $q = 1 - p$ probability of failure
- x = number of successes in n trials

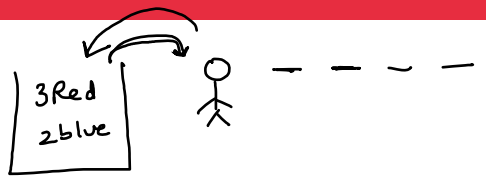
Imp

- The total number of trials is fixed at n.
- Each trial is binary, i.e., it has only two possible outcomes: success or failure.
- Probability of success is the same in all trials, denoted by p.

Expected Value

These should be the involvement of combination at various outcomes

upGrad



X : The count of Red balls a player may get

$X : \{4, 3, 2, 1, 0\}$

i) These should be fixed # of trails = 4
ii) Each trail should be binary < can get Red ball / not get Red ball
iii) $P(S) = P(\text{Red ball}) = \frac{3}{5}$

$P(x) = {}^{n}C_{x}\, p^{x} q^{n-x}$

n = # of trails
x = success in n trails
p = prob. of success
q = prob. of failure

$x = 1$
what's the prob. that out of 4 trails the person get 1 red ball

R	B	B	B
B	R	B	B
B	B	R	B
B	B	B	R

$P(x=1) = {}^{4}C_{1} \left(\frac{3}{5}\right)^{1} \left(\frac{2}{5}\right)^{4-1}$

$= {}^{4}C_{1} \left(\frac{3}{5}\right)\left(\frac{2}{5}\right)^{3} = 4 \times \frac{3}{5} \times \left(\frac{2}{5}\right)^{3} = 0.1536$

$^{n}C_{x} = \frac{n!}{x!\,(n-x)!}$

$\frac{4!}{1!\,3!} = \frac{4 \times 3!}{3!} = 4$

Expected Value

upGrad

Question-1: The 2010 U.S. Census found the chance of a household being a certain size. The data is in the table ("Households by age," 2013). What is the probability that a household will have at least 5 members?

Size of household	1	2	3	4	5	6	7 or more
Probability	26.7%	33.6%	15.8%	13.7%	6.3%	2.4%	1.5%

$P(x \geq 5) = P(5) + P(6) + P(7 \text{ or more})$

$= 0.063 + 0.024 + 0.015$

$P(x \geq 5) = 0.10$

A ● 0.33
B ● 0.10 ✓
C ● 0.26

Question-2: A discrete probability distribution of scoring runs in one throw of a ball by a particular batsman in a cricket match is given in the table. Find the missing probability.

F	0	1	2	3	4	5	6
$P(F)$	0.15	0.27	0.35	0.15	0.04	0.01	?

$P(0) + P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = 1$

$P(6) = 1 - (P(0) + P(1) + P(2) + P(3) + P(4) + P(5))$

$= 1 - (0.15 + 0.27 + 0.35 + 0.15 + 0.04 + 0.01)$

$P(6) = 0.03$

A ● 0.3
B ● 0.02
C ● 0.03 ✓

Expected Value

upGrad

Question-3: A hockey goaltender has a save percentage of 0.920. What would be the expected number of goals scored on this goaltender in a game where she faced 35 shots?

$P(\text{save}) = 0.920$

$P(\text{miss}) = 1 - 0.920$

$\#\ \text{of shots} = (1 - 0.920) \times 35$

$= 2.8$

A ● 5.2
B ● 2.8 ✓
C ● 2
D ● 2.5

Question-4: Which of the following is not a property of a Binomial Experiment?

A ● All trials are identical.
B ● Each trial has only two possible outcomes.
C ● ✓ The probability of success may change from trial to trial. ✗
D ● The purpose of the experiment is to determine the number of successes that occurs during the n trials.

Expected Value

upGrad

The probability of a man hitting the target at a shooting range is 1/4. If he shoots 10 times, what is the probability that he hits the target exactly three times?

Question-1: Choose the correct Random Variable that quantifies the experiment.

A ● X : Number of correct hits.

B ● X: Number of times he hits the target exactly 3 times.

C ● X: Number of subsequent correct hits.

X : The number of correct hits

{10, 9, 8, . . . 0}

# of trails = 10

is this exp. binary = hit / miss

Question-2: What will be the range of the random variable X?

A ● 0<=X<=3

B ● 0<=X<=10

C ● 1<=X<=10

x=3

[H H H H T H H H T T
T T T T T T T T H]

P(hit) = 1/4

$P(X=3) = {}^{10}C_3 \left(\frac{1}{4}\right)^3 \left(\frac{3}{4}\right)^7$

$\frac{10!}{3!\,7!} = \frac{10 \times 9 \times 8 \times 7!}{3 \times 2 \times 1 \times 7!}$

$= 120 \times \left(\frac{1}{4}\right)^3 \left(\frac{3}{4}\right)^7$

$= 0.25$

Question-3: What is the probability that he hits the target exactly three times?

A ● 0.15

B ● 0.25

Expected Value

upGrad

One game that is popular at some carnivals and amusement parks involves selecting a floating plastic duck at random from a pond full of ducks. In most cases, the letter S, M, or L appears on the bottom of the duck, signifying that the winner receives a small, medium, or large prize, respectively. The duck is then returned to the pond for the next game.

Although the prizes are typically toys, crafts, etc., suppose that the monetary values of the prizes are as follows: Small is $\$0.50$, Medium is $\$1.50$, and Large is $\$5.00$.

The probabilities of winning an item on 1 duck selection are as follows: Small 60%, Medium 30%, and Large 10%. What is the expected monetary value of the prizes a player can win?

$$E(x) = \sum_{i=1}^{n} x_i \, P(x = x_i)$$

X	$P(X=x_i)$
\$0.5	0.6
\$1.5	0.3
\$5.0	0.1

X: The amount a player may win

$E(x) = 0.5 \times 0.6 + 1.5 \times 0.3 + 5.0 \times 0.1$

$E(x) = 1.25$

On a long run, each player on an avg is going to win \$1.25

Expected Value

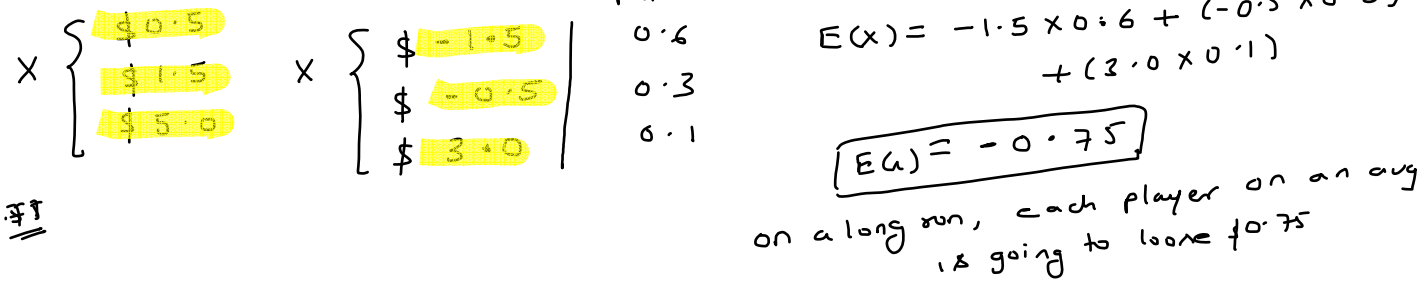
upGrad

Revisiting the charity carnival, recall that when selecting a duck, the average monetary value of the prizes you win per game is $1.25. How can the charity running the carnival make any money if it is paying out $1.25 to each player on average for each game?

To address this, in most cases a player must pay to play a game, and that is where the charity (or any other group running such a game) would earn its money.

Question: Imagine that the cost to play the game is $2.00. What are the expected net earnings for the charity? What are the expected net winnings for a player?

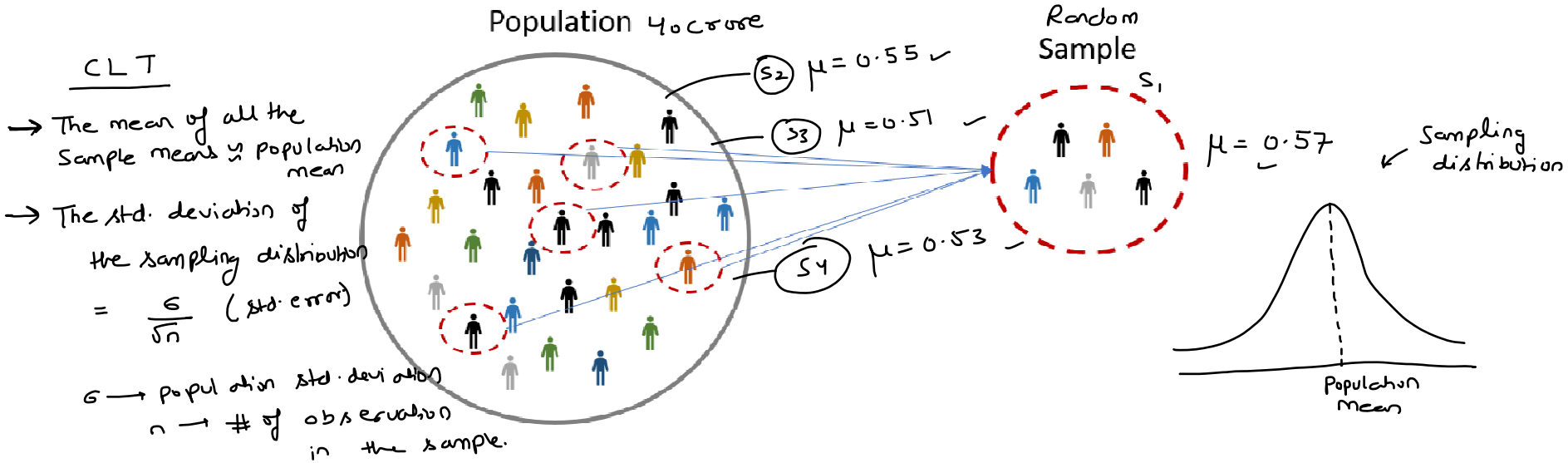
- 0.25
- -0.25
- -0.75



Center Limit Theorem

upGrad

The Problem: Let's say that, for a business application, you want to find out the average number of times people in urban India visited malls last year. That's 400 million (40 crore) people! You can't possibly go and ask every single person how many times they visited the mall. That's a costly and time-consuming process. How can you reduce the time and money spent on finding this number?



Center Limit Theorem

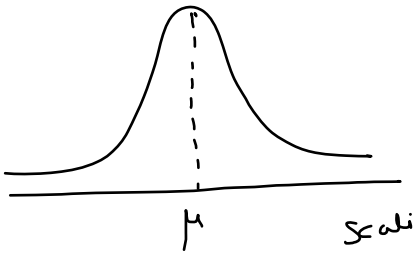
upGrad

→ avg number of people who visited the mall last year is 0.57

or

→ The avg number of people who visited the mall last year is between 0.45 — 0.55 and I am 95% confident about this.

Confidence interval



$$\text{Confidence interval} = \mu_{\bar{x}} \pm \frac{Z\sigma}{\sqrt{n}}$$

Scaling your std. devi according to confidence

Std deviation of sampling dist

Center Limit Theorem

upGrad

CLT is a statistical theory that states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population.

1. Sampling distribution's mean (μ_x) = Population mean (μ)
2. Sampling distribution's standard deviation (Standard Error) = $\frac{\sigma}{\sqrt{n}}$
3. For n > 30, the sampling distribution becomes normally distributed

This is called the "Central Limit Theorem"

Center Limit Theorem

upGrad

Calculating Population Parameters from Sample Parameters

Sample mean ($\bar{X}$)

Sample standard deviation (S)

Sample size (n)

Confidence interval (y% confidence level) = $\left(\bar{X} - \frac{Z^*S}{\sqrt{n}} , \bar{X} + \frac{Z^*S}{\sqrt{n}} \right)$

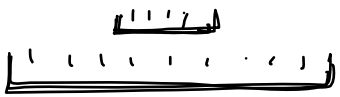
where Z^* is the Z score associated with y% confidence level

Confidence Level	Z^*
90%	±1.65
95%	±1.96
99%	±2.58

Center Limit Theorem

upGrad

A simple random sample of 50 adults womens is obtained, and each person's red blood cell count (in cells per microliter) is measured The sample mean is 4.63. The population standard deviation for red blood cell counts is 0.54. Construct the 95% confidence interval estimate for the mean red blood cell counts of adults.



$n = 50$

$\mu = 4.63$

$\sigma = 0.54$

95% $\quad z(95\%) = \pm 1.96$

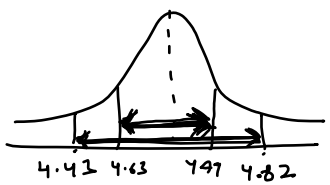
99%

$$4.63 \pm \frac{2.58 \times 0.54}{\sqrt{50}}$$

$$= 4.63 \pm 0.197$$

$$= (4.43, 4.82)$$

$$\mu \pm z\frac{\sigma}{\sqrt{n}} = 4.63 \pm \frac{1.96 \times 0.54}{\sqrt{50}}$$



$$= 4.63 \pm 0.149$$

$$= (4.481, 4.779)$$

The population mean rbc count will lie b- somewhere b/w this range and I am 95% confidence about this

$$\left[\begin{array}{l} 90\% = \pm 1.65 \\ 95\% = \pm 1.96 \\ 99\% = \pm 2.58 \end{array}\right]$$

Center Limit Theorem

upGrad

Revisiting the same example as in the previous slide:

Question: According to WHO a healthy women should have RBC count between 4.2 to 5.4 (in cells per microliter). Based on the previous calculation, what can be concluded?

- The sample of 50 adult womens is healthy.
- The sample of 50 adult womens is unhealthy.
- Can't say

Center Limit Theorem

upGrad

Researchers are concerned about the impact of students working while they are enrolled in classes, and they'd like to know if students work too much and therefore are spending less time on their classes than they should be. First, the researchers need to find out, on average, how many hours a week students are working. They know from previous studies that the standard deviation of this variable is about 5 hours

Question-1: A survey of 200 students provides a sample mean of 7.10 hours worked. What is a 95% confidence interval based on this sample?

A ● (6.10, 8.10)
B ● (6.41, 7.79)
C ● (6.57, 7.63)
D ● (7.10, 8.48)

$n = 200$

$\mu = 7.10$

$\sigma = 5$

95% $\longrightarrow \pm 1.96$

$$7.10 \pm \frac{1.96 \times 5}{\sqrt{200}}$$

$$= 7.10 \pm 0.692$$

$$= (6.41, 7.79)$$

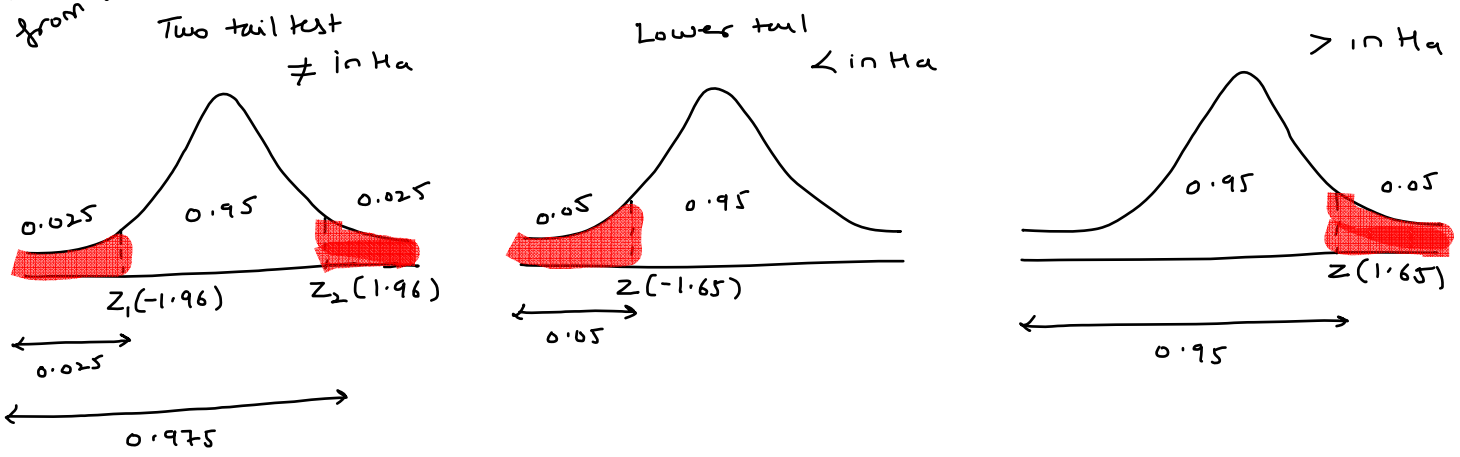
Center Limit Theorem

upGrad

Rule:
always consider the area under the curve from left side

$\alpha = 0.05$

Z-Value and Area Under Curve



Center Limit Theorem

upGrad

biology exam scores = [62, 63, 28, 27, 79, 25, 32, ...]

normally distributed.

68% observation → 64 & 76

95% → 58, 82

99% → 52, 88

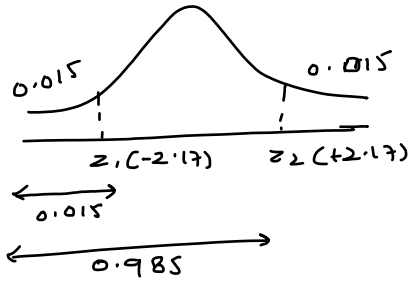
Question-1: What should be the value of z used in a 97% confidence interval?

A ● 2.17

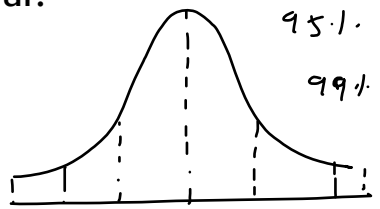
B ● 1.40

C ● 1.81

D ● 1.89



$\frac{0.03}{2} = 0.015$



68% → ± 1 SD

95% → ± 2 SD

99.7% → ± 3 SD

Question-2: A set of biology exam scores are normally distributed with a mean of 70 points and a standard deviation of 6 points. Let X represent the score on a randomly selected exam from this set.

Find P(64<X<76)

64 (−6) 70 (+6) 76

A ● 0.841

B ● 0.158

C ● 0.68

D ● Cannot say from the information provided

Formulating Null and Alternate Hypothesis

Formulate the hypothesis
- make the conclusion (Testing) / CVM / p-value

Example-1: A restaurant owner installed a new automated drink machine The machine is designed to dispense 530 mL of liquid on the medium size setting. The owner suspects that the machine may be dispensing too much in medium drinks. They decide to take a sample of 30 medium drinks to see if the average amount is significantly greater than 530 mi.

$H_0: \mu \leq 530 ml$
$H_a: \mu > 530 ml$

Rule-1: NULL: $\geq, \leq, =$; Alternate: $>, <, \neq$

Rule-2: The claim doesn't always points to a alternate hypothesis

Rule-3: NULL & Alternate hypothesis are always going to be opposite of each other without any overlap.

at least, at most, equal = null

greater, less, different = alternate.

Formulating Null and Alternate Hypothesis

Example-2: A city had an unemployment rate of 7%, percent. The mayor pledged to lower this figure and supported programs to decrease unemployment. A group of citizens wanted to test if the unemployment rate had actually decreased, so they obtained a random sample of citizens to see what proportion of the sample was unemployed.

$H_0: \mu \geq 7\%$

$H_a: \mu < 7\%$

Hypothesis Testing

upGrad

1

Formulating Null and Alternate Hypothesis

Question-1: We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0). The null and alternative hypotheses are:

A ● H0 = 2, Ha !=2 ✓
B ● H0!=2, Ha=2
C ● H0>2, Ha<2

H_0 : $\mu = 2$
H_a : $\mu \neq 2$

Question-2: In an issue of U.S. News and World Report, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U.S. students take advanced placement exams and 4.4% pass. Test if the percentage of U.S. students who take advanced placement exams is more than 6.6%. State the null and alternative hypotheses.

A ● H0>0.066, Ha = 0.066
B ● H0<=0.066, Ha>0.066 ✓
C ● H0>=0.066, Ha<0.066

H_0 : $\mu \leq 6.6\%$
H_a : $\mu > 6.6\%$

Slide 22

1 Don't add these question
Sumit Shukla, 8/1/2020

2

Formulating Null and Alternate Hypothesis

Question-3: eHealthInsurance claims that in 2011, the average monthly premium paid for individual health coverage was $183. Suppose you are suspicious that the average, or mean, cost is actually higher. State the null and alternative hypothesis you would use to test this.

A ● H0 = 183, Ha !=183
B ● H0 <= 183, Ha >183 ✓
C ● H0 < 183, Ha !=183

$H_0: \mu \leq 183$ | $\mu = 183$
$H_a: \mu > 183$ | $\mu > 183$

Question-4: A survey was conducted to get an estimate of the proportion of smokers among the graduate students. Report says 38% of them are smokers. Chatterjee doubts the result and thinks that the actual proportion is much less than this. Choose the correct choice of null and alternative hypothesis Chatterjee wants to test.

A ● Ho: p=0.38 versus Ha: p<=0.38
B ● Ho: p=0.38 versus Ha: p >0.38
C ● Ho: p>=0.38 versus Ha: p< 0.38 ✓
● None of the above.

$H_0: \mu \geq 38\%$
$H_a: \mu < 38\%$

Slide 23

2 Don't add this question
Sumit Shukla, 8/1/2020

95%

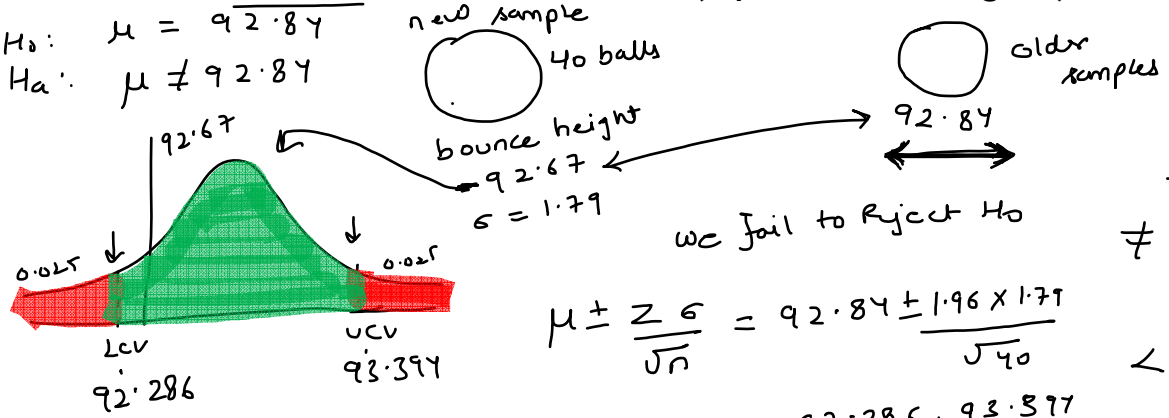
Sig. level = area of Rejection $1 - 0.05 = 0.95$

Critical Value Method

A sample of 40 new baseballs had a bounce height mean of 92.67 inches and a SD of 1.79 inches. Use a .05 sig. level to determine whether there is sufficient evidence to support the claim that the new balls have bounce heights with a mean different from 92.84 inches. (a previous test figure).

$H_0: \mu = 92.84$

$H_a: \mu \neq 92.84$



we fail to Reject H_0

$$\mu \pm \frac{z\sigma}{\sqrt{n}} = 92.84 \pm \frac{1.96 \times 1.79}{\sqrt{40}}$$

$$= 92.286, 93.394$$

we fail to reject H_0

Rule-4

$\neq$ in H_a : Two tail test
Rejection Region on both side

$<$ in H_a : Lower tail test
left side

$>$ in H_a : upper tail test
Right side.

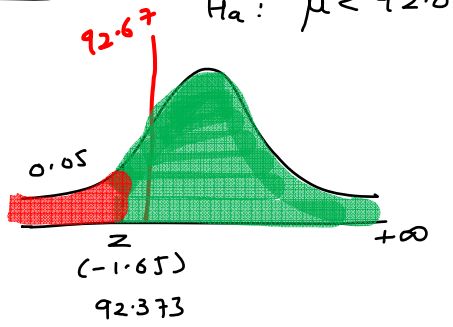
Critical Value Method

A sample of 40 new baseballs had a bounce height mean of 92.67 inches and a SD of 1.79 inches. Use a .05 sig. level to determine whether there is sufficient evidence to support the claim that the new balls have bounce heights with a mean less than 92.84 inches. (a previous test figure).

Step-1 :

$H_0 : \mu \geq 92.84$

$H_a : \mu < 92.84$



$$\mu - \frac{z \times \sigma}{\sqrt{n}}$$

$$= 92.84 - \frac{1.65 \times 1.79}{\sqrt{40}}$$

$$= 92.373$$

fail to reject H_0

Hypothesis Testing

upGrad

α

$P > 0.05 \rightarrow$ fail to Reject null

$P < 0.05 \rightarrow$ Reject null

[P is high, null will fly
P is low, null will low

P- Value Method

A sample of 40 new baseballs had a bounce height mean of 92.67 inches and a SD of 1.79 inches. Use a .05 sig. level to determine whether there is sufficient evidence to support the claim that the new balls have bounce heights with a mean different from 92.84 inches. (a previous test figure).

Step-1 : $H_0 : \mu = 92.84$

$H_a : \mu \neq 92.84$

x = test mean (old sample

μ = sample mean (new sample

step-2 : $Z = \frac{x - \mu}{\sigma/\sqrt{n}} = \frac{92.84 - 92.67}{1.79/\sqrt{40}}$

$Z = 0.6006$

step-3 : find AUC using z value

$z(0.600) = 0.7257$

Step-4 : find p-value

$P = 1 - AUC$

$P = 1 - 0.7257$

$= 0.2743$

Step-5 : If your test is two tail

$P = PX2 = 0.2743 \times 2$

$= 0.5486$

If your test is single tail

$P = P$

$\rightarrow$ fail to Reject null.

EDA

upGrad

Following table is a correlation matrix between the prices of stocks of the following companies: NVIDIA, Ford, Shell, and Alphabet.

Based on the information provided in the table, answer the following questions.

	A	B	C	D	E
1		**Nvidia**	**Ford**	**Shell**	**Alphabet**
2	**Nvidia**	1			
3	**Ford**	-0.7601	1		
4	**Shell**	0.9343	-0.6293	1	
5	**Alphabet**	0.9691	-0.7881	0.8920	1

Question-1: What can be concluded by looking at the correlation between Ford and Nvidia?

1. Both have a correlation of **0.76**
2. As the stock price of Nvidia goes up, it's highly likely the stock price of Ford will also rise.
3. As the stock price of Nvidia goes up, it's highly likely the stock price of Ford will fall.

- A ● 1 and 2
- B ● Only 1
- C ● 1 and 3

Expenditures of a Company (in Lakh Rupees) per Annum Over the given Years.

Year	Item of Expenditure				
	Salary	Fuel and Transport	Bonus	Interest on Loans	Taxes
1998	288	98	3.00	23.4	83
1999	342	112	2.52	32.5	108
2000	324	101	3.84	41.6	74
2001	336	133	3.68	36.4	88
2002	420	142	3.96	49.4	98

1710 17

Refer to the above table and answer the following questions

Question-1: The total amount of bonus paid by the company during the given period is approximately what percent of the total amount of salary paid during this period?

- 0.1%
- 0.5%
- 1% ✓

$\frac{17}{1710} \times 100 \approx 1\%$

EDA

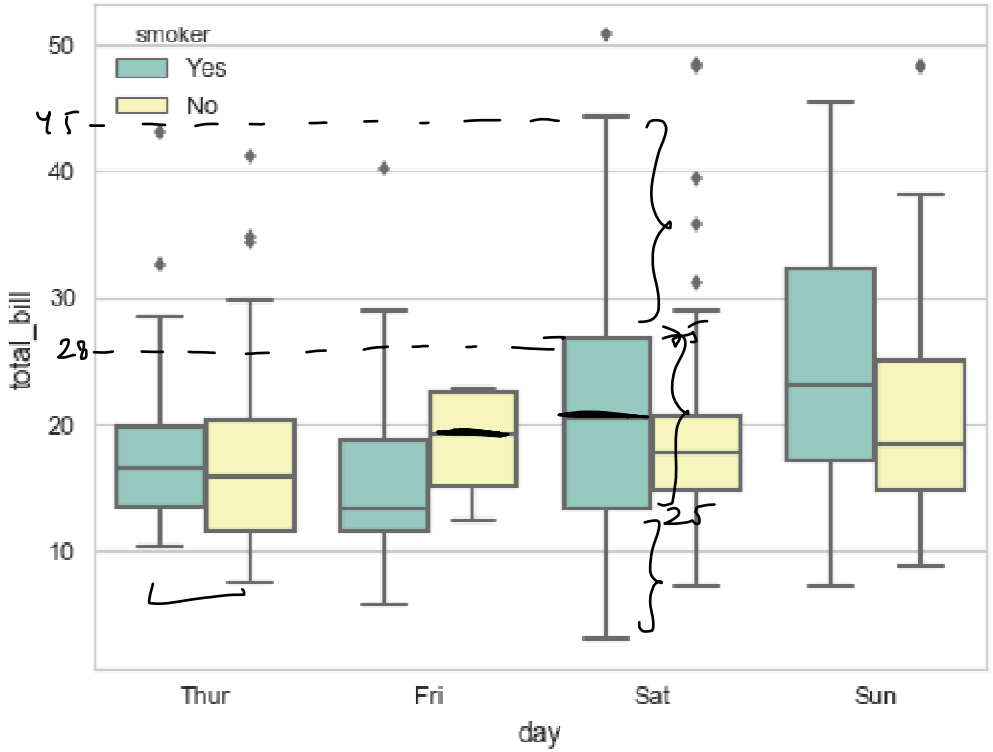
upGrad

Question-1: What is the upper bound range of the total bill for the Smokers on Saturday?

A ● 28-45 ✓
B ● 30-50
C ● 38-50

Question-2: On which particular day the median total bill for both Smokers and Non-Smokers is approximately same?

A ● Thursday ✓
B ● Friday
C ● Saturday
D ● Sunday



Python < MCQ = 5 coding

Rule:

By default indexing works from left to right side.

```
ls =  [ 1,  2,  3 , 'Sumit' , [1, 2, 3 ] ]
+ve    0    1   2      3          4
-ve   -5   -4  -3     -2         -1
```

ls [0 : 3] = [1, 2, 3]

ls [-1 : -3] = [ ]

ls [-1 : -3 : -1] = [[1,2,3], 'Sumit']

-1 included

-3 excluded

-1 reverse the direction of indexing

upGrad

#LifeKoKaroLift

Thank You!

References:

- https://www.superprof.co.uk/resources/academic/maths
- https://www.khanacademy.org/math
- YouTube: MATHRoberg
- https://www.surveygizmo.com/resources
- https://courses.lumenlearning.com/introstats1/chapter/null-and-alternative-hypotheses/
- http://www.csun.edu
- http://users.stat.ufl.edu
- https://www.ck12.org/book/CK-12-Advanced-Probability-and-Statistics-Concepts/
- https://www.indiabix.com/data-interpretation