

Statistics

Part I

Types of Studies

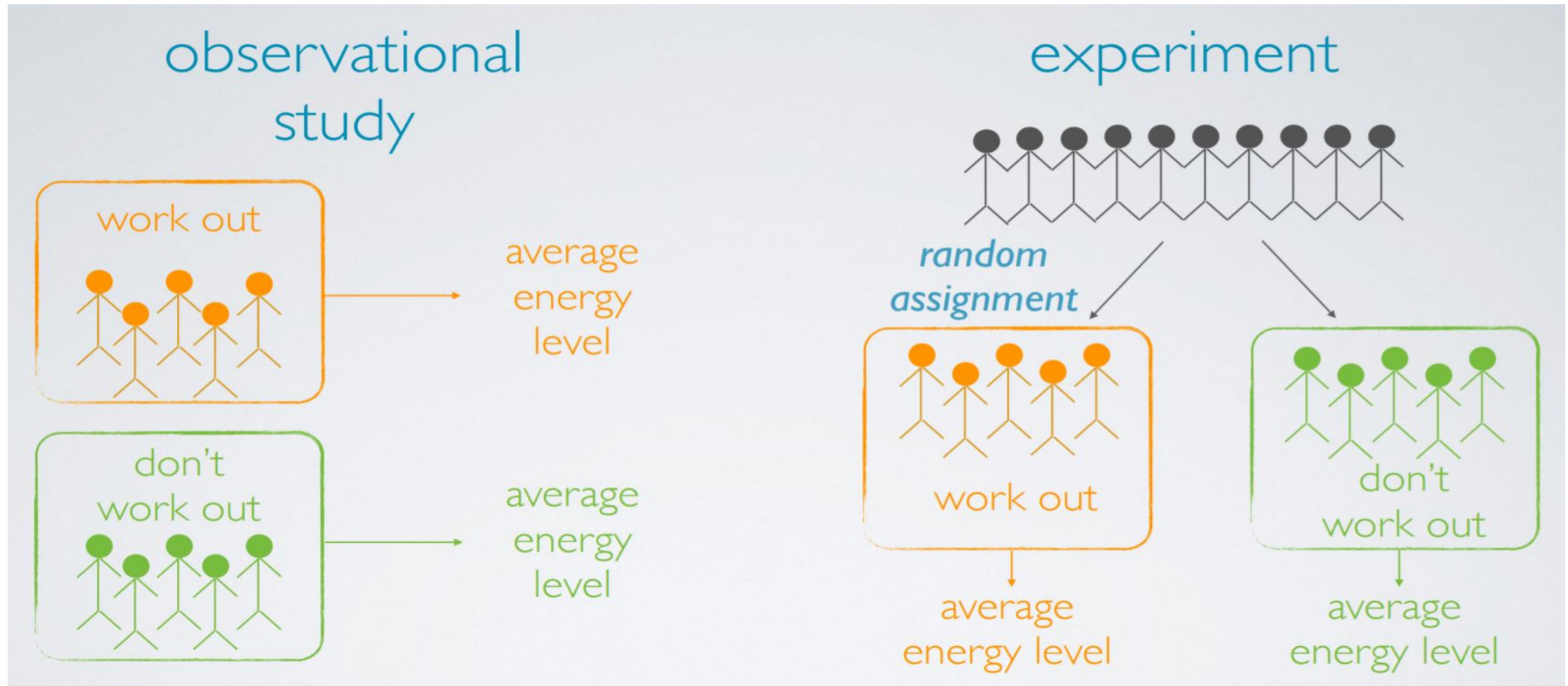
- Observational Studies:

- Researcher collects the data in such a way that he does not directly involved in how the data arise.
- Only establishes the association(correlation)between explanatory and response variables.
 - ❖ Retrospective- old data
 - ❖ Prospective – current data

- Experimental Studies

- Researcher randomly assigns subjects to various treatments.
- Therefore can establish causal connections between variables.
- Treatment and Control group.

Observational Studies and Experiment



Confounding Variable

- Confounding Variable: variables that affect both the explanatory and the response variable, and that make it seem like there is a relationship between them.
- Let a study tracked sunscreen use and skin cancer. It was found that more sunscreen used more prone to skin cancer.
- Does this mean that sunscreen causes skin cancer??
- There can be any other variable associated with the study not taken into account i.e sun exposure.
- Someone is more exposed to sun uses more sunscreen and more chances of skin cancer.
- This sun exposure is the confounding variable , correlated with both the explanatory and the response variable.
- Another example is of diet and weight and age is confounding variable.

Statistical Terms



Image credit: Wonderlane CC BY 2.0 <http://www.flickr.com/photos/wonderlane/623188866/>

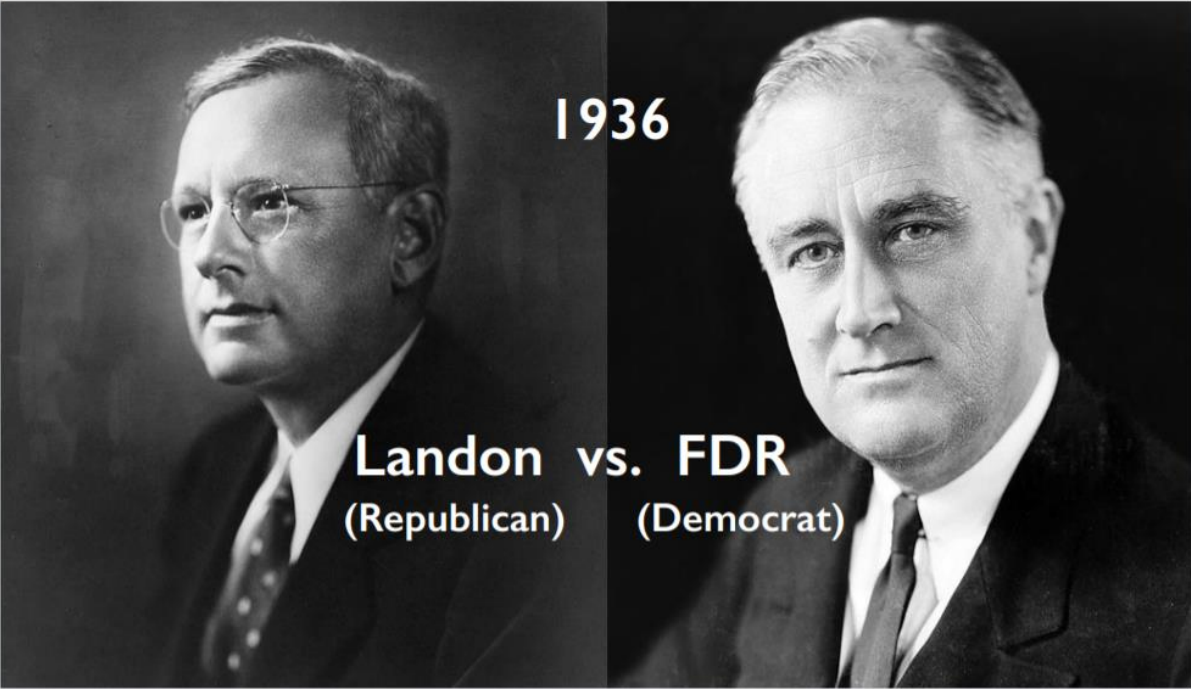
Population Vs Sample

- What is the average mercury content in swordfish in Atlantic Ocean?
- Each fish is a case.
- Instead you can take 50 fish sample.
- Exploratory Data Analysis-→ Sampling-→ Inferences. E.g Soup
- If soup sample is taken without stirring.

Sample Biasing

- Convenience Sampling : Individuals who are easily accessible are more likely to be included in the sample.
- Non Response Sampling : If only a (non-random) fraction of the randomly sampled people respond to a survey such that the sample is no longer representative of the population.
- Voluntary Sampling : Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue

Case Study



1936

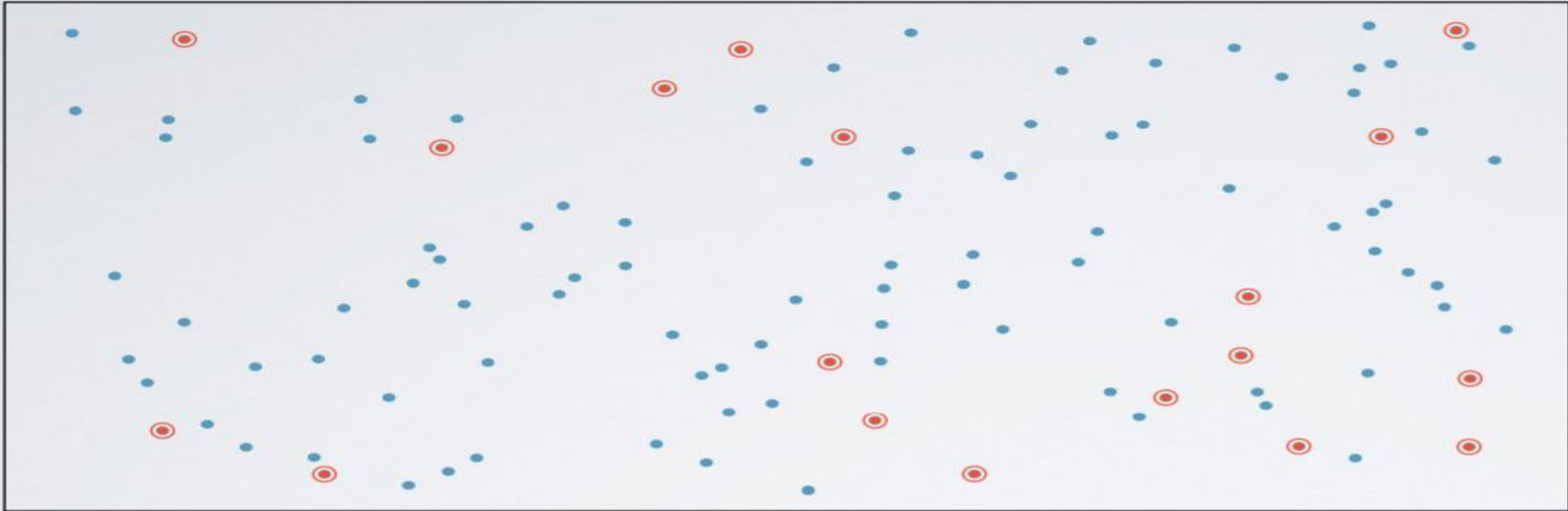
Landon vs. FDR
(Republican) (Democrat)

The Literary Digest
Election results

Lose with 43% of the votes
Win with 62% of the votes

Types of sampling

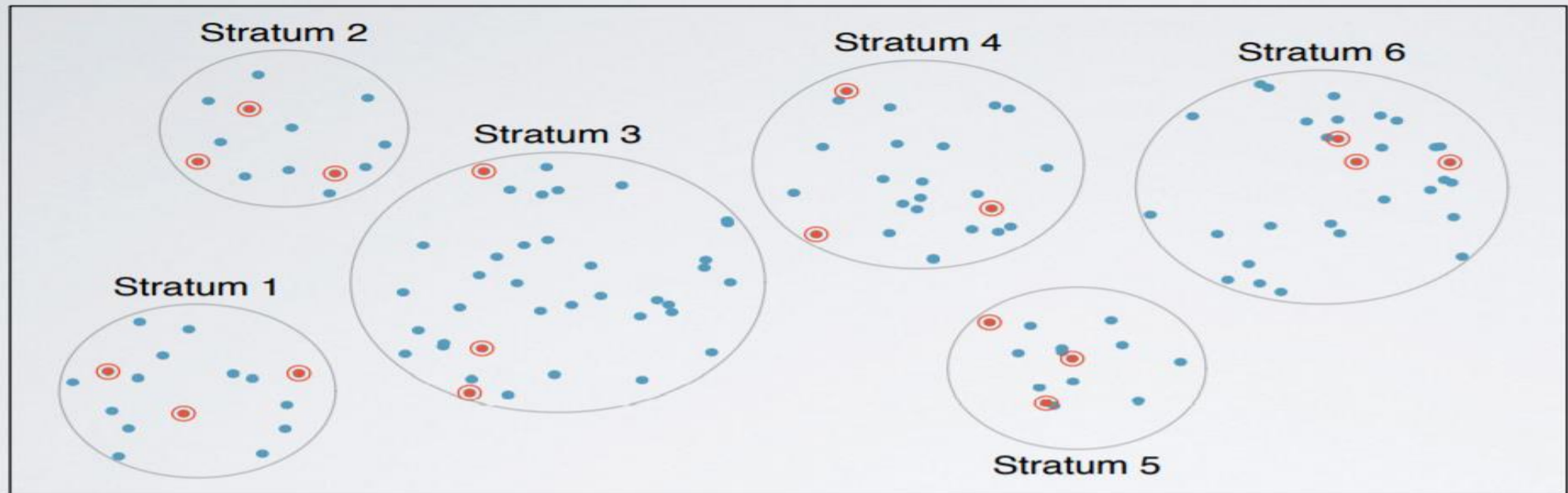
simple random sample (SRS)



each case is equally likely to be selected

Types of sampling

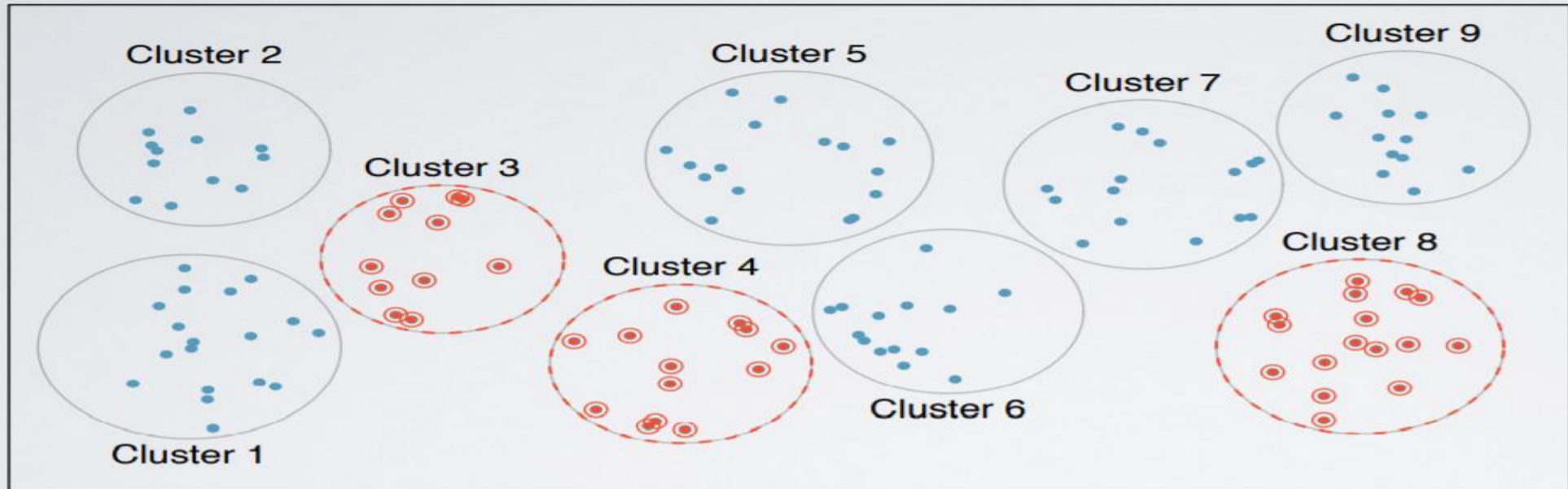
stratified sample



divide the population into homogenous **strata**,
then randomly sample from within each stratum

Types of sampling

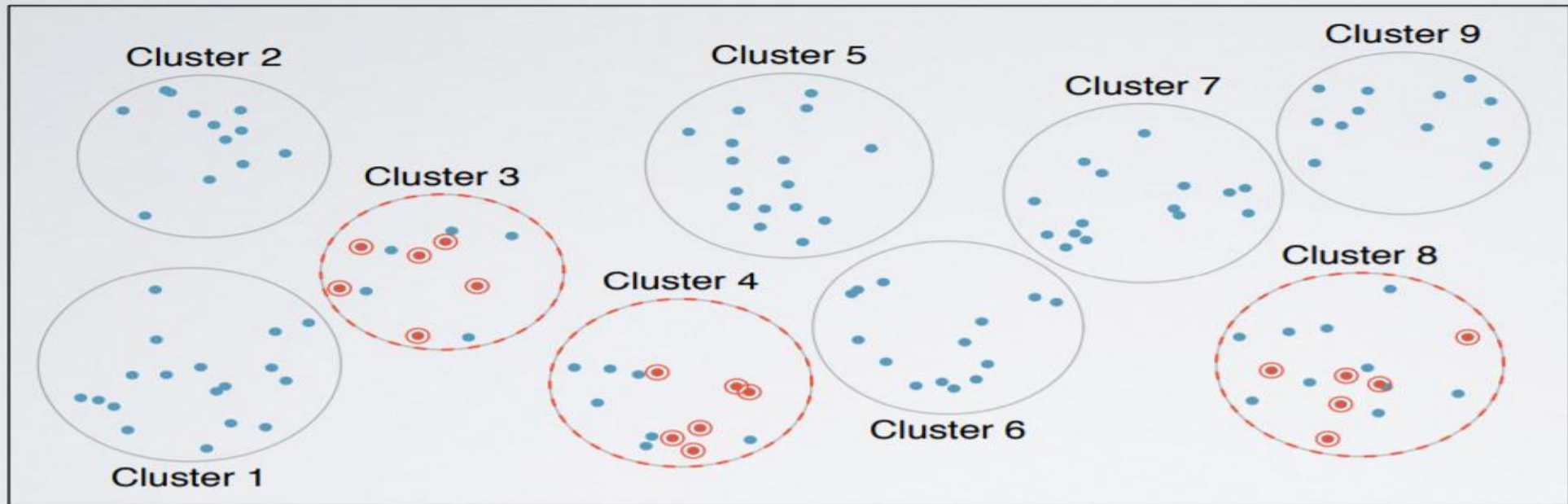
cluster sample



divide the population **clusters**,
randomly sample a few clusters,
then sample all observations within these clusters

Types of sampling

multistage sample



divide the population **clusters**,
randomly sample a few clusters,
then randomly sample within these clusters

Principles of Experimental Design

(1) control

compare treatment of interest to a control group

(2) randomize

randomly assign subjects to treatments

(3) replicate

collect a sufficiently large sample, or replicate the entire study

(4) block

block for variables known or suspected to affect the outcome

principles of
experimental design

Blocking Variable



more on blocking

- ▶ design an experiment investigating whether energy gels help you run faster:
 - ▶ treatment: energy gel
 - ▶ control: no energy gel
- ▶ energy gels might affect pro and amateur athletes differently
- ▶ block for pro status:
 - ▶ divide the sample to pro and amateur
 - ▶ randomly assign pro and amateur athletes to treatment and control groups
 - ▶ pro and amateur athletes are equally represented in both groups

Experimental Terminology

placebo

fake treatment,
often used as the
control group for
medical studies

placebo effect

showing change
despite being on
the placebo

blinding

experimental units
don't know which
group they're in

double-blind

both the experimental
units and the researchers
don't know the group
assignment

experimental
terminology

Example 1

Rahul asked for donation to his society members for Ganesh Utsav. 10% of the society members did not give any donation. Some 40% of members gave him 300/- Rs, 50% of Members gave him 500/- Rs. Select the correct statement:

- A. The mean is greater than median.
- B. The median is greater than mean
- C. The mean is same as that of median.
- D. Cannot be determined.

Correct Ans: B

Example 2

Which of the following is true?

- a. In a symmetric distribution, more than 50% data are below and less than 50% are above the mean.
 - b. In a left skewed distribution, roughly 50% of data are below and 50% are above the mean
 - c. In a right skewed distribution, less than 50% of the data are below the mean
 - d. In a left skewed distribution, less than 50% of the data are below the mean
- Ans. 'd' (Since median is the 50th percentile, and in a left skewed distribution $\text{mean} < \text{median}$, less than 50% of the data will be smaller than the mean)

Robustness

- Measures of Centre: Mean, Median.
 - Robust: Median
 - Non Robust: Mean
- Measures of spread – Range, SD ,IQR
 - Robust: IQR
 - Non Robust: Range, SD
- Robust stats are useful for describing skewed distributions or extreme observations.
- Non Robust statistics are useful for symmetric distributions

Example 3

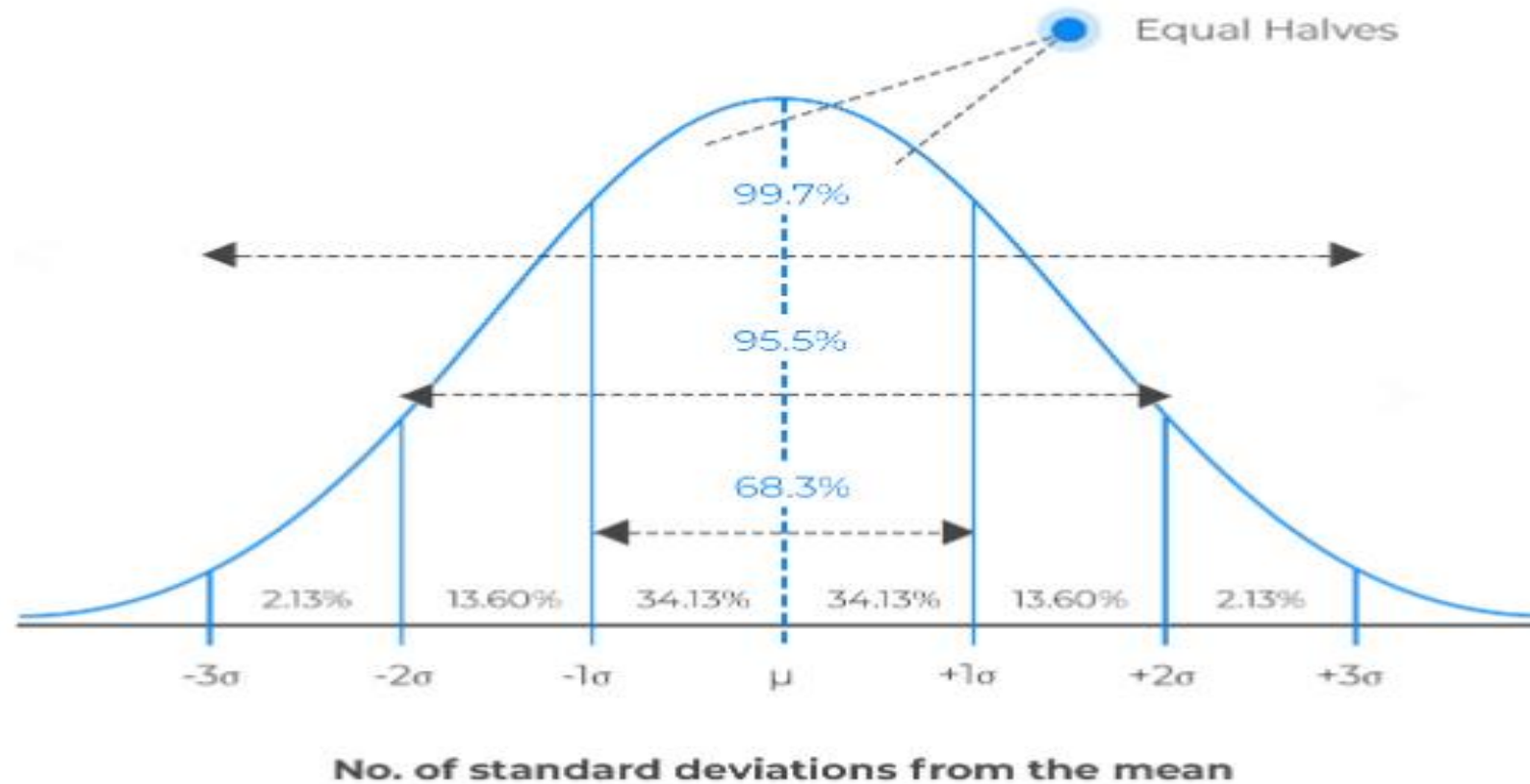
- You have collected data from a large Company with many employees having a salary of less than 1 L per month, a few managers who have salaries between 1 L to 1.5 L and a few high level executives whose salaries are beyond 1.5 L. determine the shape, these salaries would be expected to follow and decide whether median or mean would better represent a typical salary of an employee of the company.
- A. Right skewed, mean is a better measure of salary
 - B. Right skewed, median is a better measure of salary
 - C. Symmetric, mean is a better measure of salary
 - D. Symmetric, median is a better measure of salary
 - E. Left skewed, mean is a better measure of salary
 - F. Left skewed, median is a better measure of salary
- Ans. B

Normal Distribution

- Unimodal and Symmetric
- Also known as a Bell curve
- Follows very strict guidelines about how variability of data are distributed around the mean.
- Many variables are 'nearly normal'

Normal distribution curve

- Its short hand notation is $X \sim N(\mu, \sigma^2)$

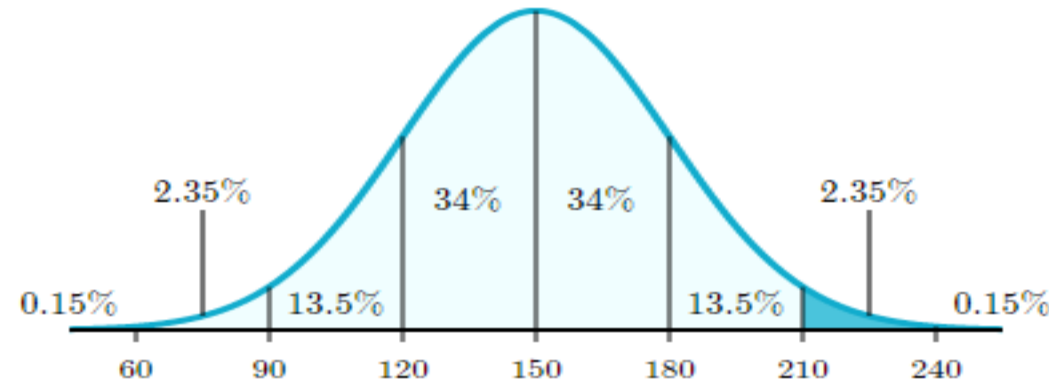


<https://analystprep.com/cfa-level-1-exam/quantitative-methods/key-properties-normal-distribution/>

Example 4

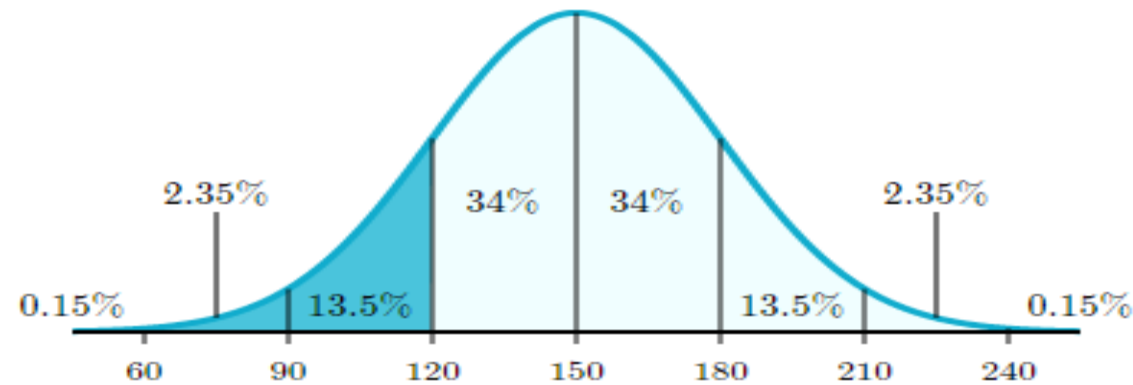
- A certain variety of pine tree has a mean trunk diameter $\mu=150\text{cm}$ and a standard deviation $\sigma=30\text{cm}$. Approximately what percent of these trees have a diameter greater than 210?

Ans: $2.35\% + 0.15\% = 2.5\%$



Example 5

- A certain variety of pine tree has a mean trunk diameter $\mu=150\text{cm}$ and a standard deviation $\sigma=30\text{cm}$. A certain section of a forest has 500 of these trees. Approximately how many of these trees have a diameter smaller than 120cm?



Ans: $0.15\% + 2.35\% + 13.5\% = 16\%$

16% of 500 = 80

Example 6

Scholastic Aptitude Test (SAT) scores are distributed nearly normally with mean 1500 and Standard Deviation 300, which of the following is false?

- a. Roughly 68% of students score between 1200 to 1800
- b. Roughly 95% of students score between 900 to 2100
- c. Roughly 99% of students score between 600 to 2400
- d. No. students can score below 600

Ans. d

Example 7

- A doctor collects a large set of heart rate measurements, that approximately follow a normal distribution. The Doctor reports only 3 statistics mean = 110 beats per minute; minimum = 65 beats per minute and the maximum = 155 beats per minute. Which of the following is most likely to be the Standard deviation?
 - a. 5
 - b. 15
 - c. 35
 - d. 90

Ans. b

Example

A college Admissions officer wants to determine which of the 2 applicants scored better in their test. Student A scored 1800 in Scholastic Aptitude Test (SAT) and Student B who scored 24 in American College Testing Examination.

SAT $N(\mu=1500, \sigma=300)$

ACT $N(\mu=21, \sigma=5)$

- It is difficult to find out the best candidate as both the exams taken were different with different mean and standard deviation.
- Hence both the exam scores should be scaled to a common scale.
- This is called as standardization.
- We find a standard Z score and then compare the performances.

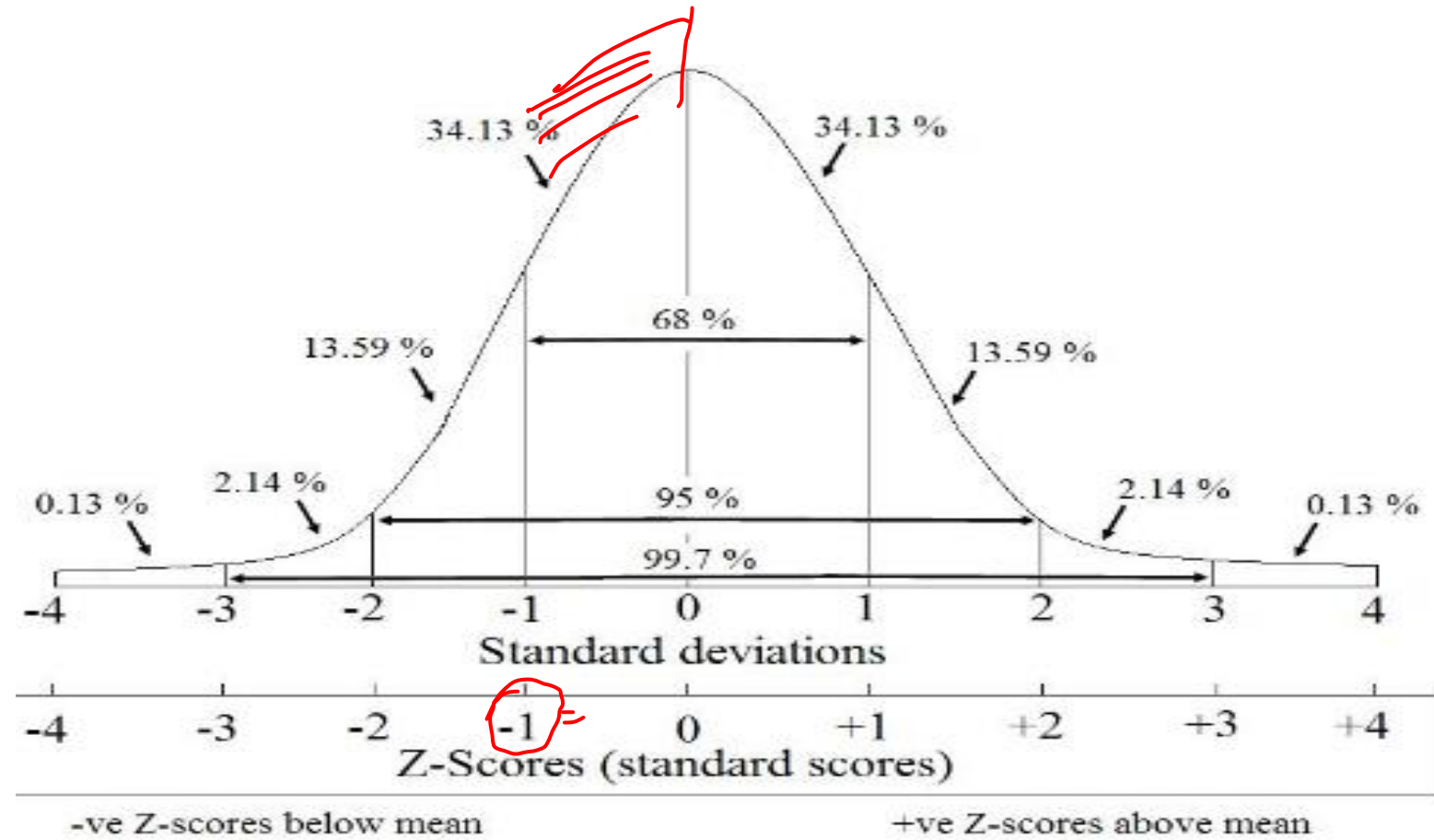
StandardiZed Score

- $Z = (\text{Observation} - \text{mean}) / \text{Standard Deviation}$

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

- The value of the z-score tells you how many standard deviations you are away from the mean. If a z-score is equal to 0, it is on the mean.
- A positive z-score indicates the raw score is higher than the mean average. For example, if a z-score is equal to +1, it is 1 standard deviation above the mean.
- A negative z-score reveals the raw score is below the mean average. For example, if a z-score is equal to -2, it is 2 standard deviations below the mean.
- Enables us to compare two scores that are from different normal distributions. The standard score does this by converting (in other words, standardizing) scores in a normal distribution to z-scores in what becomes a standard normal distribution.
- Useful in identifying Unusual observations.
- Usually $|Z| > 2$ is considered unusual.

Z score contd...



Solution with Z score

- Z score is given by

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

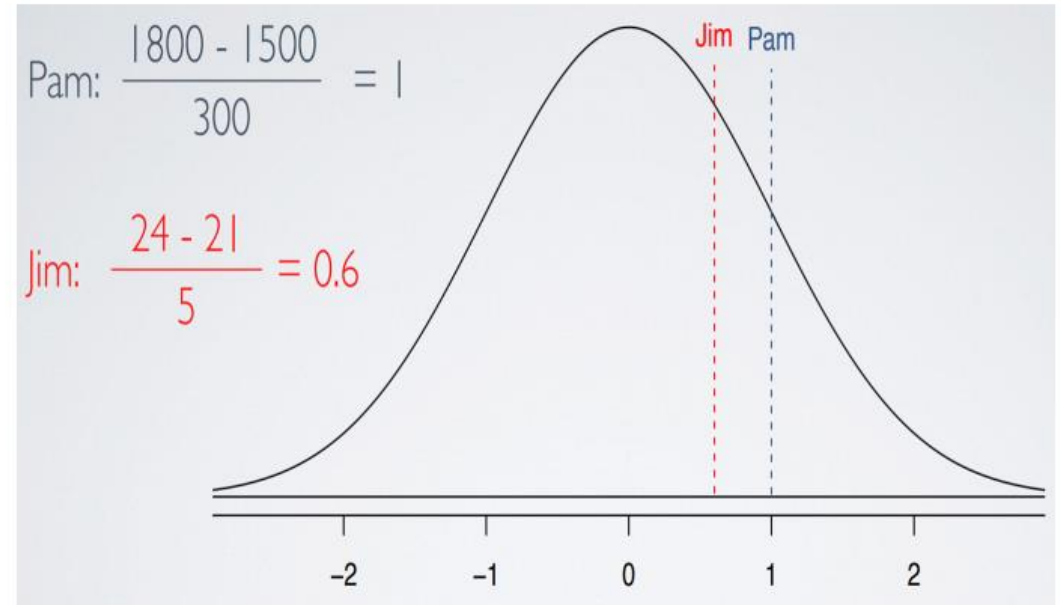
- Z score of A: $Z = (1800 - 1500) / 300$

Z score of A = 1

- Z score of B: $Z = (24 - 21) / 5$

Z score of B = 0.6

- Z score of A is greater than that of B, hence A performed better in the exam.



Example 8

Scores on a standardized test are nearly normally distributed with a mean of 100 and a standard deviation of 20. If these scores are converted to standard normal Z-scores, which of the following statements will be correct?

- a. The mean will be 0 and the median should be roughly 0 as well
- b. The mean will equal 0, but the median can not be determined
- c. The mean of the standardized scores will be 100
- d. The mean of the standardized scores will be 5

Ans. 'a'

Percentiles

- When the distribution is normal, Z-scores can also be used to calculate the percentile.
- Percentiles are defined as the percentage of observations that fall below a given data point
- Graphically, percentile is the area below the probability distribution curve, to the left of that observation.

Percentile from Z score

To find the percentile from the Z score:

- Find the value of Z score from given observed value , mean and SD

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

- Then find the corresponding percentile value from Z score from the probability distribution table(or you can use software to find it).
- For example: for a Z score of 1.32 the percentile value is 0.9066

Probability Distribution Table

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633

+ve Z.

Z	Second decimal place of Z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6702	0.6736	0.6773	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

*For Z > 3.50, the probability is greater than or equal to 0.9998.



0.1

0

Second decimal place of Z										Z
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	-3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	-3.3
0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	-3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	-3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	-3.0
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	-2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	-2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	-2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	-2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	-2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	-2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	-2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	-2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	-2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	-2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	-1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	-1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	-1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	-1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	-1.5
0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808	-1.4
0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968	-1.3
0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151	-1.2
0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357	-1.1
0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587	-1.0
0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841	-0.9
0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119	-0.8
0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420	-0.7
0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743	-0.6
0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085	-0.5
0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446	-0.4
0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821	-0.3
0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207	-0.2
0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602	-0.1
0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000	-0.0

*For $Z \leq -3.50$, the probability is less than or equal to 0.0002.

0.01

-1.2

-0.08

-1.28

$Z = -1.28$

Example 9

SAT scores are distributed normally, with mean 1500 and SD 300. John scored 1800 in his SAT. What is his percentile score?

Solu: To find the percentile score we have to use Z score table.

$$Z = (\text{observed value} - \text{mean value}) / \text{SD}$$

$$= (1800 - 1500) / 300$$

$$Z = 1$$

From probability distribution table for $Z=1$, percentile value is 0.8413. Therefore in percentage it is 84.13%.

OR

R command:

```
> pnorm (1800, mean = 1500, SD = 300)
```

```
[1] 0.8413
```

This means John scored better than 84.13% of the SAT students

Example 10

ACT scores are distributed nearly normally, with mean 21 and SD 5. Jim scored 24 in ACT. Which of the following is true?

- a. Jim's Z-score is -0.6
- b. Jim scored better than approx. 72.57% of the ACT students
- c. 72.57 % of ACT takers scored better than Jim
- d. Jim's percentile score is 60%

Ans. $Z = 0.6$ which implies that ans. is 0.7257. (ref. Probability distribution table). Hence answer is 'b'

Example 11

A friend tells you that she scored in the top 10% of the SAT (mean = 1500, SD = 300). What is the lowest score she could have gotten?

The total area under the curve is 1, the percentile score associated with cut off value for top 10% is $1 - 0.1 = 0.9$

Using the table, the value associated with 90th percentile i.e. 0.9 is 0.8997 and corresponding Z-score is 1.28.

Hence, $1.28 = (x - 1500) / 300$ i.e. $x = 1884$.

In R ,

```
>qnorm(0.90,1500,300) (observe d)
```

[1]1884.465

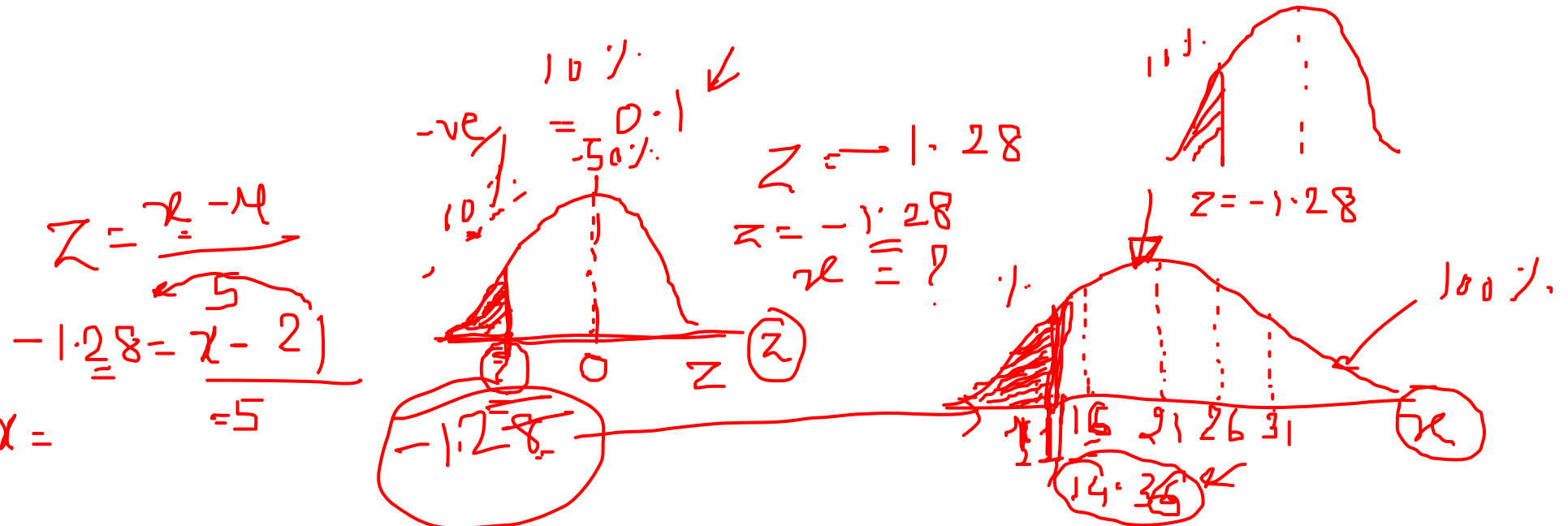
Example 12 a

ACT scores are distributed nearly normally, with mean 21 and SD 5. A friend of yours tells you that he scored in the bottom 10% in the exam. What is the highest possible score, he would have got?

- a. 14.6
- b. 27.4
- c. 12.75
- d. 29.25

Ans. 'a'

$Z = (x - 21) / 5$ From table (negative Z) we get $z = -1.28$ for 10% of value.
Therefore, $[(-1.28) * (5)] + 21 = 14.6$



Example 13

- Consider the SAT and ACT statistics as given:

SAT Statistics

	SAT	ACT
Mean	1500	21
SD	300	5

Q1: Shannon is randomly selected SAT taker.
What is the probability Shannon scores at least 1630 on her SAT?

Q2: What is the probability of Shannon scoring SAT scores less than 1630?

Example 14

The distribution of number of hours of sleep the college students get has a mean of 7 hours. 62% of the students sleep between 6 to 8 hours, 92% of the students sleep between 5 to 9 hours and 95% sleep between 4 to 10 hours. Which of the following is true?

- a. The distribution is more variable than a normal distribution with mean 7 and SD 1
- b. The distribution is less variable than a normal distribution with mean 7 and SD 1
- c. The distribution is nearly normal

Ans: a (62% instead of 68%, 92% instead of 95 and 95 instead of 99%).

Example 14

Suppose weights of checked baggage of airlines passengers follow a nearly normal distribution with mean 45 pounds and SD 3.2 pounds. Most airlines charge a fee for baggage that weigh in excess of 50 pounds. What % of airline passengers are expected to incur this fee?

If in a month, 64000 passengers visited the airport, with Rs. 2500 as excess charges, how much revenue was generated at the airport?

Example

$$Z = (50-45)/3.2 = 1.56$$

For a Z-score of 1.56, we get 0.9406 as area below the curve.

Hence complementing it, $1-0.9406 = 0.0594$ i.e. 5.94% is the expected answer

The revenue generated would be –

5.94 of 64000 = 3802 passengers

Revenue generated = Rs. 95,05,000

Binomial distribution:

If p represents probability of success, $(1-p)$ represents probability of failure, n represents number of independent trials, and k represents number of successes

$$P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$



Binomial Conditions

- For a variable to follow a binomial distribution the conditions to follow are –
 1. The trials must be independent-Head or tail
 2. The number of trials must be fixed.
 3. Every trial outcome must be classified as a success or failure.
 4. The probability of success, p , must be same for each trial- For first toss probability of head is 0.5 for second toss also it is 0.5 and so on.

Example

According to a 2013 Gallup poll, worldwide only 13% of employees are engaged at work (psychologically committed to their jobs and likely to be making positive contributions to their organizations). Among a random sample of 10 employees, what is the probability that 8 of them are engaged at work?

Solu: $p(k=8) = (n \text{ choose } k) p^k (1-p)^{(n-k)}$

$n = 10$; p (success = engaged) = 0.13; $1-p = 0.87$; $k = 8$

$10 \text{ choose } 8 = 10!/(8!(10-8)!)$

$= 45$

$P(k=8) = (10 \text{ choose } 8) 0.13^8 \times 0.87^2$
 $= 0.00000278$

Mean in Binomial Distribution

- In the example it is given that 13% employees are engaged.
- Among a random sample of 100 employees, on an average 13 peoples are engaged. i.e mean is

$$\mu = 100 \times 0.13 = 13$$

In mathematical terms,

Expected value of Binomial distribution = $\mu = np$

Standard Deviation in Binomial distribution

- We are claiming that mean is 13 . But it does not mean that every random sample of 100 employees exactly 13 will be engaged at work. Than how much do we expect the value to vary?
- This variability around the mean can be anticipated using standard deviation.

$$\sigma = (np(1-p))^{1/2}$$

$$\text{In the given case, } \sigma = (100 \times 0.13 \times 0.87)^{1/2} \\ = 3.36$$

This means that 13 out of 100 employees are expected to be engaged at work, give or take approximately 3.36 employees.

Example

- A 2012 Gallup poll survey suggests that 26.2% of Americans are obese. Which of the following is false?
 - a. Among a random sample of 1000 Americans, we can expect 262 to be obese
 - b. Random samples of 1000 Americans, where there are at most 230 are obese people, would be considered unusual
 - c. The standard deviation of number of obese Americans in random samples of 1000 is roughly 14
 - d. Random samples of 1000 Americans, where at least 300 are obese would not be considered unusual

Ans. 'd'

Solu: 26.2% obese . $p=0.262$ $n=1000$

$$\sigma = (np(1-p))^{1/2}$$

In the given case, $\sigma = (1000 \times 0.262 \times 0.738)^{1/2}$

$$= 13.9$$

$$1\sigma = 13.9$$

$$2\sigma = 27.8$$

- option a and c are true.
- Usual value are in between $\mu - 2\sigma$ to $\mu + 2\sigma$ i.e here ,between $262 - 27.8 = 234.2$ and $262 + 27.8 = 289.8$ i.e between $(234.2, 289.8)$.
- But here the value 230 is not in the usual range . Hence option b is also true.

Example-facebook Usage

- A recent study suggested that facebook users get more than they give i.e 40% make friend request but 63% received at least one request, users who pressed like button 14 times for their friends post received like 20 times, 12% tagged their friend photo but 35% were tagged themselves on a photo. These are called as power users.
- 25% of the FB users are considered as Power users. These are the ones who get more than they give.
- Average FB user has 245 friends. What is the probability that an average FB user with 245 friends has 70 or more friends who are power users.

Solu: $p=0.25, n=245, K=70$ or more

Solution Contd..

$$P(K \text{ greater than or equal to } 70) = P(k=70) + P(K=71) + P(k=72) + \dots + P(k=245)$$

- The calculations are lengthy and tedious. Hence in such cases we use the concept of Normal Approximation to Binomial Distribution.

Success-failure rule: A binomial distribution with at least 10 expected successes and 10 expected failures closely follows a normal distribution.

$$\begin{aligned} np &\geq 10 \\ n(1-p) &\geq 10 \end{aligned}$$

Normal approximation to the binomial: If the success-failure condition holds,

$$\text{Binomial}(n,p) \sim \text{Normal}(\mu, \sigma)$$

$$\text{where } \mu = np \text{ and } \sigma = \sqrt{np(1-p)}$$

- This is allowed only if

$$np \geq 10 \text{ and } n(1-p) \geq 10$$

then the binomial distribution problem can be solved as normal distribution.

In other words, a binomial distribution becomes normal distribution when we increase the value of n .

- Let us consider example where $p=0.25$ and following cases

i) $n=10$ for $K=0,1,2,3,4,5,6,7,8,9,10$ we get p with following formula

$$p(k \text{ trials}) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

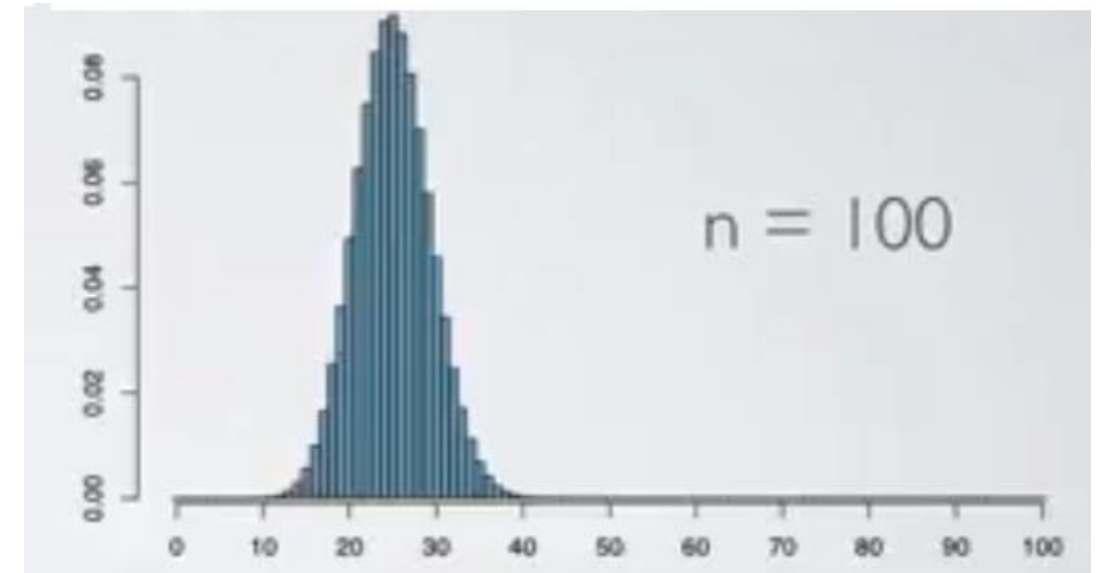
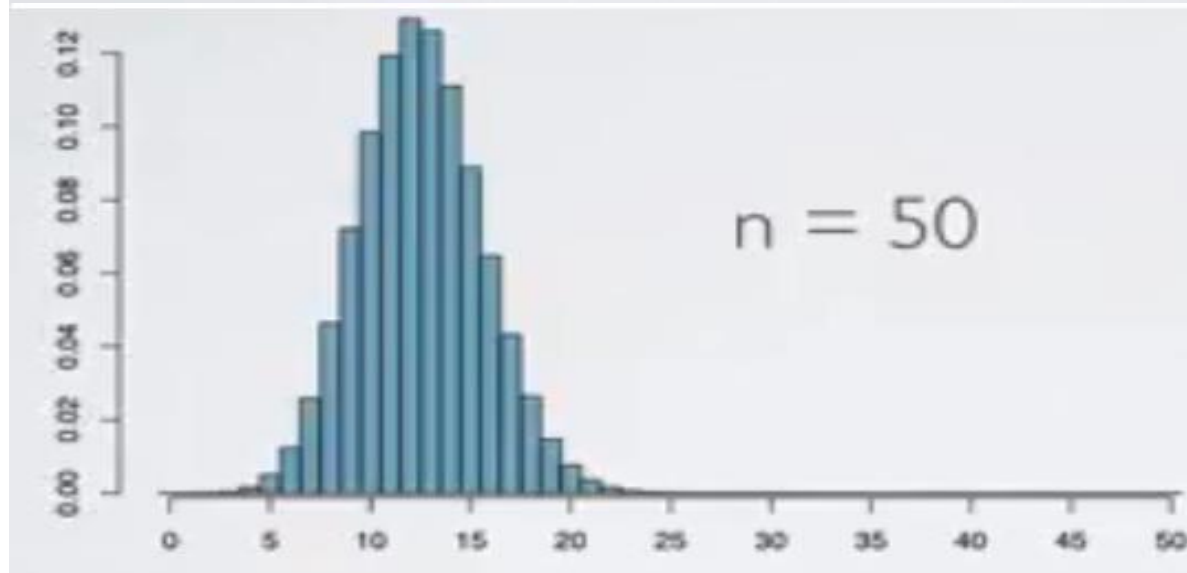
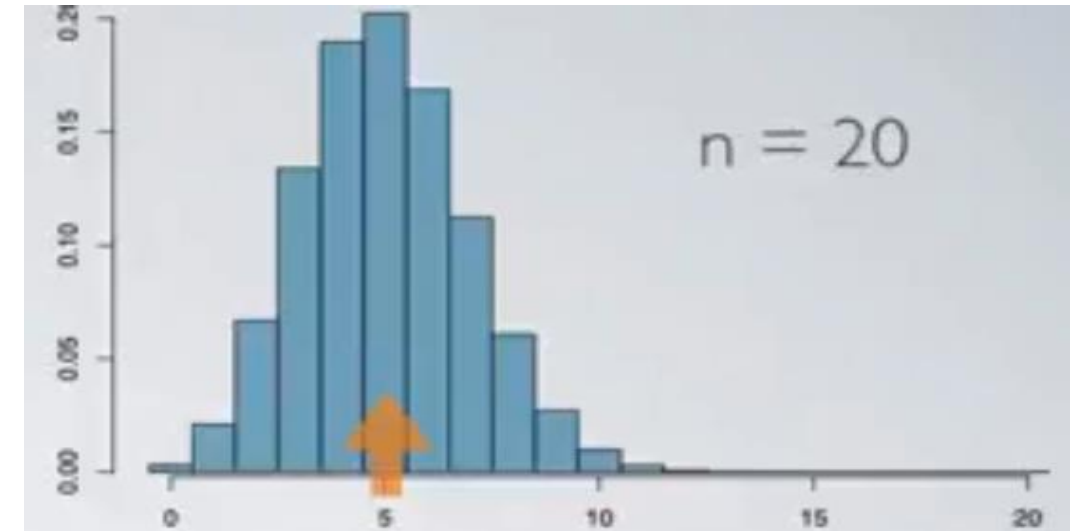
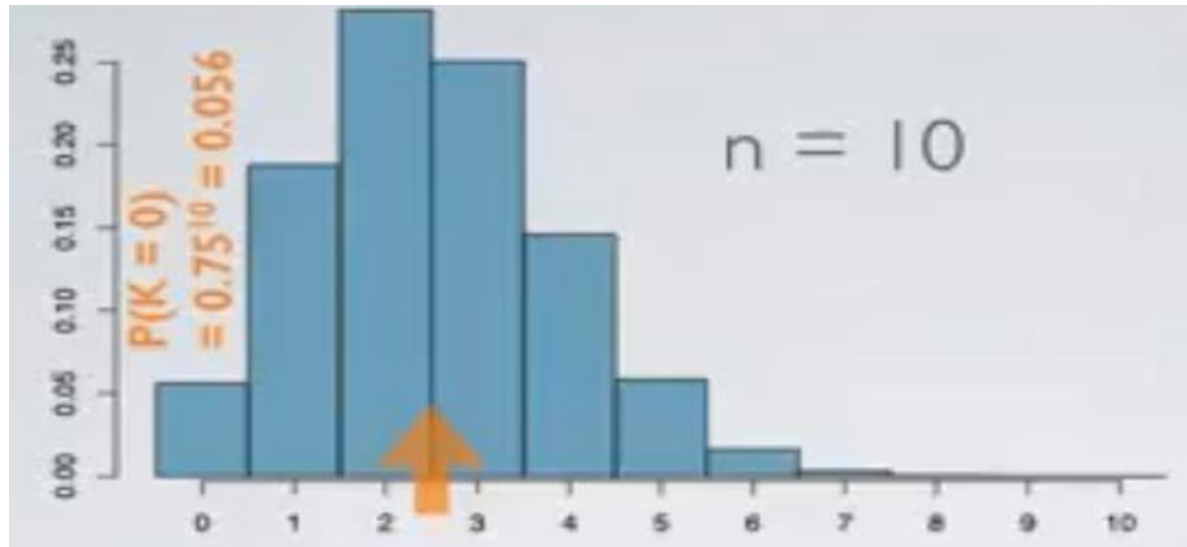
$$p(k=0) = 0.0563, p(k=1) = 0.1877, p(k=2) = 0.2815 \dots \text{And so on}$$

Histogram is plotted for this case of $n=10$

- ii) Repeating the process for $n=20$, $n=50$, $n=100$

We observe that for $n=10$ histogram is rightly skewed and as we increase the value of n we get nearly normal distribution.

Normal Approximation to Binomial Distribution



Normal Approximation to Binomial Distribution

- Keeping p constant, if we increase sample size, the $n \times p$ increases and the distribution resembles normal distribution.
- To apply normal distribution, information required can be estimated by the mean and standard deviation of the original binomial distribution.
- The rule of the thumb for the approximation is the 'Success-Failure Condition'

$$np \geq 10, n(1-p) \geq 10$$

Solu Contd...

- In the facebook example we use normal approximation

$$np = 245 \times 0.25 = 61.25$$

$$n(1-p) = 245 \times (1 - 0.25) = 183.75$$

Both are greater than 10.

- Hence using the resemblance between Normal and Binomial distribution, mean and SD can be found as

$$\mu = np = 61.25 \text{ and } \sigma = (np(1-p))^{1/2} = 6.78.$$

Solu Contd...

- With mean and SD we can find Z score and then probability.
- Observed value =70 Mean=61.25 SD=6.78
- $Z \text{ score} = (70 - 61.25) / 6.78$
 $= 1.29$
- From Table we get value=0.9015
- Therefore for greater values $1 - 0.9015 = 0.0985$ i.e 9.85%
- Thus the probability that an average FB user has 70 or more friends who are power users is 9.85%.

What is the minimum required n for a binomial distribution with $p = 0.25$ to closely follow a normal distribution?

$$n \times 0.25 \geq 10$$

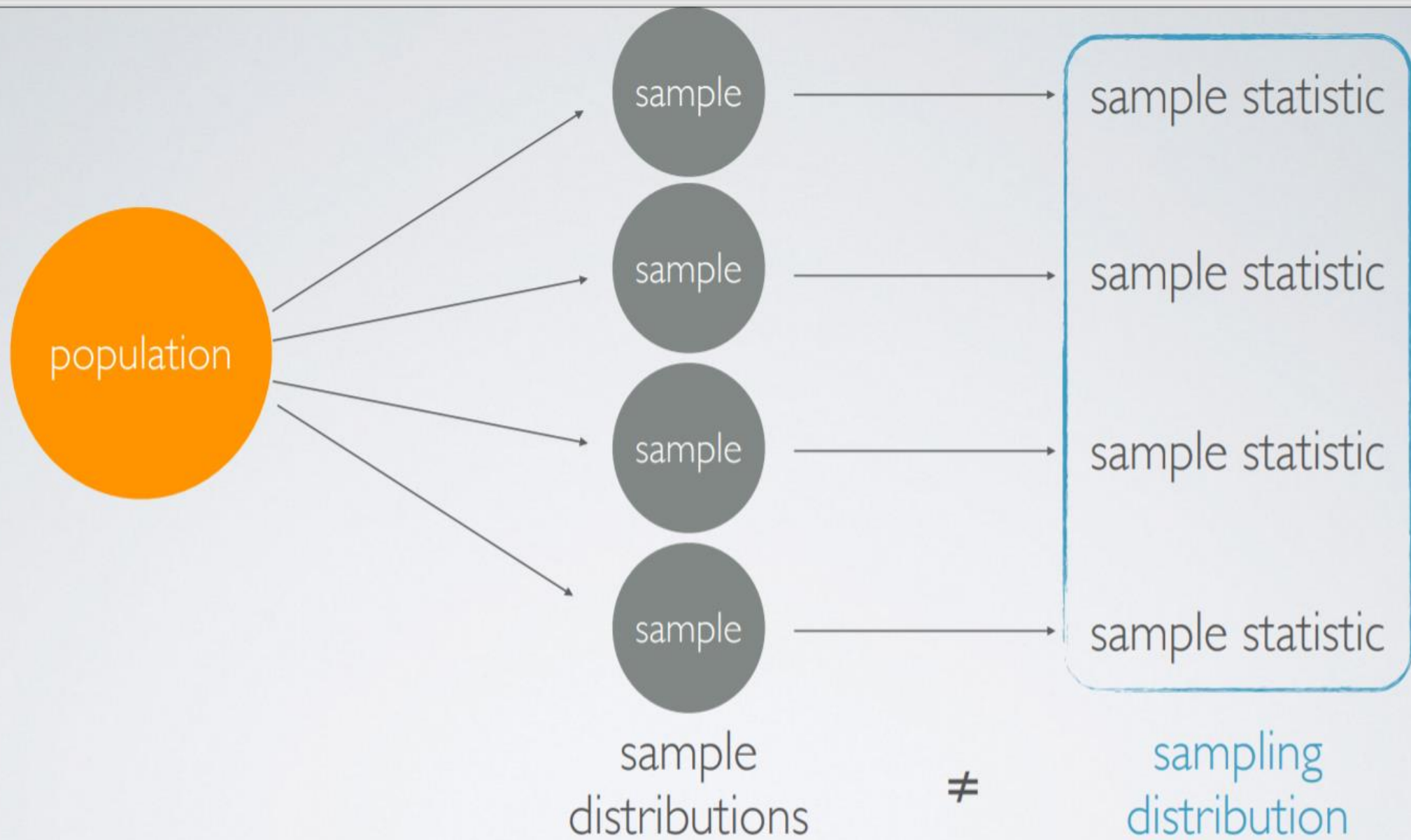
$$n \geq 10 / 0.25$$

$$n \geq 40$$

$$n \times 0.75 \geq 10$$

$$n \geq 10 / 0.75$$

$$n \geq 13.33$$



Central Limit Theorem (CLT): The distribution of sample statistics is nearly normal, centered at the population mean, and with a standard deviation equal to the population standard deviation divided by square root of the sample size.

$$\bar{x} \sim N \left(\text{mean} = \mu, SE = \frac{\sigma}{\sqrt{n}} \right)$$

Shape

center

spread