◉ Sheet ○ Slides

Number

8

Done by:

Ola Al-juneidi

corrected by:

sufian alhafez

Doctor

Mahmoud Al-Hussami

Record #9 Slide #4-slides 1-47

Probability

The concept of probability developed from the study of games of chance like cards, dice, flipping the coin, etc. A process like flipping a coin, rolling a dice or drawing a card from a deck is called a **probability experiment**. An **outcome** is a specific result of a single trial of a probability experiment.

The probability theory is the foundation for statistical inference that we will work on for testing the hypothesis in the concept of parametric techniques.

The probability distribution is a device for indicating the values that a random variable may have (the chance).

There are two categories of random variables. These are:

1) **Discrete random variables:**
   1. Binomial distribution:
      - The random variable can only assume 1 of 2 possible outcomes (0 or 1 /1 or 2...). There are a fixed number of trials and the results of the trials are independent.
      -Ex: flipping the coin (either head or tail) and counting the number of heads in 10 trials. The probability of getting a head is 0.5 (a fraction), and the sum of the probabilities must be equal to 1
   2. Poisson Distribution:
      -The random variable can assume a value between 0 and infinity.
      - Counts usually follow a Poisson distribution
      -Ex: number of ambulances needed in a city in a given night. It can be from nothing to a number that we don't know.

A discrete random variable X has a finite number of possible values. The probability distribution of X lists the values and their probabilities.

Value of X	X_1	X_2	X_3	….	X_k
Probability	p_1	P_2	p_3	….	p_k

- Every probability (p_i) is a number between 0 and 1.
- The sum of the probabilities must be 1.
- Find the probabilities of any event by adding the probabilities of the particular values that make up the event.

EXAMPLE:

The instructor in a large class gives 15% each of A's and D's, 30% each of B's and C's and 10% F's; the sum of the percentages of the marks (how many A's, how many B's and so on) will be 1 (100%).

The student's grade on a 4-point scale is a random variable X (A=4).

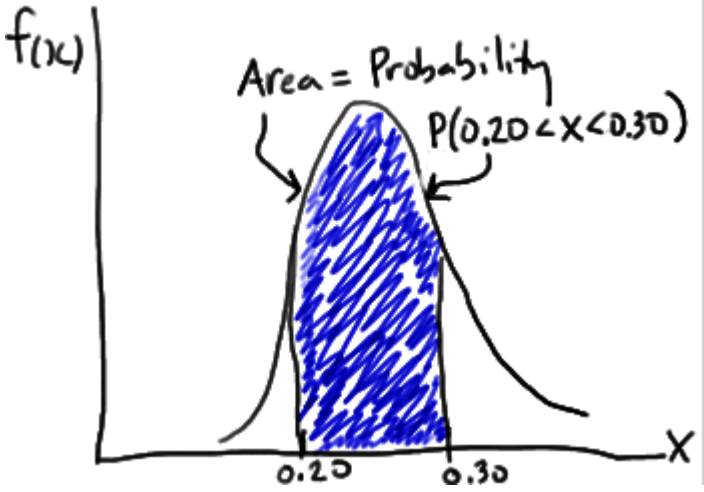
Grade	F=0	D=1	C=2	B=3	A=4
Probability	0.10	0.15	0.30	0.30	0.15

*Q: What is the probability (P) that a student selected at random will have <u>a B or better</u>?
-ANSWER: P (grade of 3 or 4) =P(X=3) + P(X=4)= 0.3 + 0.15 = 0.45

2) Continuous random variables:

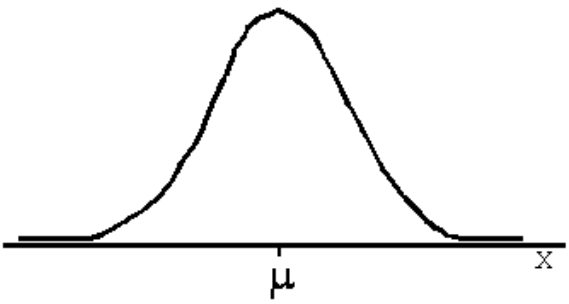
- When it follows a Binomial or a Poisson distribution the variable is restricted to taking integer values only *BUT* the continuous variable is the one with <u>fractions</u>.
- Between two values of a continuous random variable we can always find a third.
- A *histogram* is used to represent a *discrete* probability distribution and a smooth curve called the **probability density** is used to represent a <u>continuous probability distribution</u>.
- The **probability density** is a smooth curve where the frequency of occurrence of values between any two points equals the total area under the curve between the two points and the x-axis. The area under the smooth curve is equal to 1.

What does that mean?!
On the curve we have 2 points (A and B). When we calculate the area under the curve between these 2 points we get the probability of getting a value between these 2 numbers and the number we get will be a fraction. The total area under the curve is 1.



Normal Distribution

- We have what is called **normal distribution (frequency distribution)** before we standardize data, and <u>standardized</u> normal distributions (Z-distributions).
- The normal or frequency distribution is also called belt shaped curve, normal curve, or Gaussian distribution.
- A normal distribution is one that is unimodal, symmetric, and not too peaked or flat. The mean, median, and mode for a normal distribution are equal and their value divides the curve into two equal halves (mirror image). 68% of the population is one standard deviation above and one standard deviation below the mean.



- Given its name by the French mathematician Quetelet who, in the early 19th century, noted that many human attributes (e.g. height, weight, intelligence) appeared to be distributed normally.
- Inferential statistics work on the <u>reference population</u> and we want to generalize about the reference population; it is all about its mean (μ).

- Two parameters define a normal distribution; The standard deviation (σ) and the mean (μ):

  (1) The standard deviation (σ) specifies the amount of dispersion around the mean.

  (2) The normal curve is unimodal and symmetric about its mean (μ).

This creates a family of distributions depending on whatever the values of σ and μ are.

- Also called a Probability density function. The probability is interpreted as "area under the curve". The area under the whole curve (probability) = 1. Therefore 50% is to the right of μ and 50% is to the left of μ.
- The random variable takes on an infinite # of values within a given interval.
- The probability that X = any particular value is 0. Consequently, we talk about intervals. <u>The probability is = to the area under the curve</u>.

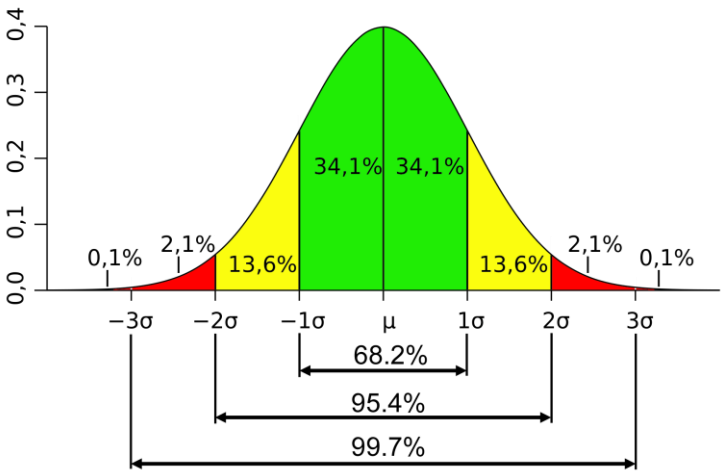
3:00-13:00

- 1 standard deviation (σ) above and below the mean is ~68% of the area under the curve. 2 standard deviations (σ) above and below the mean is ~95%
- In other words, perpendiculars of:

  (a) ±σ contain about 68%
  (b) ±2 σ contain about 95%
  (c) ±3 σ contain about 99.7%

  of the area under the curve.

  -Outside 3 standard deviations of the sample (not the general population) are outliers.



We said that a normal distribution is determined by μ and σ. This creates a family of distributions depending on whatever the values of μ and σ are.

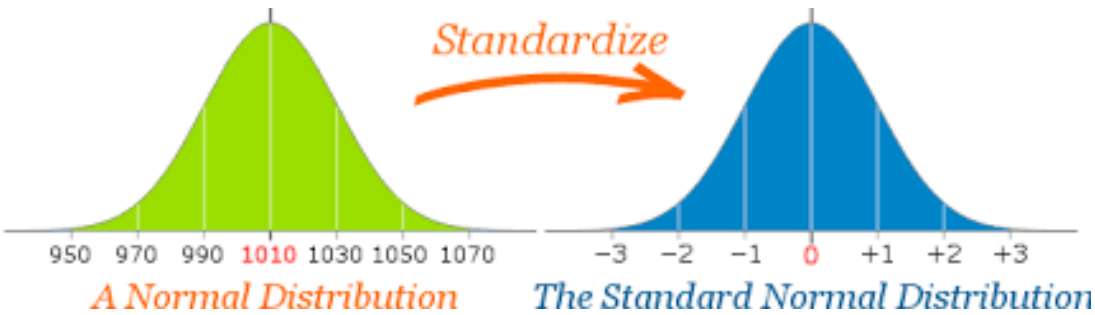
To make it easier, we can convert it to a **standardized normal distribution**.

Standard Z score

To convert the distribution to a standardized normal distribution we calculate the z for each value where:

$$z = \frac{(x - \mu)}{\sigma}$$

Given the values of μ and σ we can convert a value of x to a value of z and find its probability using the table of normal curve areas. Then μ will be equal to 0 and σ=1.



Importance of Normal Distribution to Statistics

- Although most distributions are not exactly normal, most variables tend to have approximately normal distribution.
- Many inferential statistics assume that the populations are distributed normally.
- The normal curve is a probability distribution and is used to answer questions about the likelihood of getting various particular outcomes when sampling from a population.

Why Do We Like The Normal Distribution So Much?

There is nothing "special" about standard normal scores

- These can be computed for observations from any sample/population of continuous data values
- The score measures how far an observation is from its mean in standard units of statistical distance

But, if distribution is not normal, we may not be able to use Z-score approach.

*Q: Is every variable normally distributed?
A: Absolutely not
*Q: Then why do we spend so much time studying the normal distribution?
A: Some variables are normally distributed; a bigger reason is the "**Central Limit Theorem**"?!

We said that we want to test the hypothesis. We work on inferential statistics, and the first assumption in the parametric techniques is that the dependent variable is normally distributed. However, the trait in the reference population may or may not be normally distributed. So, there is another concept that we work on in the inferential statistics which is the **Central Limit Theorem**.

Central Limit Theorem

EXAMPLE: height of males

-We have a population, and in this population the dependent variable may be normally distributed or not. If we take a sample (sample1) and calculated its mean, can its mean be equal to the mean of the population? It may or may not be equal to it. Assume that μ (the mean of the population) is unknown. In sample1 the mean was 160cm. Then we take another sample and its mean is 165cm. If we continued and took an infinite number of samples and plot their means, the means will be normally distributed (according to this theory) even if the mean of the means is NOT equal to the mean of the population. Actually, the mean of the means is equal to the population mean. This theory is used as a prerequisite assumption for the dependent variable.

The reference population may or may not be normally distributed. But if you take different samples and plot their means on a curve, they will be normally distributed as this theory states.

To get rid of the variety between the samples we calculate the standard error of the mean= $\sigma/\sqrt{n}$, where n is the sample size.

RECAP

The Central Limit Theorem describes the characteristics of the "population of the means" which has been created from the means of an infinite number of random population samples of size (N), all of them drawn from a given "parent population."

It predicts that regardless of the distribution of the parent population:

- The mean of the population of means is always equal to the mean of the parent population from which the population samples were drawn.
- The standard deviation of the population of means is always equal to the standard deviation of the parent population divided by the square root of the sample size (N)
- The distribution of means will increasingly approximate a normal distribution as the size N of samples increases.

13:00-23:00

A consequence of Central Limit Theorem is that if we average measurements of a particular quantity, the distribution of our average tends toward a normal one.

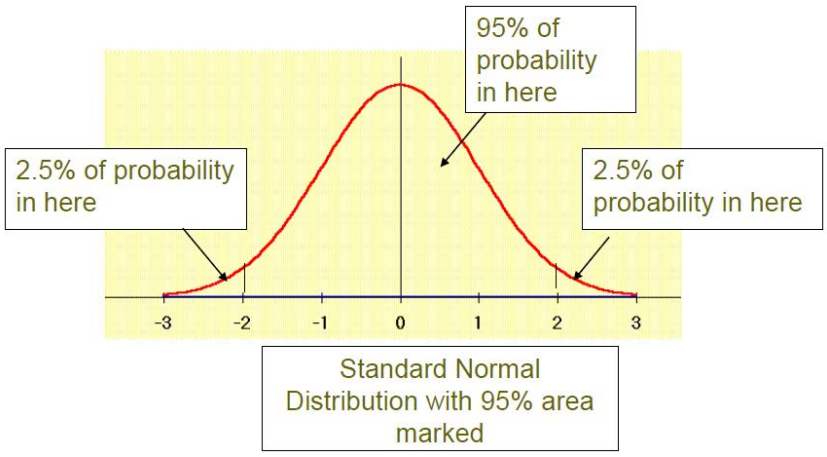
In addition, if a measured variable is actually a combination of several other uncorrelated variables, all of them "contaminated" with a random error of any distribution, our measurements tend to be contaminated with a random error that is normally distributed as the number of these variables increases.

Thus, the Central Limit Theorem explains the ubiquity of the famous bell-shaped "Normal distribution" (or "Gaussian distribution") in the measurements domain.

We said that 1 standard deviation (or z score) above the mean, the area (probability) will be 34% of the total area. Now what if we want *half* of it? Here we have to use the z table which is found calculated in books for the positive and negative (right and left of the curve). (You can find the table in the slide)

Why do we care about the z table? To calculate the <u>probabilities</u> (p). The p value has to do with the significance and to accept or reject the null hypothesis which we'll learn later on.

Above 2 z scores we have 2.5% from each side with a total of 5% (look at the figure)

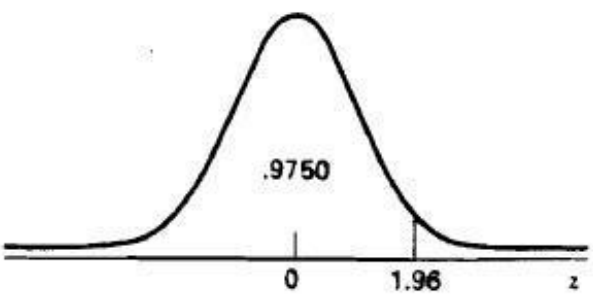


Calculating Probabilities

Probability calculations are always concerned with finding the probability that the variable assumes any value in an interval between two specific points a and b. The probability that a continuous variable assumes the a value between a and b is the area under the graph of the density between a and b.

EXAMPLES: use the table in slide 25 to answer these questions. When z is positive use the positive table, and when it's negative use the negative one.

1. What is the probability that z < 1.96?
    a. First sketch a normal curve
    b. Draw a line for z = 1.96
    c. Find the area in the table
    d. The answer is the area to the left of the line P(z < 1.96) =0.975



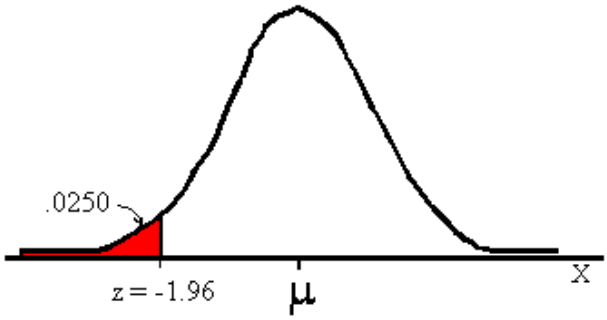
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z
0.00	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359	0.00
0.10	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753	0.10
0.20	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141	0.20
0.30	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517	0.30
0.40	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879	0.40
0.50	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224	0.50
0.60	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549	0.60
0.70	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852	0.70
0.80	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133	0.80
0.90	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389	0.90
1.00	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621	1.00
1.10	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830	1.10
1.20	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015	1.20
1.30	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177	1.30
1.40	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319	1.40
1.50	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441	1.50
1.60	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545	1.60
1.70	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633	1.70
1.80	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706	1.80
1.90	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767	1.90
2.00	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817	2.00
2.10	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857	2.10
2.20	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890	2.20
2.30	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916	2.30
2.40	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936	2.40
2.50	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952	2.50
2.60	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964	2.60
2.70	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974	2.70
2.80	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981	2.80
2.90	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986	2.90
3.00	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990	3.00
3.10	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993	3.10
3.20	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995	3.20
3.30	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997	3.30
3.40	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998	3.40
3.50	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	3.50
3.60	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	3.60
3.70	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	3.70
3.80	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	3.80

2. What is the probability that z < -1.96?

   Draw the sketch then find the answer from the table

   The answer is the area to the left of the line P(z < -1.96) = .0250

z	−0.09	−0.08	−0.07	−0.06	−0.05	−0.04	−0.03	−0.02	−0.01	0.00	z
−3.80	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	−3.80
−3.70	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	−3.70
−3.60	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002	−3.60
−3.50	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	−3.50
−3.40	.0002	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	−3.40
−3.30	.0003	.0004	.0004	.0004	.0004	.0004	.0004	.0005	.0005	.0005	−3.30
−3.20	.0005	.0005	.0005	.0006	.0006	.0006	.0006	.0006	.0007	.0007	−3.20
−3.10	.0007	.0007	.0008	.0008	.0008	.0008	.0009	.0009	.0009	.0010	−3.10
−3.00	.0010	.0010	.0011	.0011	.0011	.0012	.0012	.0013	.0013	.0013	−3.00
−2.90	.0014	.0014	.0015	.0015	.0016	.0016	.0017	.0018	.0018	.0019	−2.90
−2.80	.0019	.0020	.0021	.0021	.0022	.0023	.0023	.0024	.0025	.0026	−2.80
−2.70	.0026	.0027	.0028	.0029	.0030	.0031	.0032	.0033	.0034	.0035	−2.70
−2.60	.0036	.0037	.0038	.0039	.0040	.0041	.0043	.0044	.0045	.0047	−2.60
−2.50	.0048	.0049	.0051	.0052	.0054	.0055	.0057	.0059	.0060	.0062	−2.50
−2.40	.0064	.0066	.0068	.0069	.0071	.0073	.0075	.0078	.0080	.0082	−2.40
−2.30	.0084	.0087	.0089	.0091	.0094	.0096	.0099	.0102	.0104	.0107	−2.30
−2.20	.0110	.0113	.0116	.0119	.0122	.0125	.0129	.0132	.0136	.0139	−2.20
−2.10	.0143	.0146	.0150	.0154	.0158	.0162	.0166	.0170	.0174	.0179	−2.10
−2.00	.0183	.0188	.0192	.0197	.0202	.0207	.0212	.0217	.0222	.0228	−2.00
−1.90	.0233	.0239	.0244	.0250	.0256	.0262	.0268	.0274	.0281	.0287	−1.90
−1.80	.0294	.0301	.0307	.0314	.0322	.0329	.0336	.0344	.0351	.0359	−1.80
−1.70	.0367	.0375	.0384	.0392	.0401	.0409	.0418	.0427	.0436	.0446	−1.70
−1.60	.0455	.0465	.0475	.0485	.0495	.0505	.0516	.0526	.0537	.0548	−1.60

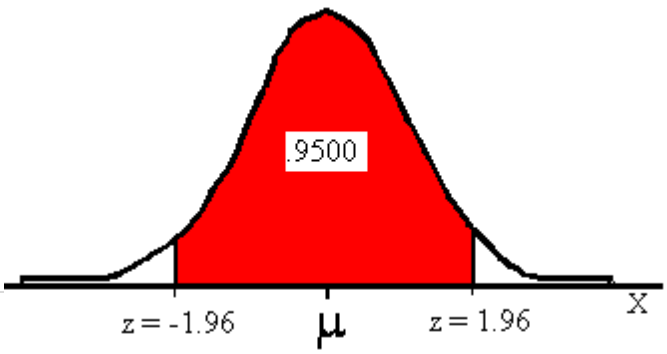


3. What is the probability that -1.96 < z < 1.96?

   (1) Sketch a normal curve

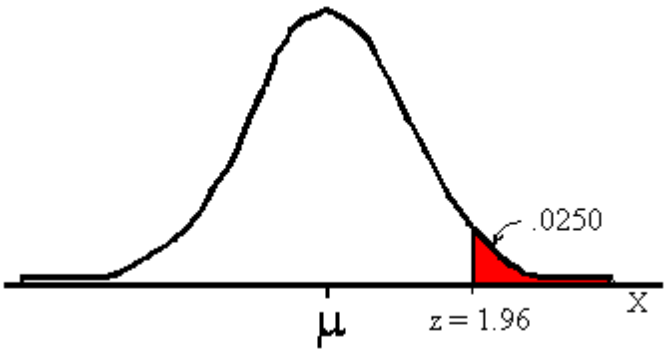
   (2) Draw lines for lower z = -1.96, and upper z = 1.96

   (3) Find the area in the table corresponding to each value



(4) The answer is the area between the values.
Subtract lower from upper:
P(-1.96 < z < 1.96) = .9750 - .0250 = .9500

4. What is the probability that z > 1.96?
   (1) Sketch a normal curve
   (2) Draw a line for z = 1.96
   (3) Find the area in the table
   (4) The answer is the area to the right of the line. It is found by subtracting the table value from 1.0000: P(z > 1.96) =1.0000 - .9750 = .0250



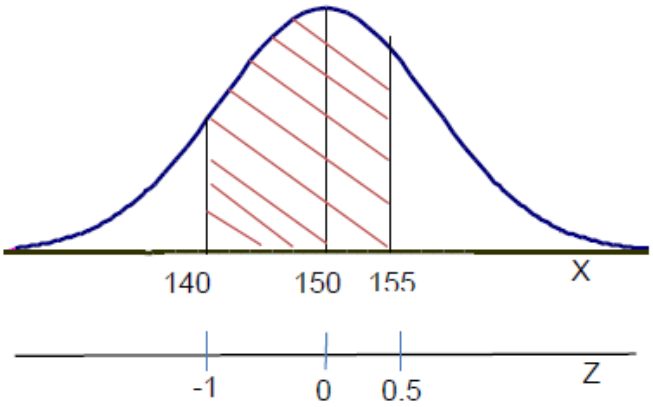
EXAMPLE: weight

If the weight of males is N.D. (normally distributed) with μ=150 and σ=10, what is the probability that a randomly selected male will weigh between 140 lbs and 155 lbs? [Important Note: Always remember that the probability that X is equal to any one particular value is zero, P(X=value) =0, since the normal distribution is continuous.]
** We calculate the area according to the z score, so we first have to calculate z for these values.

<u>Solution:</u>

a. Z = (**140** – 150)/ 10 = -1.00 s.d. (standard deviation) from mean
   Area under the curve = .3413 (from Z table)
b. Z = (**155** – 150) / 10 =+.50 s.d. from mean
   Area under the curve = .1915 (from Z table)

Answer: .3413 + .1915 = .5328



23:00-34:30

EXAMPLE: salary

Suppose that the average salary of college graduates is N.D. with μ=$40,000 and σ=$10,000. (we are not concerned in caluclating d/e, slides 48-50 are not included)

a) What proportion of college graduates will earn $24,800 or less?

Always DRAW a picture of the distribution to help you solve these problems. First draw the curve and assign μ on it. Then you have to calculate z for the values.

After that you find the proportion from the given table.

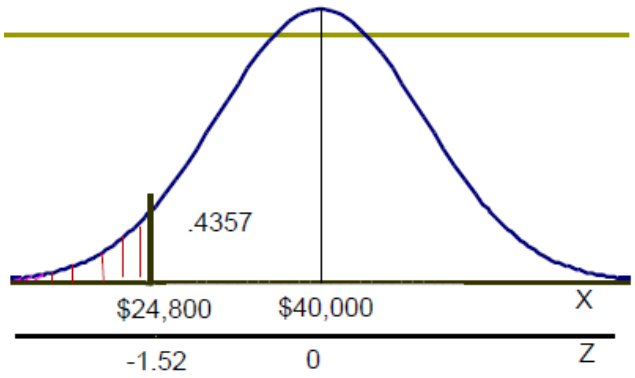
<u>Solution:</u>

-Convert the $24,800 to a Z-score:
Z = ($24,800 - $40,000)/$10,000 = -1.52.

-First Find the area between 0 and -1.52 in the Z-table. From the Z table, that area is .4357.

-Then, the area from -1.52 to - ∞ is:
0.5000- 0.4358 = 0.0643

Answer: 6.43% of college graduates will earn less than $24,800.

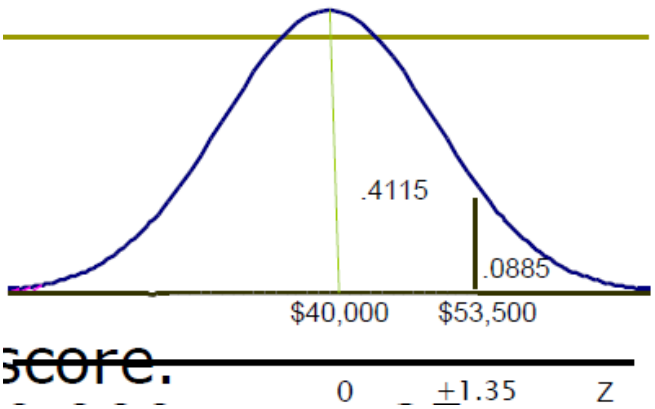


b) What proportion of college graduates will earn $53,500 or more?

<u>Solution:</u>

- Convert the $53,500 to a Z-score. Z = ($53,500 - $40,000)/$10,000 = +1.35.
- Find the area between 0 and +1.35 in the Z-table: .4115 is the table value.
- When you DRAW A PICTURE you see that you need the area in the tail:
  .5 - .4115 = .0885.

Answer: .0885. Thus, 8.85% of college graduates will earn $53,500 or more.



c) What proportion of college graduates will earn between $45,000 and $57,000?

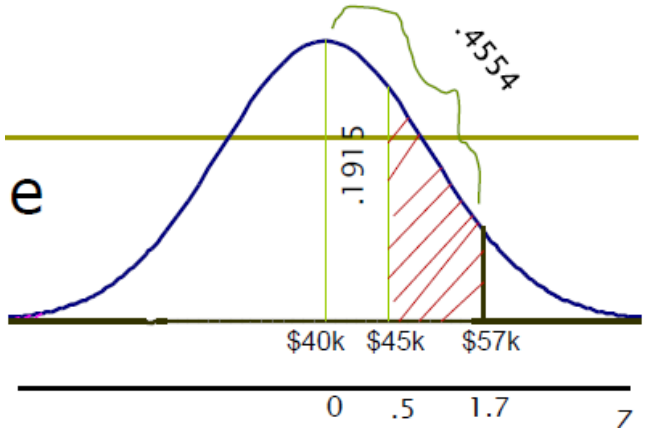
- Z = $45,000 – $40,000 / $10,000 = .50
  Z = $57,000 – $40,000 / $10,000 = 1.70

- From the table, we can get the area under the curve between the mean (0) and .5; we can get the area between 0 and 1.7. From the picture we see that neither one is what we need.

What do we do here? Subtract the small piece from the big piece to get exactly what we need.

Answer: 0.4554 – 0.1915 =0 .2639

So 26.36% of college graduates will earn between $45,000 and $57,000



EAMPLE: What's the probability of getting a math SAT score of 575 or less, μ=500 and σ=50?

$$Z = \frac{575-500}{50} = 1.5$$

i.e., A score of 575 is 1.5 standard deviations above the mean

To look up Z= 1.5 in standard normal chart (or enter into SAS), no problem! = .9332

EXAMPLE: IQ

If IQ is ND with a mean of 100 and a S.D. of 10, what percentage of the population will have:

(a) IQs ranging from 90 to 110?

<u>Solution:</u>

Z = (90 – 100)/10 = -1.00

Z = (110 -100)/ 10 = +1.00

Area between 0 and 1.00 in the Z-table is .3413;

Area between 0 and -1.00 is also .3413 (Z-distribution is symmetric)

Answer is .3413 + .3413 = .6826.

(b) IQs ranging from 80 to 120?

Solution:

Z = (80 – 100)/10 = -2.00

Z = (120 -100)/ 10 = +2.00

Area between =0 and 2.00 in the Z-table is .4772;

Area between 0 and -2.00 is also .4772 (Z-distribution is symmetric).

Answer is .4772 + .4772 = .9544.

34:30-44:30

"The one who falls and gets up is much stronger than the one who never fell"

THE END