

Lecture 3

The Normal Distribution

Section A: The Standard Normal Distribution Defined

2

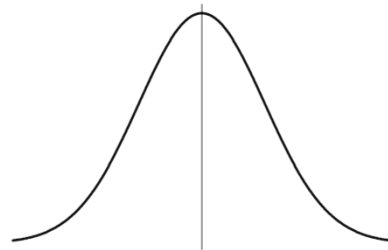
Learning Objectives

- Upon completion of this lecture you will be able to:
 - Describe the basic properties of the normal curve
 - Describe how any normal distribution is completely defined by its mean and standard deviation
 - Recite the 68-96-99.7% rule for the normal distribution with regards to standard deviations
 - Feel comfortable (or be on your way to feeling comfortable) working with standard normal tables

3

The Normal Distribution

- The Normal Distribution is a theoretical probability distribution that is perfectly symmetric about its mean (and median and mode), and has a “bell” like shape



4

The Normal Distribution

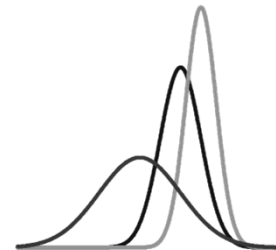
- The Normal Distribution is also called the “Gaussian Distribution” in honor of its inventor Carl Friedrich Gauss



5

The Normal Distribution

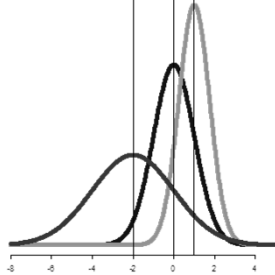
- Normal Distributions are uniquely defined by two quantities: a mean (μ), and standard deviation (σ)
- There are literally an infinite number of possible normal curves, for every possible combination of (μ) and (σ)



6

The Normal Distribution

- Normal Distributions are uniquely defined by two quantities: a mean (μ), and standard deviation (σ)
- There are literally an infinite number of possible normal curves, for every possible combination of (μ) and (σ)



7

The Normal Distribution

- This function defines the normal curve for any given (μ) and (σ)

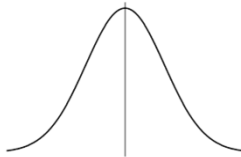
$$\text{"Proportion of values = } x\text{"} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

8

The Normal Distribution

- All Normal distributions, regardless of mean and standard deviation values, have the same structural properties:

- Mean = median = mode
- Values are symmetrically distributed around the mean
- Values "closer" to the mean are more frequent than values "further" from the mean



9

The Normal Distribution

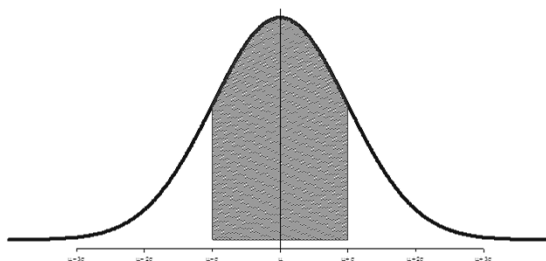
- All Normal distributions, regardless of mean and standard deviation values, have the same structural properties:

- The entire distribution of values described by a normal distribution can be completely specified by knowing just the mean and standard deviation
- Since all normal distributions have the same structural properties, we can use a reference distribution, called the standard normal distribution, to elaborate on some of these properties
- In Section B, we'll show that any normal distribution can be easily rescaled to this standard normal distribution

10

The 68-95-99.7 Rule for the Normal Distribution

- 68% of the observations fall within one standard deviation of the mean



11

The 68-95-99.7 Rule for the Normal Distribution

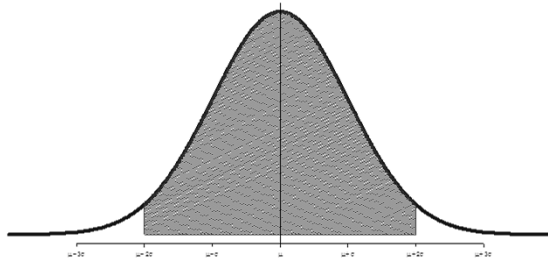
- There are several ways to state this. For data whose distribution is approximately normal:

- 68% of the observations fall within one standard deviation of the mean
- The probability that any randomly selected value is within one standard deviation of the mean is 0.68 or 68%

12

The 68-95-99.7 Rule for the Normal Distribution

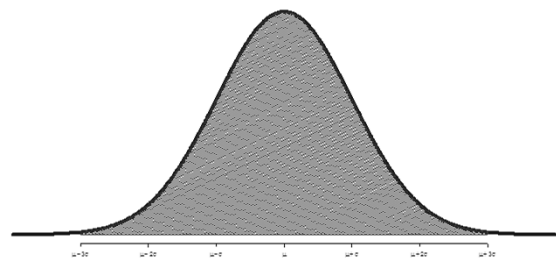
- 95% of the observations fall within two standard deviations of the mean (truthfully, within 1.96)



13

The 68-95-99.7 Rule for the Normal Distribution

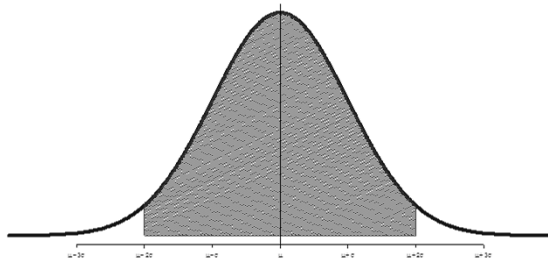
- 99.7% of the observations fall within three standard deviations of the mean



14

The 68-95-99.7 Rule for the Normal Distribution

- 95% of the observations fall within two standard deviations of the mean (truthfully, within 1.96) : Let's consider this for a moment



15

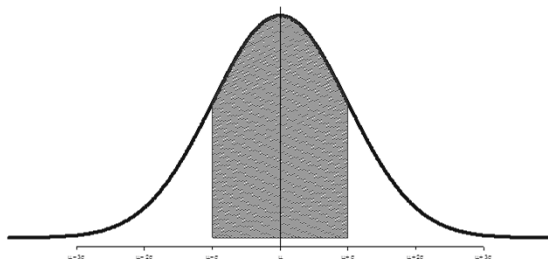
The Normal Distribution

- The middle 95% of values fall between
- 2.5% of the values are smaller than (and hence 97.5% are greater than)
- 97.5% of the values are smaller than (and hence 2.5% are greater than)

16

The 68-95-99.7 Rule for the Normal Distribution

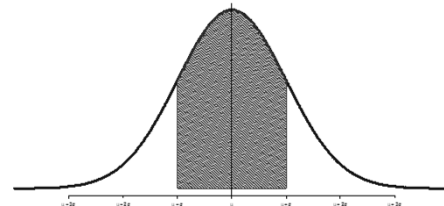
- Again, we know that 68% of the observations in a normal distribution are in the interval $(\mu - \sigma, \mu + \sigma)$



17

The 68-95-99.7 Rule for the Normal Distribution

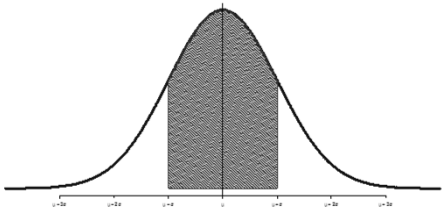
- What percentage of observations following a normal distribution are more than 1 standard deviation above the mean?
(can also be phrased: "What is the probability that an individual observation is more than one standard deviation above the mean in a normal distribution?")



18

The 68-95-99.7 Rule for the Normal Distribution



- What percentage of observations following a normal distribution are more than 1 standard deviation away from the mean in either direction (i.e. more than 1 standard deviation above the mean or less than 1 standard deviation below the mean?)





The 68-95-99.7 Rule for the Normal Distribution

- Where did this rule come from: in other words, how do I know these relationships?
- What about the percentages under the curve for other standard deviation distances from the mean?
- All of the information I quoted, and much more can be found in a "standard normal table"

Fraction of Observations Under Standard Normal

Within Z SDs of the mean		More than Z SDs above the mean	More than Z SDs below the mean
Z			
1.0	68.27%	15.87%	31.73%
2.0	95.45%	2.28%	4.55%
2.5	98.76%	0.62%	1.24%
3.0	99.73%	0.13%	0.27%

Fraction of Observations Under Standard Normal

Within Z SDs of the mean		More than Z SDs above the mean	More than Z SDs below the mean
Z			
1.0	68.27%	15.87%	31.73%
2.0	95.45%	2.28%	4.55%
2.5	98.76%	0.62%	1.24%
3.0	99.73%	0.13%	0.27%

Standard Normal Tables

- Exhibit A¹

STANDARD NORMAL TABLE (Z)

Entries in the table give the area under the curve between the mean and a standard deviation above the mean. For example, for $z = 1.25$ the area under the curve between the mean (0) and $z = 1.25$ is 0.2443.

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	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7122	0.7156	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7421	0.7453	0.7484	0.7515	0.7546
0.7	0.7577	0.7607	0.7637	0.7667	0.7696	0.7725	0.7754	0.7782	0.7811	0.7839
0.8	0.7868	0.7896	0.7924	0.7952	0.7979	0.8006	0.8033	0.8060	0.8086	0.8113
0.9	0.8139	0.8166	0.8192	0.8218	0.8244	0.8269	0.8294	0.8319	0.8344	0.8369
1.0	0.8394	0.8419	0.8444	0.8469	0.8493	0.8518	0.8542	0.8566	0.8590	0.8613
1.1	0.8638	0.8661	0.8684	0.8708	0.8729	0.8750	0.8770	0.8790	0.8810	0.8829
1.2	0.8849	0.8868	0.8887	0.8906	0.8925	0.8943	0.8961	0.8979	0.8996	0.9013
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	0.9237	0.9251	0.9266	0.9281	0.9295	0.9309	0.9323
1.5	0.9338	0.9352	0.9366	0.9379	0.9393	0.9406	0.9419	0.9432	0.9444	0.9457
1.6	0.9469	0.9481	0.9493	0.9505	0.9516	0.9527	0.9538	0.9549	0.9560	0.9570
1.7	0.9581	0.9591	0.9601	0.9611	0.9621	0.9631	0.9641	0.9651	0.9661	0.9671
1.8	0.9681	0.9691	0.9700	0.9709	0.9718	0.9727	0.9736	0.9745	0.9754	0.9763
1.9	0.9772	0.9781	0.9790	0.9798	0.9806	0.9814	0.9822	0.9830	0.9838	0.9846
2.0	0.9854	0.9861	0.9868	0.9876	0.9883	0.9890	0.9897	0.9903	0.9909	0.9915
2.1	0.9920	0.9926	0.9931	0.9936	0.9941	0.9945	0.9949	0.9953	0.9957	0.9961
2.2	0.9965	0.9969	0.9973	0.9977	0.9980	0.9984	0.9987	0.9990	0.9993	0.9996
2.3	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
2.4	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
2.5	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
2.6	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
2.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
2.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
2.9	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.0	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

¹http://classes.engr.oregonstate.edu/cce/winter2012/ce492/Modules/08_specifications_qa/normal_distribution.htm

Standard Normal Tables

- Exhibit A

STANDARD NORMAL TABLE (Z)

Entries in the table give the area under the curve between the mean and a standard deviation above the mean. For example, for $z = 1.25$ the area under the curve between the mean (0) and $z = 1.25$ is 0.2443.

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	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7122	0.7156	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7421	0.7453	0.7484	0.7515	0.7546
0.7	0.7577	0.7607	0.7637	0.7667	0.7696	0.7725	0.7754	0.7782	0.7811	0.7839
0.8	0.7868	0.7896	0.7924	0.7952	0.7979	0.8006	0.8033	0.8060	0.8086	0.8113
0.9	0.8139	0.8166	0.8192	0.8218	0.8244	0.8269	0.8294	0.8319	0.8344	0.8369
1.0	0.8394	0.8419	0.8444	0.8469	0.8493	0.8518	0.8542	0.8566	0.8590	0.8613
1.1	0.8638	0.8661	0.8684	0.8708	0.8729	0.8750	0.8770	0.8790	0.8810	0.8829
1.2	0.8849	0.8868	0.8887	0.8906	0.8925	0.8943	0.8961	0.8979	0.8996	0.9013
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	0.9237	0.9251	0.9266	0.9281	0.9295	0.9309	0.9323
1.5	0.9338	0.9352	0.9366	0.9379	0.9393	0.9406	0.9419	0.9432	0.9444	0.9457
1.6	0.9469	0.9481	0.9493	0.9505	0.9516	0.9527	0.9538	0.9549	0.9560	0.9570
1.7	0.9581	0.9591	0.9601	0.9611	0.9621	0.9631	0.9641	0.9651	0.9661	0.9671
1.8	0.9681	0.9691	0.9700	0.9709	0.9718	0.9727	0.9736	0.9745	0.9754	0.9763
1.9	0.9772	0.9781	0.9790	0.9798	0.9806	0.9814	0.9822	0.9830	0.9838	0.9846
2.0	0.9854	0.9861	0.9868	0.9876	0.9883	0.9890	0.9897	0.9903	0.9909	0.9915
2.1	0.9920	0.9926	0.9931	0.9936	0.9941	0.9945	0.9949	0.9953	0.9957	0.9961
2.2	0.9965	0.9969	0.9973	0.9977	0.9980	0.9984	0.9987	0.9990	0.9993	0.9996
2.3	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
2.4	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
2.5	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
2.6	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
2.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
2.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
2.9	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.0	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

Lecture 3: Statistical Reasoning for Public Health: Estimation, Inference, & Interpretation

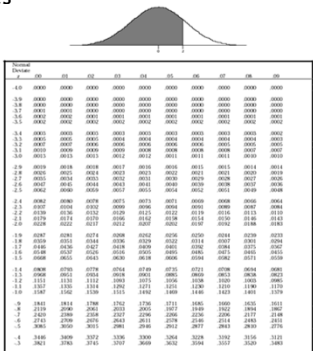
Standard Normal Tables

Exhibit A

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0200	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2421	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2968	0.2996	0.3025	0.3053	0.3081	0.3109	0.3136
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4494	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817

Standard Normal Tables

Exhibit B²



²http://classes.engr.oregonstate.edu/cce/winter2012/ce492/Modules/08_specifications_qa/normal_distribution.htm

Standard Normal Tables

Exhibit B

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-4.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-2.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-1.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
0.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
2.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
3.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
4.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

Standard Normal Tables

Exhibit B

Normal Deviate	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379

Summary

Section B: Applying the Principles of the Normal Distribution to Sample Data to Estimate Characteristics of Population Data

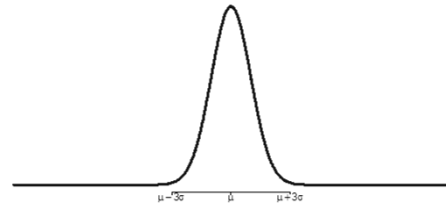
Learning Objectives

- Upon completion of this lecture you will be able to:
 - Create ranges containing a certain percentage of observations in an (approximately normal) distribution using only an estimate of the mean and standard deviation
 - Figure out how far any individual data point is from the mean of its distribution in standardized units (compute a z-score)
 - Convert z-scores to statements about relative proportions/probabilities for values that have an (approximately) normal distribution

31

The Normal Distribution

- The normal distribution is a theoretical probability distribution: no real data is perfectly described by this distribution
- For example, in a true normal distribution, the tails go onto to *negative and positive infinity*, respectively



32

The Normal Distribution

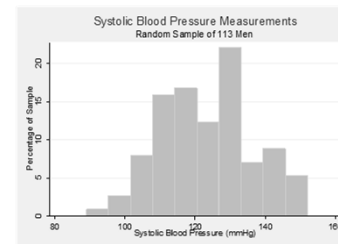
- However, the distributions of some data will be well approximated by a normal distribution: in such situations we can use the properties of the normal curve to characterize aspects of the data distribution

33

Example 1

- Example 1: Blood pressure measurements from a random sample of 113 adult men taken from a clinical population

Estimate of μ : $\bar{x} = 123.6$ mmHg ; Estimate of σ : $s = 12.9$ mmHg

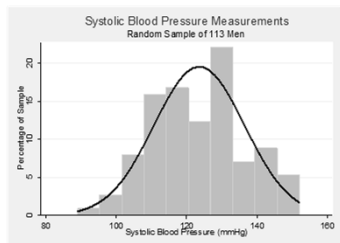


34

Example 1: Blood Pressure Data

- Example 1: Systolic blood pressure (SBP) measurements from a random sample of 113 adult men taken from a clinical population

Estimate of μ : $\bar{x} = 123.6$ mmHg ; Estimate of σ : $s = 12.9$ mmHg

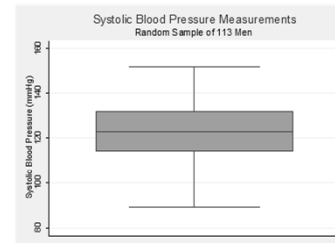


35

Example 1: Blood Pressure Data

- Example 1: Systolic blood pressure (SBP) measurements from a random sample of 113 adult men taken from a clinical population

Estimate of μ : $\bar{x} = 123.6$ mmHg ; Estimate of σ : $s = 12.9$ mmHg



36

Example 1

- Using only the sample mean and standard deviation, and assuming normality, let's estimate the 2.5th and 97.5th percentiles SBP in this population

2.5th %ile: $\bar{x} - 2s = 123.6 - (2 \times 12.9) = 97.8 \text{ mmHg}$

97.5th %ile: $\bar{x} + 2s = 123.6 + (2 \times 12.9) = 149.4 \text{ mmHg}$

Based on this sample data, we estimate that most (95%) of the men in this clinical population have systolic blood pressures between 97.8 and 149.4 mmHg.

(Note: the observed 2.5th and 97.5th percentiles of the 113 sample value are 100.7 mmHg and 151.2 mmHg respectively)

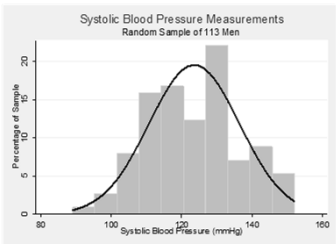
Example 1

- Suppose you want to use the results from this sample of 113 men from a clinic to evaluate individual male patients relative to the population of all such patients.

For example, suppose a patient in your clinic has a SBP measurement of 130 mmHg. What proportion of men at the clinic have SBP measurements greater than this patient?

Example 1

- We can use the sample mean and SD, and the assumption of normality to estimate this percentage:



Example 1

- If we translate this measurement of 130 mmHg to units of standard deviation, we can find out how many standard deviations this person's SBP is above the sample mean. To do this:

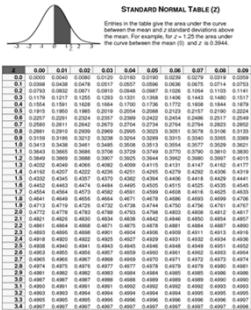
Take $\frac{\text{individual value} - \text{mean}}{\text{SD}} = \frac{(130.0 - 123.6)\text{mmHg}}{12.9 \text{ mmHg/SD}}$

$= \frac{6.4 \text{ mmHg}}{12.9 \text{ mmHg/SD}} \approx 0.5 \text{ SD}$

- Now the same question can be rephrased as "What percentage of observations in a normal curve are more than 0.5 SD above it's mean?"

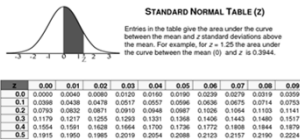
Example 1

- "What percentage of observations in a normal curve are greater than 0.5 SD above it's mean?"



Example 1

- "What percentage of observations in a normal curve are greater than 0.5 SD above it's mean?"



Example 1

- Another way to interpret this is as an (estimated) probability: the probability that any males in the population has a blood pressure measurement more than .5 standard deviations above the mean is .31 or 31%
- So ultimately, this estimates that 31% of the males in this population have blood pressures greater than 130 mmHg (ie: using only the mean and sd, we have estimated the 69th percentile to be 130 mmHg)

43

Example 1

- Another way to interpret this is as an (estimated) probability: the probability that any males in the population has a blood pressure measurement more than .5 standard deviations above the mean is .31 or 31%
- So ultimately, this estimates that 31% of the males in this population have blood pressures greater than 130 mmHg (ie: using only the mean and sd, we have estimated the 69th percentile to be 130 mmHg)

Just for context/comparison: the 70% percentile of the observed 113 values is 130 mmHg

44

z-score

- The type of computation we did to convert the SBP value of 130 to the number of SDs above (or below) the sample mean is sometime called a z-score
- There is nothing special about a z-score; it is simply a measure of the relative distance (and direction) of a single observation in a data distribution relative to the mean of the distribution; this distance is converted to units of standard deviation
- This is akin to converting kilometers to miles, or dollars to rupees

45

z-score: parallel example

- You are an American apartment hunting in an unnamed European city. You wish to find an apartment within walking distance (+/- 1.5 miles) of the large organic supermarket, which is on Main Boulevard (E/W). You are only considering apartments on Main Blvd.

The supermarket is 2 km west of the main city square. You are interested in 3 apartments:

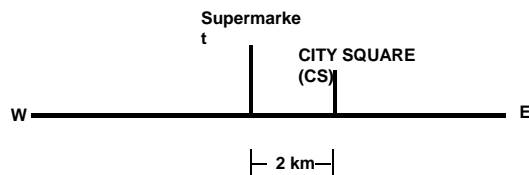
Apt 1 is 6 km west of the city square

Apt 2 is .75 km west of the city square

Apt 3 is 1 km *east* of the city square

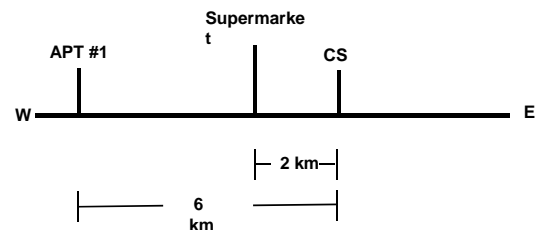
z-score: parallel example

- Picture



z-score: parallel example

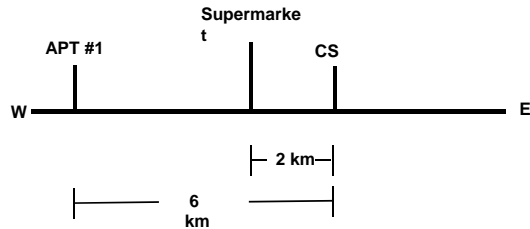
- Picture



z-score: parallel example

- How far is APT1 from Supermarket?:

“raw distance” = $6 - 2 = 4$ km



z-score: parallel example

- How many miles is 4 km?

well, $1 \text{ km} \approx 1.6 \text{ miles}$

$$\text{So, } 4 \text{ km} = \frac{4 \text{ km}}{1.6 \text{ km/mile}} \approx 2.5 \text{ miles}$$

z-score: parallel example

- Distance for each, converted to miles

$$\text{Apt 1: } \frac{\text{Apt1 from TC} - \text{Super from TC}}{\text{miles/km}} = \frac{(6-2)\text{km}}{1.6\text{km/mile}} \approx 2.5 \text{ mile}$$

$$\text{Apt 2: } \frac{(0.75-2)\text{km}}{1.6\text{km/mile}} = \frac{-1.25 \text{ km}}{1.6\text{km/mile}} \approx -0.78 \text{ miles}$$

$$\text{Apt 3: } \frac{(-1-2)\text{km}}{1.6\text{km/mile}} = \frac{-3 \text{ km}}{1.6\text{km/mile}} \approx -1.88 \text{ miles}$$

z-score

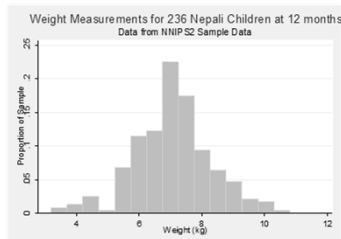
- In some sense, the z-score is the “statistical mile”
- It allows to convert observations from different distributions with different measurement scales to comparable units
- When dealing with data that follow an (approximately) normal distribution these z-scores tell us everything we need to know about the relative positioning of individual observations in the distribution of all observations
- We can compute z-scores for data arising from any type of distribution; however, for data from non-normal distributions, and it will inform us about the relative position
 - However, with non-normal data this may not translate into correct percentile information

52

Example 2: Weight Data

- Data on 236 Nepali children one year old at the time of the NNIPS2 (Nepal Nutritional Intervention Project-Sarlahi) Study

Estimate of μ : $\bar{x} = 7.1 \text{ kg}$; Estimate of σ : $s = 1.2 \text{ kg}$

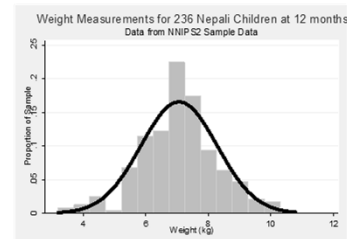


53

Example 2

- Data on 236 Nepali children one year old at the time of the NNIPS2 (Nepal Nutritional Intervention Project-Sarlahi) Study

Estimate of μ : $\bar{x} = 7.1 \text{ kg}$; Estimate of σ : $s = 1.2 \text{ kg}$



54

Example 2

- Using only the sample mean and standard deviation, and assuming normality, let's estimate a range of weights for most (95%) Nepali children who were 12 months old

2.5th %ile: $\bar{x} - 2s =$

97.5th %ile: $\bar{x} + 2s =$

55

Example 2

- Using only the sample mean and standard deviation, and assuming normality, let's estimate a range of weights for most (95%) Nepali children who were 12 months old

2.5th %ile: $\bar{x} - 2s = 7.1 - (2 \times 1.2) = 4.7 \text{ kg}$

97.5th %ile: $\bar{x} + 2s = 7.1 + (2 \times 1.2) = 9.5 \text{ kg}$

Based on this sample data, we estimate that most (95%) of Nepali children who were 12 months had weights between 4.7 kg and 9.5 kg

(Note: the empirical 2.5th and 97.5th percentile of the 113 sample value are 4.4 kg and 9.7 kg respectively)

56

Example 2

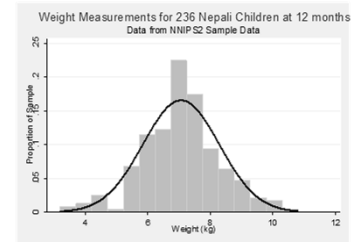
- Suppose a mother brings her child to a pediatrician for the 12 check up, and wants to evaluate where the child's weight is relative to the population of 12 months olds in Nepal
- Her child is 5 kg

How does this child compare in weight to the weight of all 12 month olds in Nepal?

57

Example 2

- Suppose a mother brings her child to a pediatrician for the 12 check up, and wants to evaluate where the child's weight is relative to the population of 12 months olds in Nepal: Her child is 5 kg



58

Example 2

- If we translate this measurement of 5 kg to units of standard deviation, we can find how where this child's weight compares the mean of all such children

Take $\frac{\text{individual value} - \text{mean}}{\text{SD}} =$

- The original question can be asked as "What percentage of observations in a normal curve are more than 1.75 SDs below it's mean"

59

Example 2

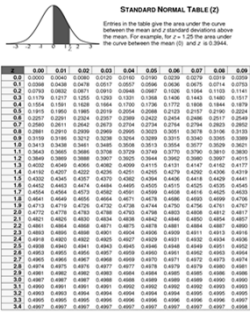
- If we translate this measurement of 5 kg to units of standard deviation, we can find how where this child's weight compares the mean of all such children

Take $\frac{\text{individual value} - \text{mean}}{\text{SD}} = \frac{(5 - 7.1)\text{kg}}{1.2 \text{ kg/SD}}$
 $= \frac{-2.1\text{kg}}{1.2 \text{ kg/SD}} \approx -1.75 \text{ SD}$

60

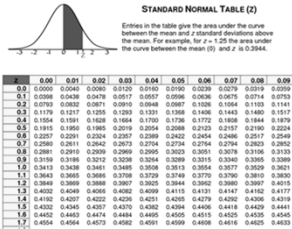
Example 2

- “What percentage of observations in a normal curve are more than 1.75 SDs below it’s mean?”
(ie, have a z-score < -1.75)



Example 2

- “What percentage of observations in a normal curve are more than 1.75 SDs below it’s mean?”
(ie, have a z-score < -1.75)



Example 2

- “What percentage of observations in a normal curve are more than 1.75 SDs below it’s mean?”
(ie, have a z-score < -1.75)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633

Example 2

- “What percentage of observations in a normal curve are more than 1.75 SDs below it’s mean?”
(ie, have a z-score < -1.75)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633

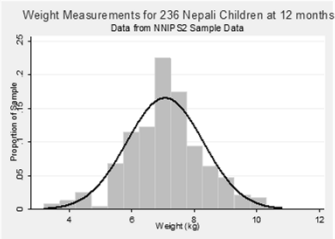
Note: the observed 5th percentile of these 236 measurements is 5 kg.

Example 2

- We could answer a broader question about the child who weighed 5 kg as well: “What percentage of 12 month old Nepali children have weights more extreme (unusual) than this child?”
 - What percentage of weights are farther than 1.75 SDs from the mean in either direction (above or below: ie $z < -1.75$ or $z > 1.75$, sometimes expressed as $|z| > 1.75$)
- Note: the above question can also be phrased “What is the probability that a 12 month old Nepali child will have a weight measurement more than 1.75 SD from the mean of all such children (above or below)?”

Example 2

- What percentage of weights are farther than 1.75 SDs from the mean in either direction (above or below: ie $z < -1.75$ or $z > 1.75$, sometimes expressed as $|z| > 1.75$)



Summary

- The normal distribution is a theoretical probability distribution, which can be completely defined by two characteristics: the mean and standard deviation
- No real world data has a perfect normal distribution; however, some continuous measures are reasonably approximated by a normal distribution

67

Summary

- When dealing with samples from populations of (approximately) normally distributed data, the distribution of sample values will also be approximately normal. We can use the sample mean and standard deviation estimates, (\bar{x} and s) to:
 - Create ranges containing a certain percentage of observations or in or in other words:
 - Estimate the probability that an observed data point falls within a certain range of values
 - Figure out how far any individual data point is from the mean of its distribution in standardized unit (compute a z-score)
 - Convert z-scores to statements about relative proportions/probabilities (and hence percentiles) for values that have an (approximately) normal distribution

68

Section C : What Happens When We Apply the Properties of the Normal Distribution to Data Not Approximately Normal: A Warning

69

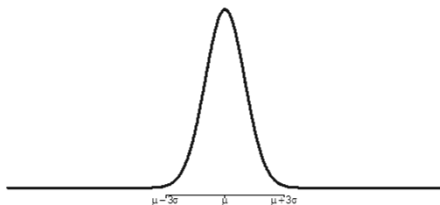
Learning Objectives

- Upon completion of this lecture you will be able to:
 - Describe situations in which using only the mean and standard deviation of a distribution of values to characterize the entire distribution will not work well
 - Realize that z-scores are nothing “special”; z-scores are just a (standardized) measure of distance
 - Understand the z-scores do not necessarily align with the corresponding percentiles for a normal distribution for data that does not follow a normal distribution
 - Choose the right approach to estimating ranges for individual values, and computing percentage greater (or less) than a specific value using non-normal data distributions

70

The Normal Distribution

- The normal distribution is a theoretical probability distribution: no real data is perfectly described by this distribution
- For example, in a true normal distribution, the tails go onto *negative and positive infinity, respectively*



71

The Normal Distribution

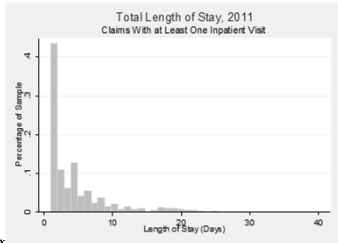
- However, the distributions of some data will be well approximated by a normal distribution: in such situations we can use the properties of the normal curve to characterize aspects of the data distribution
- But: the distributions of much data WILL NOT be well approximated by a normal distribution: in such situations using the properties of the normal curve to characterize aspects of the data distribution will yield invalid results

72

Example 1

- Example 1: Length of stay claims at Heritage Health with an inpatient stay of at least one day in 2011¹ (12,928 claims)

Estimate of μ : $\bar{x} = 4.3$ days ; Estimate of σ : $s = 4.9$ days



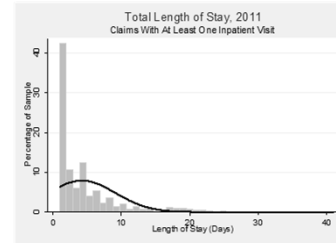
¹ <http://inclass.kaggle.com/>

73

Example 1

- Example 1: Length of stay claims at Heritage Health with an inpatient stay of at least one day in 2011

Estimate of μ : $\bar{x} = 4.3$ days ; Estimate of σ : $s = 4.9$ days

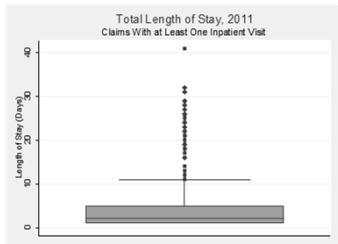


74

Example 1

- Example 1: Length of stay claims at Heritage Health with an inpatient stay of at least one day in 2011

Estimate of μ : $\bar{x} = 4.3$ days ; Estimate of σ : $s = 4.9$ days



75

Example 1

- Using only the sample mean and standard deviation, and assuming normality, let's estimate the 2.5th and 97.5th percentiles length of stay of in this population

2.5th %ile: $\bar{x} - 2s = 4.3 - 2 \times 4.9 = -5.5$ days

97.5th %ile: $\bar{x} + 2s = 4.3 + (2 \times 4.9) = 14.1$

Based on this sample data, we estimate that most (95%) of the persons making claims in this healthcare population had length of stays between -5.5 and 14.1 days in 2011. (???????)

(Note: the empirical 2.5th and 97.5th percentile of the 12,298 sample values are 1 day and 20 days respectively)

76

Example 1

- In this example, using the properties of the normal curve to estimate an interval containing the "middle 95%" of length of stay values for the claims population yields useless results
- Better to take the observed 2.5th and 97.5th percentiles of the sample data and report these as an estimate of the "middle 95%"

"Based on this sample data, we estimate that most (95%) of the persons making claims in this healthcare population had length of stays between 1 and 21 days in 2011."

77

Example 1

- Suppose we wish to use these data to estimate the proportion of the claims population with total length of stay of greater than 5 days.

78

Example 1

- If we translate this measurement of 5 days to units of standard deviation, we can find where 5 days is relative to the sample mean length of stay. To do this, first find the “z-score”:

$$\begin{aligned} \text{Take } \frac{\text{individual value} - \text{mean}}{\text{SD}} &= \frac{(5.0 - 4.3) \text{ days}}{4.9 \text{ days/SD}} \\ &= \frac{0.7 \text{ days}}{4.9 \text{ days/SD}} \approx 0.14 \text{ SD} \end{aligned}$$

79

Example 1

- If we translate this measurement of 5 days to units of standard deviation, we can find where 5 days is relative to the sample mean length of stay. To do this, first find the “z-score”:

$$\begin{aligned} \text{Take } \frac{\text{individual value} - \text{mean}}{\text{SD}} &= \frac{(5.0 - 4.3) \text{ days}}{4.9 \text{ days/SD}} \\ &= \frac{0.7 \text{ days}}{4.9 \text{ days/SD}} \approx 0.14 \text{ SD} \end{aligned}$$

I'll let you verify that the probability that an of getting an observation that is greater than 0.14 SD above the mean of a normal distribution is 0.44 or 44%.

80

Example 1

- However, if we look at some percentiles of the sample data:

Percentile	Value
2.5 th	1 day
10 th	1 day
25 th	1 day
50 th	2 days
75 th	5 days
90 th	10 days
97.5 th	20 days

81

Example 1

- However, if we look at some percentiles of the sample data:

Percentile	Value
2.5 th	1 day
10 th	1 day
25 th	1 day
50 th	2 days
75 th	5 days
90 th	10 days
97.5 th	20 days

82

Example 1

- To get more clarity, let's add in the 60th and 70th, and 80th percentiles

Percentile	Value
2.5 th	1 day
10 th	1 day
25 th	1 day
50 th	2 days
60 th	4 days
70 th	4 days
75 th	5 days
80 th	6 days
90 th	10 days
97.5 th	20 days

83

Example 1

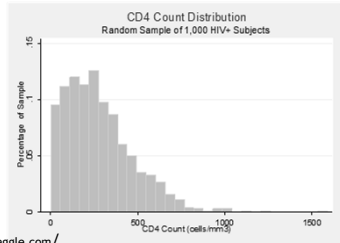
- Based on these analyses, we estimate that approximately 25% of the claims had total length of stay greater than 5 days
- This percentage is a lot smaller than the estimate of 44% we got using the mean and standard deviation to compute a z-score

84

Example 2

- Example 2: CD4 counts for a random sample of 1,000 HIV+ positive patients from a citywide clinical population²

Estimate of μ : $\bar{x} = 280 \text{ cell/mm}^3$; Estimate of σ : $s = 198 \text{ cells/mm}^3$



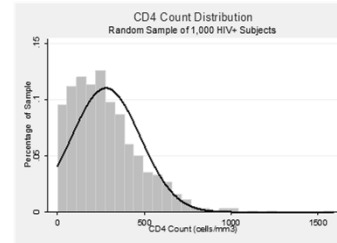
¹ <http://inclass.kaggle.com/>

85

Example 2

- Example 2: CD4 counts for a random sample of 1,000 HIV+ positive patients from a citywide clinical population

Estimate of μ : $\bar{x} = 280 \text{ cell/mm}^3$; Estimate of σ : $s = 198 \text{ cells/mm}^3$

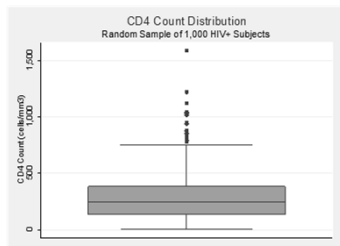


86

Example 2

- Example 2: CD4 counts for a random sample of 1,000 HIV+ positive patients from a citywide clinical population

Estimate of μ : $\bar{x} = 280 \text{ cell/mm}^3$; Estimate of σ : $s = 198 \text{ cells/mm}^3$



87

Example 2

- Using only the sample mean and standard deviation, and assuming normality, let's estimate the 2.5th and 97.5th percentiles of CD4 counts in this population

$$2.5^{\text{th}} \text{ tile: } \bar{x} - 2s = 280 - (2 \times 198) = -116 \text{ cells/mm}^3$$

$$97.5^{\text{th}} \text{ tile: } \bar{x} + 2s = 280 + (2 \times 198) = 676 \text{ cells/mm}^3$$

Based on this sample data, we estimate that most (95%) of population of HIV+ persons had CD4 counts between -116 and 676 cells/mm³.

(Note: the empirical 2.5th and 97.5th percentile of the 12,298 sample value are 11 and 722 cells/mm³ respectively)

88

Example 2

- Historically, guidelines about when to start Anti-retroviral therapy (ART) have changed as a function of CD4 count (and have varied by country)
- At one point, it was recommended that ART be initiated for those with CD4 counts < 350 cells/mm³
- Based on our sample of 1,000 let's estimate the proportion of HIV+ subjects in the population with CD4 count < 350 cells/mm³

89

Example 2

- If we translate 350 cells/mm³ to units of standard deviation, we can find where this value is relative to the sample mean CD4 count. To do this, first find the "z-score":

$$\begin{aligned} \text{Take } \frac{\text{individual value} - \text{mean}}{\text{SD}} &= \frac{(350 - 280) \text{ cells/mm}^3}{198 (\text{cells/mm}^3)/\text{SD}} \\ &= \frac{70 \text{ cells/mm}^3}{198 (\text{cells/mm}^3)/\text{SD}} \approx 0.35 \text{ SDs} \end{aligned}$$

90

Example 2

- If we translate 350 cells/mm³ to units of standard deviation, we can find where this value is relative to the sample mean CD4 count. To do this, first find the “z-score”:

- Take
$$\frac{\text{individual value} - \text{mean}}{\text{SD}} = \frac{(350 - 280) \text{ cells/mm}^3}{198 (\text{cells/mm}^3)/\text{SD}}$$
$$= \frac{70 \text{ cells/mm}^3}{198 (\text{cells/mm}^3)/\text{SD}} \approx 0.35 \text{ SDs}$$

I'll let you verify that the proportion of observations that are less than 0.35 standard deviation ($z < 0.35$) in a normal curve is approximately 0.64 or 64%.

91

Example 2

- If we examine the empirical percentiles of the CD4 data, 350 corresponds to the 70th percentile. So if we used the properties of the normal distribution with these data, we would underestimate the proportion of persons qualifying for ART (using the cutoff of 350) by roughly 6%.

92