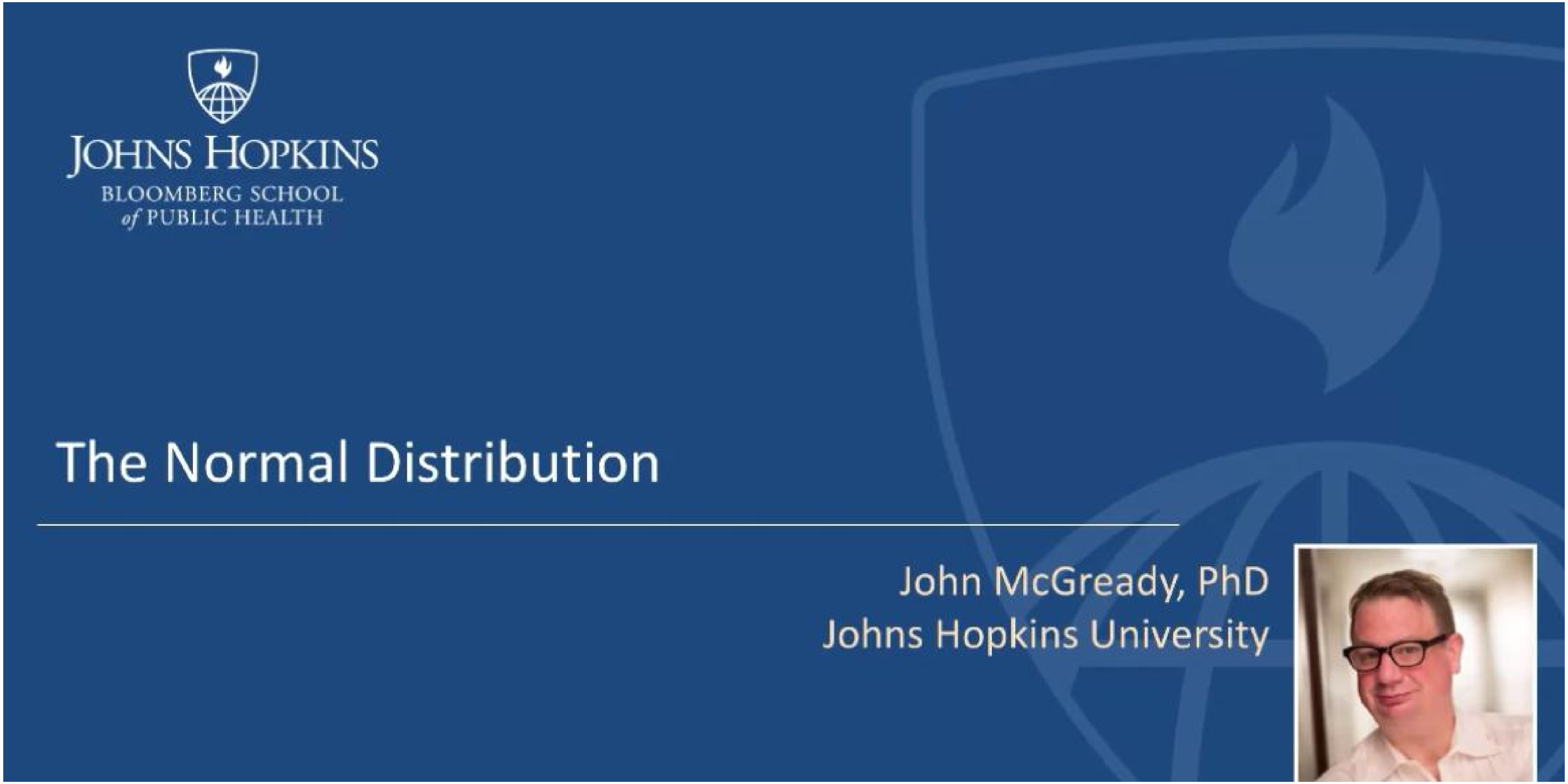


JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

The Normal Distribution

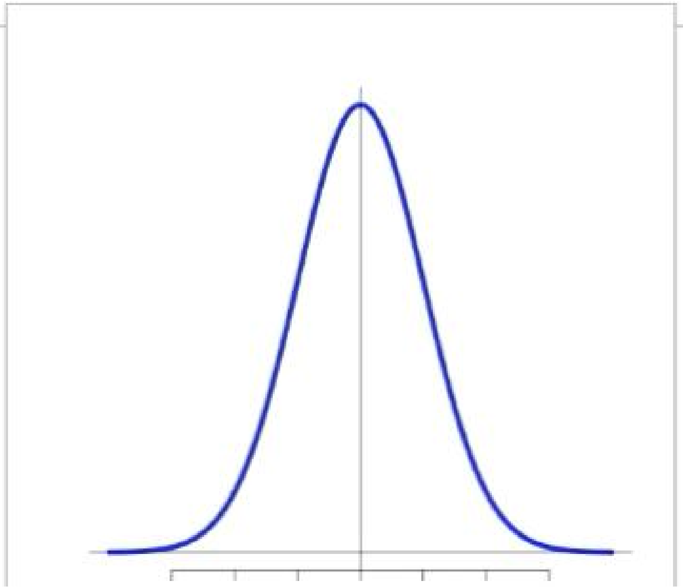
John McGready, PhD
Johns Hopkins University

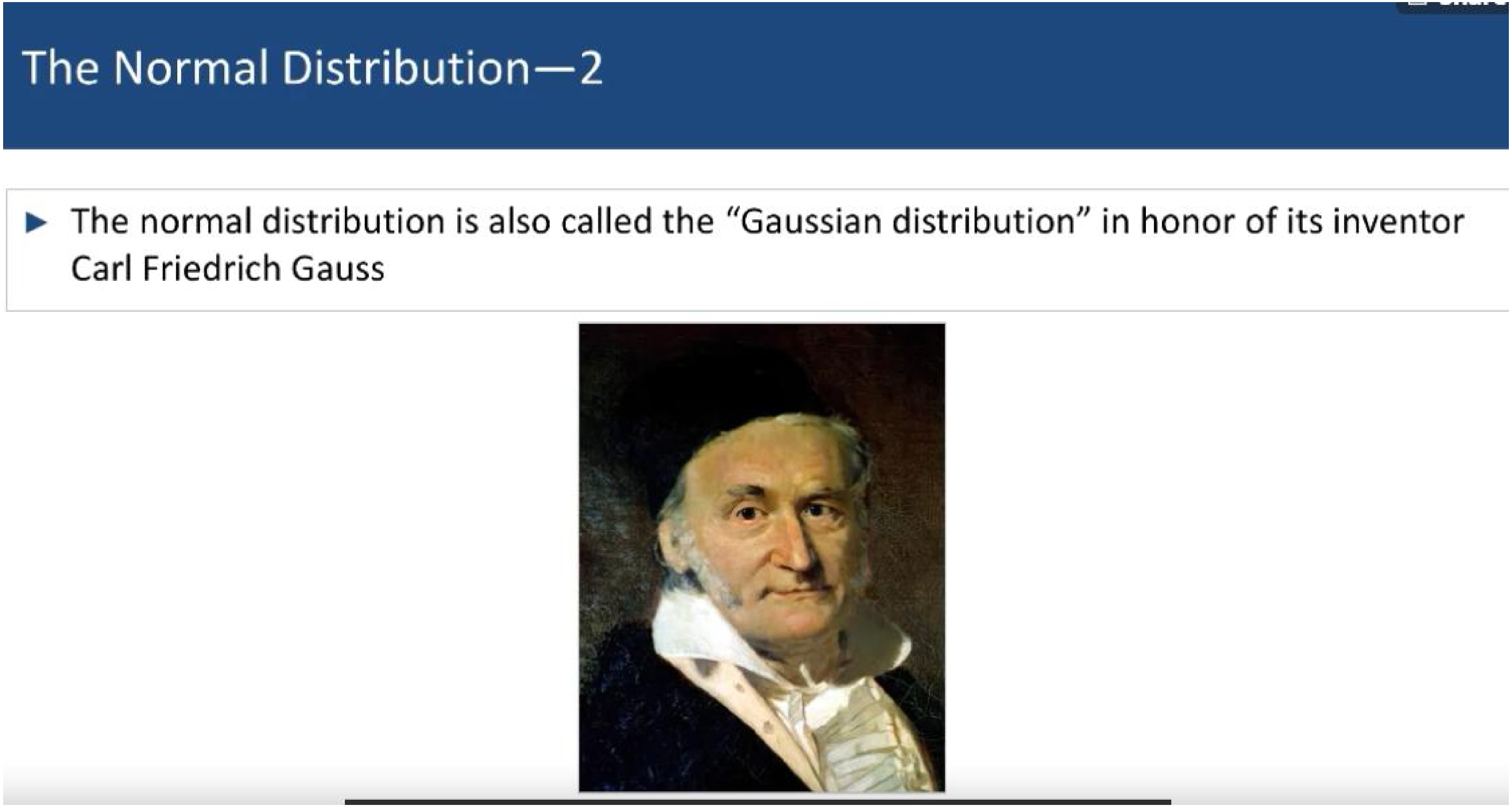
Learning Objectives

- Upon completion of this lecture, you will be able to:
  - Describe the basic properties of the normal curve
  - Describe how the normal distribution is completely defined by its mean and standard deviation
  - Recite the 68–95–99.7% rule for the normal distribution with regards to standard deviations

The Normal Distribution—1

- The normal distribution is a theoretical probability distribution that is perfectly symmetric about its mean (and median and mode)
  - A “bell”-like shape



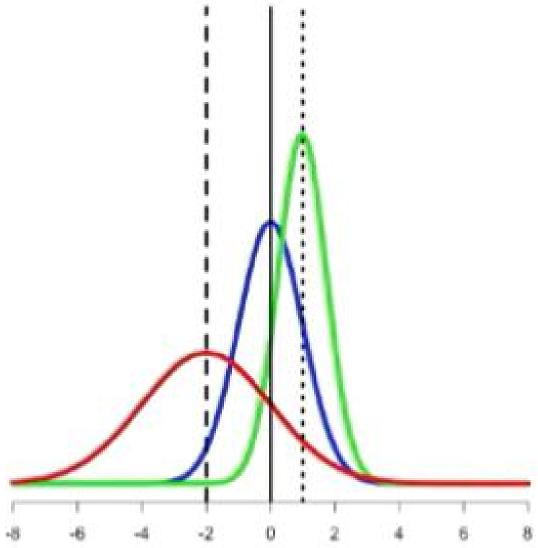


The Normal Distribution—2

- The normal distribution is also called the "Gaussian distribution" in honor of its inventor Carl Friedrich Gauss

Defining Quantities for any 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)



Underlying Formula for Normal Distribution

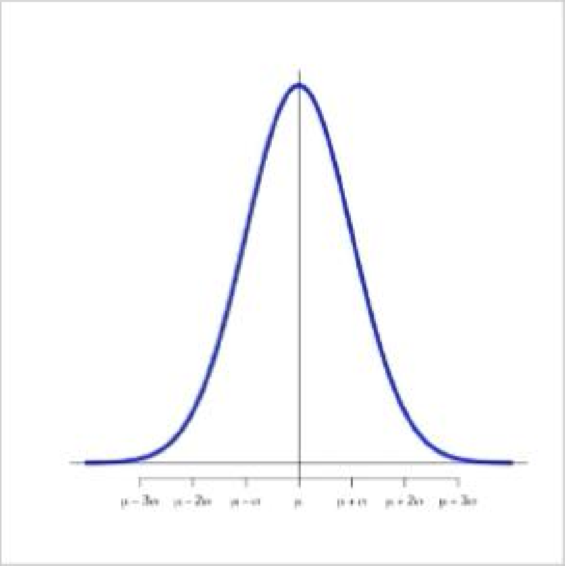
- This function defines the normal curve for any given (μ) and (σ)

- The proportion of values falling between a and b under a normal curve is given by:

$$\int_{a}^{b} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$$

Structural Properties of the Normal Distribution (Curve)—1

- 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 "farther" from the mean

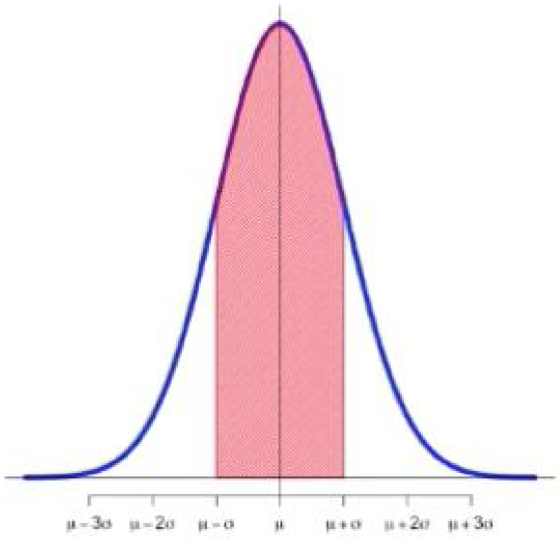


Structural Properties of the Normal Distribution (Curve)—2

- 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 the next section, we'll show that any normal distribution can be easily rescaled to this standard normal distribution

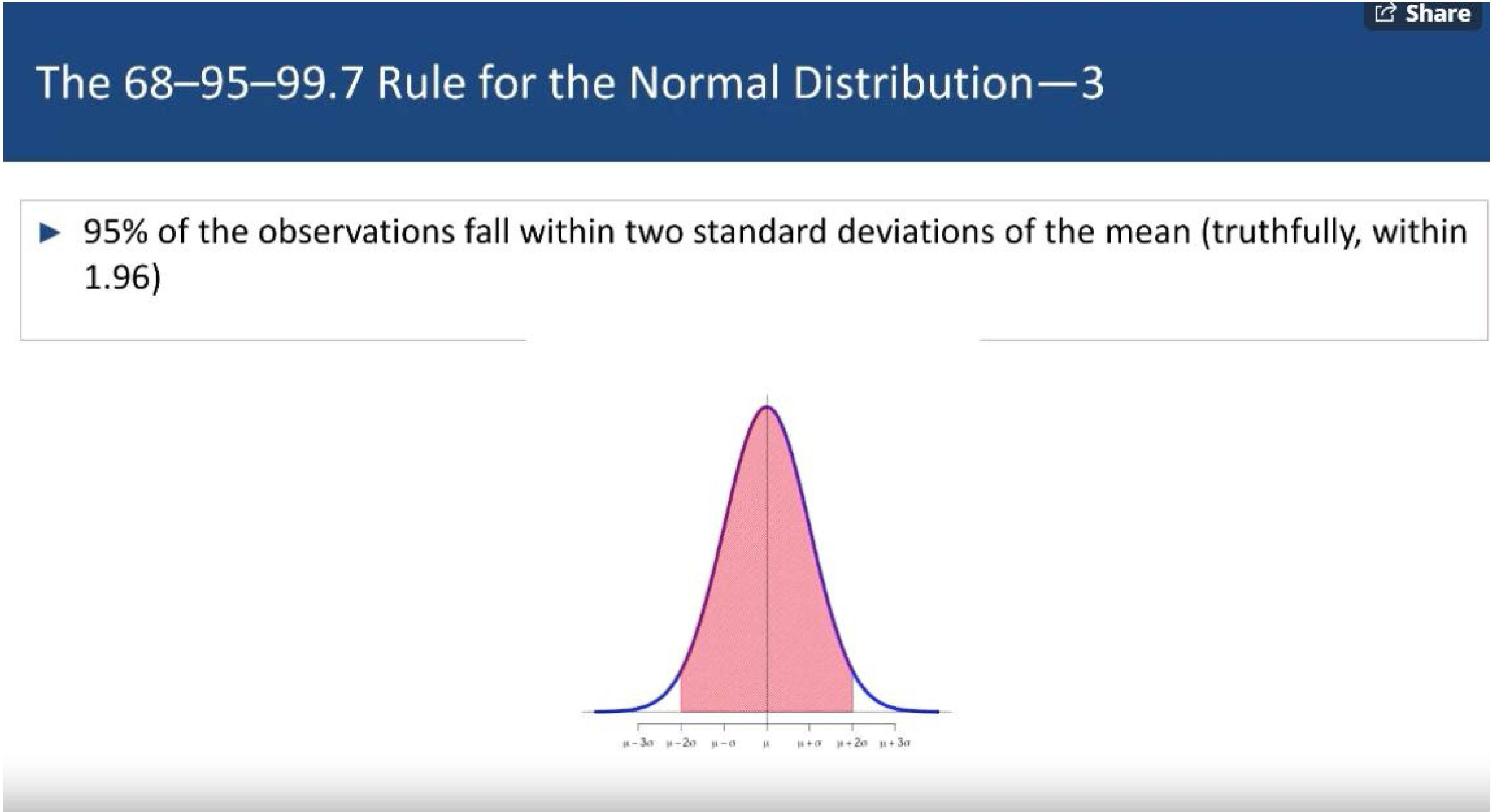
The 68–95–99.7 Rule for the Normal Distribution—1

- 68% of the observations in a normal distribution fall within one standard deviation of the mean



The 68–95–99.7 Rule for the Normal Distribution—2

- 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%

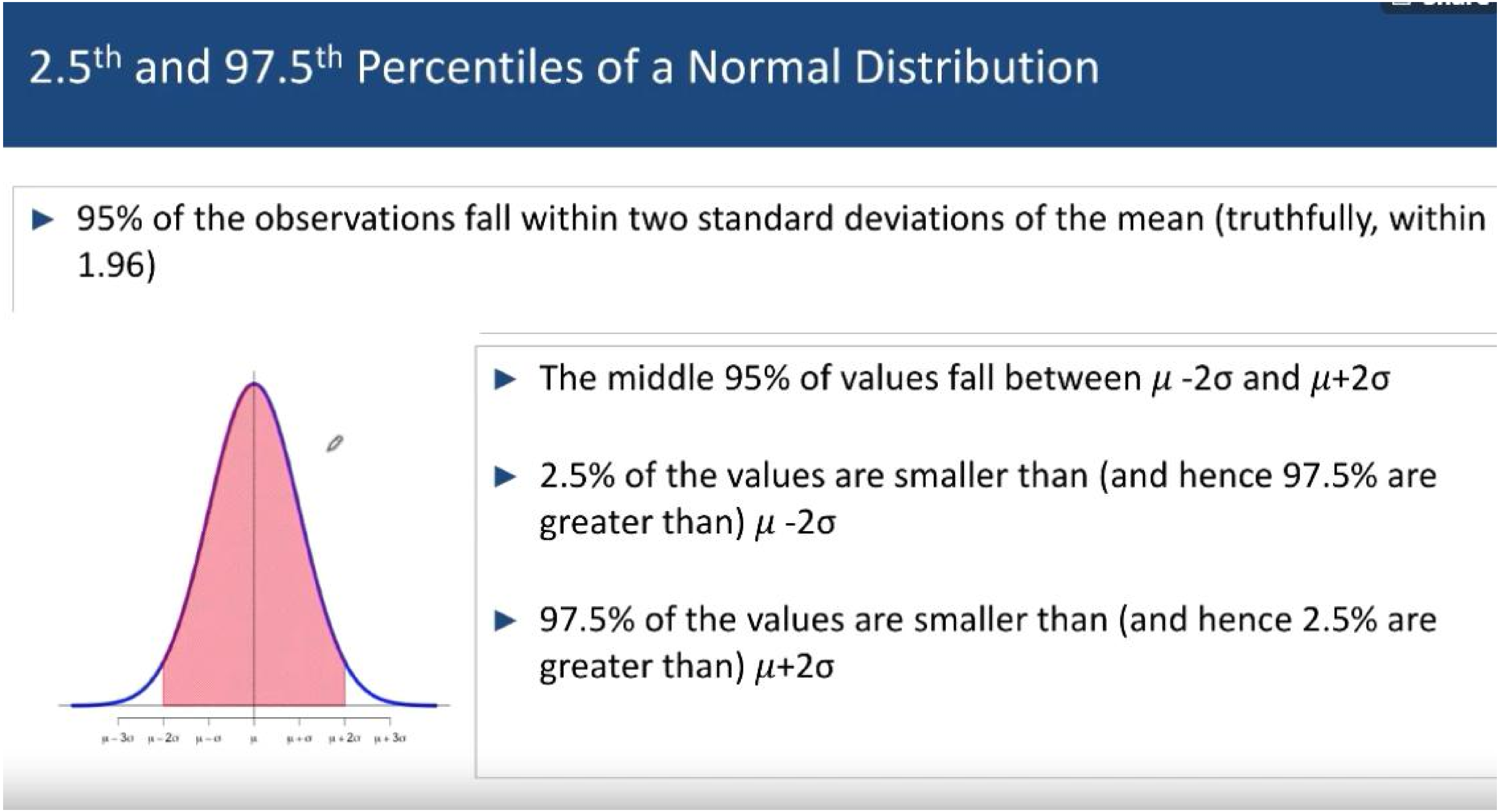


The 68–95–99.7 Rule for the Normal Distribution—3

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

The 68–95–99.7 Rule for the Normal Distribution—4

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



2.5th and 97.5th Percentiles of a Normal Distribution

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

- The middle 95% of values fall between μ -2σ and μ+2σ
- 2.5% of the values are smaller than (and hence 97.5% are greater than) μ -2σ
- 97.5% of the values are smaller than (and hence 2.5% are greater than) μ+2σ

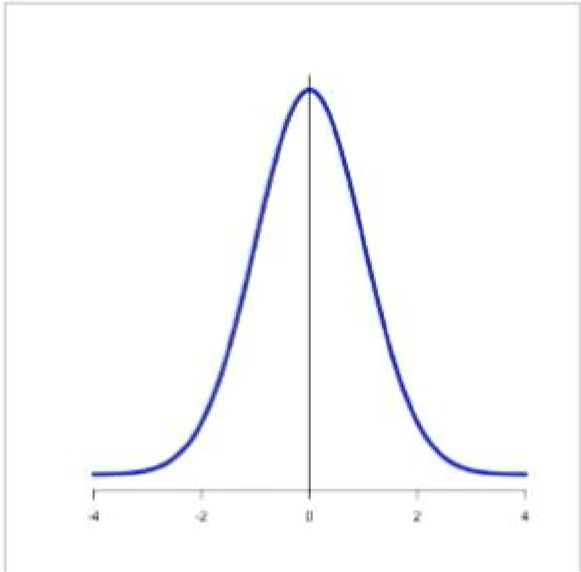
Percentage of Observations Under the Normal Distribution

- 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”

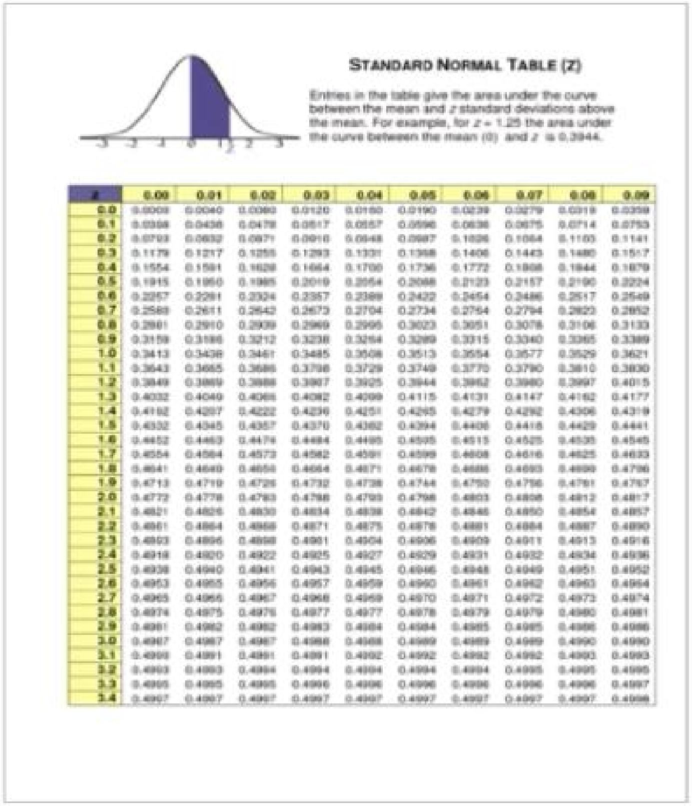
The Standard Normal Distribution

- The standard normal distribution is a normal distribution with mean μ =0, and standard deviation σ=1

- Any normal distribution with mean μ and standard deviation σ can be rescaled to a standard normal distribution.



Percentage of Observations Under the Normal Distribution: Exhibit A

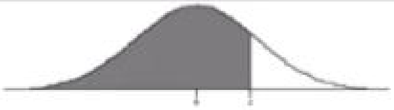


STANDARD NORMAL TABLE (Z)

Entries in the table give the area under the curve between the mean and z standard deviations above the mean. For example, for z = 1.25 the area under the curve between the mean (0) and z is 0.3944.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	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	0.2422	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	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
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	0.3513	0.3554	0.3577	0.3529	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	0.4495	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
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

Percentage of Observations Under the Normal Distribution: Exhibit B



Normal Deviate z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-4.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.8	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.7	.0001	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483

Percentage of Observations Under the Normal Distribution—2

- In this class, I will only have you find relevant percentages under a normal curve for some early activities, and this will be done easily using R

- Generally speaking, I only want you to be familiar with the 68–95–99.7 rule

- Such computations will be wrapped into other analyses later in the course and completely handled by a computer

Using R to Compute Normal Curve Percentages—1

- We can use R as a calculator, i.e., an automatic standard normal table
- The relevant command that "looks up" values in a standard normal table is:
- For converting any standard deviation value (above or below the mean), *z*, to a corresponding proportion under a normal curve, the syntax is:

```
pnorm(z)
```

Using R to Compute Normal Curve Percentages—2

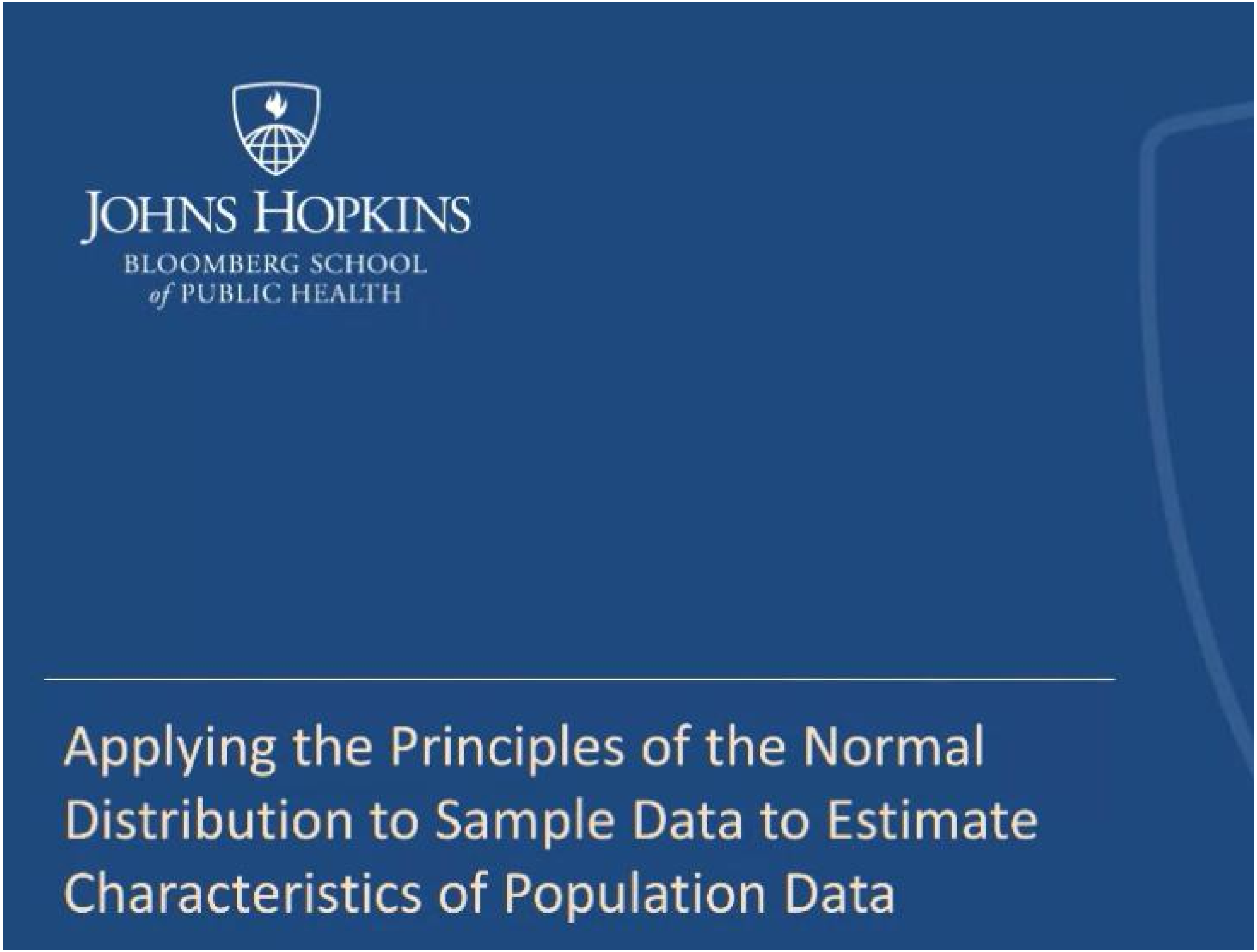
- As with any print version of a standard normal table, it is important to know what information pnorm(z) returns

- The normal distribution is a theoretical probability distribution that is symmetric and "bell-shaped"

- There are literally an infinite number of normal distributions, and each can be completely specified by only two quantities: the mean and standard deviation

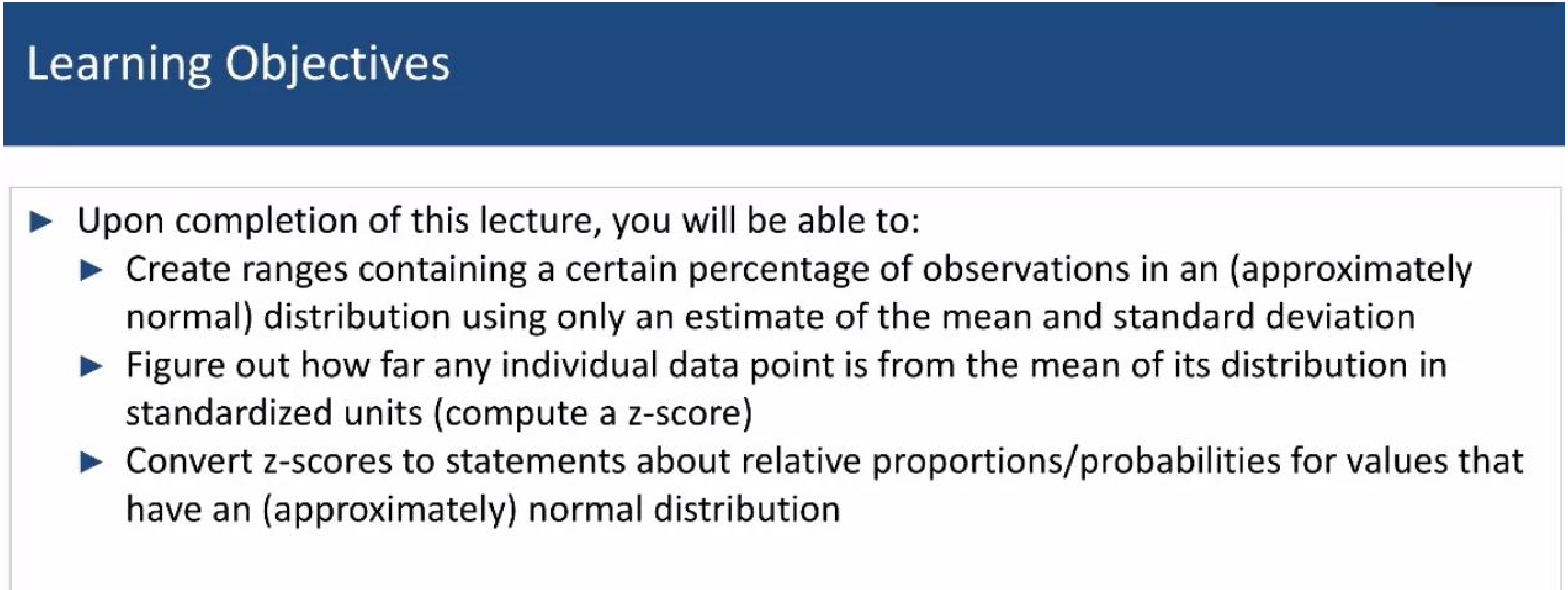
- For all normal distributions
  - 68% of observations described by a normal distribution fall within 1 sd of the mean
  - 95% of observations described by a normal distribution fall within 2 sds of the mean
  - 99.7% of observations described by a normal distribution fall within 3 sds of the mean

- Other such percentages can be found using a standard normal table (available via R)



JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

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



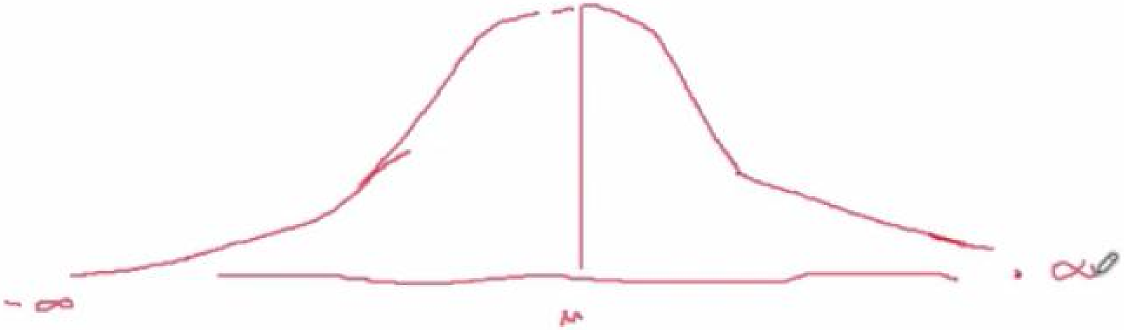
Learning Objectives

- 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

The Normal Distribution, Generally Speaking—1

- 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 on to negative and positive infinity, respectively



The Normal Distribution, Generally Speaking—2

- 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

Using Sample Estimates and Properties of Normal Distribution to Estimate Percentiles

- 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

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

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

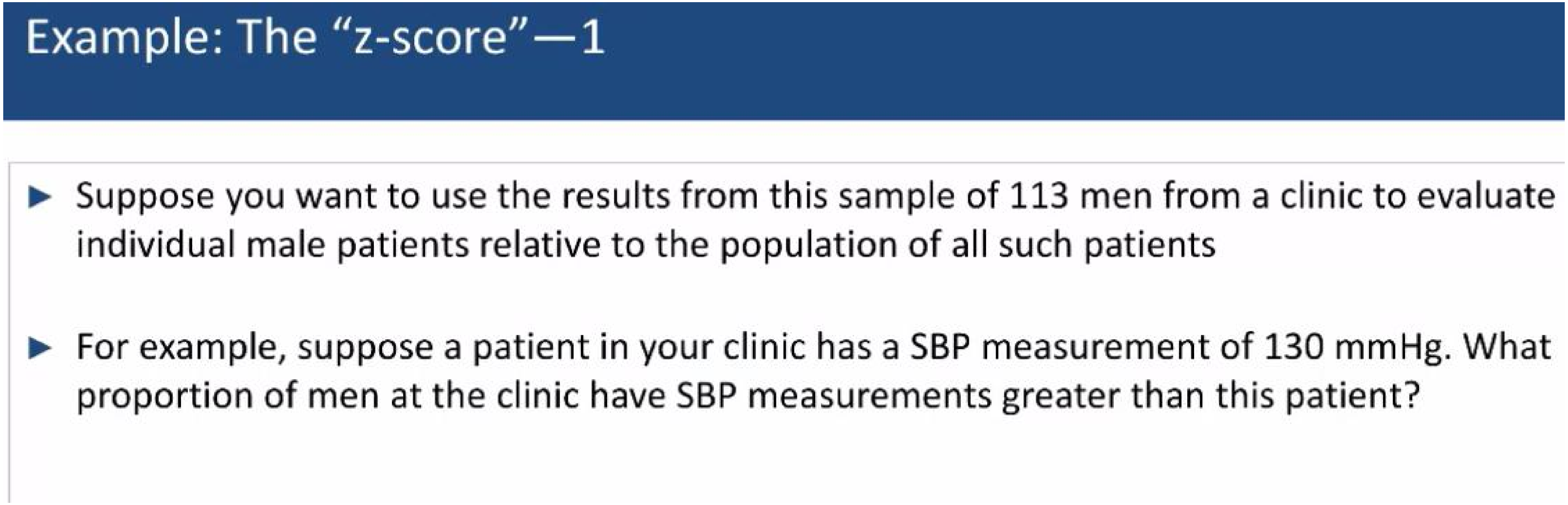
- Based on this sample data, we estimate that *most (95%) of the men* in this clinical population have systolic blood pressures *between 97.8 and 149.4 mmHg*

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

Example: The “z-score”—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: The “z-score”—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: The “z-score”—3

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

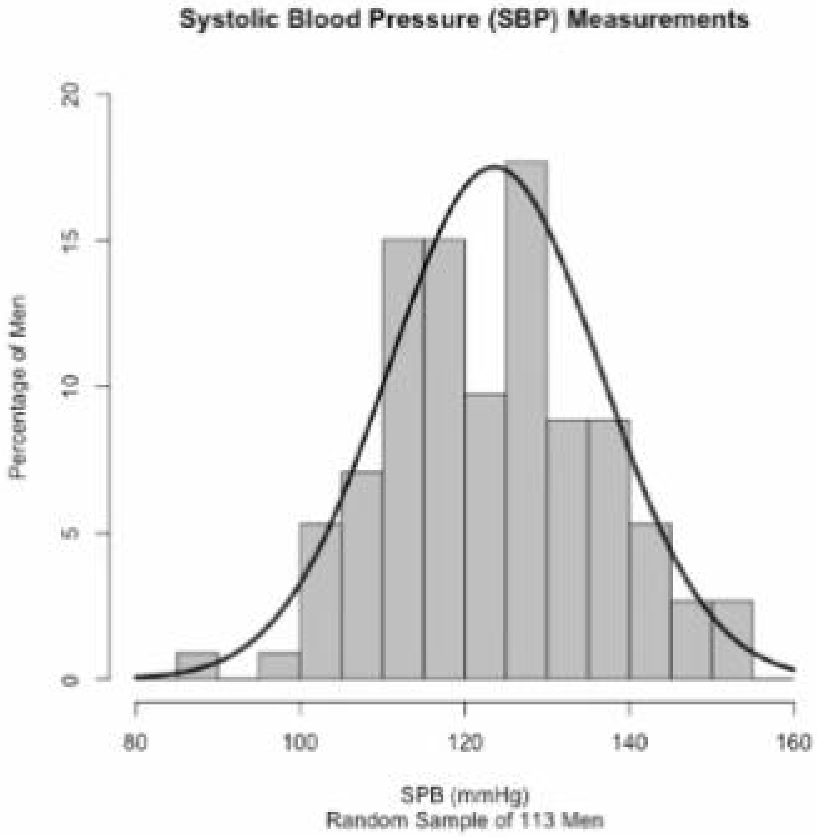
$$\text{Take } \frac{individual\ observation - \bar{x}}{s} = \frac{130\ \ - 123.6}{12.9} = \frac{6.4 \text{ mmHg}}{12.9\ \text{ mmHg}/s}$$
$$\approx 0.5\ standard\ deviations$$

- Now, the same question can be rephrased as, “What percentage of observations in a normal curve are more than 0.5 SD above its mean?”

Example: The "z-score"—2

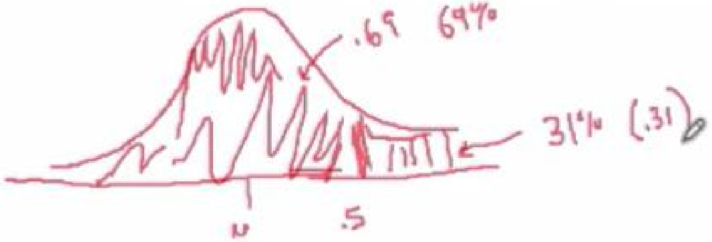
- We want to figure out:

$$\bar{x} = 123.6 \text{ mmHg}$$
$$s = 12.9 \text{ mmHg}$$
$$\hat{m} = 123.0 \text{ mmHg}$$

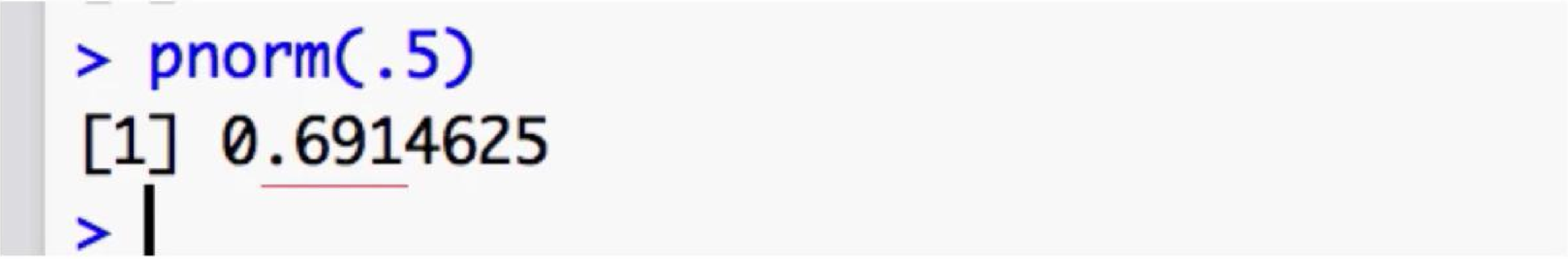


Example: The “z-score”—4

- “What percentage of observations in a normal curve are greater than 0.5 SD above its mean?”



- Using pnorm in R:



```
> pnorm(.5)
[1] 0.6914625
> 
```

Example: The “z-score” —5

- So 69% of the observations described by a standard normal curve are less than or equal to 0.5 standard deviations above the mean of 0

- Hence, the remaining 31% are more than 0.5 standard deviations above 0

- In terms of the original question posed, this means that an estimated 31% of the males in this population have blood pressures greater than 130 mmHg (i.e., using only the mean and sd, we have estimated the 69th percentile to be 130 mmHg)

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

Example: The “z-score” —6

- 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%

The "z-score," Generally Speaking—1

- 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 sometimes 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

The "z-score": A Parallel Example—1

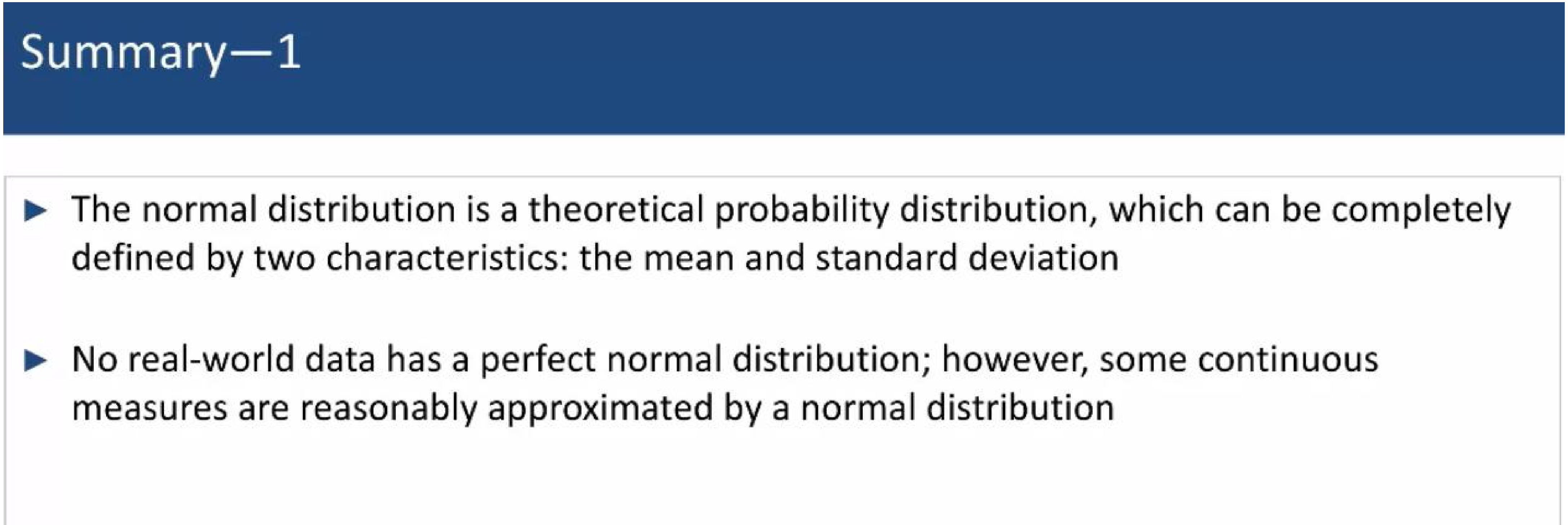
- You are an American who is apartment hunting in an unnamed European city. You wish to find an apartment within walking distance (+/- 1.5 miles) of the large organic supermarket, which is on Main Boulevard (E/W). You are only considering apartments on Main Blvd.

- 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

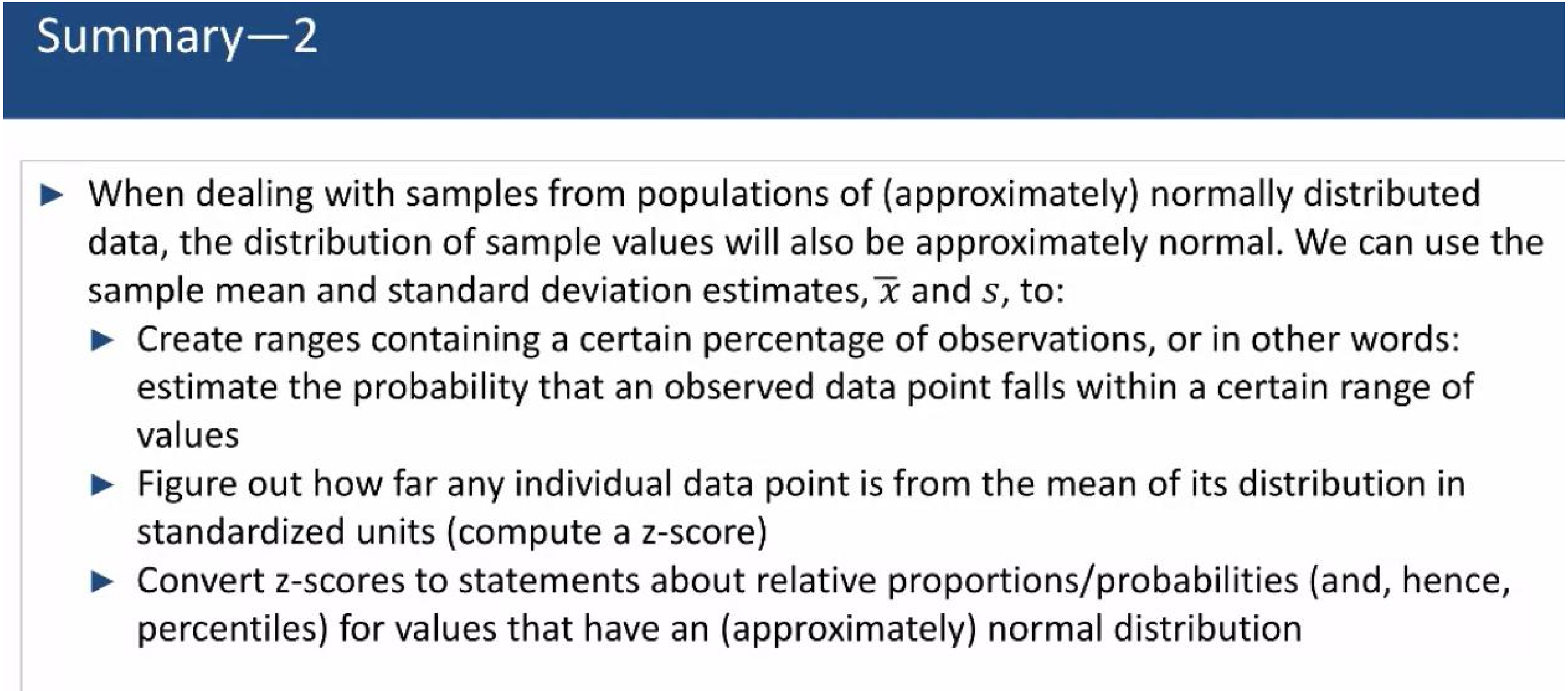
Apt 3 is 1 km *east* of the city square



Summary—1

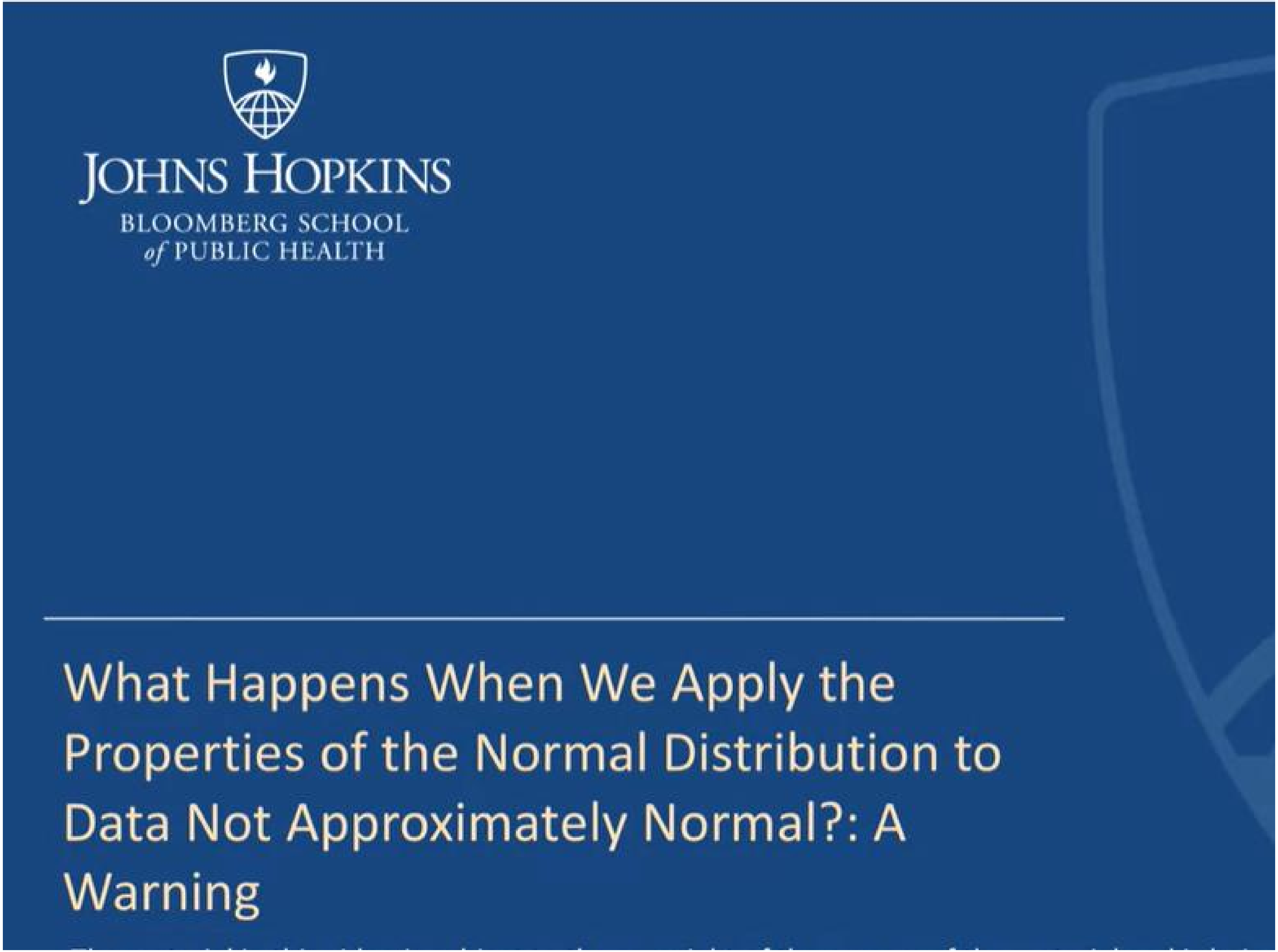
- The normal distribution is a theoretical probability distribution, which can be completely defined by two characteristics: the mean and standard deviation

- No real-world data has a perfect normal distribution; however, some continuous measures are reasonably approximated by a normal distribution



Summary—2

- 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, $\overline{x}$ and s, to:
  - Create ranges containing a certain percentage of observations, 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 units (compute a z-score)
  - Convert z-scores to statements about relative proportions/probabilities (and, hence, percentiles) for values that have an (approximately) normal distribution

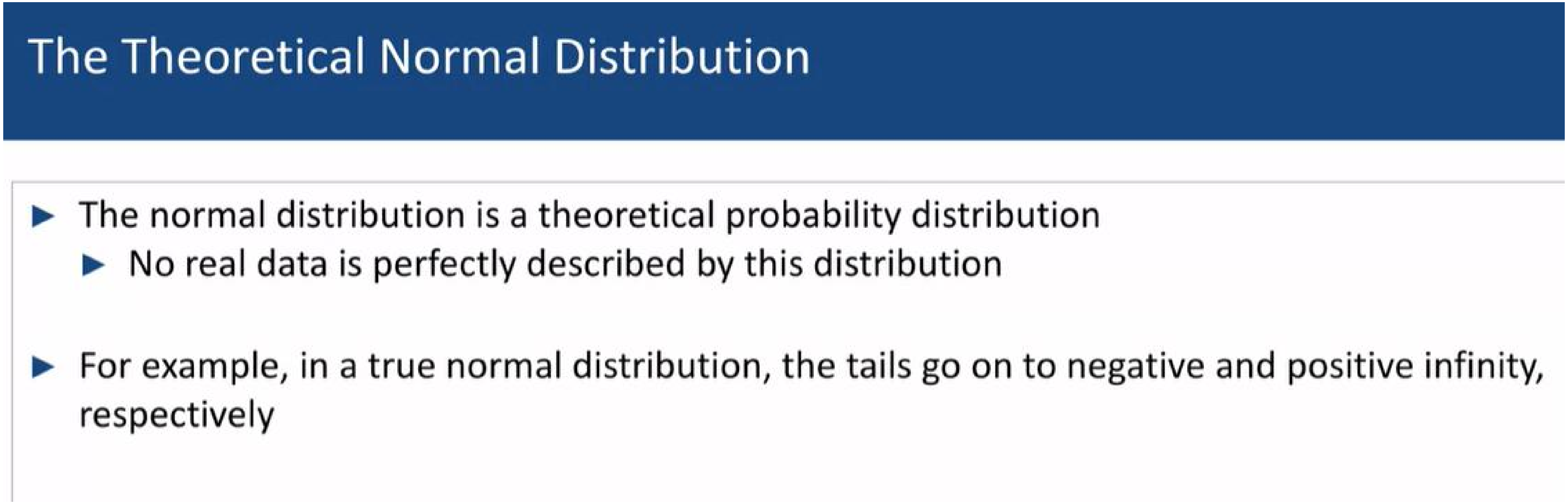


JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

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

Learning Objectives

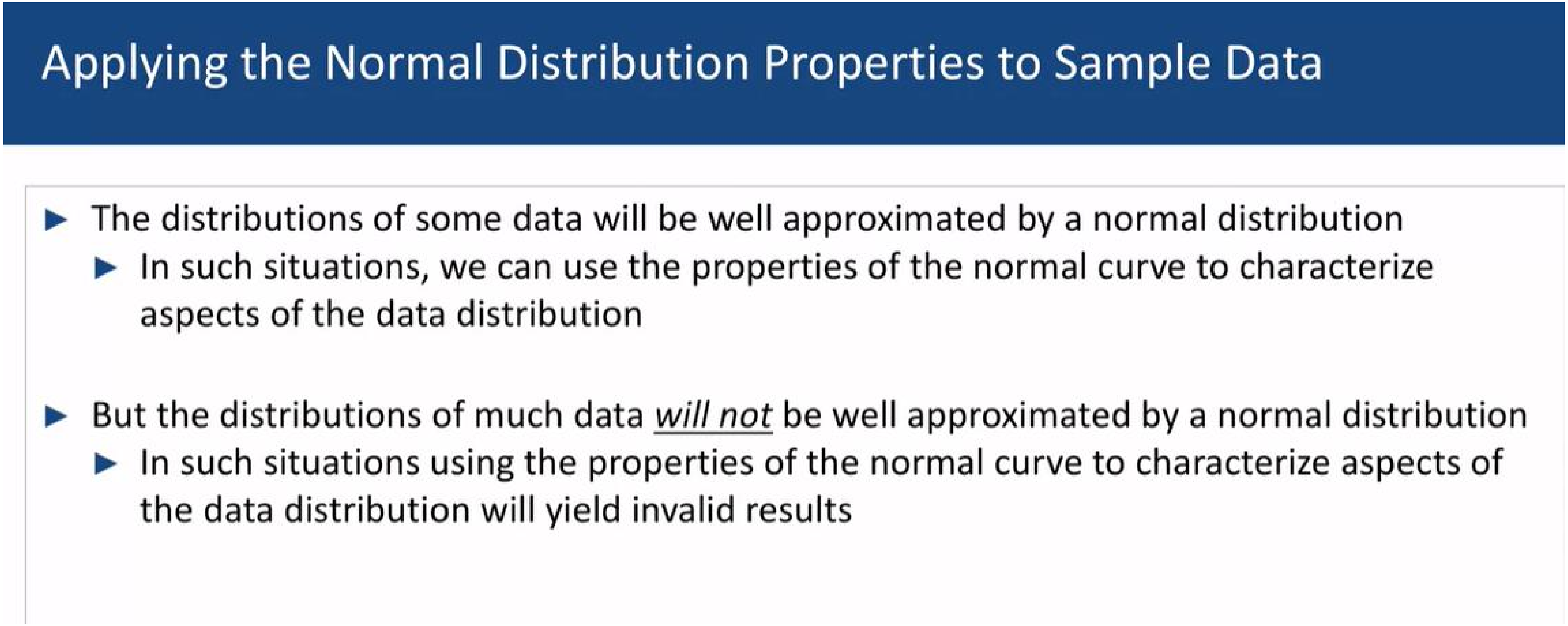
- Upon completion of this lecture, you will be able to:
  - Describe situations in which using only the mean and standard deviation of a distribution of values to characterize the entire distribution will not work well
  - Realize that z-scores are nothing “special”; z-scores are just a (standardized)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



The Theoretical Normal Distribution

- The normal distribution is a theoretical probability distribution
  - No real data is perfectly described by this distribution

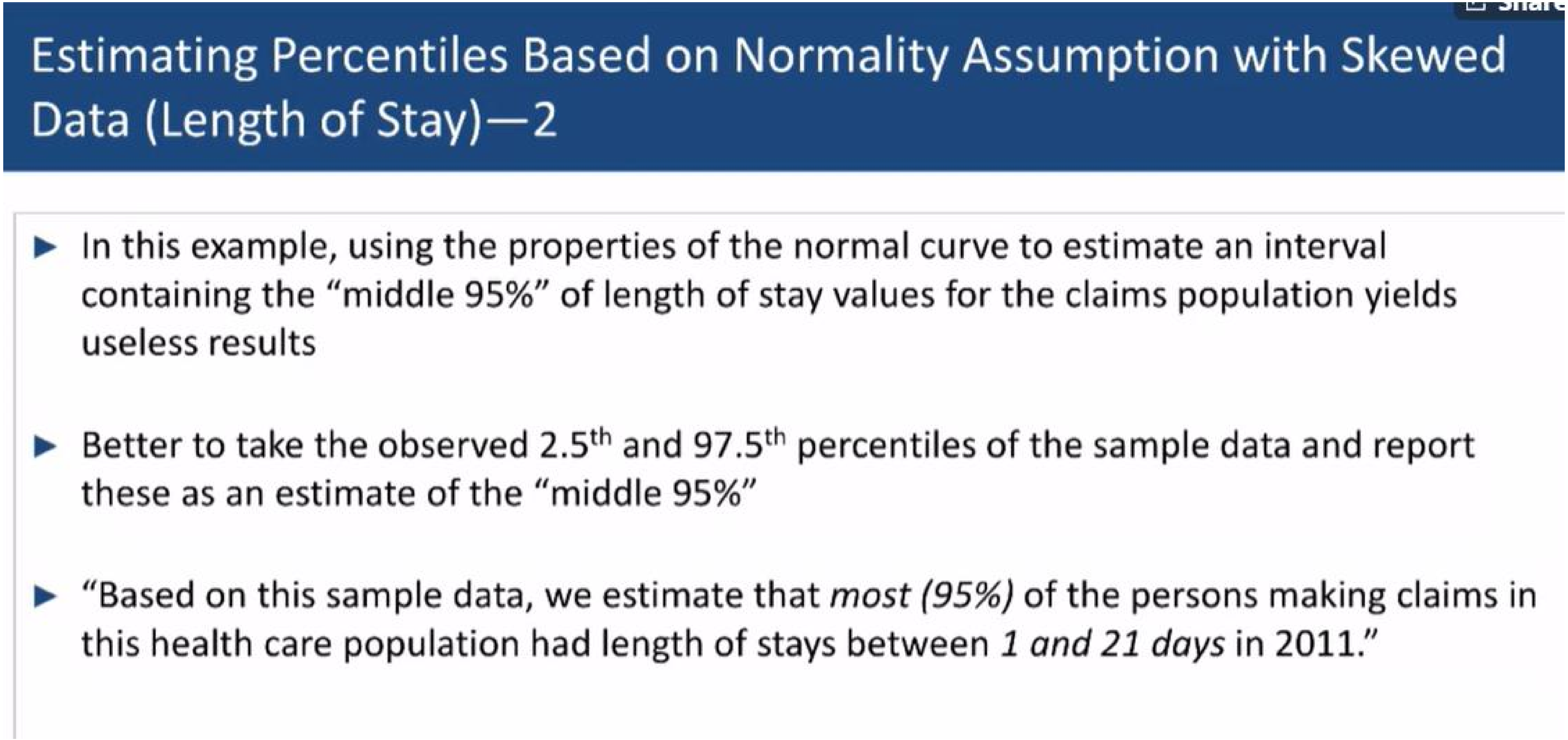
- For example, in a true normal distribution, the tails go on to negative and positive infinity, respectively



Applying the Normal Distribution Properties to Sample Data

- 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

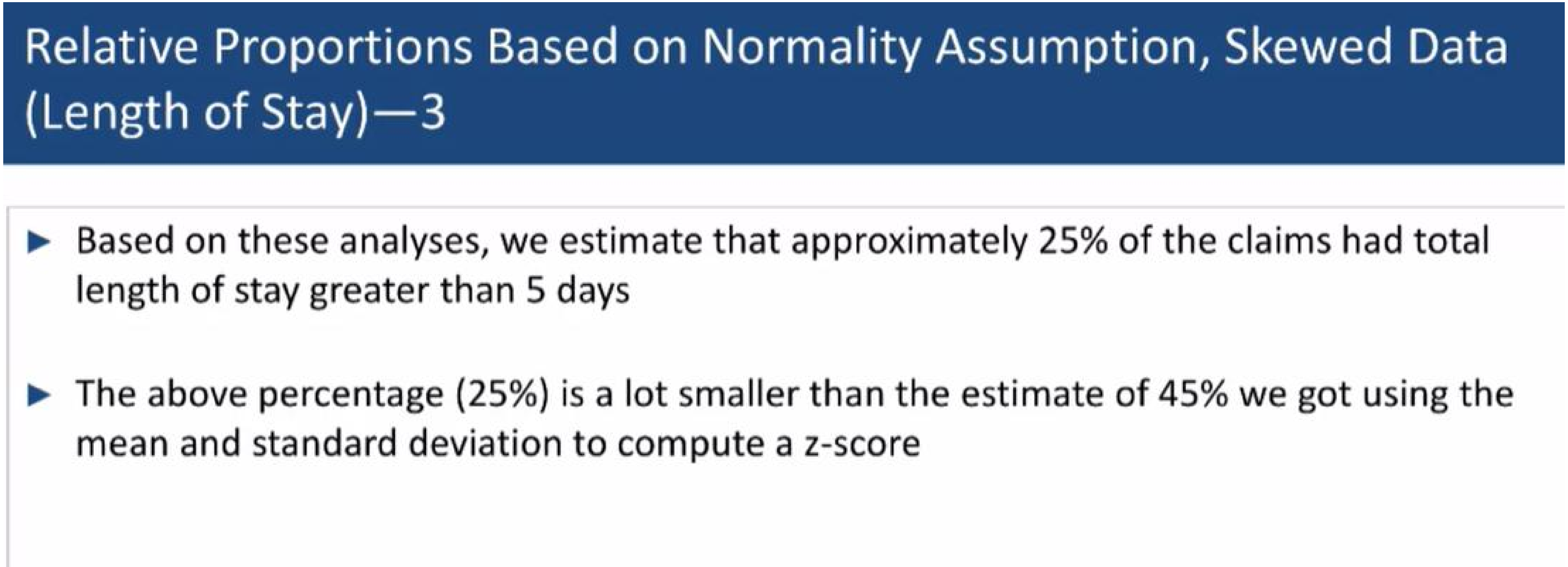


Estimating Percentiles Based on Normality Assumption with Skewed Data (Length of Stay)—2

- 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%”

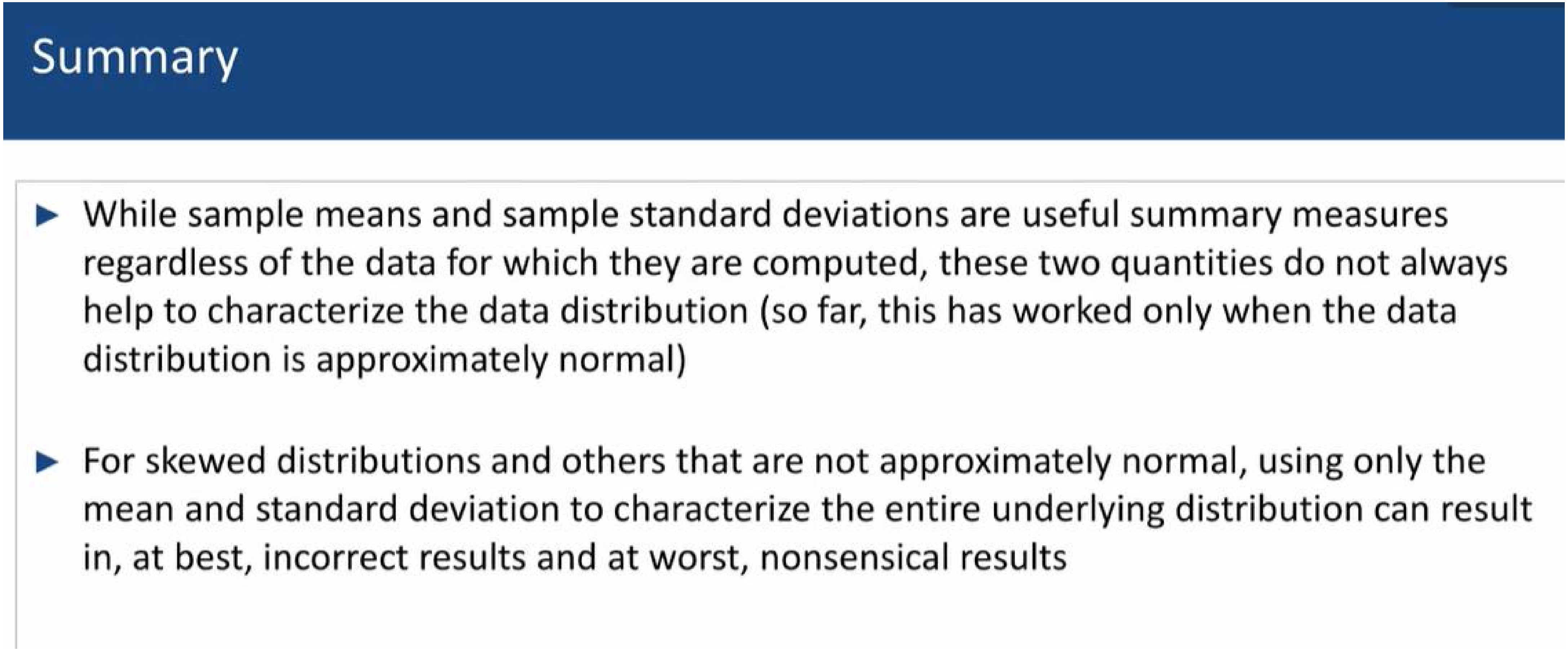
- “Based on this sample data, we estimate that *most (95%)* of the persons making claims in this health care population had length of stays between *1 and 21 days* in 2011.”



Relative Proportions Based on Normality Assumption, Skewed Data (Length of Stay)—3

- Based on these analyses, we estimate that approximately 25% of the claims had total length of stay greater than 5 days

- The above percentage (25%) is a lot smaller than the estimate of 45% we got using the mean and standard deviation to compute a z-score



Summary

- While sample means and sample standard deviations are useful summary measures regardless of the data for which they are computed, these two quantities do not always help to characterize the data distribution (so far, this has worked only when the data distribution is approximately normal)

- For skewed distributions and others that are not approximately normal, using only the mean and standard deviation to characterize the entire underlying distribution can result in, at best, incorrect results and at worst, nonsensical results