Chapter 5: The Role of Probability Pt. 2

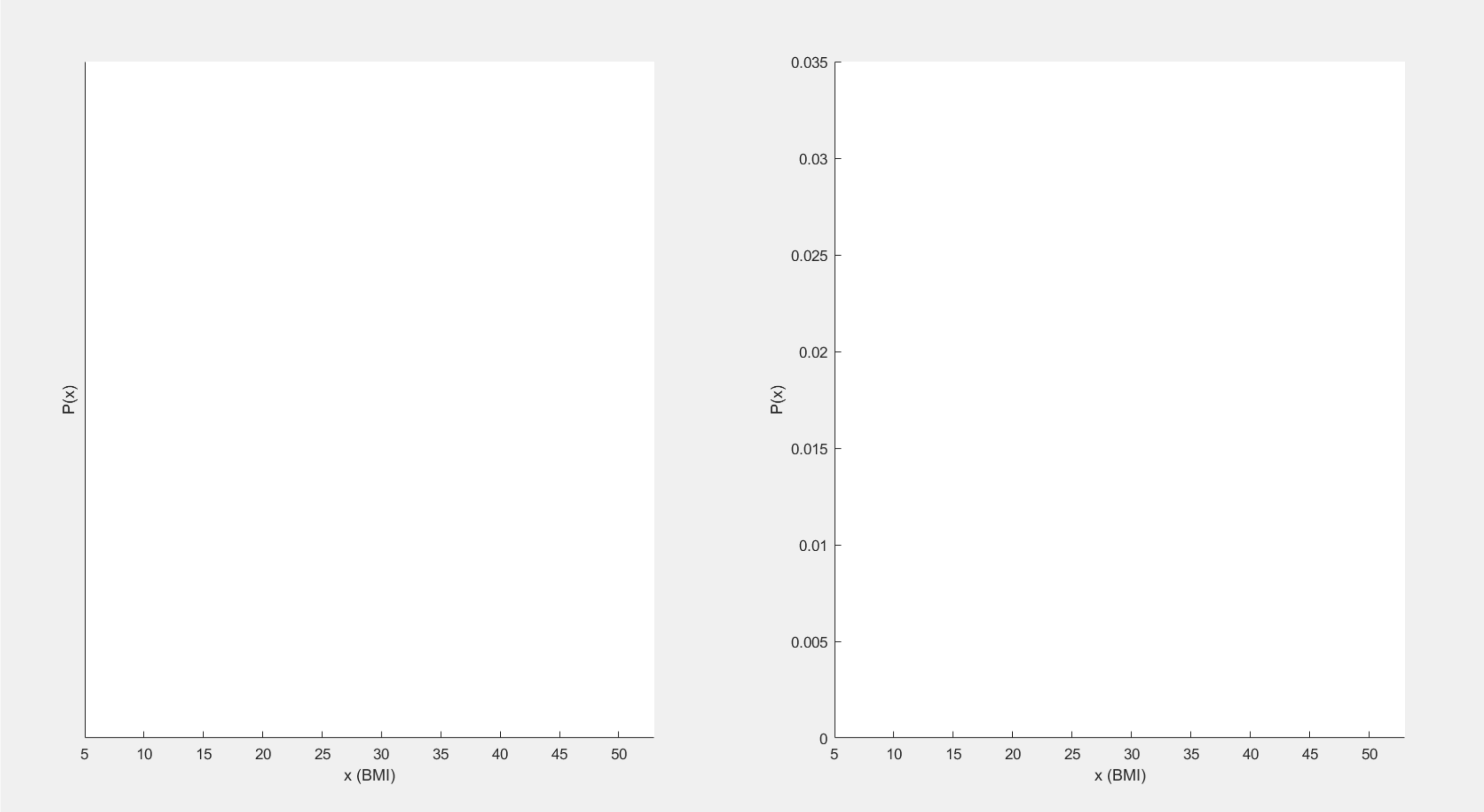
Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

5.6 Probability Models – Normal Distribution

$$BMI = 703.03 \times \frac{Weight\ in\ pounds}{(Height\ in\ inches)^2}$$

The mean BMI for males aged 60 is μ=29 kg/m^2 with standard deviation σ=6 kg/m^2 (with a normal distribution).

Demonstrate the build up of normal distribution from individual observations.



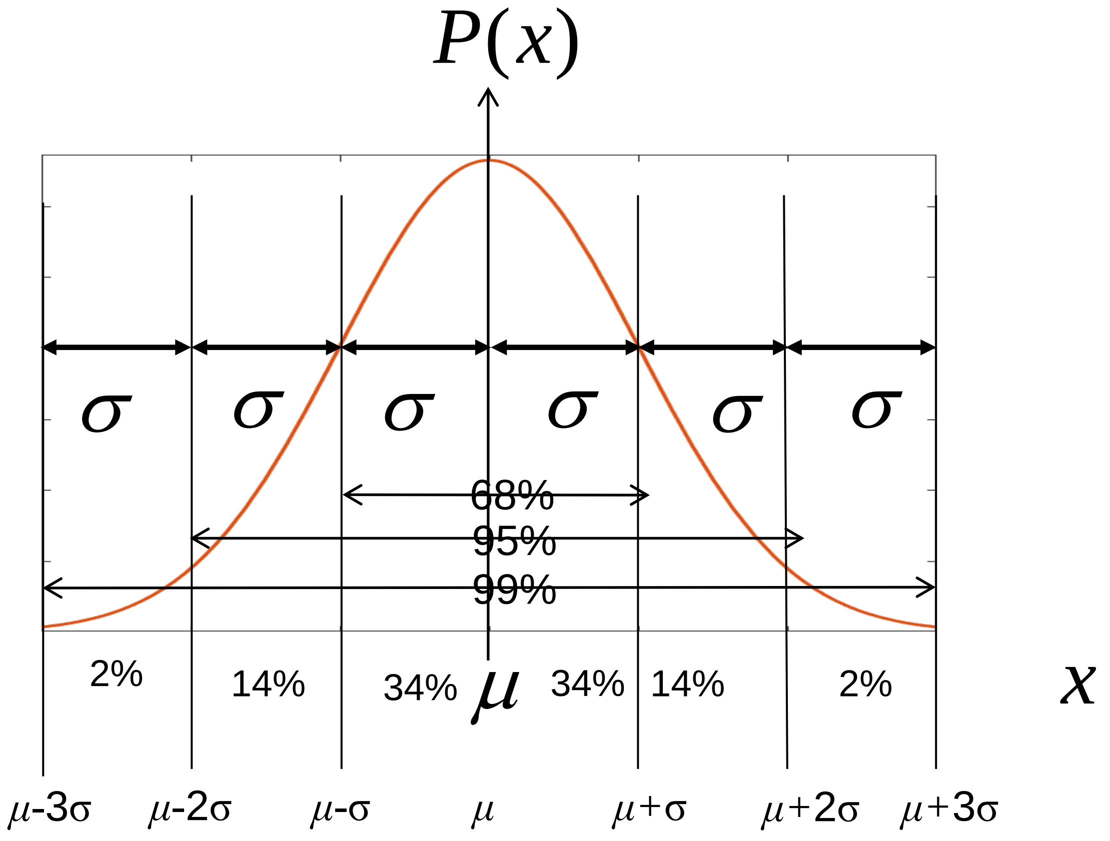
5.6 Probability Models – Normal Distribution

The normal distribution is often used for continuous outcomes.

You may know it as the bell curve or Gaussian distribution.

Its functional form is

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Total Area Under Curve = 1

Symmetric about the mean.

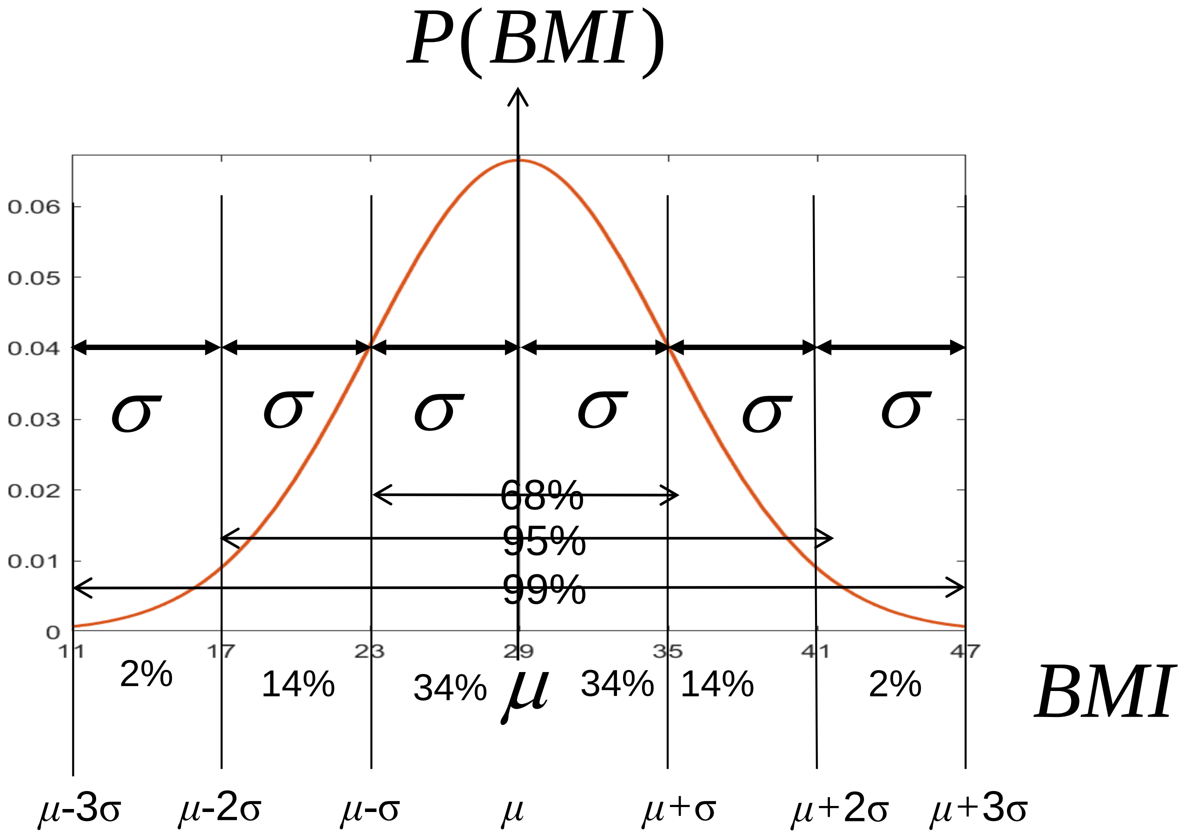
mean = median = mode.

mean μ & variance σ^2

5.6 Probability Models – Normal Distribution

The mean BMI for males aged 60 is μ=29 kg/m^2 with standard deviation σ=6 kg/m^2 (with a normal distribution). Its functional form is

$$P(x) = \frac{1}{6\sqrt{2\pi}} e^{-\frac{(x-29)^2}{2(6)^2}}$$



Symmetric about the mean.

mean = median = mode.

mean μ & variance σ^2

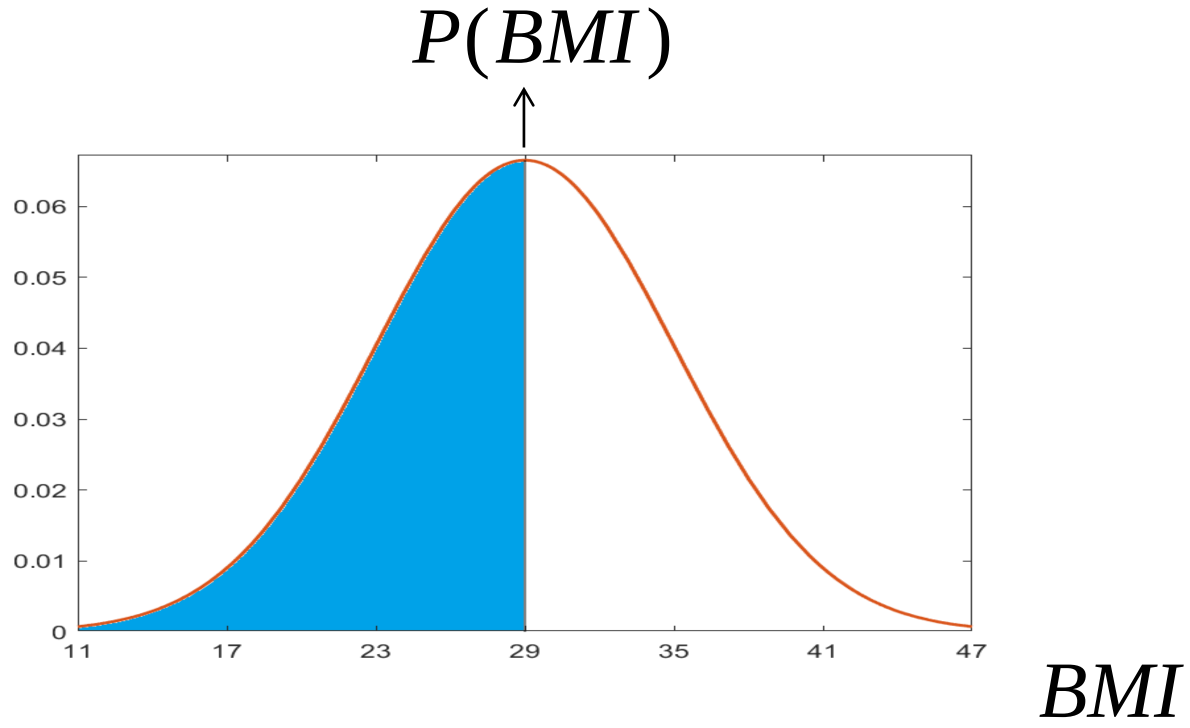
Total Area Under Curve = 1

5.6 Probability Models – Normal Distribution

The mean BMI for males aged 60 is μ=29 kg/m^2 with standard deviation σ=6 kg/m^2 (with a normal distribution).

What is the probability that a male has a BMI<29?

$P(BMI<29)=0.5$



Total Area Under Curve = 1

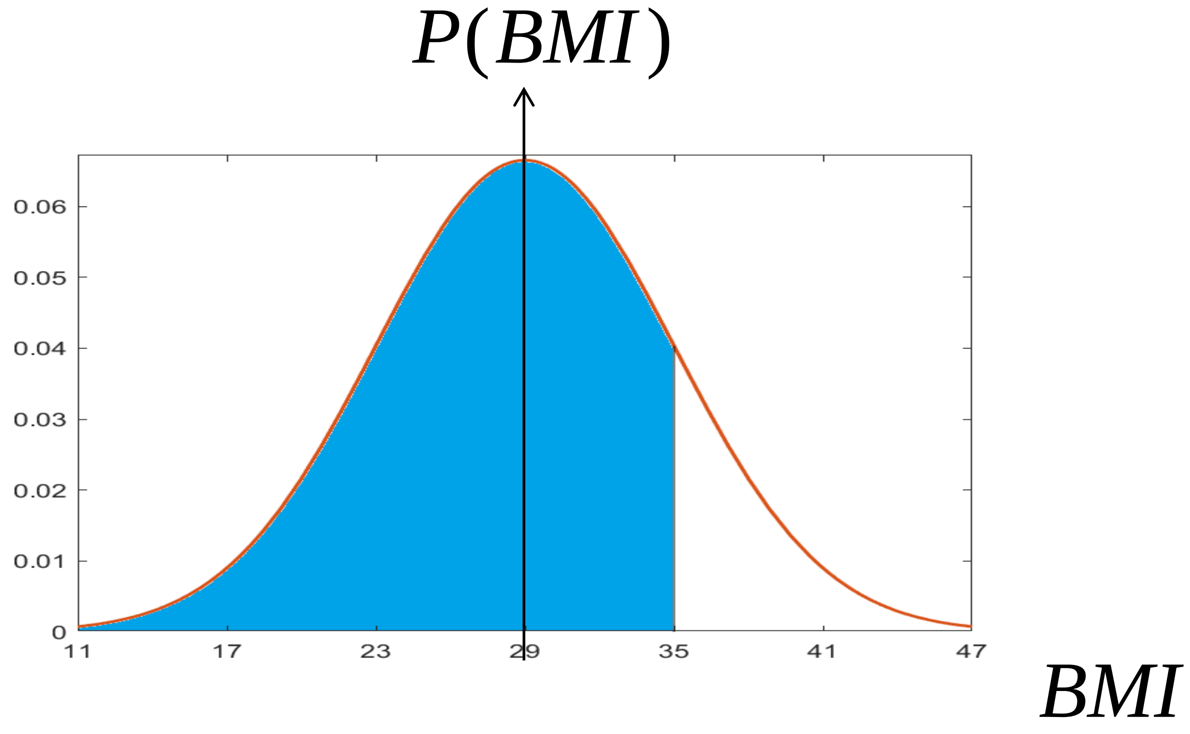
5.6 Probability Models – Normal Distribution

The mean BMI for males aged 60 is $\mu=29\ \mathrm{kg/m^2}$ with standard deviation $\sigma=6\ \mathrm{kg/m^2}$ (with a normal distribution).

What is the probability that a male has a $BMI<35$?

$P(BMI<35)=?$

Larger than 0.5 but less than 1.



Total Area Under Curve = 1

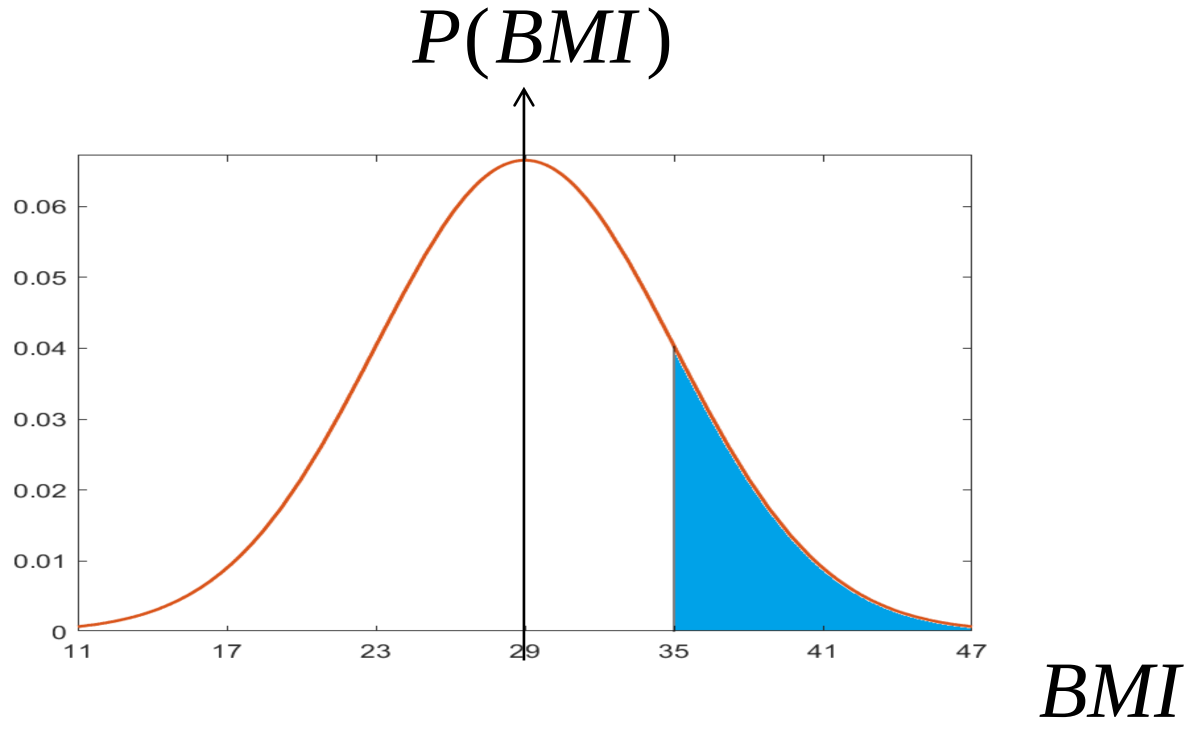
5.6 Probability Models – Normal Distribution

The mean BMI for males aged 60 is $\mu=29$ kg/m^2 with standard deviation $\sigma=6$ kg/m^2 (with a normal distribution).

What is the probability that a male has a $BMI>35$?

$P(35<BMI)=?$

Less than 0.5.



Total Area Under Curve = 1

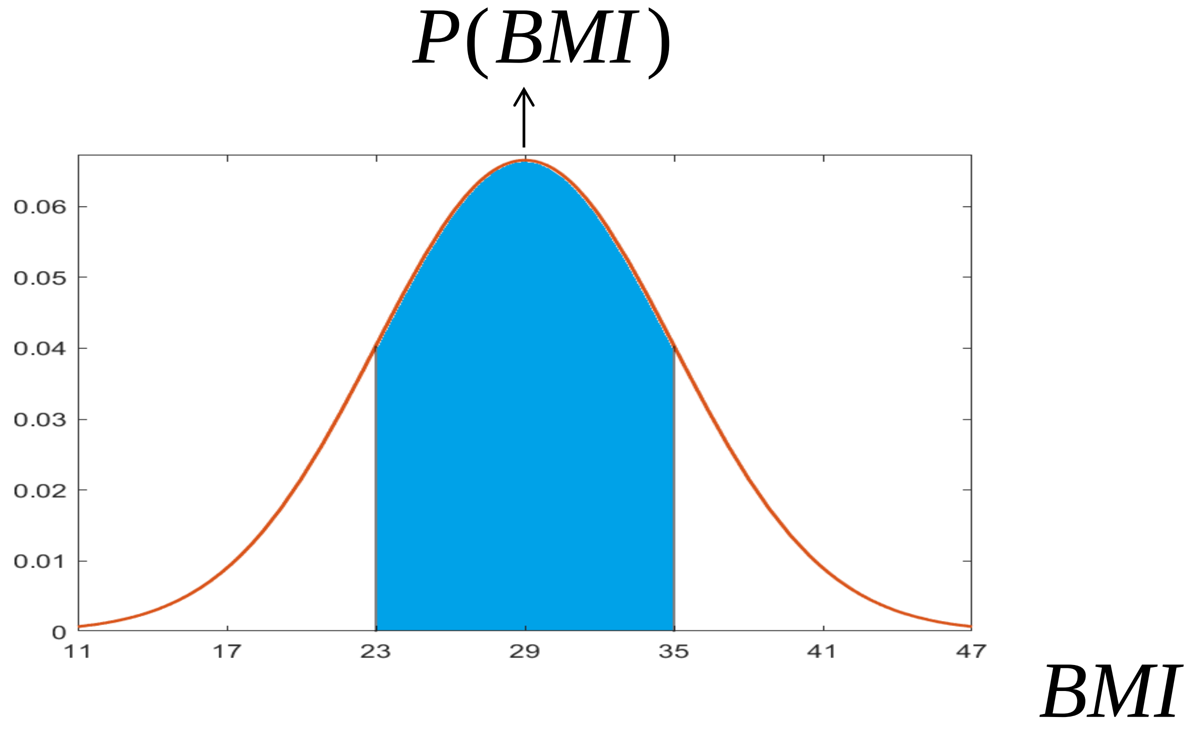
5.6 Probability Models – Normal Distribution

The mean BMI for males aged 60 is μ=29 kg/m^2 with standard deviation σ=6 kg/m^2 (with a normal distribution).

What is the probability that a male has a $23<BMI<35$?

$P(23<BMI<35)=?$

Less than 1.



Total Area Under Curve = 1

5.6 Probability Models – Normal Distribution

The mean BMI for males aged 60 is μ=29 kg/m^2 with standard deviation σ=6 kg/m^2 (with a normal distribution).

What is the probability that a male has a $23<BMI<35$?

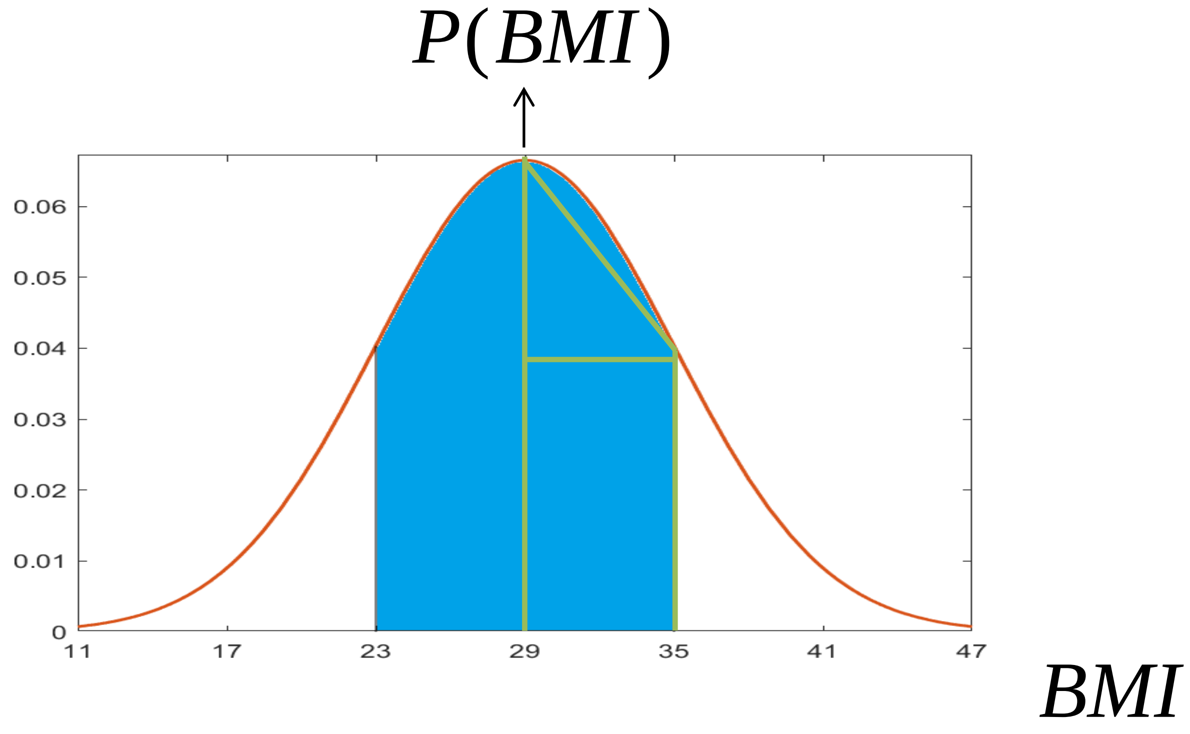
$P(23<BMI<35)=?$

Less than 1.

Twice area rectangle and triangle?

$$A \approx 2(lW + \frac{ab}{2}) = 2(0.04*6 + \frac{.0265*6}{2}) = 0.64$$

We slightly undercounted, it’s around 0.68.



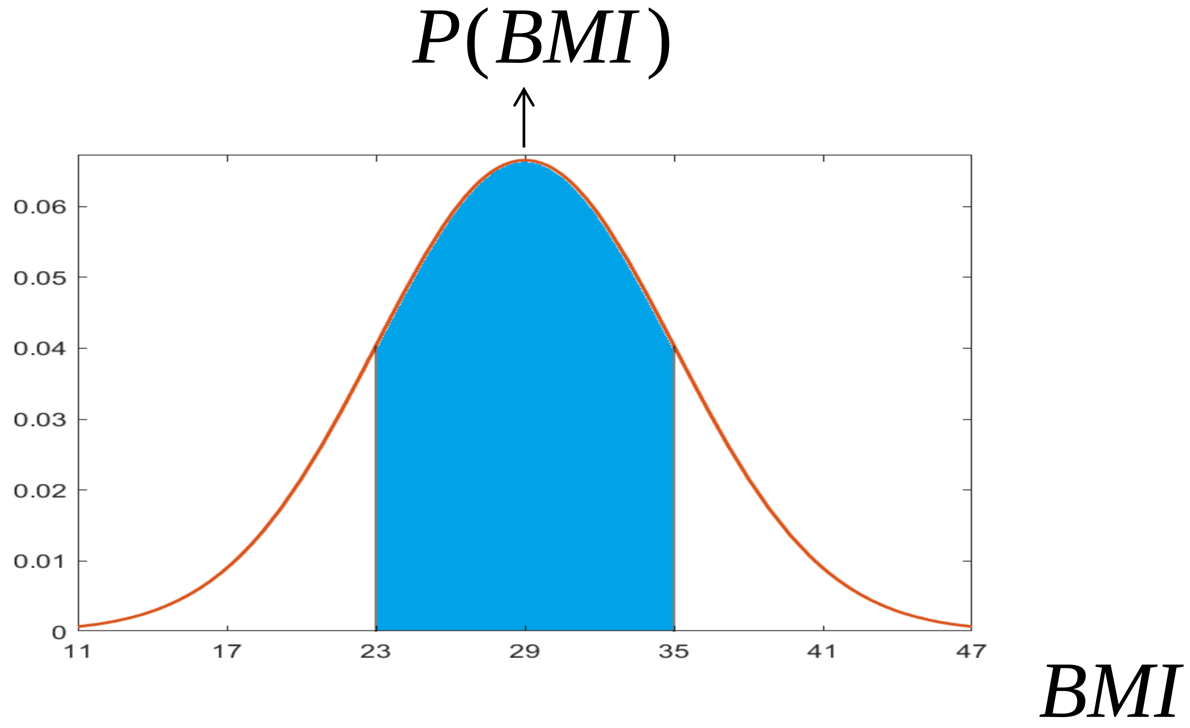
Total Area Under Curve = 1

5.6 Probability Models – Normal Distribution

The mean BMI for males aged 60 is μ=29 kg/m^2 with standard deviation σ=6 kg/m^2 (with a normal distribution).

Normally in math we do something called an integral. $x=BMI$

$$A = P(23 < x < 35) = \int_{23}^{35} \frac{1}{6\sqrt{2\pi}} e^{-\frac{(x-29)^2}{2(6)^2}} dx$$

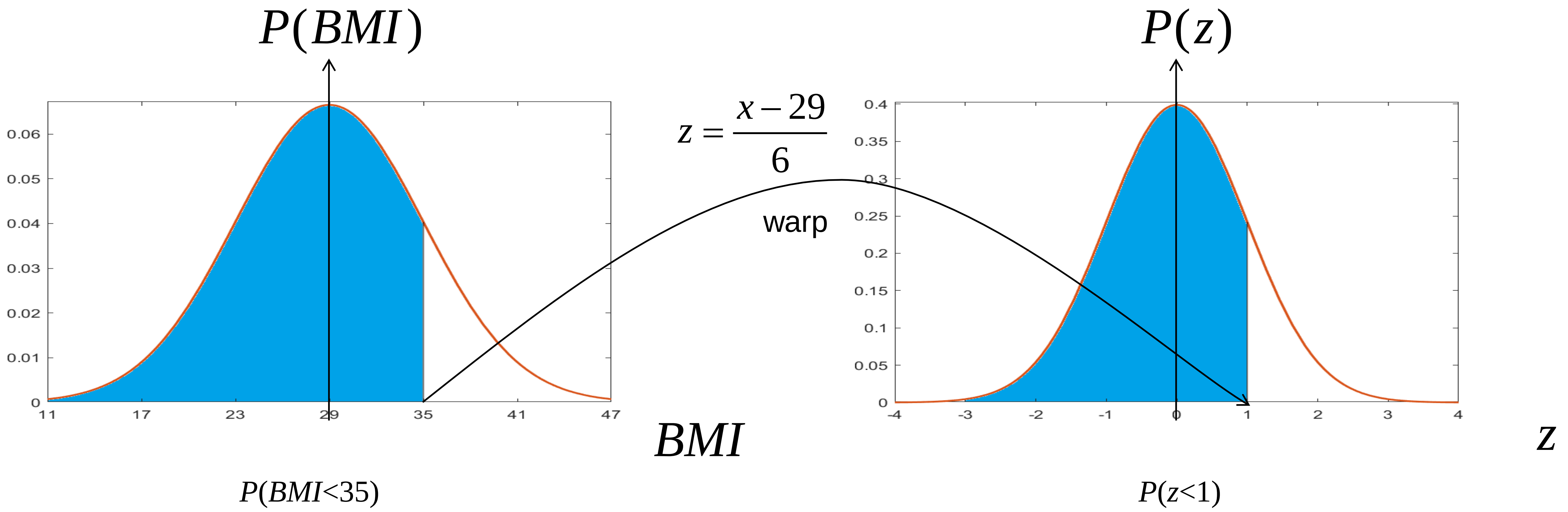


But we are not doing Calculus and even if we know Calculus, we can't integrate $P(x)$!

Total Area Under Curve = 1

5.6 Probability Models – Normal Distribution

We need to convert from the BMI (x) axis to a new "z" axis, $z = \dfrac{x-\mu}{\sigma}$.



Area under curve on z axis same as area under curve on x axis.

Total Area Under Curve = 1

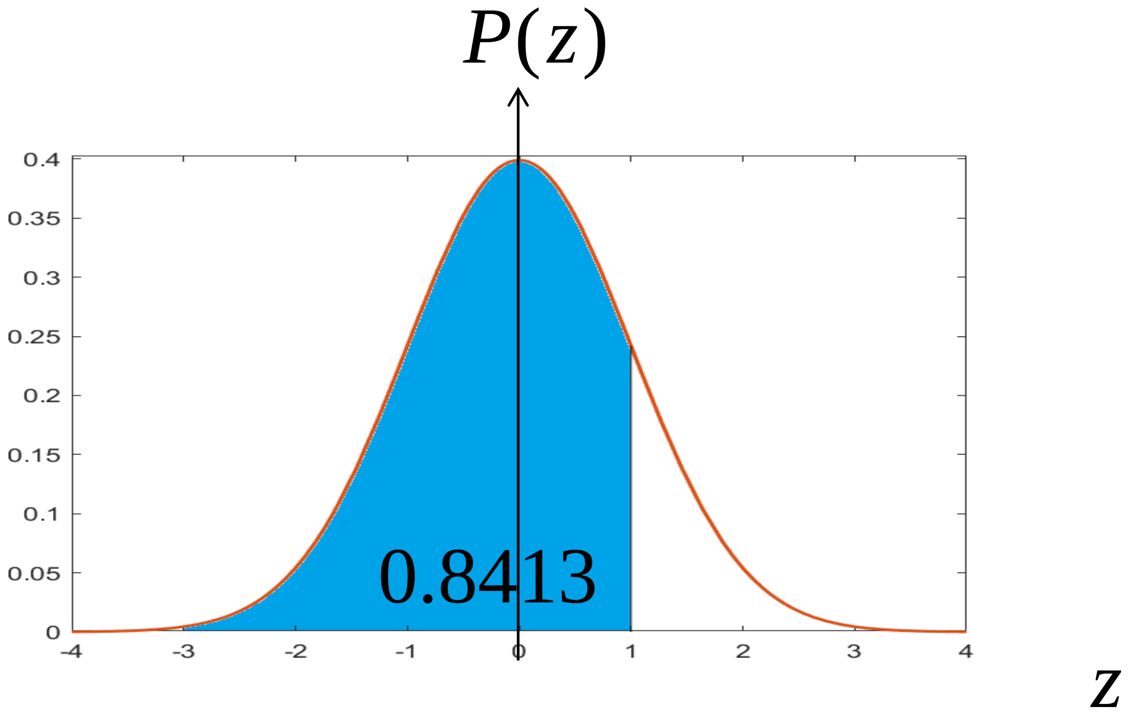
5.6 Probability Models – Normal Distribution

Now we have the z axis, we look up the area in a table.

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

z_i	.00	.01	.02	.03	.04	.05	.06	.07	.08	.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

A



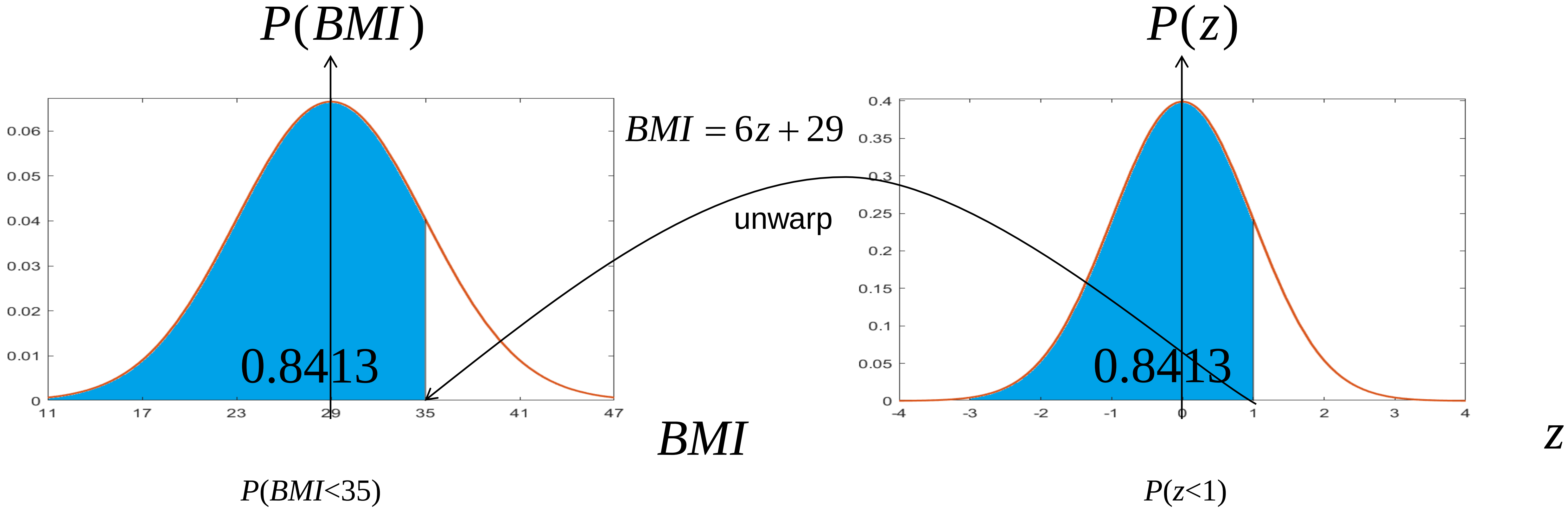
$P(z<1)$

Area under curve on z axis same as area under curve on x axis.

Total Area Under Curve = 1

5.6 Probability Models – Normal Distribution

We need to convert from the BMI (x) axis to a new "z" axis, $z = \dfrac{x-\mu}{\sigma}$.

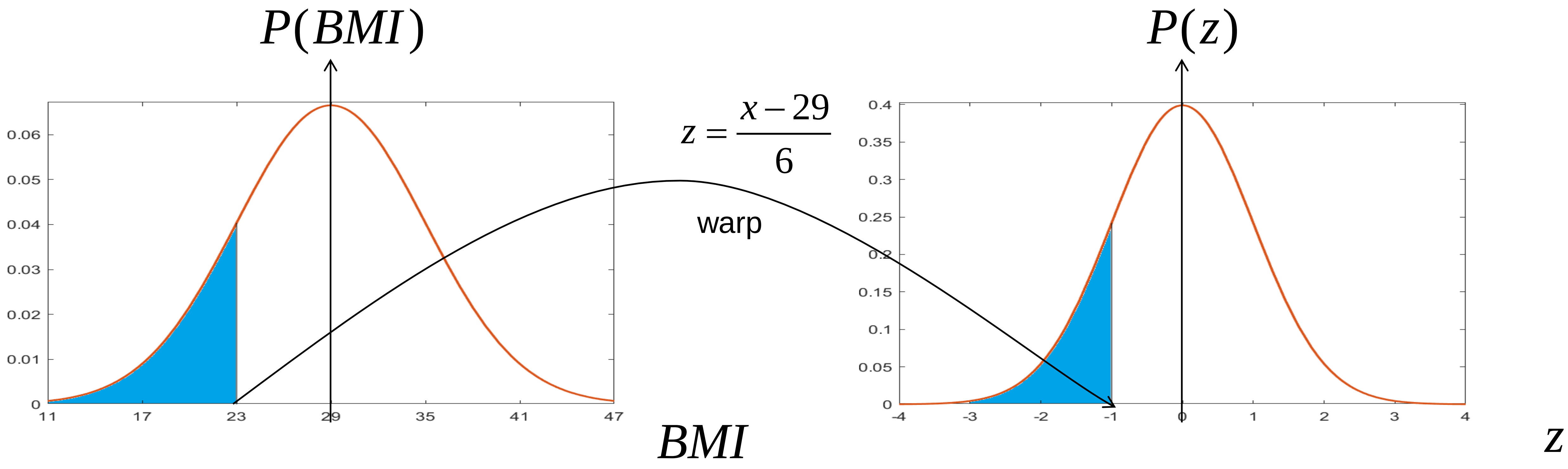


Area under curve on z axis same as area under curve on x axis.

Total Area Under Curve = 1

5.6 Probability Models – Normal Distribution

We need to convert from the BMI (x) axis to a new "z" axis, $z = \dfrac{x-\mu}{\sigma}$.



$P(BMI<23)$ $P(z<1)$

Area under curve on z axis same as area under curve on x axis.

Total Area Under Curve = 1

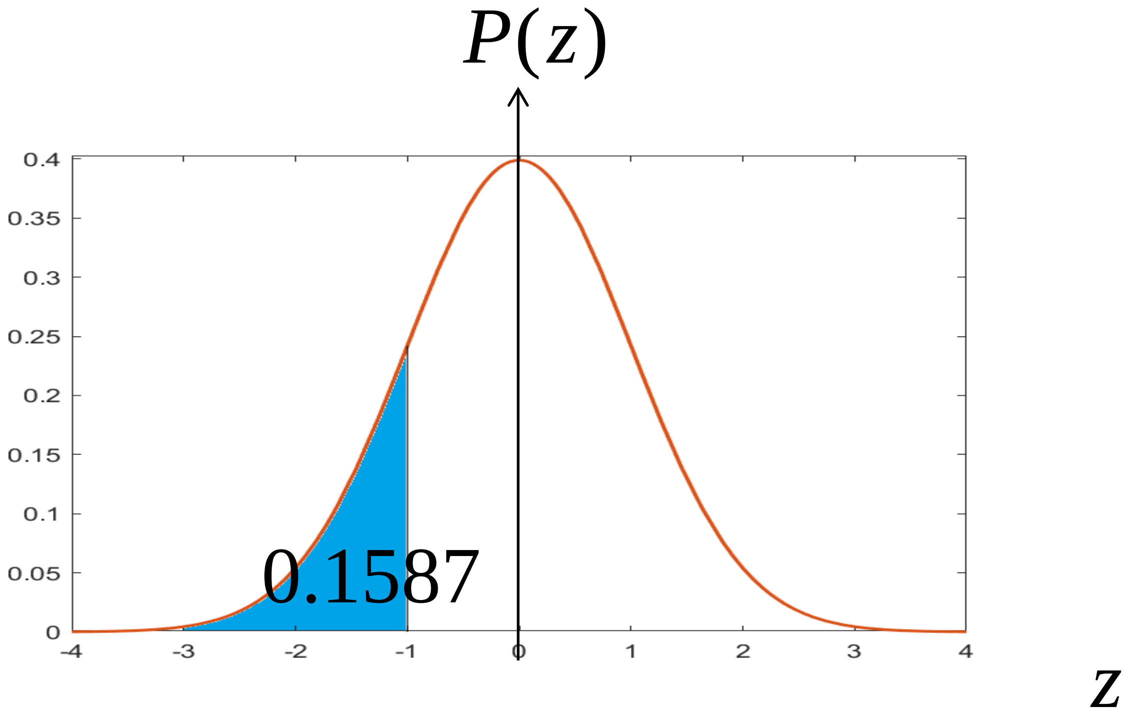
5.6 Probability Models – Normal Distribution

Now we have the z axis, we look up the area in a table.

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

A

z_i	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641



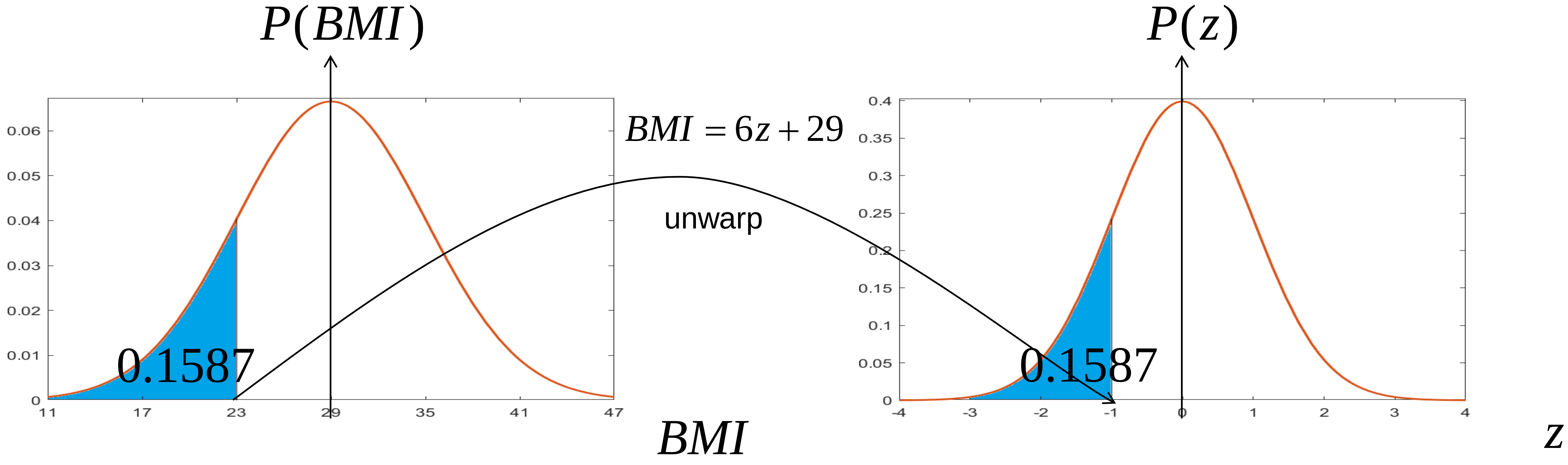
$P(z<1)$

Area under curve on z axis same as area under curve on x axis.

Total Area Under Curve = 1

5.6 Probability Models – Normal Distribution

We need to convert from the BMI (x) axis to a new "z" axis, $z = \dfrac{x-\mu}{\sigma}$.



$P(BMI<23)$

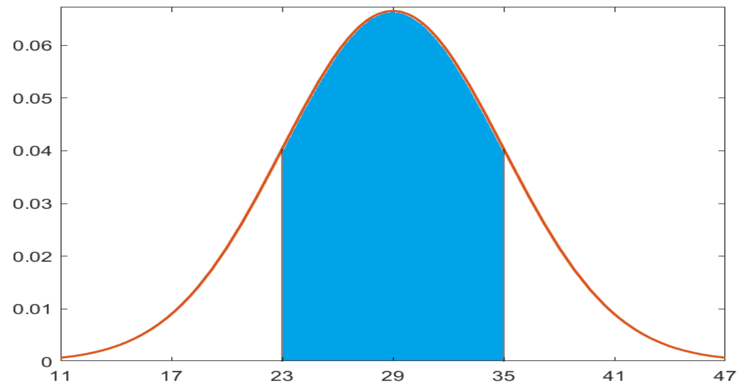
$P(z<1)$

Area under curve on z axis same as area under curve on x axis.

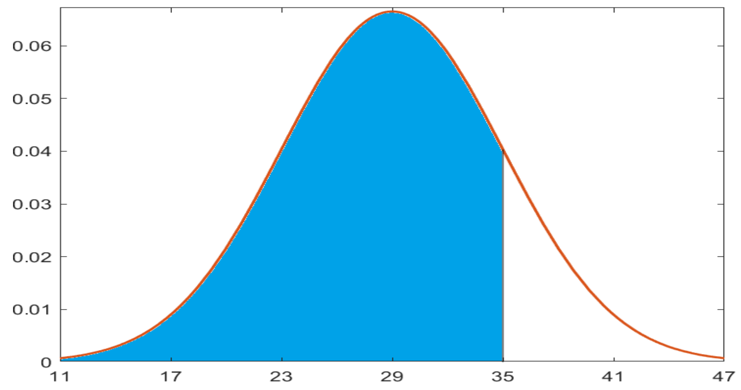
Total Area Under Curve = 1

5.6 Probability Models – Normal Distribution

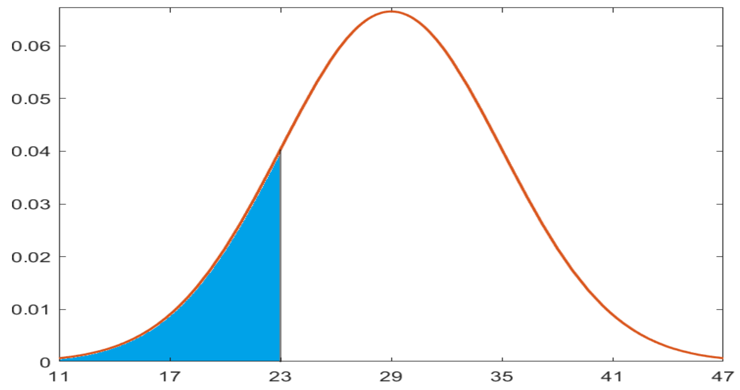
The mean BMI for males aged 60 is μ=29 kg/m^2 with standard deviation σ=6 kg/m^2 (with a normal distribution).



=



−



$0.6826 = 0.8413 - 0.1587$

5.6 Probability Models – Sampling Distributions

In science one major thing that we do is to take a random sample of data $x_1,\ldots,x_n$, and average the observations, $\bar{X}$.

Is something called the **Sampling Distribution** (of the sample means). The **Sampling Distribution** says, if we take a random sample $x_1,\ldots,x_n$, from a population with mean μ and standard deviation σ and average the observations $\bar{X}$, then $\bar{X}$ has a mean $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$. So by averaging, we've reduced our standard deviation!

5.6 Probability Models – Sampling Distributions

In science one major thing that we do is to take a random sample of data $x_1,\ldots,x_n$, and average the observations, $\bar{X}$.

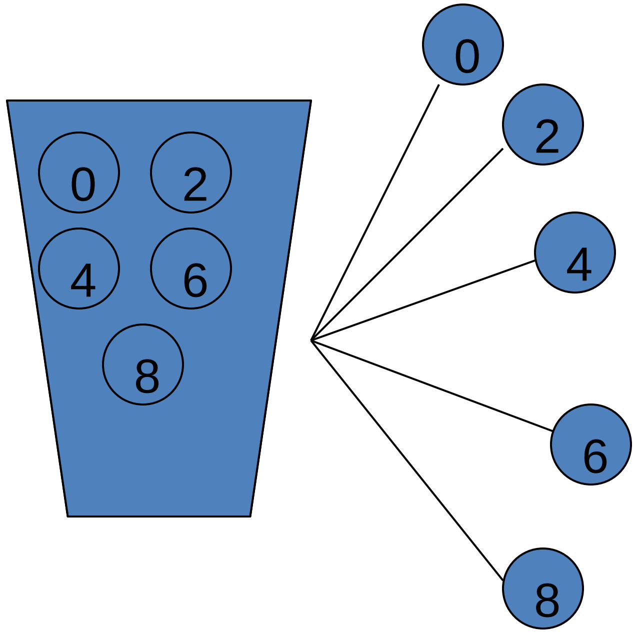
Above & beyond the Sampling Distribution is the **Central Limit Theorem**. The **Central Limit Theorem** (**CLT**) says, that if n is large, i.e. $n>30$, then $\bar{X}$ has an approximately normal distribution with mean $\mu_{\bar{X}}=\mu$ and standard deviation $\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}}$ no matter what original distribution the data $x_1,\ldots,x_n$ came from.

This is **HUGE**, meaning we can use our old friend the normal distribution.

5.6 Probability Models – Sampling Distributions

Example: N=5 balls in bucket, select n=1 with replacement.

Population data values:
0, 2, 4, 6, 8.



5 possible values

S={0, 2, 4, 6, 8}

$x = 0,$ occurs one time

$x = 2,$ occurs one time

$x = 4,$ occurs one time

$x = 6,$ occurs one time

$x = 8,$ occurs one time

Prob. of each value = 1/5 = 0.2

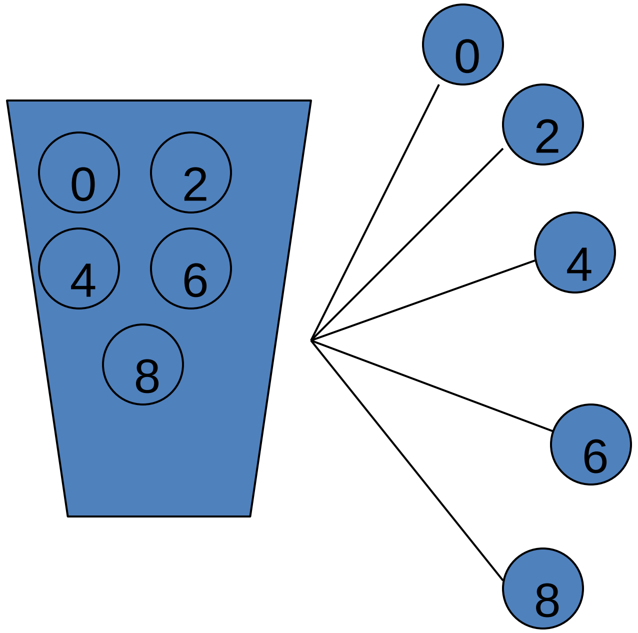
The **Sampling Distribution** says, if we take a random sample $x_1,\ldots,x_n$, from a population with mean μ and standard deviation σ and average the observations , then the average has a mean μ and standard deviation $\sigma/\sqrt{n}$. So by averaging, we've reduced our standard deviation!

The **Central Limit Theorem** (**CLT**) says, that if n is large, i.e. n>30, then the average has an approximately normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ no matter what original distribution the data $x_1,\ldots,x_n$ came from.

5.6 Probability Models – Sampling Distributions

Example: N=5 balls in bucket, select n=1 with replacement.

Population data values:
0, 2, 4, 6, 8.



5 possible values

S={0, 2, 4, 6, 8}

$$x = 0, \text{ occurs one time}$$
$$x = 2, \text{ occurs one time}$$
$$x = 4, \text{ occurs one time}$$
$$x = 6, \text{ occurs one time}$$
$$x = 8, \text{ occurs one time}$$

Prob. of each value = 1/5 = 0.2

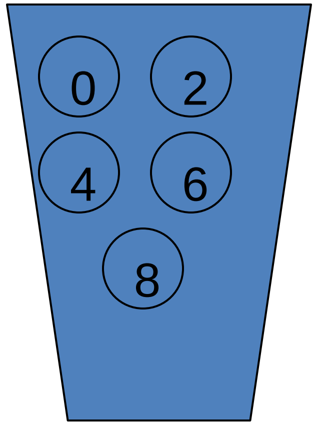
The **Sampling Distribution** says, if we take a random sample $x_1,\ldots,x_n$, from a population with mean μ and standard deviation σ and average the observations , then the average has a mean μ and standard deviation $\sigma/\sqrt{n}$. So by averaging, we've reduced our standard deviation!

The **Central Limit Theorem** (**CLT**) says, that if n is large, i.e. $n>30$, then the average has an approximately normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ no matter what original distribution the data $x_1,\ldots,x_n$ came from.

5.6 Probability Models – Sampling Distributions

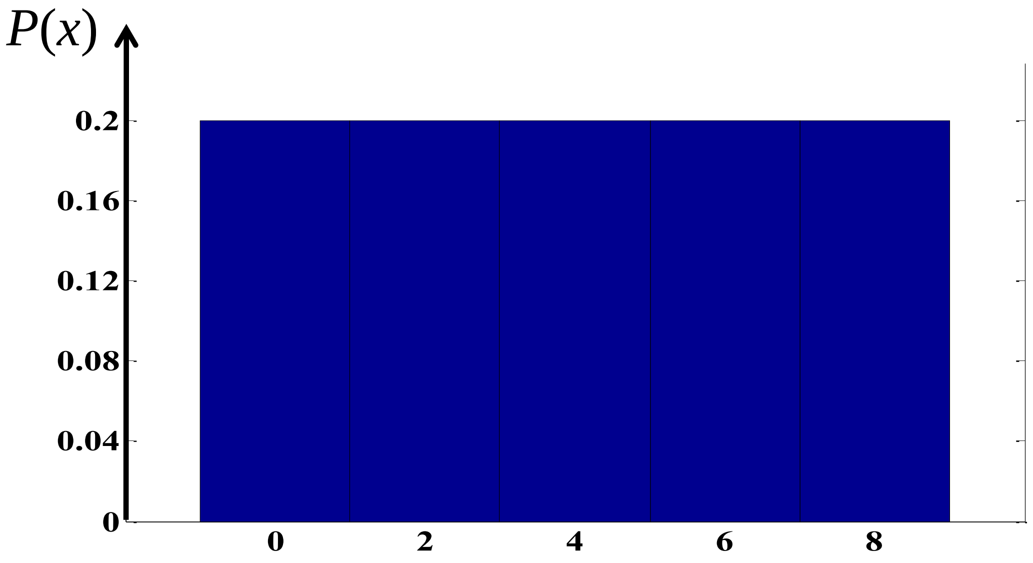
Example: N=5 balls in bucket, select n=1 with replacement.

Population data values:
0, 2, 4, 6, 8.



5 possible values

x	$P(x)$
0	1/5
2	1/5
4	1/5
6	1/5
8	1/5



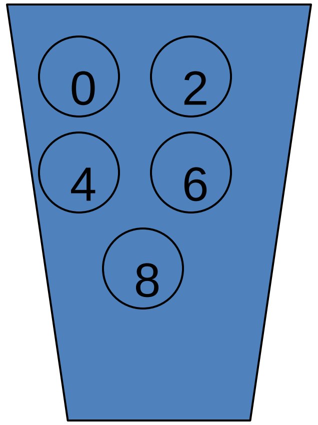
The **Sampling Distribution** says, if we take a random sample $x_1,\ldots,x_n$, from a population with mean μ and standard deviation σ and average the observations , then the average has a mean μ and standard deviation $\sigma/\sqrt{n}$. So by averaging, we've reduced our standard deviation!

The **Central Limit Theorem** (**CLT**) says, that if n is large, i.e. n>30, then the average has an approximately normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ no matter what original distribution the data $x_1,\ldots,x_n$ came from.

5.6 Probability Models – Sampling Distributions

Example: N=5 balls in bucket, select n=1 with replacement.

Population data values:
0, 2, 4, 6, 8.



5 possible values

x	$P(x)$
0	1/5
2	1/5
4	1/5
6	1/5
8	1/5

$$\mu = \sum [xP(x)]$$

$$= 0(1/5) + 2(1/5) + 4(1/5)$$
$$+ 6(1/5) + 8(1/5)$$
$$= 4$$

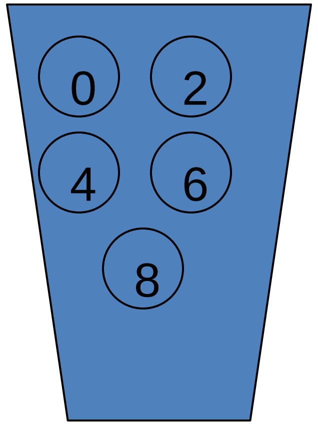
The **Sampling Distribution** says, if we take a random sample $x_1,\ldots,x_n$, from a population with mean μ and standard deviation σ and average the observations , then the average has a mean μ and standard deviation $\sigma/\sqrt{n}$. So by averaging, we've reduced our standard deviation!

The **Central Limit Theorem** (**CLT**) says, that if n is large, i.e. n>30, then the average has an approximately normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ no matter what original distribution the data $x_1,\ldots,x_n$ came from.

5.6 Probability Models – Sampling Distributions

Example: N=5 balls in bucket, select n=1 with replacement.

Population data values: 0, 2, 4, 6, 8.



5 possible values

x	$P(x)$
0	1/5
2	1/5
4	1/5
6	1/5
8	1/5

$$\sigma^2 = \sum[(x-\mu)^2 P(x)]$$

$$= (0-4)^2(1/5) + (2-4)^2(1/5)$$

$$= +(4-4)^2(1/5) + (6-4)^2(1/5)$$

$$+ (8-4)^2(1/5)$$

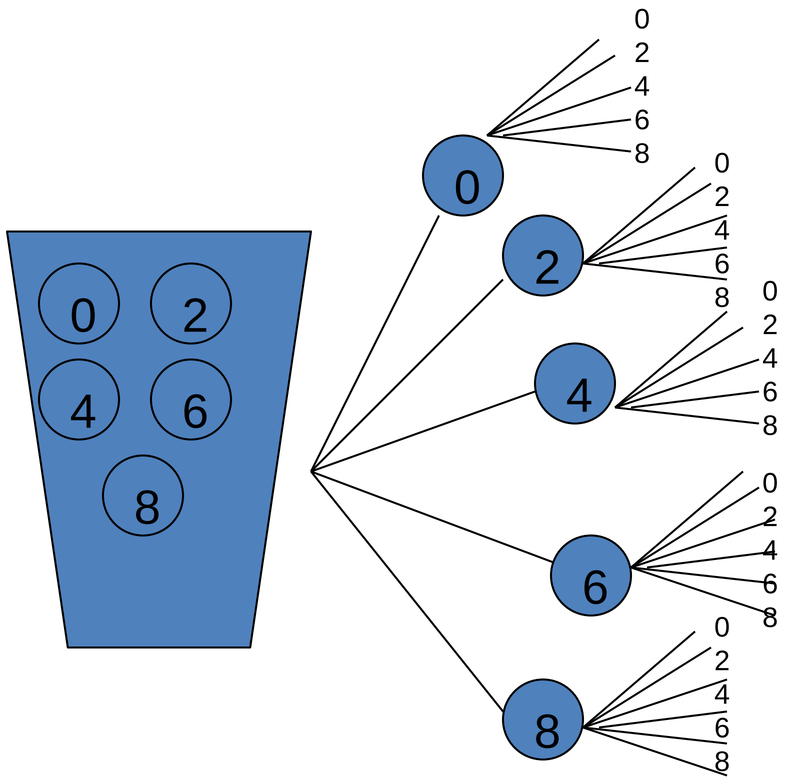
$$= 8 \quad \longrightarrow \quad \sigma = \sqrt{8} = 2\sqrt{2}$$

The **Sampling Distribution** says, if we take a random sample $x_1,\ldots,x_n$, from a population with mean μ and standard deviation σ and average the observations , then the average has a mean μ and standard deviation $\sigma/\sqrt{n}$. So by averaging, we've reduced our standard deviation!

The **Central Limit Theorem** (**CLT**) says, that if n is large, i.e. $n>30$, then the average has an approximately normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ no matter what original distribution the data $x_1,\ldots,x_n$ came from.

5.6 Probability Models – Sampling Distributions

Example: N=5 balls in bucket, select n=2 with replacement.



(0,0)	(2,0)	(4,0)	(6,0)	(8,0)
(0,2)	(2,2)	(4,2)	(6,2)	(8,2)
(0,4)	(2,4)	(4,4)	(6,4)	(8,4)
(0,6)	(2,6)	(4,6)	(6,6)	(8,6)
(0,8)	(2,8)	(4,8)	(6,8)	(8,8)

25 possible samples

The **Sampling Distribution** says, if we take a random sample $x_1,\ldots,x_n$, from a population with mean μ and standard deviation σ and average the observations , then the average has a mean μ and standard deviation $\sigma/\sqrt{n}$. So by averaging, we've reduced our standard deviation!

The **Central Limit Theorem** (**CLT**) says, that if n is large, i.e. n>30, then the average has an approximately normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ no matter what original distribution the data $x_1,\ldots,x_n$ came from.

5.6 Probability Models – Sampling Distributions

Example: N=5 balls in bucket, select n=2 with replacement.

Population data values:
0, 2, 4, 6, 8.

There are 25 possible samples.

(0,0)	(2,0)	(4,0)	(6,0)	(8,0)
(0,2)	(2,2)	(4,2)	(6,2)	(8,2)
(0,4)	(2,4)	(4,4)	(6,4)	(8,4)
(0,6)	(2,6)	(4,6)	(6,6)	(8,6)
(0,8)	(2,8)	(4,8)	(6,8)	(8,8)

Each sample has mean $\bar{X}$.

0	1	2	3	4
1	2	3	4	5
2	3	4	5	6
3	4	5	6	7
4	5	6	7	8

The **Sampling Distribution** says, if we take a random sample $x_1,\ldots,x_n$, from a population with mean μ and standard deviation σ and average the observations , then the average has a mean μ and standard deviation $\sigma/\sqrt{n}$. So by averaging, we've reduced our standard deviation!

The **Central Limit Theorem** (**CLT**) says, that if n is large, i.e. n>30, then the average has an approximately normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ no matter what original distribution the data $x_1,\ldots,x_n$ came from.

5.6 Probability Models – Sampling Distributions

Example: N=5 balls in bucket, select n=2 with replacement.

Population data values:
0, 2, 4, 6, 8.

There are 25 possible samples.

0	1	2	3	4
1	2	3	4	5
2	3	4	5	6
3	4	5	6	7
4	5	6	7	8

$\bar{x}=0,$ occurs one time

$\bar{x}=1,$ occurs two times

$\bar{x}=2,$ occurs three times

$\bar{x}=3,$ occurs four times

$\bar{x}=4,$ occurs five times

$\bar{x}=5,$ occurs four times

$\bar{x}=6,$ occurs three times

$\bar{x}=7,$ occurs two times

$\bar{x}=8,$ occurs one time

Prob. of each samples
mean = 1/25 = 0.04

The **Sampling Distribution** says, if we take a random sample $x_1,\ldots,x_n$, from a population with mean μ and standard deviation σ and average the observations , then the average has a mean μ and standard deviation $\sigma/\sqrt{n}$. So by averaging, we've reduced our standard deviation!

The **Central Limit Theorem** (**CLT**) says, that if n is large, i.e. n>30, then the average has an approximately normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ no matter what original distribution the data $x_1,\ldots,x_n$ came from.

5.6 Probability Models – Sampling Distributions

Example: N=5 balls in bucket, select n=2 with replacement.

Population data values:
0, 2, 4, 6, 8.

There are 25 possible samples.

0	1	2	3	4
1	2	3	4	5
2	3	4	5	6
3	4	5	6	7
4	5	6	7	8

$$P(\bar{x}=0)=1/25$$
$$P(\bar{x}=1)=2/25$$
$$P(\bar{x}=2)=3/25$$
$$P(\bar{x}=3)=4/25$$
$$P(\bar{x}=4)=5/25$$
$$P(\bar{x}=5)=4/25$$
$$P(\bar{x}=6)=3/25$$
$$P(\bar{x}=7)=2/25$$
$$P(\bar{x}=8)=1/25$$

Prob. of each samples
mean = 1/25 = 0.04

The **Sampling Distribution** says, if we take a random sample $x_1,\ldots,x_n$, from a population with mean μ and standard deviation σ and average the observations , then the average has a mean μ and standard deviation $\sigma/\sqrt{n}$. So by averaging, we've reduced our standard deviation!

The **Central Limit Theorem** (**CLT**) says, that if n is large, i.e. $n>30$, then the average has an approximately normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ no matter what original distribution the data $x_1,\ldots,x_n$ came from.

5.6 Probability Models – Sampling Distributions

Example: N=5 balls in bucket, select n=2 with replacement.

$$P(\bar{x}=0)=1/25$$
$$P(\bar{x}=1)=2/25$$
$$P(\bar{x}=2)=3/25$$
$$P(\bar{x}=3)=4/25$$
$$P(\bar{x}=4)=5/25$$
$$P(\bar{x}=5)=4/25$$
$$P(\bar{x}=6)=3/25$$
$$P(\bar{x}=7)=2/25$$
$$P(\bar{x}=8)=1/25$$

Represent this probability distribution with a histogram.

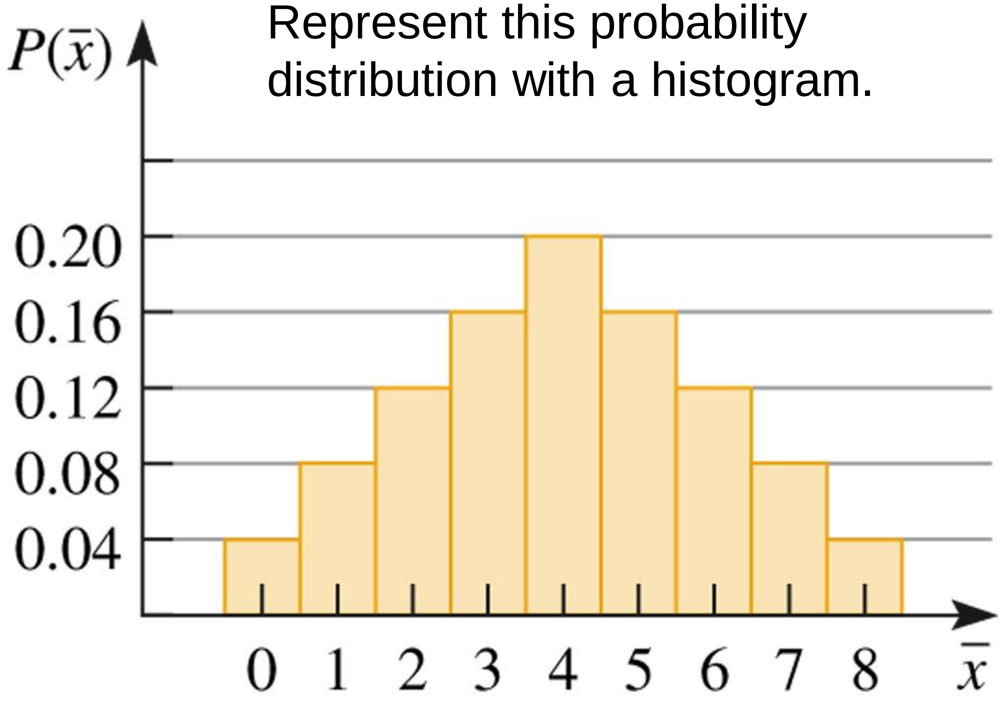


Figure from Johnson & Kuby, 2012.

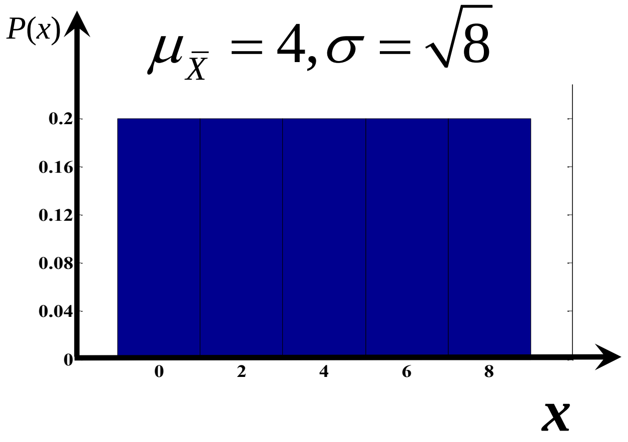
The **Sampling Distribution** says, if we take a random sample $x_1,\ldots,x_n$, from a population with mean μ and standard deviation σ and average the observations , then the average has a mean μ and standard deviation $\sigma/\sqrt{n}$. So by averaging, we've reduced our standard deviation!

The **Central Limit Theorem** (**CLT**) says, that if n is large, i.e. $n>30$, then the average has an approximately normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ no matter what original distribution the data $x_1,\ldots,x_n$ came from.

5.6 Probability Models – Sampling Distributions

Example: N=5 balls in bucket, selecting increasing n with replacement.

$\bar{X}$ from n=1 distribution



$\bar{X}$ from n=2 distribution

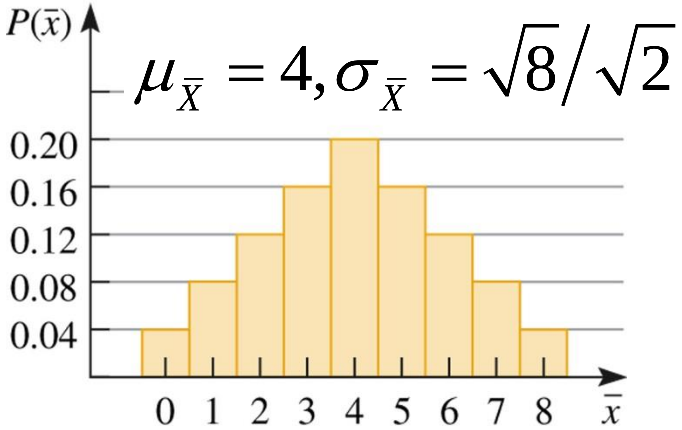
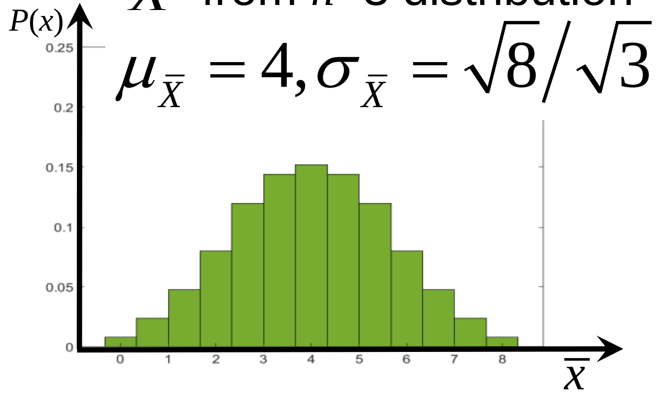
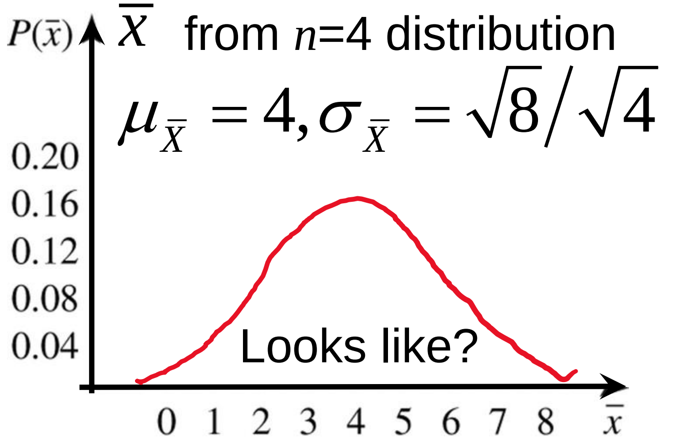


Figure from Johnson & Kuby, 2012.

$\bar{X}$ from n=3 distribution



$\bar{X}$ from n=4 distribution



n large? $\longrightarrow$

$\mu_{\bar{X}} = 4$

$\sigma_{\bar{X}} = \sqrt{8}/\sqrt{n}$

Looks like?

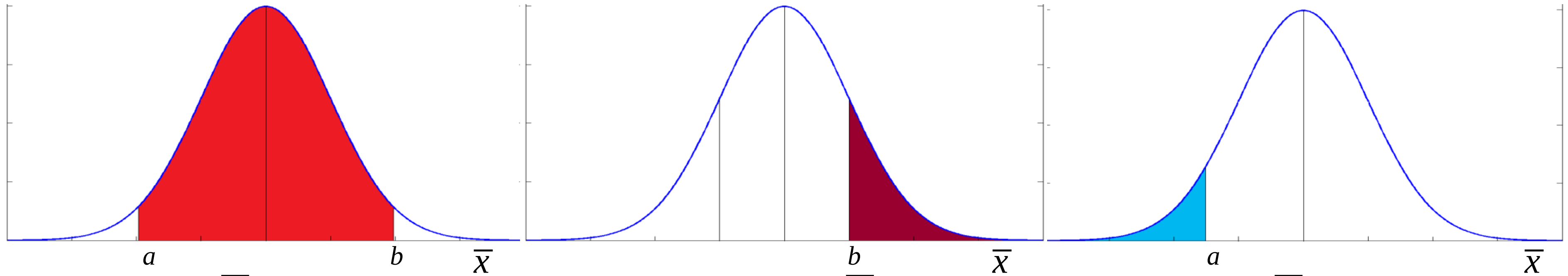
The **Sampling Distribution** says, if we take a random sample $x_1,\ldots,x_n$, from a population with mean μ and standard deviation σ and average the observations , then the average has a mean μ and standard deviation $\sigma/\sqrt{n}$. So by averaging, we've reduced our standard deviation!

The **Central Limit Theorem** (**CLT**) says, that if n is large, i.e. n>30, then the average has an approximately normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ no matter what original distribution the data $x_1,\ldots,x_n$ came from.

5.6 Probability Models – Normal Distribution

Now that we know that $\bar{X}$ has a normal distribution when n is large, we can find probabilities (areas) for finding a random mean by converting to a z and using the tables.

$$x_1,\ldots,x_n$$
$$\mu_{\bar{X}} = \mu$$
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$



$P(a < \bar{X} < b)$ ← same area → $P(c < z < d)$

$P(b < \bar{X})$ ← same area → $P(d < z)$

$P(\bar{X} < a)$ ← same area → $P(z < c)$

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \quad c = \frac{a - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \quad d = \frac{b - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

Example:
What is probability that sample mean $\bar{X}$ from a random sample of n=16 heights is greater than 70″ when μ =67 and σ =4?

Questions?

Homework 5 Part II

Read Chapter 5.

Problems # 10, 19