→ rescaling converting units

Scaling Data: multiplying (or dividing) every data value by a constant.

Example: $\{1, 2, 3, 4\}$ $\xrightarrow{5y}$ $\{5, 10, 15, 20\}$

When we multiply (or divide) every value in a data set a constant c,

a) the measures of **position** are multiplied (or divided) by c. $c \geq 0$

- mean
- median
- Q_1
- Q_3
- min
- max

all percentiles

$c \geq 0$

min < Q_1 < Q_3 < max

For $c < 0$,

$c(\max) < cQ_3 < cQ_1 < c(\min)$

→ reverses order

b) the measures of **spread** are multiplied (or divided) by c. → $c \geq 0$

- standard deviation
- IQR
- range

Variance multiply by c^2

For $c < 0$, multiply by $|c|$

c) the overall shape does not change. → $c > 0$

Example: $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

rounded

	Mean	Variance	Std. dev.	Median	Range	Min	Max	Q_1	Q_3	IQR
y	5.5	9.17	3.03	5.5	9	1	10	3	8	5
$3y$	16.5	82.5	9.08	16.5	27	3	30	9	24	15
$-2y$	-11	36.67	6.06	-11	18	-20	-2	-16	-6	10

c^2

$\{-20, -18, -16, -14, -12, -10, -8, -6, -4, -2\}$

Example: $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

	Mean	Variance	Std. dev.	Median	Range	Min	Max	Q_1	Q_3	IQR
y	5.5	9.17	3.03	5.5	9	1	10	3	8	5
$2y-1$	10	36.67	6.06	10	18	1	19	5	15	10

$\times 2^2$

original
$\bar{y} = 5.5$
$s = 3.03$

$\xrightarrow{\text{scale by } 2}$

$\bar{y} = 11$
$s = 6.06$

$\xrightarrow{\text{shift by } -1}$

$\bar{y} = 10$
$s = 6.06$

Example: For our Stat 151 class, the mean height is 168.84 cm and the standard deviation is 9.88 cm. What are the mean and standard deviation in inches? 1 cm = 0.39 in $\quad h_{in} = 0.39\, h_{cm}$

height in cm
$\bar{y} = 168.84$ cm
$s = 9.88$ cm

$\xrightarrow{\text{scale by } 0.39}$

height in inches
$\bar{y} = 65.85$ in
$s = 3.85$ in

Example: The mean temperature in a US city on January 1, 2020 was 25 ° F with a standard deviation of 5 ° F. What are the mean and standard deviation in ° C? $\quad {}^\circ C = \dfrac{{}^\circ F - 32}{1.8}$

temp in °F
$\bar{y} = 25\,^\circ F$
$s = 5\,^\circ F$

$\xrightarrow{\text{shift by } -32}$

$\bar{y} = -7$
$s = 5$

$\xrightarrow{\text{scale by } \frac{1}{1.8}}$

temp in °C
$\bar{y} = -3.89\,^\circ C$
$s = 2.78\,^\circ C$

Section 5.1: Standardizing with z-Scores

We can **standardize** a value in a data set by expressing its distance from the mean in standard deviations.

unit

These standardized values are called ***z*-scores**.

***z*-score:** the number of standard deviations above or below the mean that a data value lies.

The z-score of a value y in a data set (sample) with mean $\bar{y}$ and standard deviation s is

$$z = \frac{y - \bar{y}}{s}$$

$z > 0 \Rightarrow$ above mean

$z = 0 \Rightarrow$ equal mean

$z < 0 \Rightarrow$ below mean

- gives a measure of the relative standing of a value in a data set.
- gives us a way to compare data values which have been measured on different scales, with different units and magnitudes.

Example:

A data set has a mean of $\bar{y} = 60$ and a standard deviation of $s = 10$.

a) What is the z-score for $y = 65$?

$$z = \frac{65 - 60}{10} = \frac{5}{10} = 0.5$$

b) What is the z-score for $y = 91$?

$$z = \frac{91 - 60}{10} = \frac{31}{10} = 3.1 \leftarrow \text{outlier}$$

Example:

$n = 38$

At the 2012 Olympic games, Canadian Jessica Zelinka placed 7th in the women's heptathlon. Jessica's time in the 100 m hurdles was 12.65 seconds and her distance in the shot put was 14.81 m.

$z < 0$ better →

Event	Mean	Standard Deviation
100 m Hurdles	13.55 seconds	0.47 seconds
Shot Put	13.63 m	1.15 m

← $z > 0$ better

a) In which event did Jessica perform better?

100 m Hurdles: $z = \frac{12.65 - 13.55}{0.47} = -1.91$

on "better" side of mean in both events

Shot put: $z = \frac{14.81 - 13.63}{1.15} = 1.03$

Jessica's 100m hurdles is better since her 100m hurdles time lies farther away from the 100m hurdle mean than her shot put distance lies from the shot put mean.

b) Canadian Brianne Theisen placed 11th in the event. Her z-score in the 100 m hurdles was $z = -0.53$ and her z-score in the shot put was $z = -0.64$.

$z = \frac{y - \bar{y}}{s}$

i) What was her time in the 100 m hurdles?

$y = 0.47(-0.53) + 13.55 = 13.3$ seconds

$sz = y - \bar{y}$

$y = sz + \bar{y}$

ii) What was her distance in the shot put?

$y = 1.15(-0.64) + 13.63 = 12.89$ m

If you know the z-score of a value y in a data set with mean $\bar{y}$ and standard deviation s, then you can find the data value using the formula:

$$y = sz + \bar{y}$$

Distribution of z-Scores

Converting data values into z-scores is shifting them by the mean and then rescaling them by the standard deviation. z - distribution

Standardizing a data set into z-scores:

- changes the center (mean).
- changes the spread (standard deviation).
- does **not** change the **shape** of the distribution of a variable.

$$z = \frac{y - \bar{y}}{s}$$

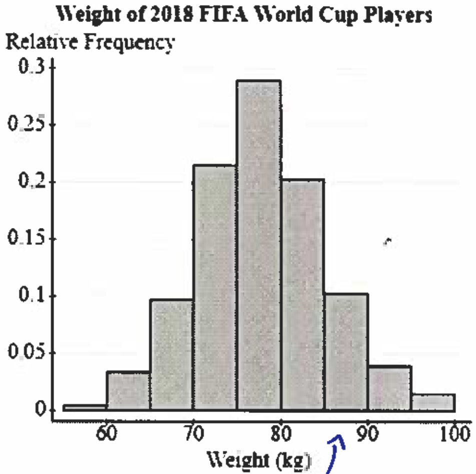
The z-distribution has:

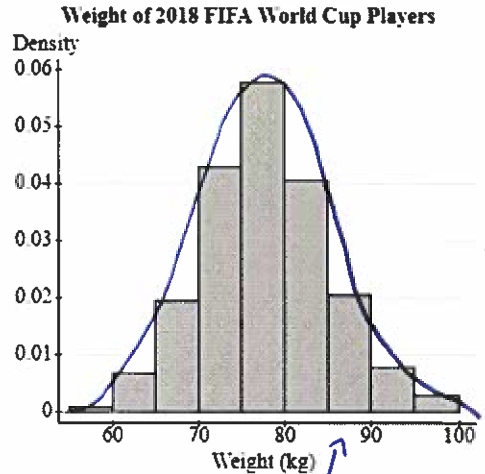
- mean 0
- standard deviation 1

$$\begin{matrix} \bar{y} = \bar{y}_0 \\ s = s_0 \end{matrix} \xrightarrow[\text{by } -\bar{y}_0]{\text{shift}} \begin{matrix} \bar{y} = 0 \\ s = s_0 \end{matrix} \xrightarrow[\text{by } \frac{1}{s_0}]{\text{scale}} \begin{matrix} \bar{y} = 0 \\ s = 1 \end{matrix}$$

Section 5.3: Density Curves and the Normal Model

$n = 736$





$A = bh$

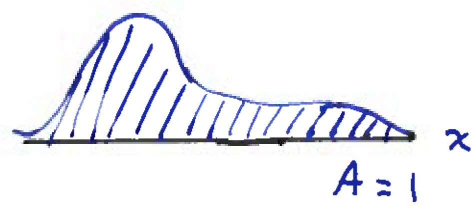
(Source: StatCrunch)

rel. frequency is height $= 0.1$

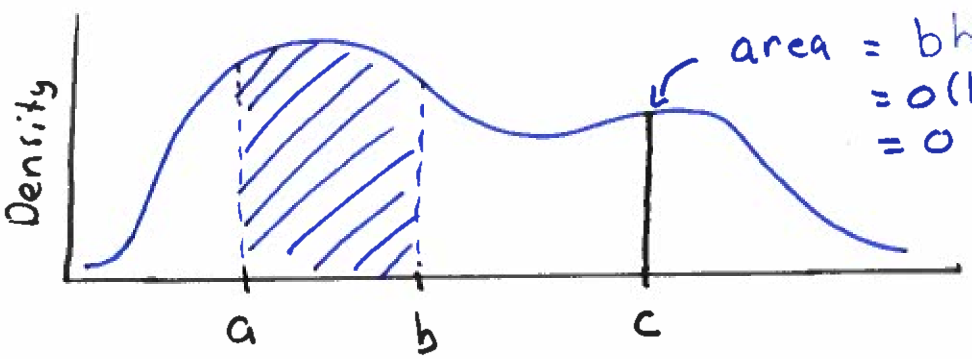
rel. frequency is area $= 5 \times 0.02 = 0.1$

The distribution of a continuous quantitative variable can be modeled by a **density curve**, where relative frequencies are represented by areas under the curve.

Properties of Density Curves



- Density curves are always positive or 0.
- The total area under a density curve (above x-axis) is 1.
- The proportion of data values that fall within an interval is equal to the area under the curve, above the x-axis, and within that interval.



area $= bh = 0(h) = 0$

$[a,b]$ or (a,b) variable

capital letters

Note: The **probability distribution** of a **continuous random variable** uses a density curve. The **probability** that an outcome falls within an interval is given by the area under the curve, above the x-axis, and within that interval.

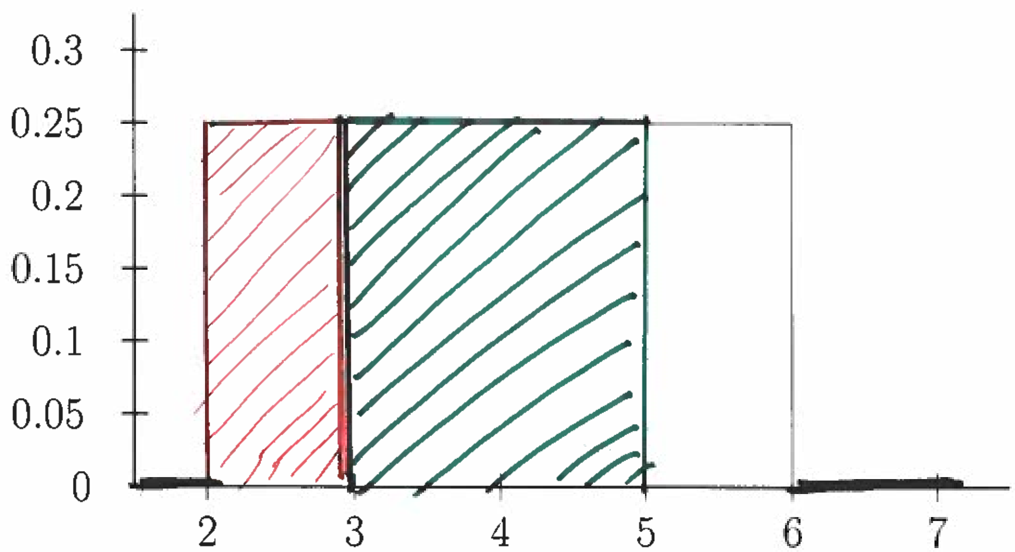
Notation: $P(a < X < b)$ is the area under a (probability) density curve between a and b.

Note: $P(X = a) = 0$.

$P(a \leq X \leq b) = P(a < X < b)$

↳ for continuous not discrete.

Example: Uniform Distribution



a) What is the total area under the curve?

$A = bh = 4(0.25) = 1$ → density curve

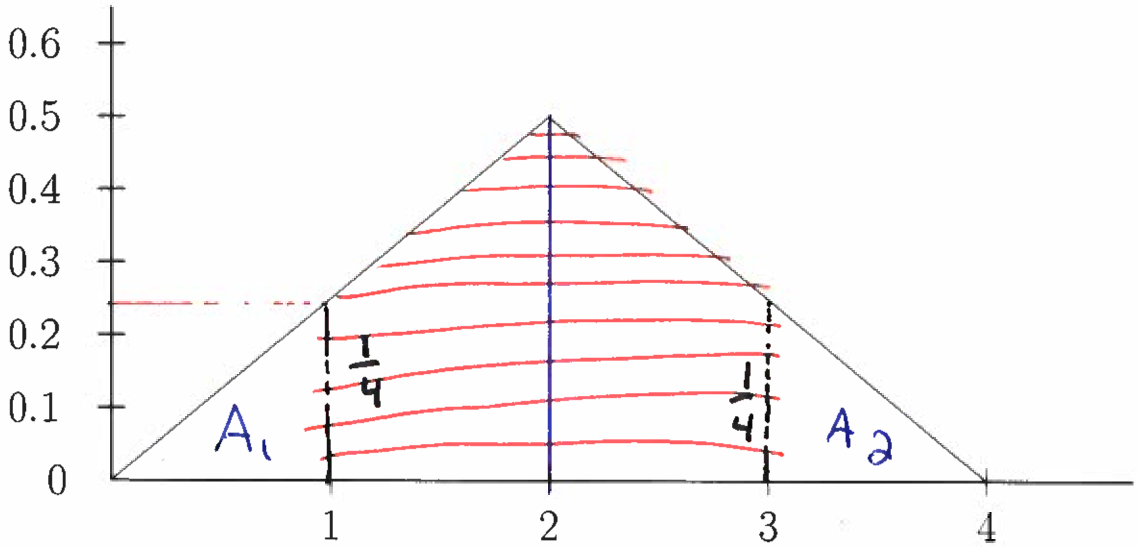
b) What is $P(X \leq 3)$?

$P(X \leq 3) = bh = 1(0.25)$
$= 0.25$

c) What is $P(3 < X < 5)$?

$P(3 < X < 5)$
$= bh$
$= 2(0.25)$
$= 0.5$

Example:



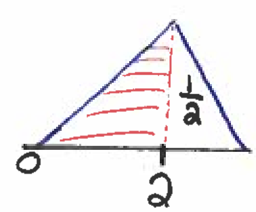
a) What is the total area under the curve?

$$A = \frac{bh}{2} = \frac{4(0.5)}{2} = 2(0.5) = 1$$

$\longrightarrow$ density curve

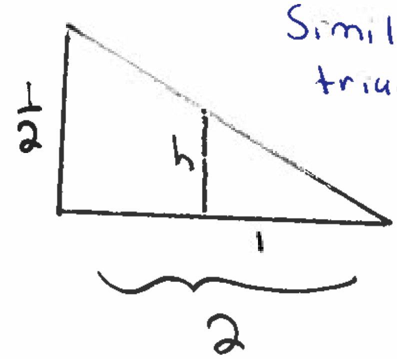
b) What is $P(X < 2)$?

$$P(X<2) = \frac{1}{2}(2)\left(\frac{1}{2}\right) = \frac{1}{2}$$



c) What is $P(1 < X < 3)$?

$$P(1<X<3)$$
$$= 1 - \left(P(X<1) + P(X>3)\right)$$
$$= 1 - \left(\frac{1}{2}(1)\left(\frac{1}{4}\right) + \frac{1}{2}(1)\left(\frac{1}{4}\right)\right)$$
$$= 1 - \left(\frac{1}{8} + \frac{1}{8}\right)$$
$$= 1 - \frac{1}{4}$$
$$= \frac{3}{4}$$



Similar triangles

$$\frac{\frac{1}{2}}{2} = \frac{h}{1}$$

$$h = \frac{1}{4}$$

Normal Models

→ weight, height

Normal curves are important density curves in statistical theory.

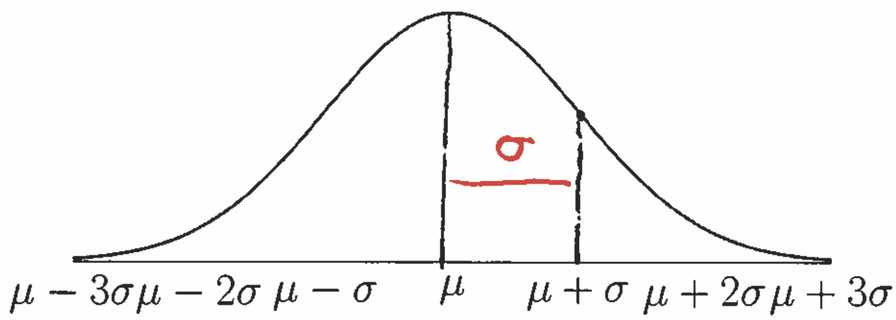
They are symmetric, unimodal, and bell-shaped curves.

There is a normal curve for every pair of mean μ and standard deviation σ

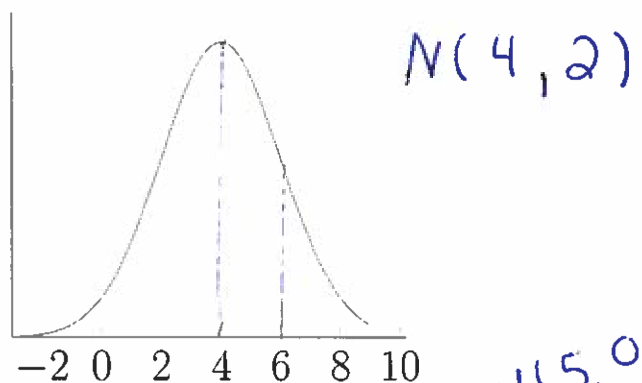
- μ is the centre of the curve.
- σ is the spread of the curve.

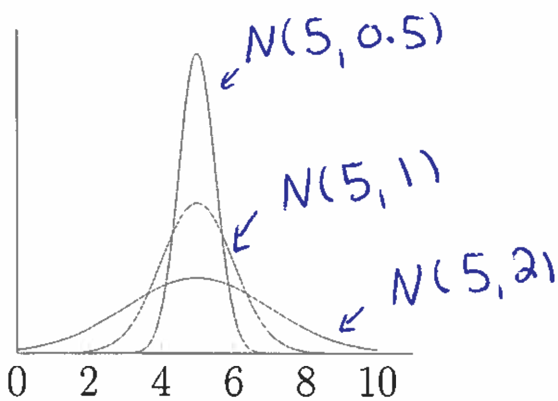
parameters of the model

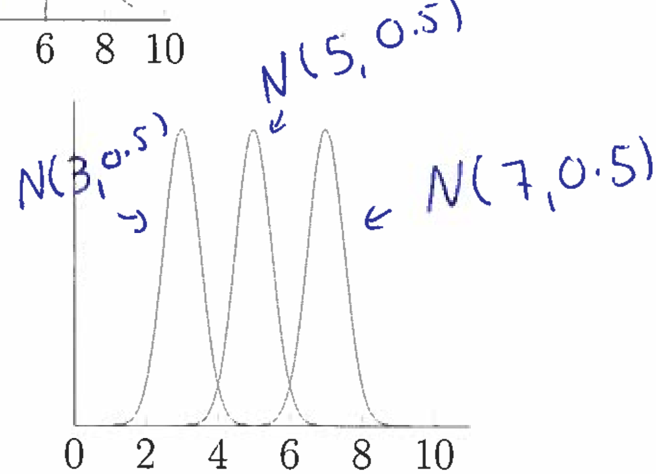
$\sigma > 0$



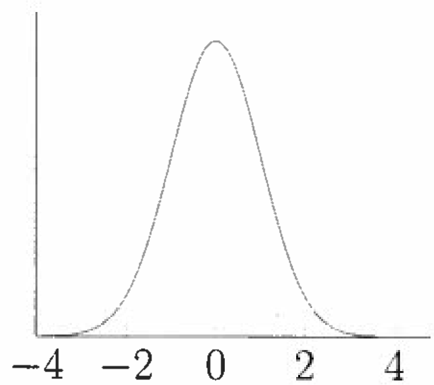
Notation: $N(\mu, \sigma)$ represents a Normal model with mean μ and standard deviation σ







Note: $N(0,1)$ is called the **standard Normal model** (or the standard Normal distribution).



$N(0,1)$ is the distribution of z-scores and so it is also called the z-distribution.

→ assume normal

If we model data with a Normal model $N(\mu, \sigma)$, we can standardize to $N(0,1)$. The standardized values are again called z-scores. The z-score of a value y is

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

Warning: → Standardizing doesn't change distribution shape

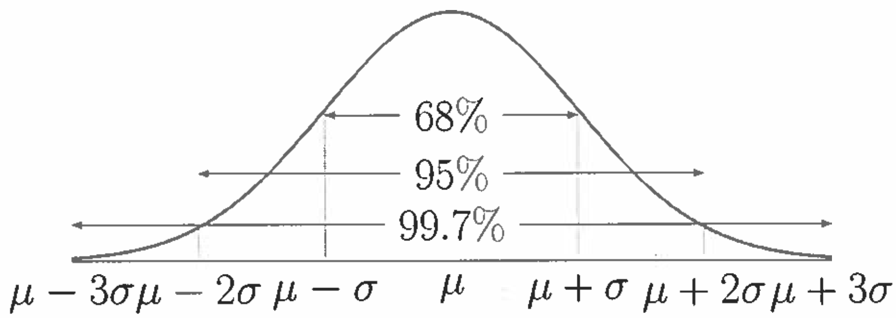
Normal models should only be used for data whose distribution is close to a **normal distribution**, that is, the shape of the data's distribution (histogram) is unimodal and fairly symmetric (**Nearly Normal Condition**).

↑ check this

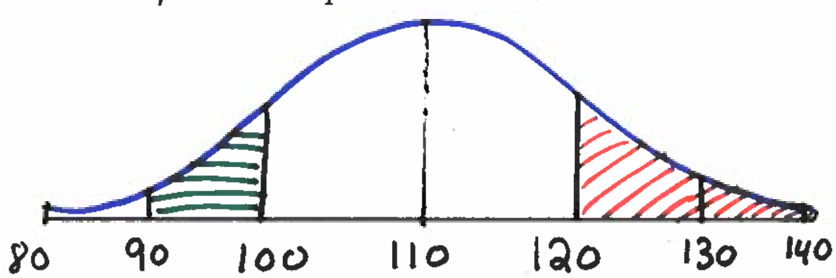
68-95-99.7 Rule for Normal Models (Empirical Rule)

In a Normal model, approximately:

- 68% of the values fall within one standard deviation of the mean.
- 95% of the values fall within two standard deviations of the mean.
- 99.7% of the values fall within three standard deviations of the mean.



Example: Suppose the U of A student IQs are (approximately) normally distributed with mean $\mu = 110$ points and standard deviation of $\sigma = 10$ points.

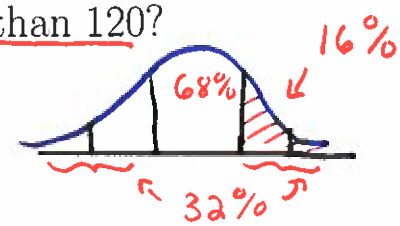


a) In what interval would you expect the central 95% of IQs to be found?

[90, 130]

b) What percent of students will have an IQ of more than 120?

16%



c) What percent of students will have an IQ of between 90 and 100?

16% - 2.5% = 13.5%