The Normal Distribution

Corinne Riddell (Instructor: Alan Hubbard and Tomer Altman)

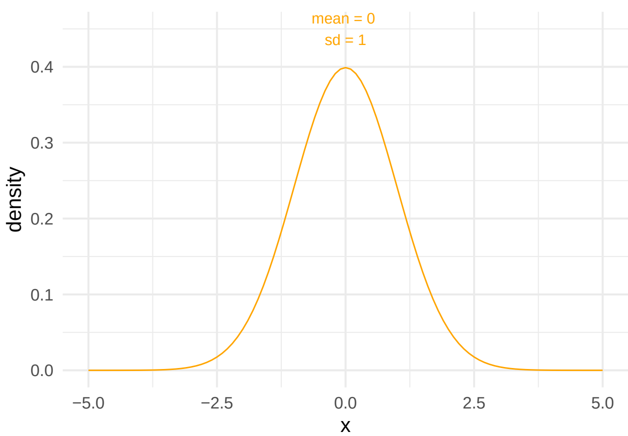
October 2, 2024

Learning objectives for today

- Learn about the Normal distribution centered at μ with a standard deviation of σ
- Learn about the standard Normal distribution where $\mu = 0$ and $\sigma = 1$ and compute z-scores
- Calculate cumulative probabilities below or above a given value for any specified Normal distribution using R
- Perform simple calculations by hand (using the 68-95-99.7 rule)

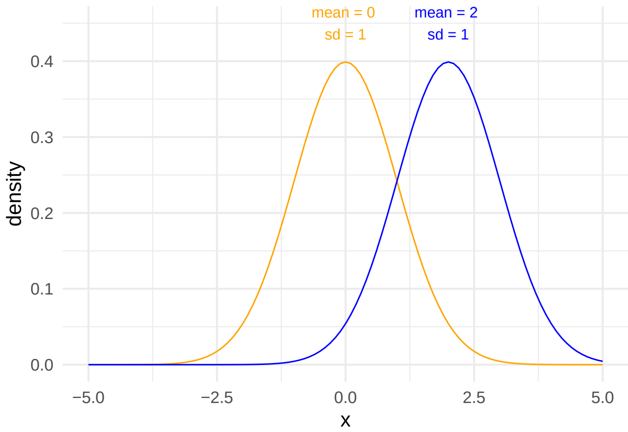
The Normal Distribution

- Here is the Normal distribution with mean of 0 (μ) and standard deviation of 1 (σ).
- It is:
  - symmetric
  - centered at μ



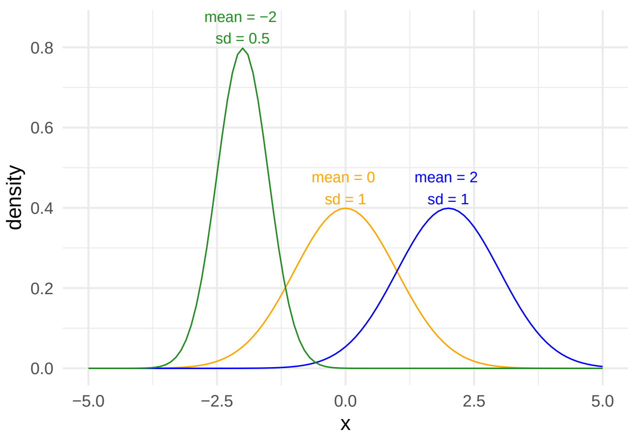
The Normal Distribution

- Let's add another Normal distribution, this one centered at 2, with the same standard deviation



The Normal Distribution

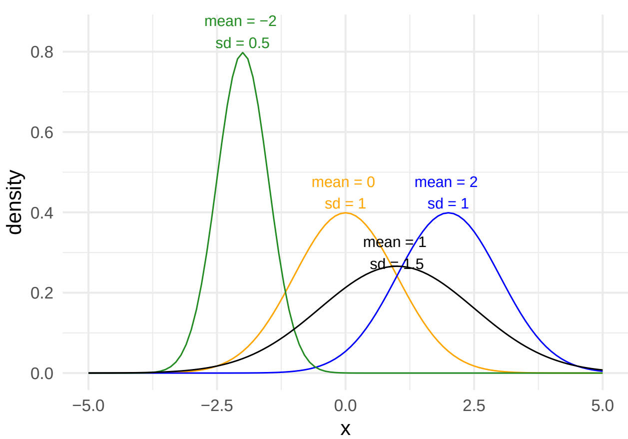
- Let's add a third Normal distribution, this one centered at -2, with a standard deviation of 0.5
- Notice how the distribution is narrowed (i.e., the spread is reduced)
- Why is the distribution "taller"?



The Normal Distribution

- Can you guess what a Normal distribution with $\mu = 1$ and $\sigma = 1.5$ would look like compared to the others?

The Normal Distribution



Properties of the Normal distribution

- the density can be drawn by knowing just two parameters , the mean (μ) and SD (σ): $f(x) = \phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
- the mean μ can be any value, positive or negative
- the standard deviation σ must be a positive number
- the mean is equal to the median (both $= \mu$)
- the standard deviation captures the spread of the distribution
- the area under the Normal distribution is equal to 1 (i.e., it is a density function)

The 68-95-99.7 rule for all Normal distributions

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

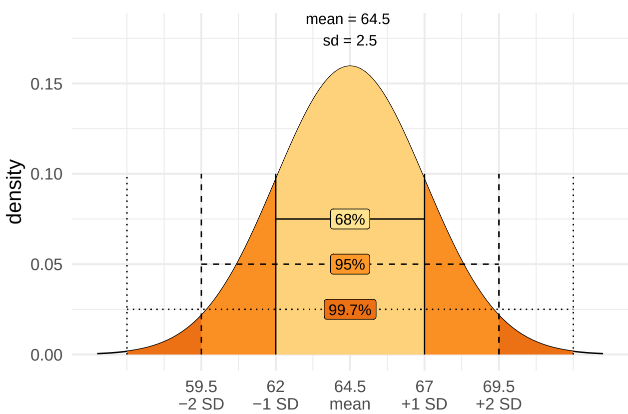
Written probabilistically:

- $P(\mu - \sigma < X < \mu + \sigma) \approx 68\%$
- $P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 95\%$
- $P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 99.7\%$

Calculations using the 68-95-99.7 rule

Example 11.1 from Baldi & Moore on the heights of young women. The distribution of heights of young women is approximately Normal, with mean $\mu = 64.5$ inches and standard deviation $\sigma = 2.5$ inches.

- We use notation to represent when a random variable follows a specific distribution. For example, letting H represent the random variable for the height of a young woman, we can then write $H \sim N(64.5, 2.5)$, to say that the random variable H follows a Normal distribution with a mean of 64.5 and a standard deviation of 2.5.

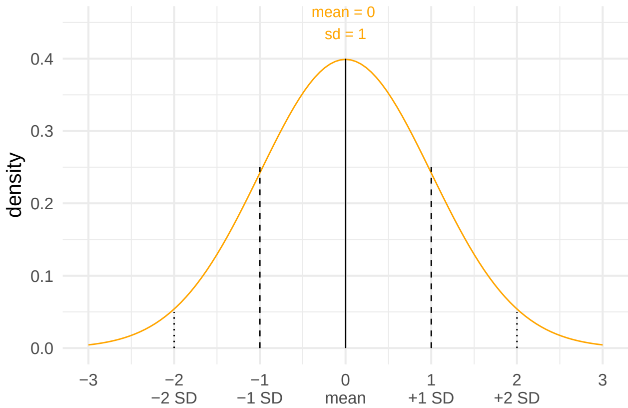


Calculations using the 68-95-99.7 rule

- What calculations could you do with these data alone?
- $P(62 < H < 67) =?$
- $P(H > 62) =?$

The standard Normal distribution

- The standard Normal distribution is the Normal distribution with $\mu = 0$ and $\sigma = 1$.
- We write: $N(0, 1)$ to denote this distribution
- $X \sim N(0, 1)$, implies that the random variable X is Normally distributed.



Standardizing Normally distributed data

- Any random variable that follows a Normal distribution can be standardized. This means we can transform its distribution from being centred at μ with a standard deviation of σ to another Normal distributuin with $\mu = 0$ and standard deviation of $\sigma = 1$
- If x is an observation from a distribution that has a mean μ and a standard deviation σ, the standardized value of x is calculated in the following way:

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

- A standardized value is often called a **z-score**
- Interpretation: z is the number of standard deviations that x is above or below the mean of the data.
- We standardize values so that we can have this interpretation, which is agnostic to the underlying mean, standard deviation, and units of measure. Standardizing Normally-distributed data is a quick way to determine if a specific value is much higher or lower than the average value.

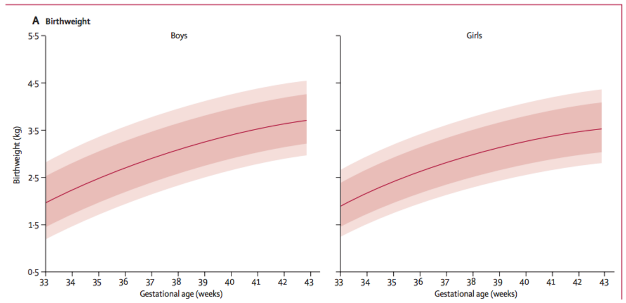
Standardizing Normally distributed data

Source: Intergrowth 21st Century

Standardizing Normally distributed data

In this image, the solid red line shows the average birthweight as a function of gestational age for boys and girls.

What is the approximate average birthweight in kilograms for a boy delivered at 33 weeks?



Reference

Standardizing Normally distributed data

The International Newborn Standards

INTERGROWTH-21st

UNIVERSITY OF OXFORD

Birth weight (Boys)

Gestational age (weeks+days)	z scores						
	-3	-2	-1	0	1	2	3
33+0	0.63	1.13	1.55	1.95	2.39	2.88	3.47
33+1	0.67	1.17	1.59	1.99	2.43	2.92	3.51
33+2	0.71	1.21	1.63	2.03	2.47	2.96	3.55
33+3	0.75	1.25	1.67	2.07	2.50	2.99	3.59
33+4	0.79	1.29	1.71	2.11	2.54	3.03	3.62
33+5	0.83	1.33	1.75	2.15	2.58	3.07	3.66
33+6	0.87	1.37	1.79	2.18	2.62	3.11	3.70
34+0	0.91	1.40	1.82	2.22	2.65	3.14	3.73
34+1	0.95	1.44	1.86	2.26	2.69	3.18	3.77
34+2	0.98	1.48	1.90	2.29	2.73	3.21	3.80
34+3	1.02	1.51	1.93	2.33	2.76	3.25	3.84
34+4	1.05	1.55	1.97	2.36	2.80	3.28	3.87
34+5	1.09	1.58	2.00	2.40	2.83	3.32	3.91
34+6	1.12	1.62	2.04	2.43	2.86	3.35	3.94
35+0	1.16	1.65	2.07	2.47	2.90	3.38	3.97

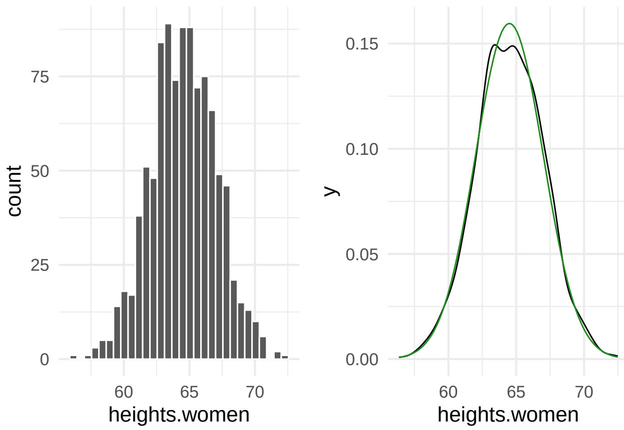
- Birthweight z-scores for boys
- How does this relate to what you see on the previous slide?

Simulating Normally distributed data in R

Suppose that we measured 1000 heights for young women:

```r
# students, rnorm() is important to know!
# this line of code generates 1000 rows of data from a Normal distribution with
# the specified mean and sd.
heights.women <- rnorm(n = 1000, mean = 64.5, sd = 2.5)

# this line of code puts this variable into a data frame
heights.women <- data.frame(heights.women)
```

We can plot the histogram of the heights, and see that they roughly follow from a Normal distribution. The green curve is a Normal distribution, and the black curve is the density plot based on the actual data:



Standardizing Normally distributed data in R

To standardize these data, we can apply the formula to compute the z-score:

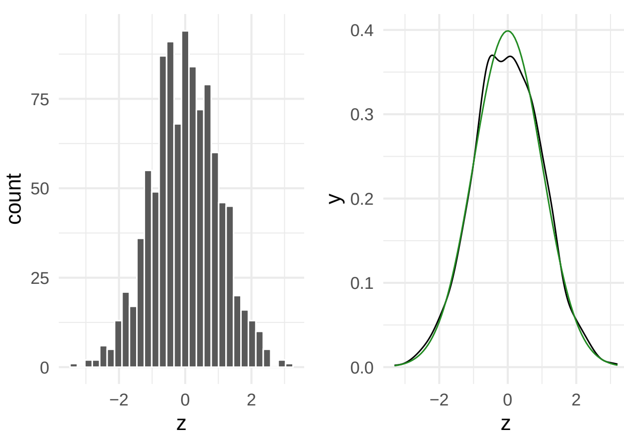
```
heights.women <- heights.women %>% mutate(mean = mean(heights.women),
                                          sd = sd(heights.women),
                                          z = (heights.women - mean)/sd)

head(heights.women)
```

```
##   heights.women     mean       sd          z
## 1      69.02117 64.54829 2.476356  1.8062350
## 2      61.99512 64.54829 2.476356 -1.0310192
## 3      62.90538 64.54829 2.476356 -0.6634372
## 4      63.69186 64.54829 2.476356 -0.3458427
## 5      65.76453 64.54829 2.476356  0.4911391
## 6      60.66542 64.54829 2.476356 -1.5679779
```

What would the distribution of the standardized heights look like?

Standardizing Normally distributed data in R

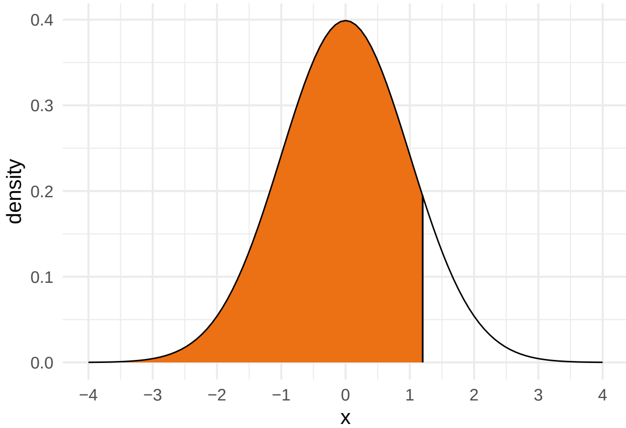
How are these plots different from the previous ones? Hint: look at the x axis.



Finding Normal probabilities

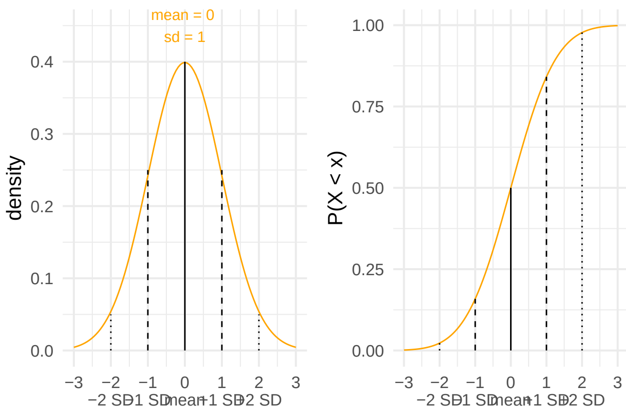
- A **cumulative probability** for a value x in a distribution is the proportion of observations in the distribution that lie at or below x.

- Here is the cumulative probability for x=1.2



Plot of Cumulative Standard Normal Distribution

- There are different ways to display a distribution such as the density and the cumulative distribution
- The cumulative distribution can be shown as a graph of the probability of being below a value on the X-axis.



Finding Normal probabilities

- Recall that 100% of the sample space for the random variable X lies under its probability density function.
- What is the amount of the area that is below x = 1.2?
- To answer this question we use the `pnorm()` function. (Think: the **p** in `pnorm` stands for probability):

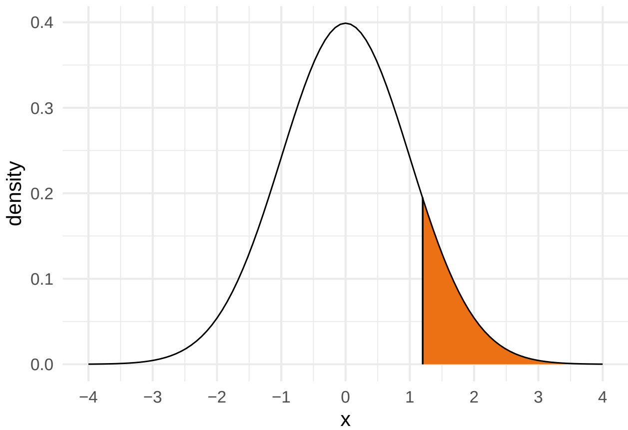
```
pnorm(q = 1.2, mean = 0, sd = 1)
```

```
## [1] 0.8849303
```

This says that approximately 88% of the probability lies below 1.2.

Finding Normal probabilities

What if we wanted the reverse: P(x>1.2)?




```r
1 - pnorm(q = 1.2, mean = 0, sd = 1)
```

```
## [1] 0.1150697
```

Alternatively:

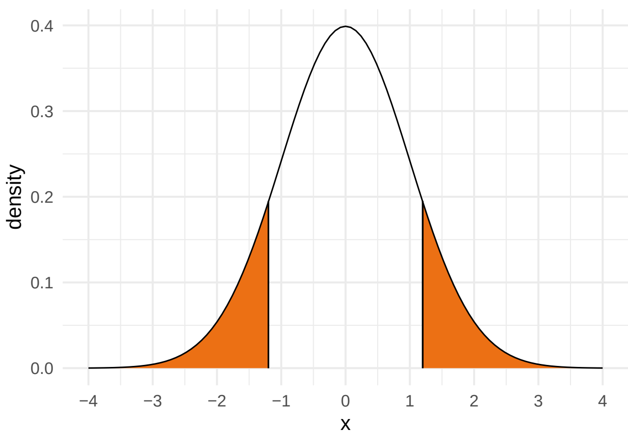
```r
pnorm(q = 1.2, mean = 0, sd = 1, lower.tail = F)
```

```
## [1] 0.1150697
```

So, 11.51% of the data is above x=1.2.

Finding Normal probabilities

What if we wanted two "tail" probabilities?: $P(x < -1.2 \text{ or } x > 1.2)$?



Finding Normal probabilities

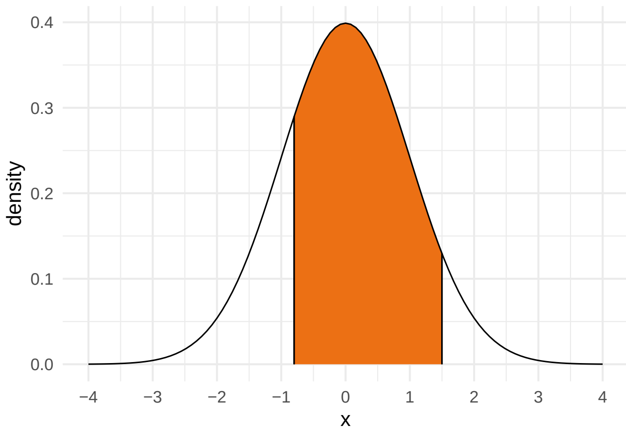
The trick: find one of the tails and then double the area because the distribution is symmetric:

```r
pnorm(q = -1.2, mean = 0, sd = 1)*2
```

```
## [1] 0.2301393
```

Finding Normal probabilities

What if we wanted a range in the middle?: $P(-0.8 < x < 1.5)$?



Finding Normal probabilities

```
# step 1: calculate the probability *below* the upper bound (x=1.5)
pnorm(q = 1.5, mean = 0, sd = 1)
```

```
## [1] 0.9331928
```

```
# step 2: calculate the probability *below* the lower bound (x = -0.8)
pnorm(q = -0.8, mean = 0, sd = 1)
```

```
## [1] 0.2118554
```

```
# step 3: take the difference between these probabilities to get what's left in
# the middle
pnorm(q = 1.5, mean = 0, sd = 1) - pnorm(q = -0.8, mean = 0, sd = 1)
```

```
## [1] 0.7213374
```

Thus, 72.13% of the data is in the range -0.8 < x < 1.5.

Your turn

To diagnose osteoporosis, bone mineral density is measured. The WHO criterion for osteoporosis is a BMD score below -2.5. Women in their 70s have a much lower BMD than younger women. Their BMD ~ N(-2, 1). What proportion of these women have a BMD below the WHO cutoff?

Hint: you do not need to find a z-score!

```
#to fill in during class
pnorm(q=-2.5, mean = -2, sd = 1)
```

```
## [1] 0.3085375
```

Recap of functions used

- `rnorm(n = 100, mean = 2, sd = 0.4)`, to generate Normally distributed data from the specified distribution
- `pnorm(q = 1.2, mean = 0, sd = 2)`, to calculate the cumulative probability below a given value