L14: The normal distribution

Learning objectives for today

- Learn about a Normal distribution centered at μ with a standard deviation of σ
- Learn about the standard Normal ($\mu = 0$ and $\sigma = 1$)
- Perform simple calculations by hand using the 68-95-99.7 rule
- Calculate probabilies above, below or within given values in a normal distribution
- Calculate the quantile for a specified cummulative probability for any normal distribution
- Understand and compute Z-scores

Statistics is everywhere

Height and success

What type of problem is this

Remembering back to our PPDAC framework. . . .

What kind of questions are these studies asking?

Counterfactual

Frayling, one of the study authors said:

> *"if you took the same man - say a 5ft 10in man and make him 5ft 7in - and sent him through life, he would be about £1,500 worse off per year"*

Were there important confounders?

This study was published in BMJ (Height, body mass index, and socioeconomic status: Mendelian randomization study in UK Biobank BMJ 2016; 352 doi: https://doi.org/10.1136/bmj.i582 (Published 08 March 2016))

They used mendelian randomization to avoid confounding - and put a DAG in their paper to illustrate

DAG

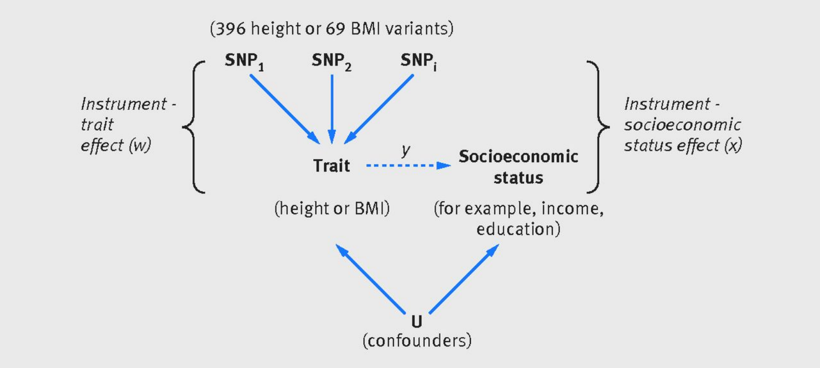


Figure 2: Study DAG

Proposed mechansim

One hypothesized mechanism for the influence of height on success is that those who are perceived as taller than normal experience positive social bias.

But wait. . . . what does “taller than normal” mean?

Three relevant Definitions of Normal (Mirriam-Webster)

1: Conforming to a type, standard , or regular pattern

2: of, relating to, or characterized by average intelligence or development

3: relating to, involving, or being a normal curve or normal distribution

Probability Distributions

Empirical vs Theoretical

- probabilities calculated from a finite amount of observed data are called Empirical probabilities
- probabilities based on theoretical functions are called Theoretical Probability Distributions

Empirical distributions

What was the empirical distribution we used to visualize a continuous variable in part I of the class?

What did we look at when summarizing continuous variables visually and numerically?

How might we summarize height?

Empirical distribution of height

Height is a continuous variable so we summarize height in a sample with:

- Measure of central tendancy
  - mean
  - median
  - mode
- Measure of variability/spread
  - standard deviation
  - IQR
- Visually we would look to see if the distribution is
  - Unimodal
  - symmetric

Why are we interested in theoretical distributions?

- Theoretical distributions apply probability theory to describe the behavior of a random variable
- For continuous variables (like height) this allows us to predict the probability associated with a range of values
- They help us answer questions about what is "normal" if by "normal" we essentially mean what is expected

The Normal Distribution

The Normal distribution

Figure 3: Carl Friedrick Gauss

Function for Normal Distribution

The underlying function which generates a probability distribution for a Normal distribution is:

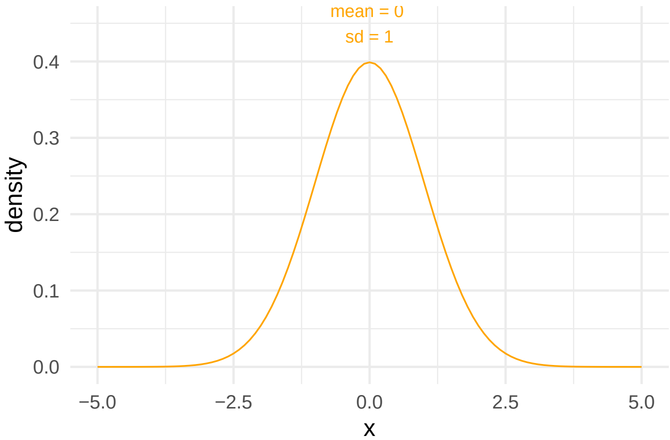
$$f(x) = \frac{1}{\sqrt{2\pi\sigma}}^{\left(-\frac{1}{2\sigma^2}\times(x-\mu)^2\right)}$$

You do NOT need to remember this However you should know that the distribution is defined by two parameters, the mean and standard deviation, and that all the values under the curve must total to 1 (the probability space here is the area under the curve)

The Normal Distribution N(0,1)

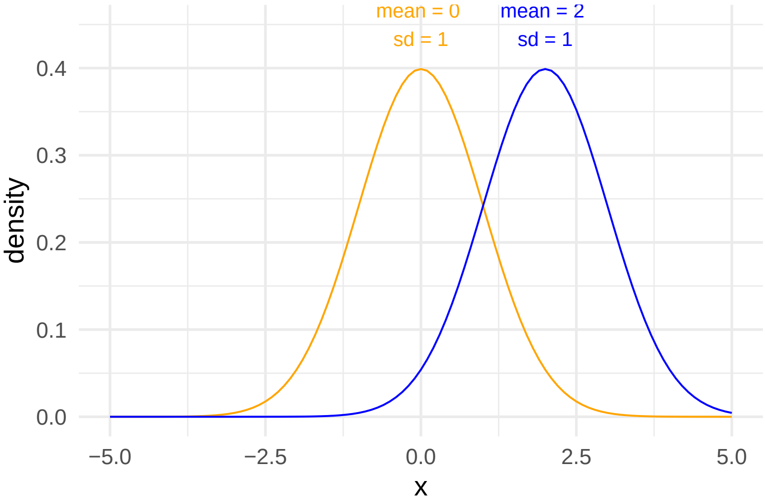
- Here is a Normal distribution with mean $(\mu) = 0$, standard deviation $(\sigma) = 1$

.



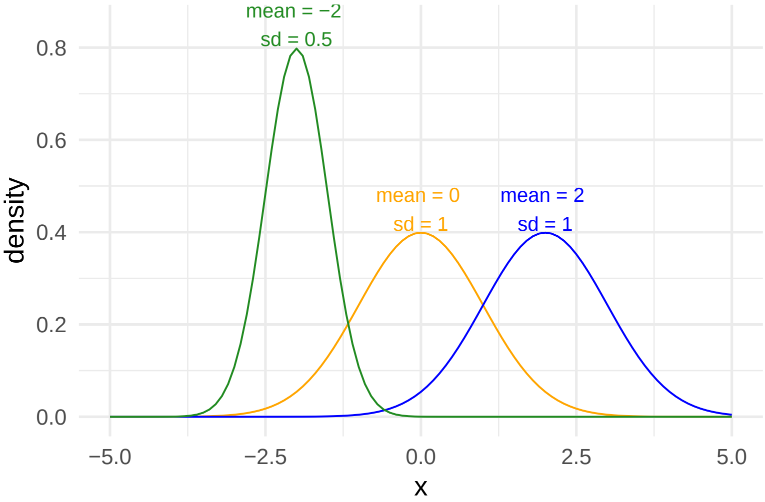
The Normal Distribution N(0,1) and N(2,1)

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



The Normal Distribution N(O,1) and N(2,1) and N(-2,0.5)

- Let's add a third Normal distribution, this one centered at -2, with a standard deviation of 0.5



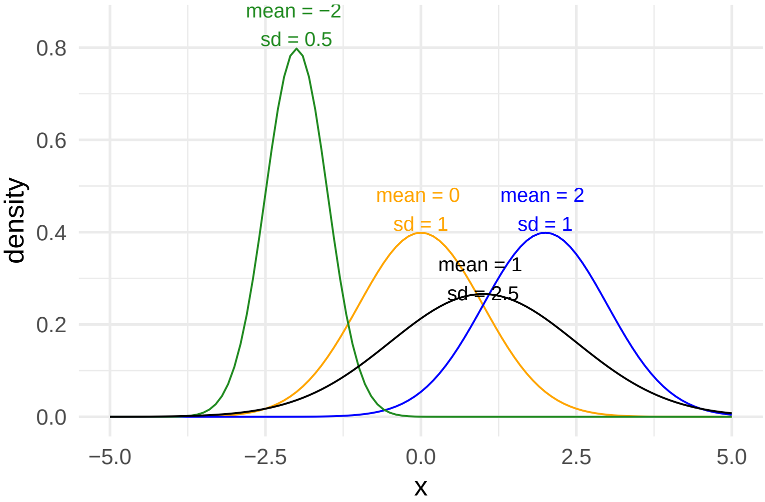
The Normal Distribution

- Notice what happens when we make the standard deviation smaller (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 $\sigma = 2.5$ would look like compared to the others?

The Normal Distribution



Properties of the Normal distribution

- 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 height of the curve at each point represents the probability of observing that value (however we never calculate probabilities for an exact value, only for ranges... why?)
- the area under the Normal distribution is equal to 1 (i.e., it is a density function)
- a Normal distribution is completely determined by its μ and σ

The 68-95-99.7 rule

The 68-95-99.7 rule for approximation in 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:

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

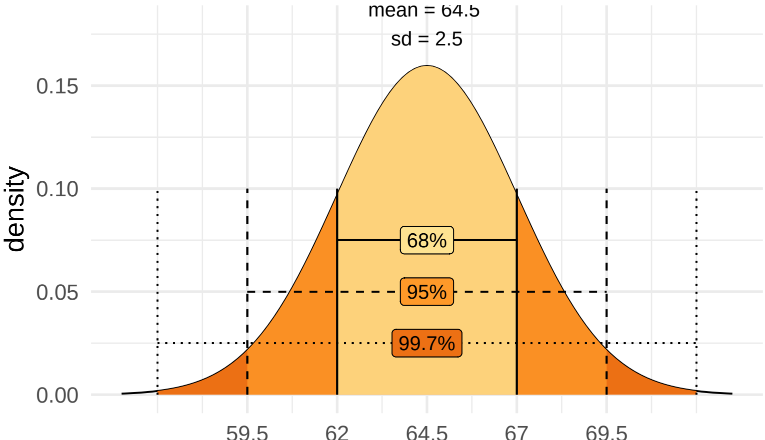
Calculations using the 68-95-99.7 rule

Remembering that the normal curve is symmetric we then also know:

- $P(X < \mu - \sigma) \approx 16\%$
- $P(X > \mu + \sigma) \approx 16\%$
- $P(X < \mu - 2\sigma) \approx 2.5\%$
- $P(X > \mu + 2\sigma) \approx 2.5\%$
- $P(X < \mu - 3\sigma) \approx .15\%$
- $P(X > \mu - 3\sigma) \approx .15\%$

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. - i.e., $H \sim N(64.5, 2.5)$, where H is defined as the height of a young woman



Calculations using the 68-95-99.7 rule

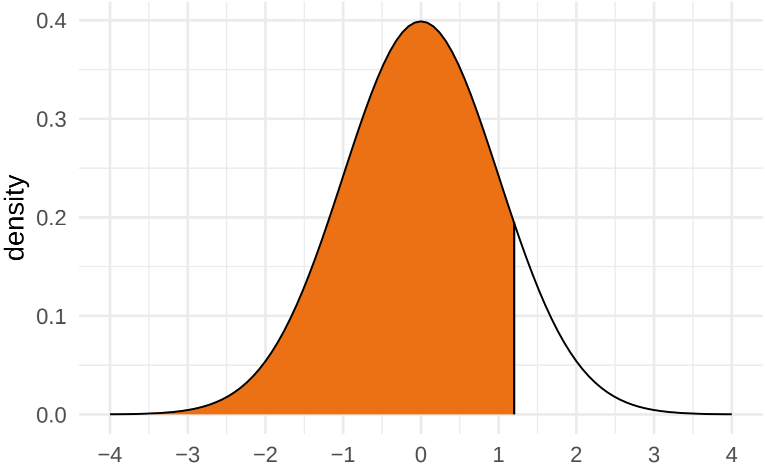
$\mu = 64.5$ inches and standard deviation $\sigma = 2.5$ inches

- What calculations could you do with just the μ and σ values and this rule?
- $\mathrm{P}(62 < \mathrm{H} < 67) = ?$
- $\mathrm{P}(\mathrm{H} > 62) = ?$
- Thinking back to the article about height - who would we consider "taller than normal" in these data?

Finding Normal probabilities

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



Finding Normal probabilities

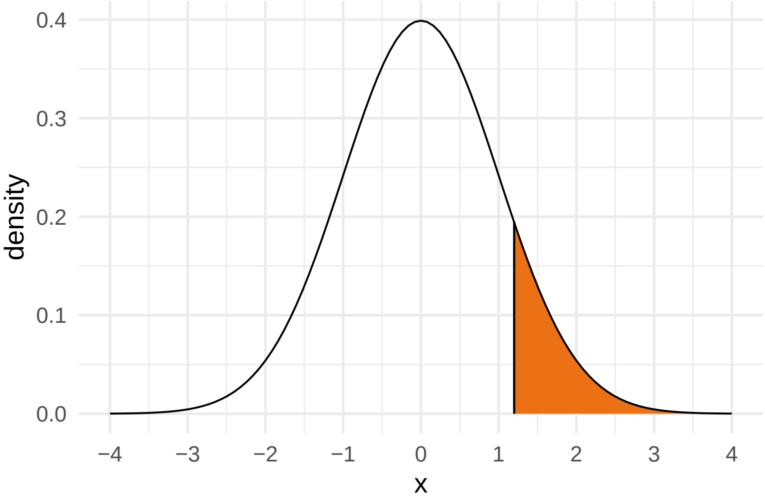
- Recall that 100% of the sample space for the random variable x lies under the probability density function.
- What is the amount of the area that is below $x = 1.2$?
- To answer this question we use the `pnorm()` function:

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

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

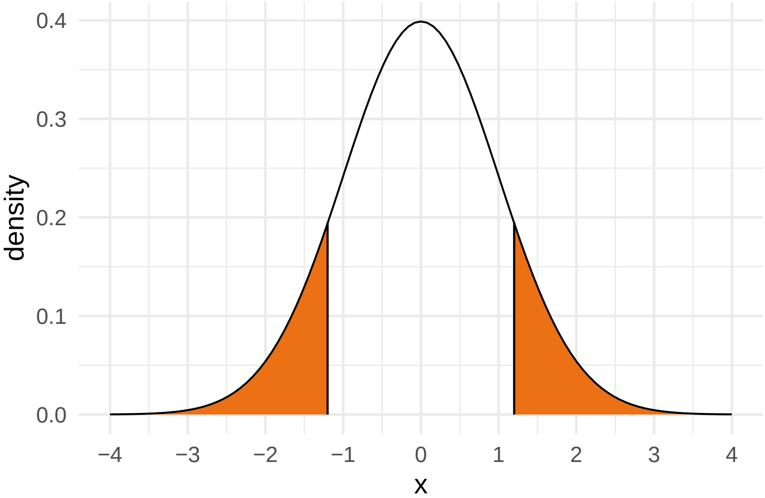
Finding Normal probabilities

What if we wanted the reverse: P(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

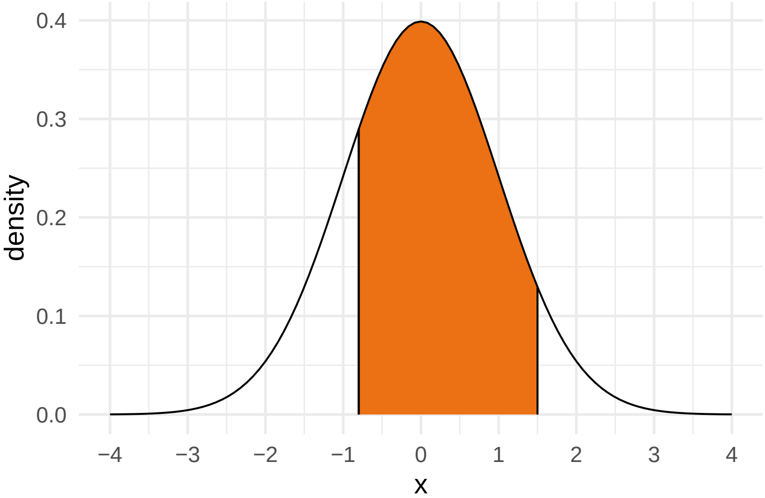
The trick: find one of the tails and then double the area because the distribution is symmetric:

```
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

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

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

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

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

```r
# step 3: take the difference between these probabilities
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 $\sim$ N(-2, 1). What proportion of these women have a BMD below the WHO cutoff?

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

Finding Normal percentiles

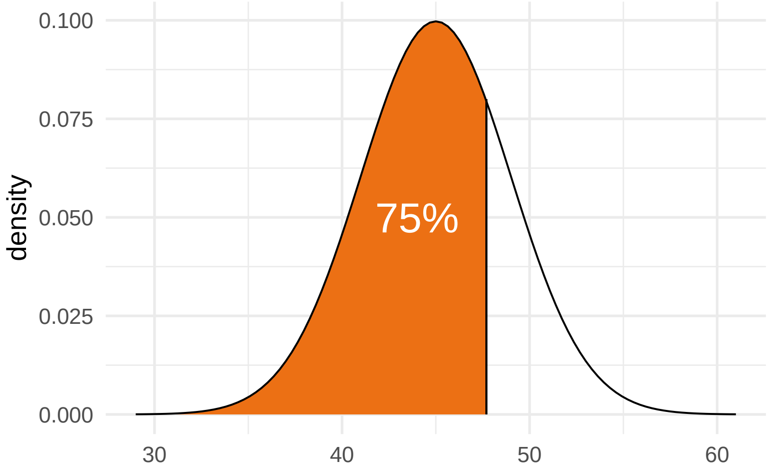
Finding Normal percentiles

Recap: so far, we have calculated the *probability* using `pnorm()` given specific values for x.

Sometimes we want to go in the opposite direction: We might be given the probability within some range and tasked with finding the corresponding x-values.

Finding Normal percentiles

Example: The hatching weights of commercial chickens can be modeled accurately using a Normal distribution with mean $\mu = 45$ grams and standard deviation $\sigma = 4$ grams. What is the third quartile of the distribution of hatching weights?



Finding Normal percentiles using the `qnorm()` function

Example: The hatching weights of commercial chickens can be modeled accurately using a Normal distribution with mean $\mu = 45$ grams and standard deviation $\sigma = 4$ grams. What is the third quartile of the distribution of hatching weights?

```
qnorm(p = 0.75, mean = 45, sd = 4)
```

```
## [1] 47.69796
```

Thus, 75% of the data is below 47.7 for this distribution.

The standard normal and Z scores

Convert to a standard

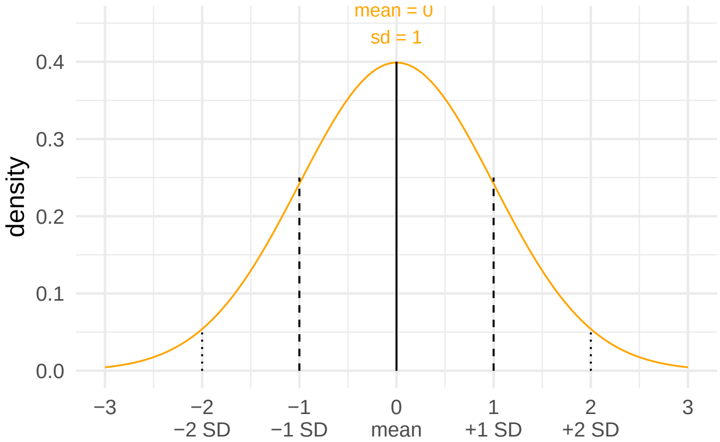
A normal distribution can have an infinite number of possible values for its mean and sd.

It would be impossible to tabulate the area associated with each and every normal curve

So instead of doing the impossible, we convert to the Standard Normal Distribution

The standard Normal distribution

- The standard Normal distribution N(0,1) has $\mu = 0$ and $\sigma = 1$.
- $X \sim N(0,1)$, implies that the random variable X is Normally distributed.



Standardizing Normally distributed data

- Any random variable that follows a Normal distribution can be standardized
- If x is an observation from a distribution that has a mean μ and a standard deviation σ,

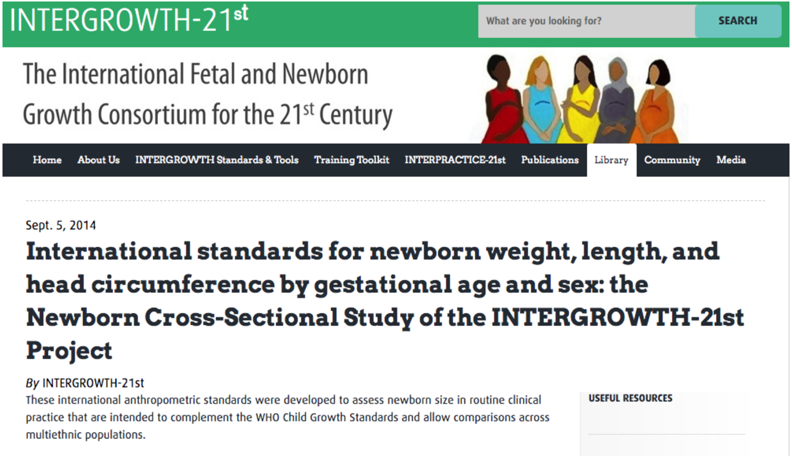
$$z = \frac{x - \mu}{\sigma}$$

What's the Z

By converting our variable of interest X to Z we can use the probabilities of the standard normal probability distribution to estimate the probabilities associated with X.

- 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.

Standardizing Normally distributed data


INTERGROWTH-21st

What are you looking for? SEARCH

The International Fetal and Newborn Growth Consortium for the 21st Century

Home | About Us | INTERGROWTH Standards & Tools | Training Toolkit | INTERPRACTICE-21st | Publications | Library | Community | Media

Sept. 5, 2014

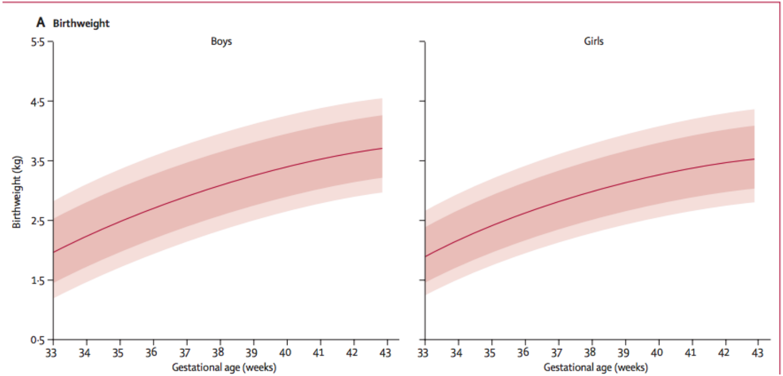
International standards for newborn weight, length, and head circumference by gestational age and sex: the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project

By INTERGROWTH-21st

These international anthropometric standards were developed to assess newborn size in routine clinical practice that are intended to complement the WHO Child Growth Standards and allow comparisons across multiethnic populations.

USEFUL RESOURCES

Standardizing Normally distributed data



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

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

Simulating Normally distributed data in R

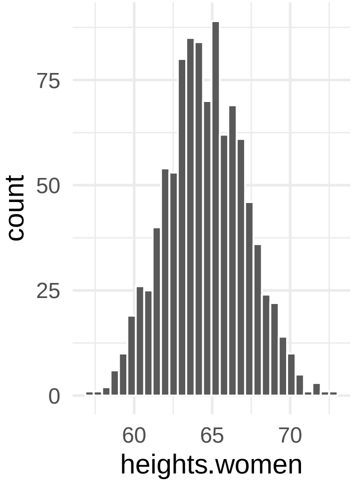
Simulating Normally distributed data in R

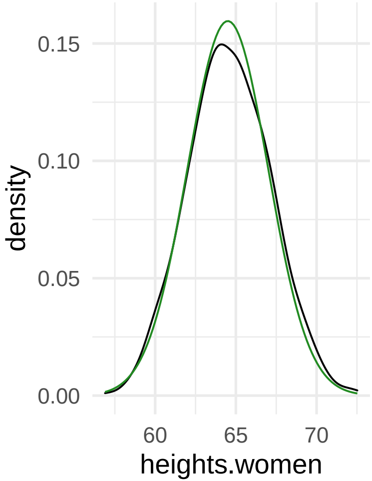
Suppose that we measured 1000 heights for young women:

```
#students, rnorm() is important to know!
heights.women <- rnorm(n = 1000, mean = 64.5, sd = 2.5)
heights.women <- data.frame(heights.women)
```

Simulating Normally distributed data in R

We can plot the histogram of the heights, and see that they roughly follow from a Normal distribution:





Standardizing Normally distributed data in R

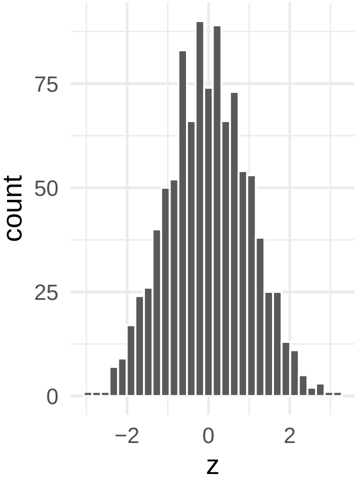
To standard these data, we can apply the formula to compute the z-value:

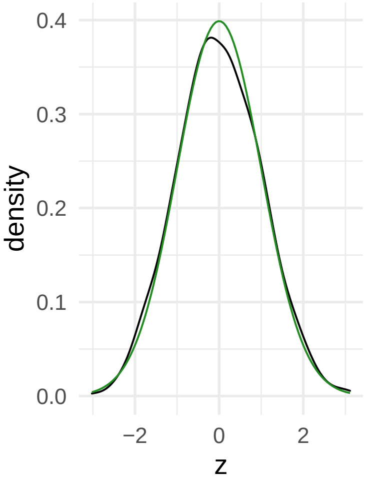
```
heights.women <- heights.women %>% mutate(mean = mean(heights.women),
                                          sd = sd(heights.women),
                                          z = (heights.women - mean)/sd)

head(heights.women)
```

```
##   heights.women     mean       sd          z
## 1      65.24195 64.59848 2.548518  0.2524863
## 2      62.95365 64.59848 2.548518 -0.6454076
## 3      62.60659 64.59848 2.548518 -0.7815887
## 4      59.90113 64.59848 2.548518 -1.8431720
## 5      69.95418 64.59848 2.548518  2.1014955
## 6      63.70314 64.59848 2.548518 -0.3513181
```

What would the distribution of the standardized heights look like?

Standardizing Normally distributed data in R





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
- `qnorm(p = 0.75, mean = 0, sd = 1)` to calculate the x-value for which some percent of the data lies below it

Comic Relief