Biostats Lecture 4: Estimators & Their Distributions

Public Health 783

Ralph Trane
University of Wisconsin–Madison

Fall 2019

Follow-up on todays lab

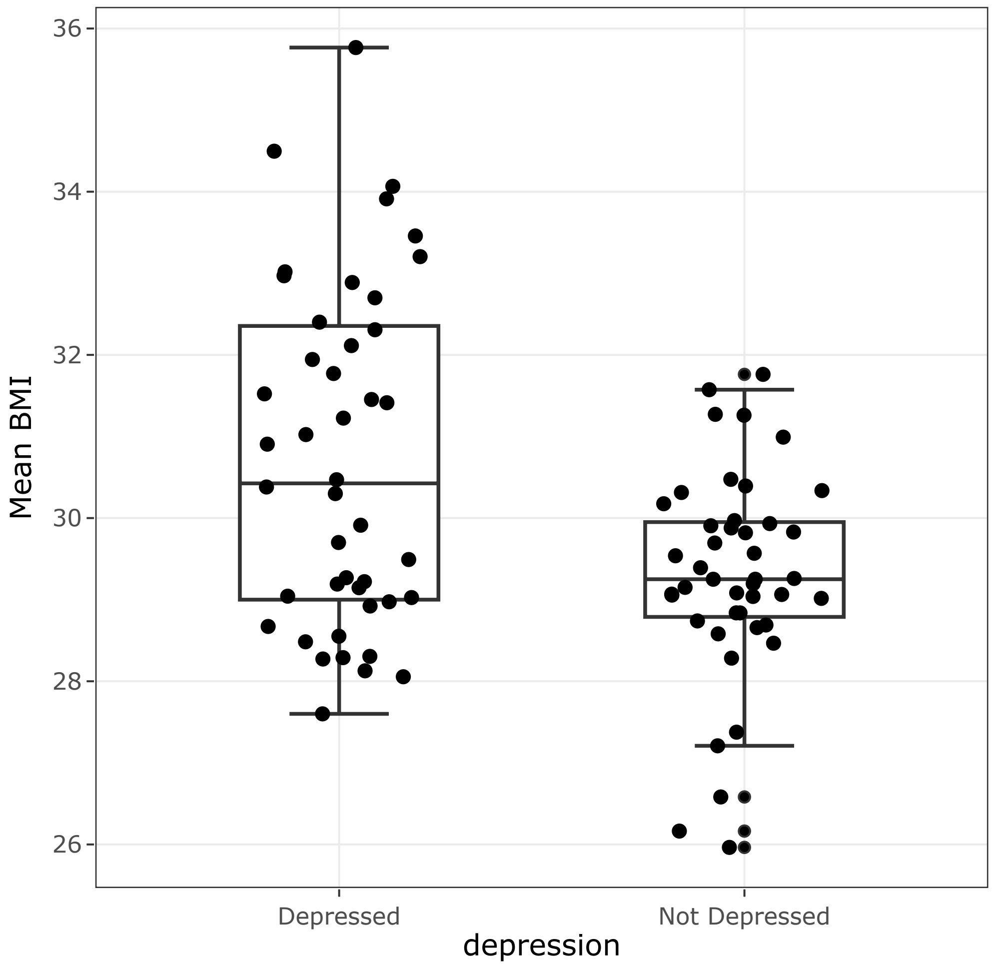
Let's just consider the relationship between BMI and depression.

You all had a sample of 50 subjects. I performed all 44 surveys on your behalves. (In reality, I gave you random samples from the SHOW population, sampled completely at random among subjects with complete data.)

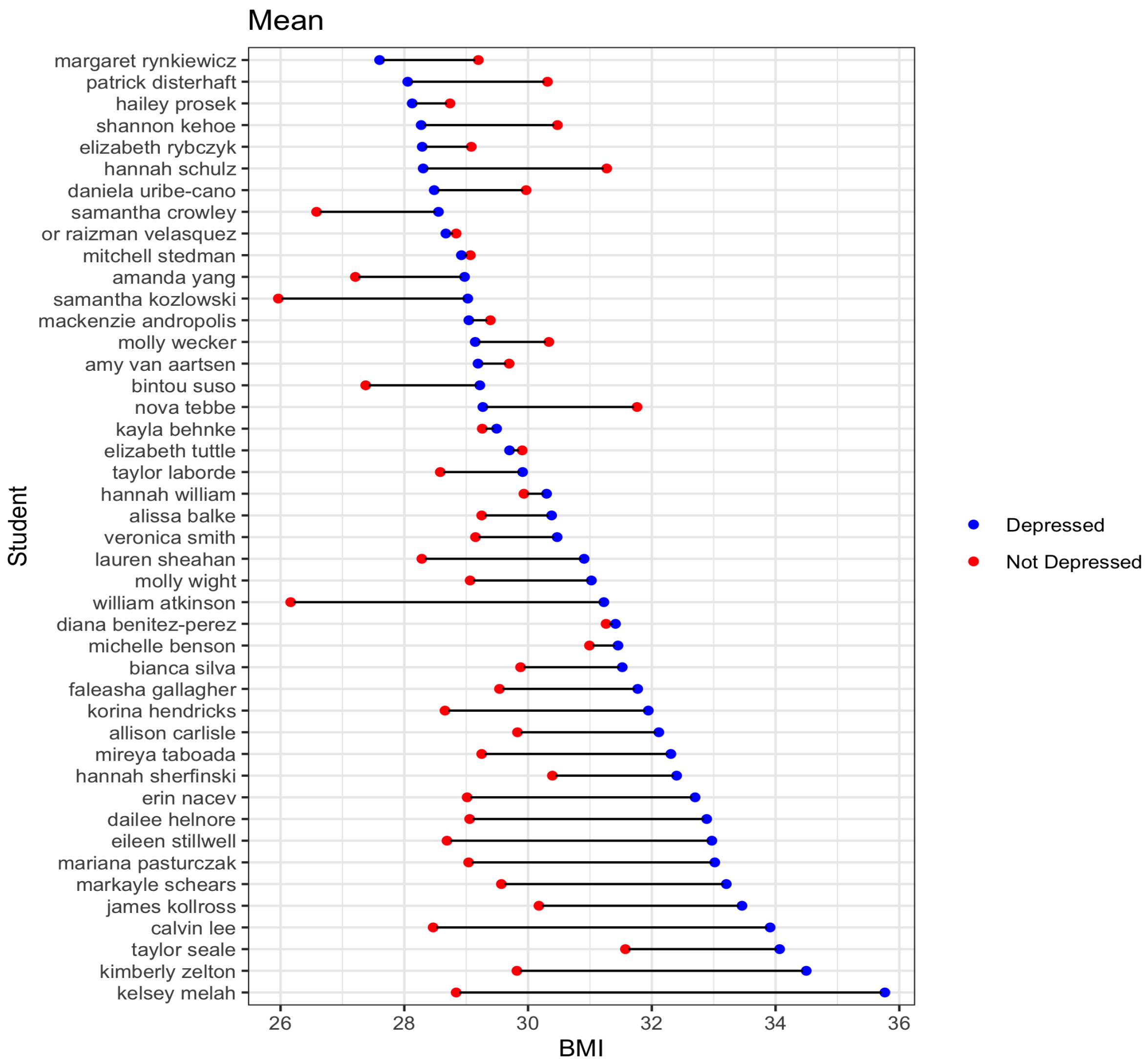
You all calculated a bunch of things, including the mean BMI in each of the two groups (depressed vs. not depressed)

As the almighty lecturer, I have access to all the samples…

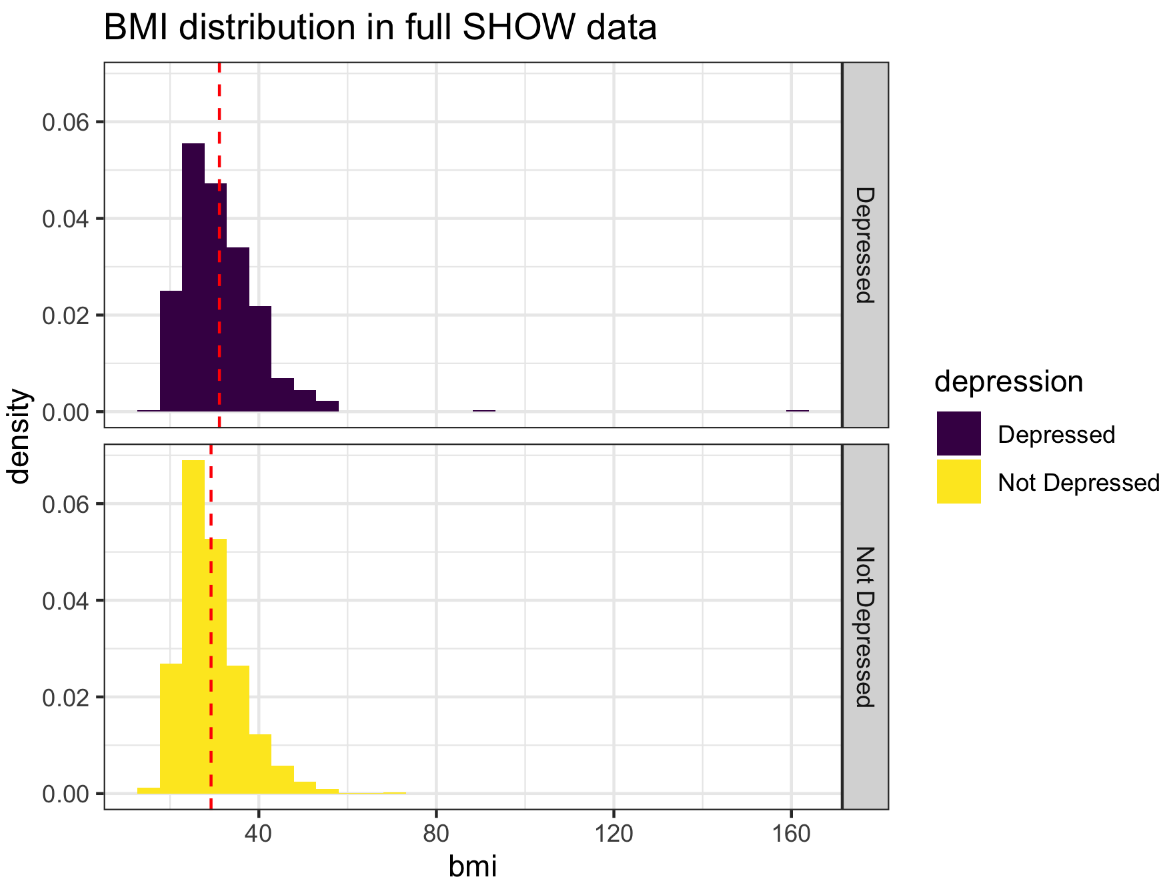
Estimators & Their Distributions



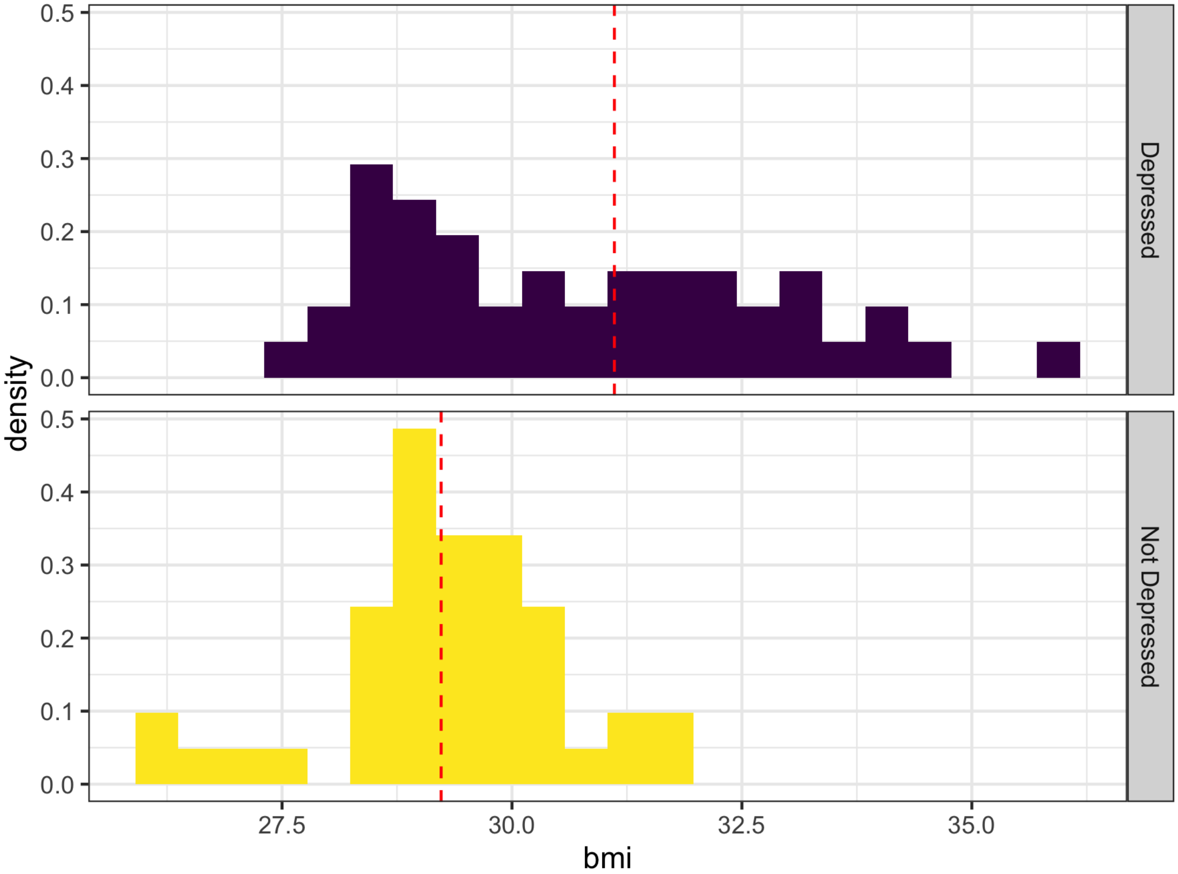
Estimators & Their Distributions



Estimators & Their Distributions



Estimators & Their Distributions



Motivational Spiel

Examples of Estimators

Parameter of Interest (most commonly used symbol)	Estimator Name	Notation and Formula
Mean of a feature μ	Sample average	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Variance of a feature σ^2	Sample variance	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
Standard deviation σ	Sample standard deviation	$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$
Probability of random individual having a disease π	Proportion in sample with disease	$P = \frac{1}{n}\sum_{i=1}^{n} X_i$
Proportion of individuals with disease π	Proportion in sample with disease	$P = \frac{1}{n}\sum_{i=1}^{n} X_i$

Example: Estimating Relative Risk

Show 10 entries Search:

	id	height	hip	waste	weight	race	marital	g
1	1509	198.5	132	131	165.2	[3] Hispanic	[3] Divorced	[1]
2	2865	150.25	126.4	104.25	89.9	[1] Non-hispanic white	[1] Married	[2] Fe
3	3112	158.45	125.2	112.95	97.7	[1] Non-hispanic white	[3] Divorced	[2] Fe

Estimators & Their Distributions

To find the relative risk, create 2 by 2 contingency table:

Show 10 entries Search:

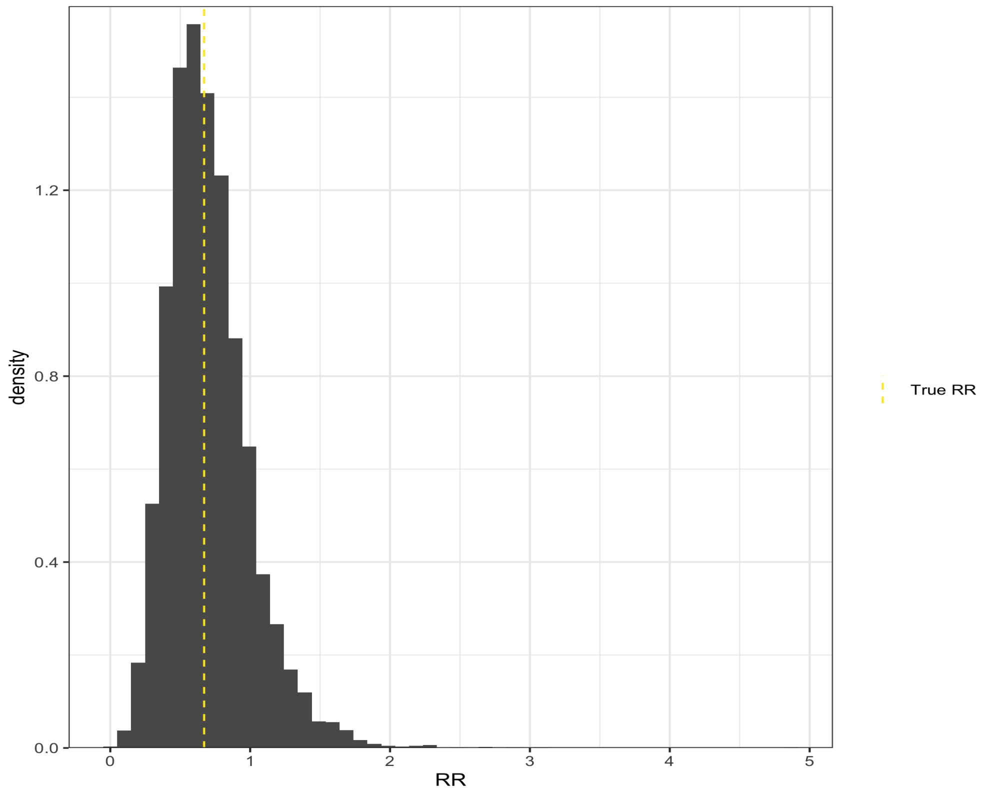
	depression_severity_binary	Female	Male	Total
1	0	28	28	56
2	1	13	6	19
3	Total	41	34	75

Showing 1 to 3 of 3 entries Previous 1 Next

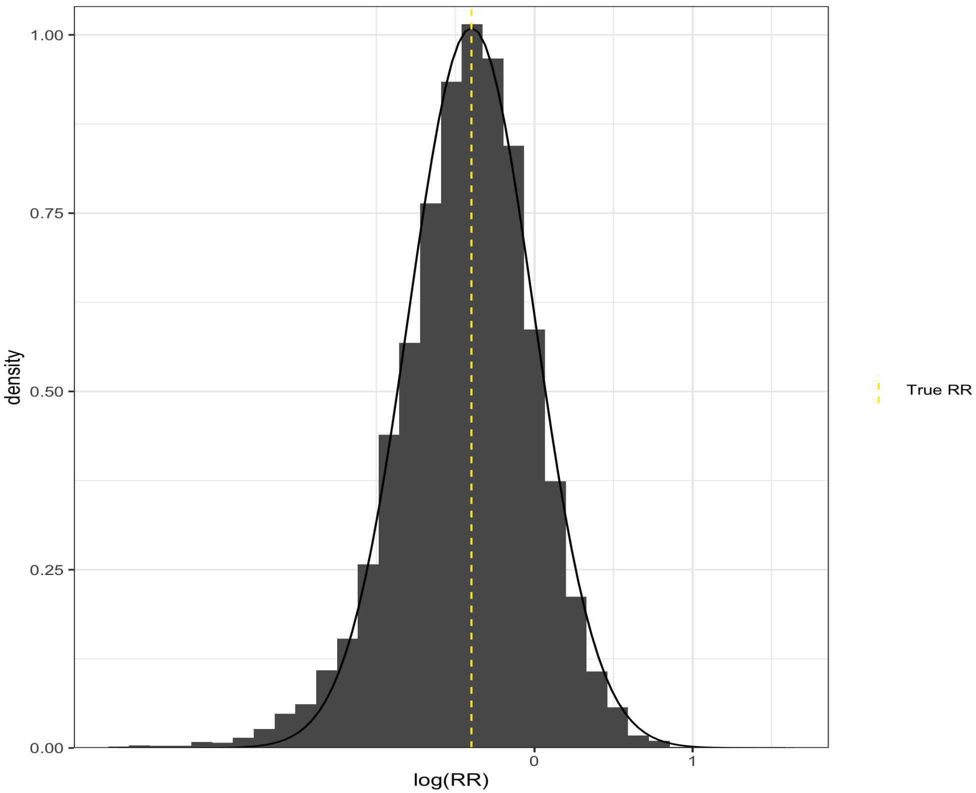
The relative risk is then calculated as

$$\begin{aligned}\frac{\text{proportion of males with severe depression}}{\text{proportion of women with severe depression}} &= \frac{6/34}{13/41} \\ &\approx 0.56.\end{aligned}$$

Estimators & Their Distributions



Estimators & Their Distributions



Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be a simple random sample from a population with mean μ and variance σ^2 (i.e. $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$ for all i). Then, as long as n is large enough, the *average* $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately $N(\mu, \sigma^2/n)$.

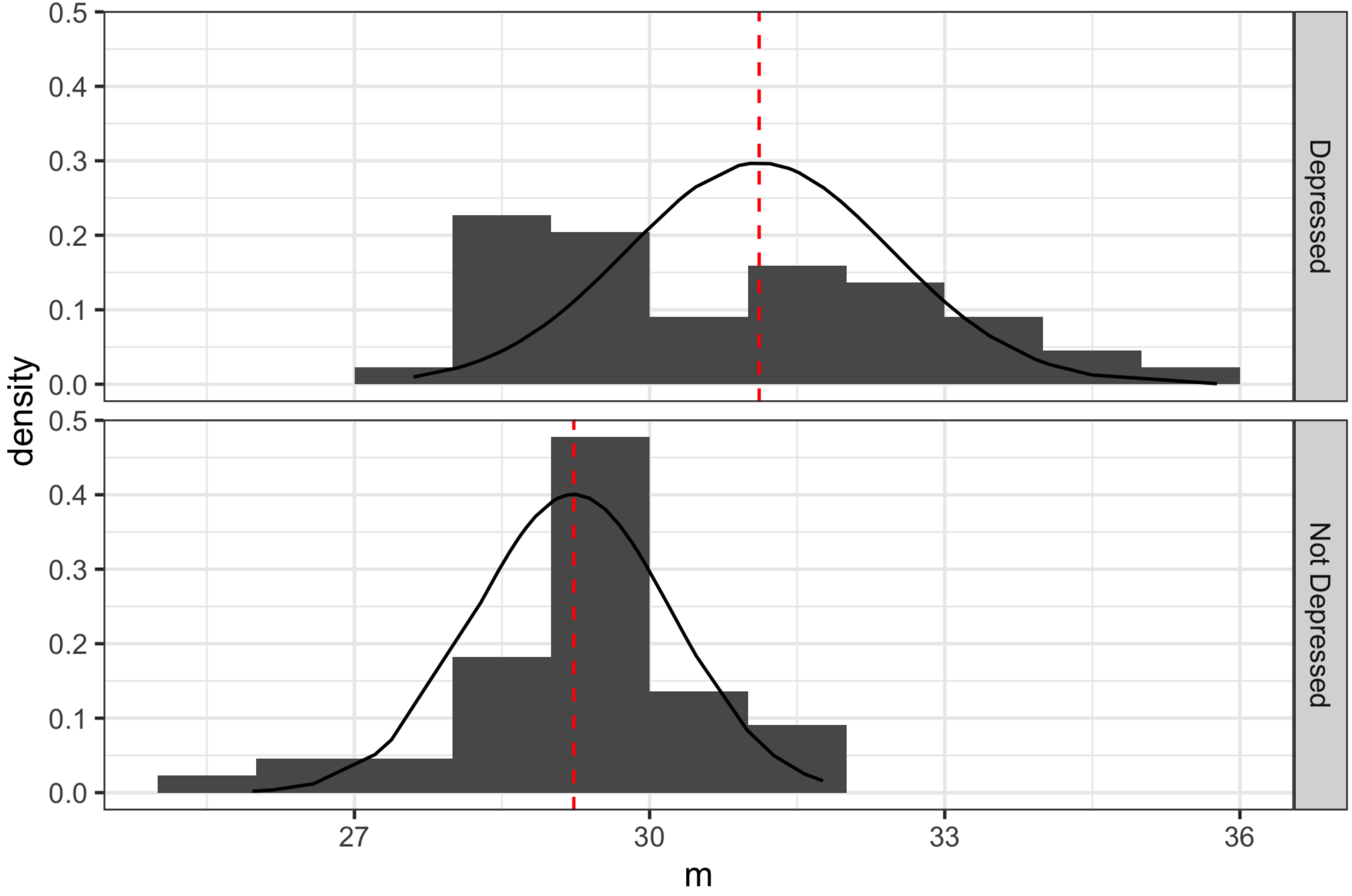
In words

- it almost doesn't matter where you start, if you take a sample, and form an average, then the outcome will be Normal!
- plus, you'll be centered around the true value!!
- PLUS, the more samples you use for your average, the smaller variance you'll have!!!

Central Limit Theorem

Example 1: mean BMI

Central Limit Theorem



Example 2: estimating proportion

How do we estimate the proportion of people with a disease in a population?

$$p = \frac{\text{number of people with disease}}{\text{total number of people}} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

where $x_i = 1$ if individual i has the disease, and $x_i = 0$ otherwise.

p is just an average! So as long as n is "large enough", we can utilize the CLT when calculating probabilities:

The prevalence of diabetes in the adult population of Wisconsin is approximately 10.6%. That is, if we randomly select an adult in Wisconsin, the probability of that adult having diabetes is approximately 10.6%. What is the probability that less than 10 individuals in a sample of 50 have diabetes? I.e. the proportion of individuals with diabetes is less than 0.2.

Example: estimating proportion

Binomial problem. $X_1, \ldots, X_{50}$ independent random variables indicating whether each of the 50 adults have diabetes (1) or not (0). Each random variable is a $\text{Bernoulli}(0.106)$ random variable. Let $Y = X_1 + \ldots + X_{50}$. We are interested in $P(Y < 10)$.

Example: estimating proportion

To calculate probabilities, we need to find the distribution of the random variable. Y is binomially distributed with $n = 50$ and $p = 0.106$. This distribution looks like this. We are interested in the area shaded red.

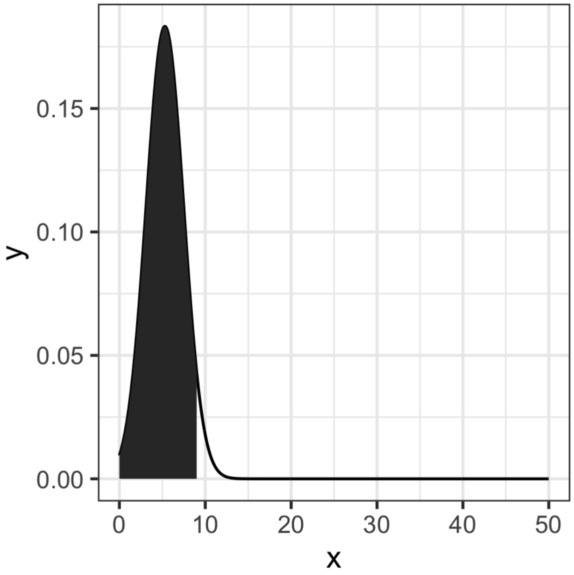
Example: estimating proportion

Can calculate directly:

$$P(Y < 10) = \sum_{i=0}^{9} P(Y = i) = \sum_{i=0}^{9} \binom{50}{i} 0.106^i (1 - 0.106)^{50-i} = 0.96505$$

Example: estimating proportion

Or, using a normal approximation, we can find the area under the normal curve:



This area is 0.95541.