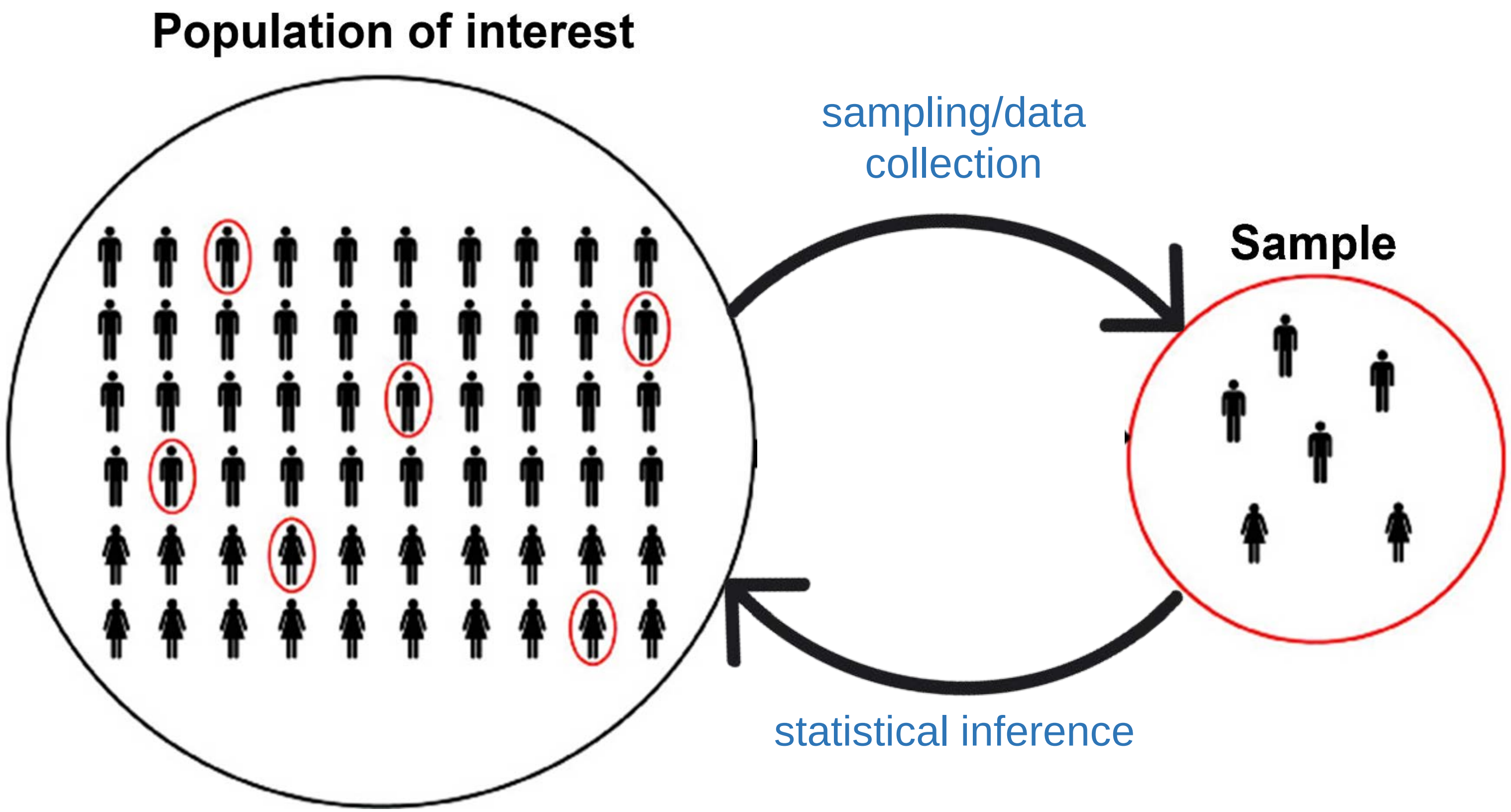
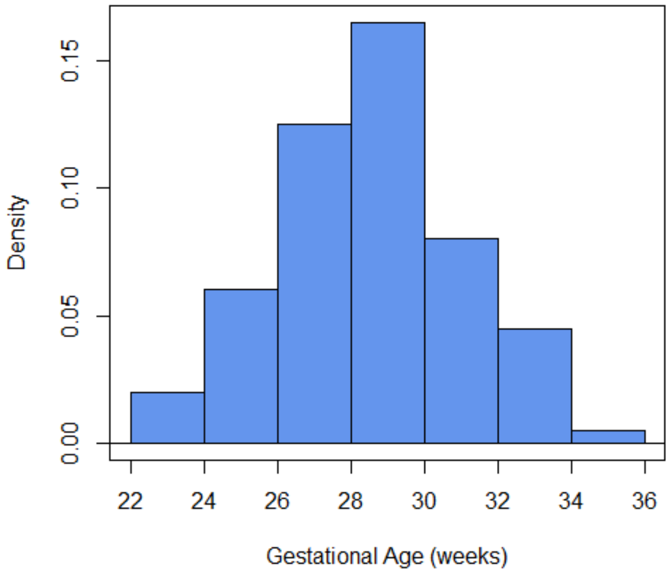
Normal distribution
Central Limit Theorem

Population vs. Sample

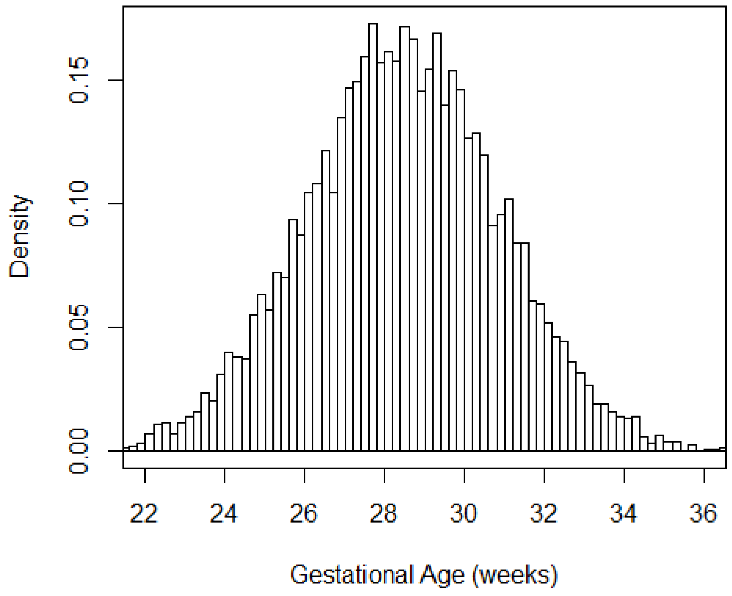


Population Distribution

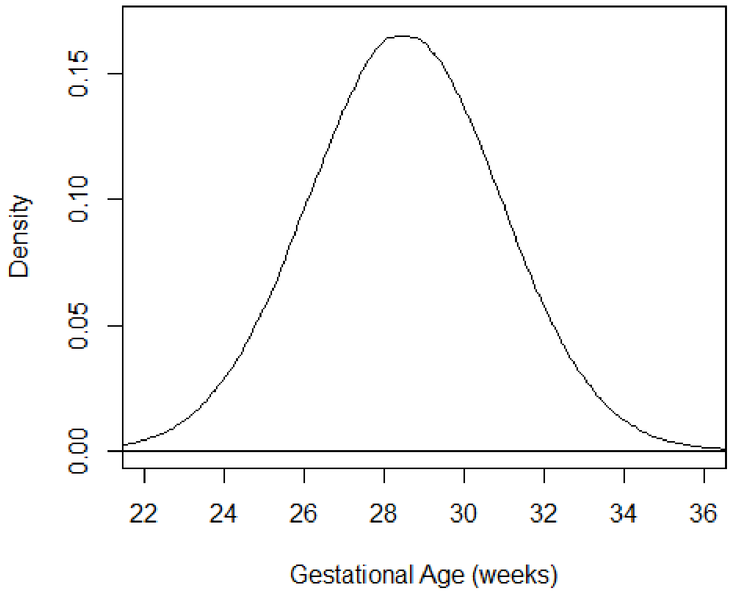
Our sample (n=100)



Large sample & narrow bars

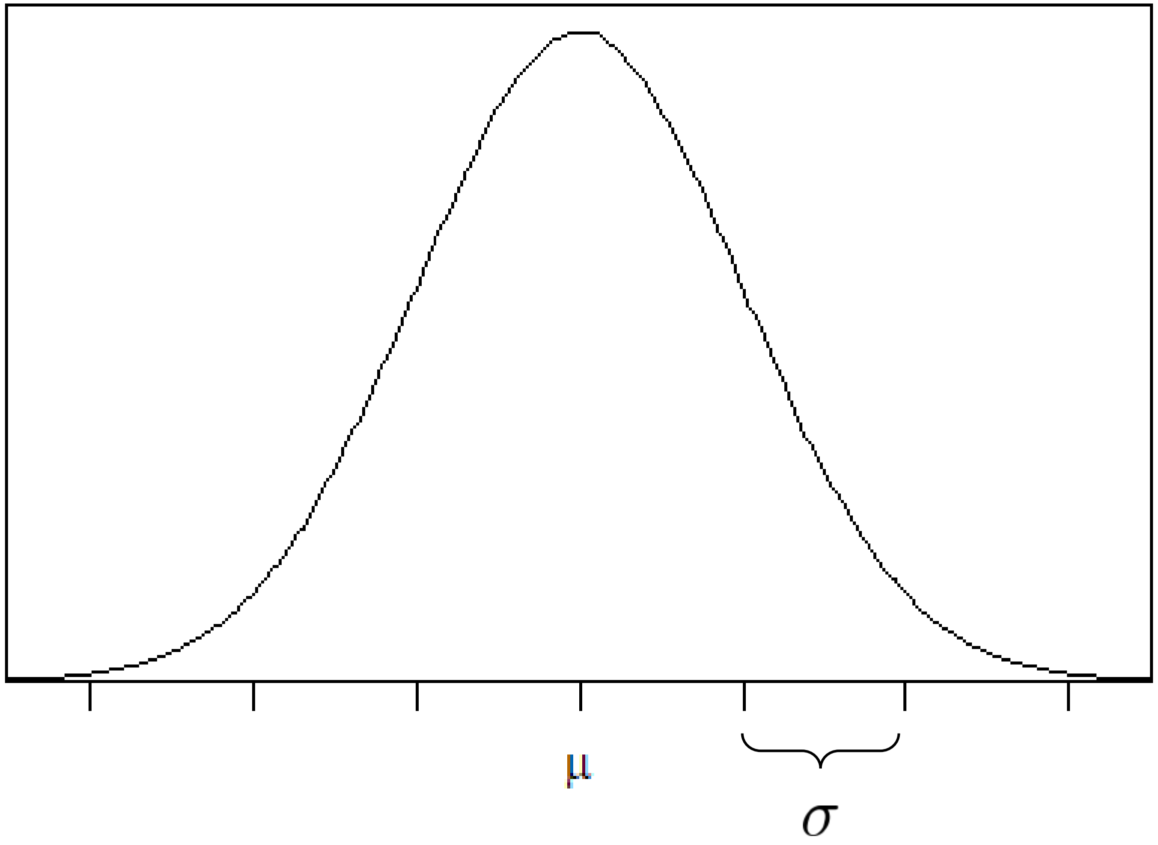


Population distribution



Normal Distribution

- Bell-shaped
- Centered at the mean, μ
- Spread determined by standard deviation, σ
  - Remember, we expect >99% of observations to fall within 3 SD of the mean



“Normally Distributed”

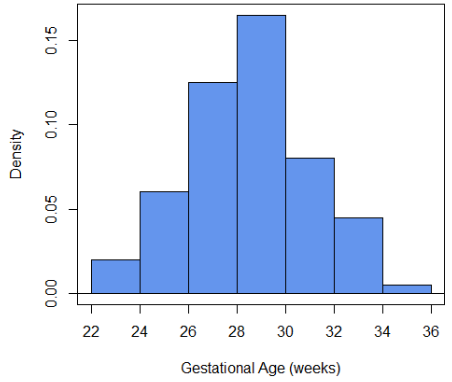
When we talk about a variable being “normally distributed”, we mean that the distribution of values in the population follows a bell-shaped curve.

Who Cares?

- Most widely used distribution in statistics
- Many continuous variables (blood pressure, weight, height, serum cholesterol level, IQ score, etc.) are normally distributed
- When a variable is normally distributed in the population, the math underlying our basic statistical tests works out really nicely
- Even when a variable is NOT normally distributed, a large enough sample size still gives us normality in the **sampling distribution**

Sampling Distribution

Sampling distribution $\neq$ Distribution of the sample



Sampling Distribution

Definition:

- Suppose we took a sample of size n and computed the sample mean, calling it $\bar{x}^{(1)}$.
- Now suppose we took a different sample of size n and computed its mean, calling it $\bar{x}^{(2)}$.
- If we repeated this process many times, say 1000 times, we would have 1000 sample means ($\bar{x}^{(1)}, \bar{x}^{(2)}, \dots, \bar{x}^{(1000)}$).
- The distribution of those 1000 sample means is called the **sampling distribution of the mean**.

Central Limit Theorem

If n is large enough, the sampling distribution of the mean will be approximately normal, regardless of the underlying population distribution that the samples were drawn from.

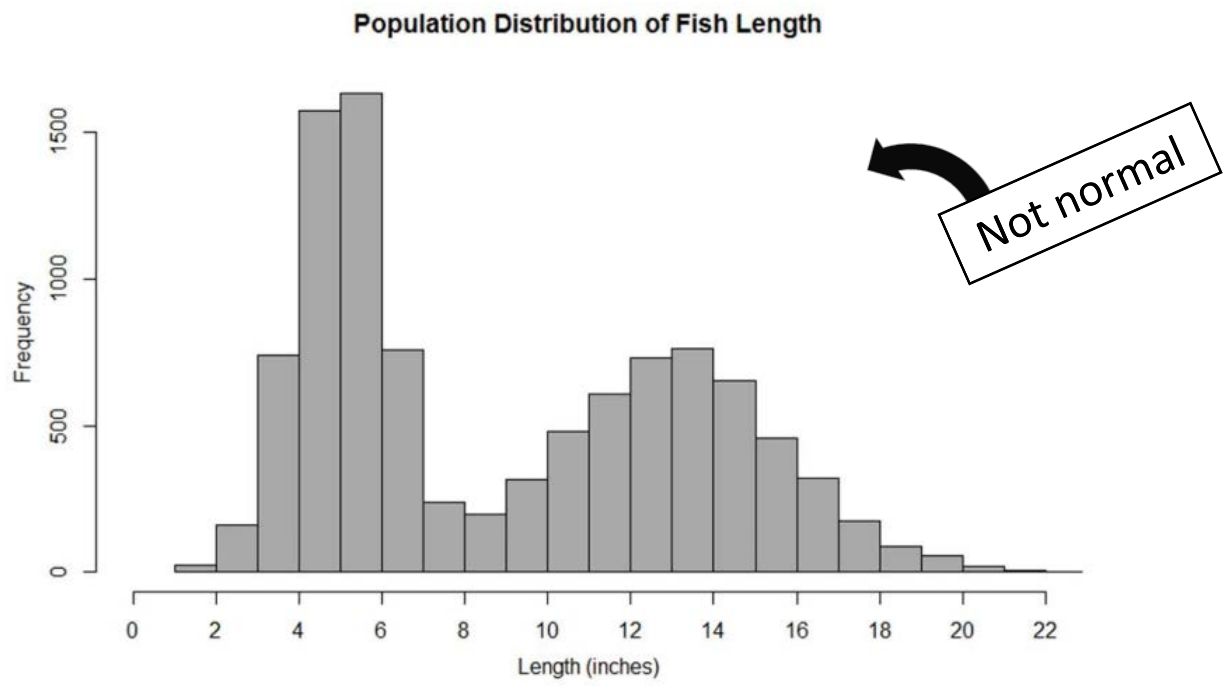
Example: Fishing

- Suppose there is a lake with 10,000 fish in it. These fish are primarily of two different species – one that is small and one that is large. We are interested in investigating the mean length of fish in the lake.

Example: Fishing

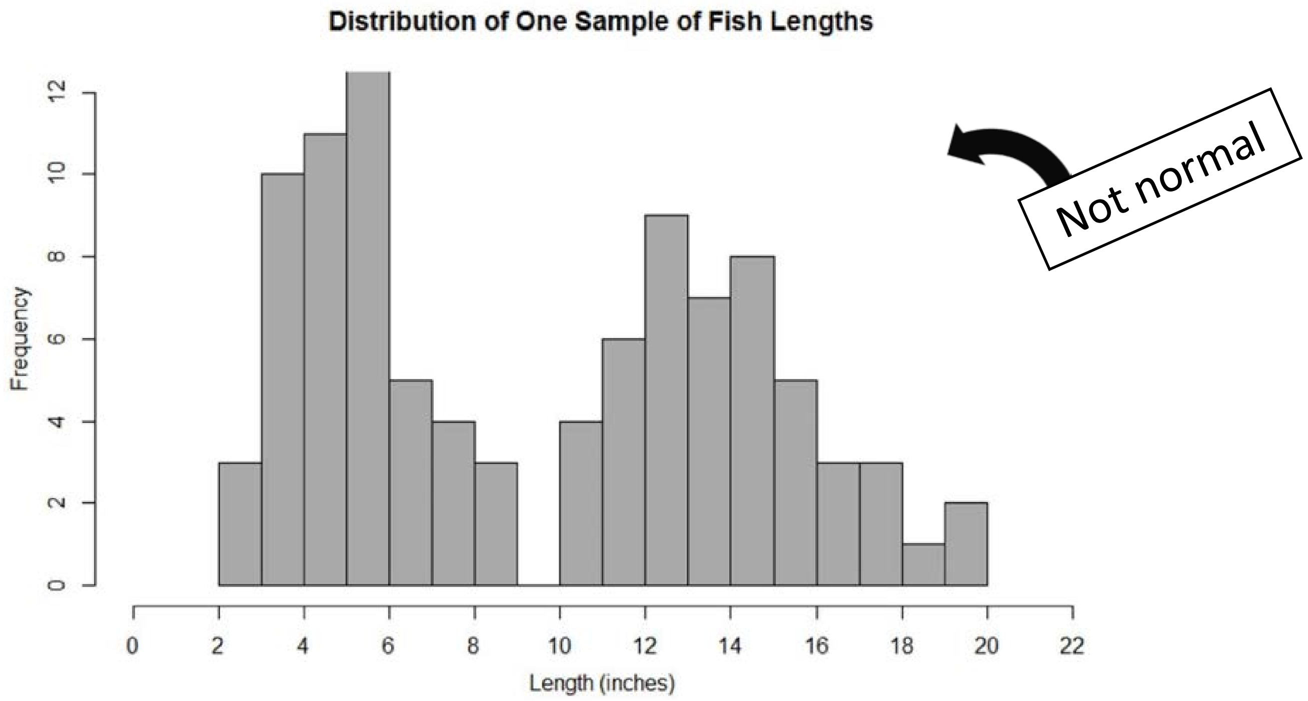
- Suppose there is a lake with 10,000 fish in it. These fish are primarily of two different species – one that is small and one that is large. We are interested in investigating the mean length of fish in the lake.

- If we could somehow measure the length of every fish in the lake, we would find the following **population** distribution of fish length for all 10,000 fish:



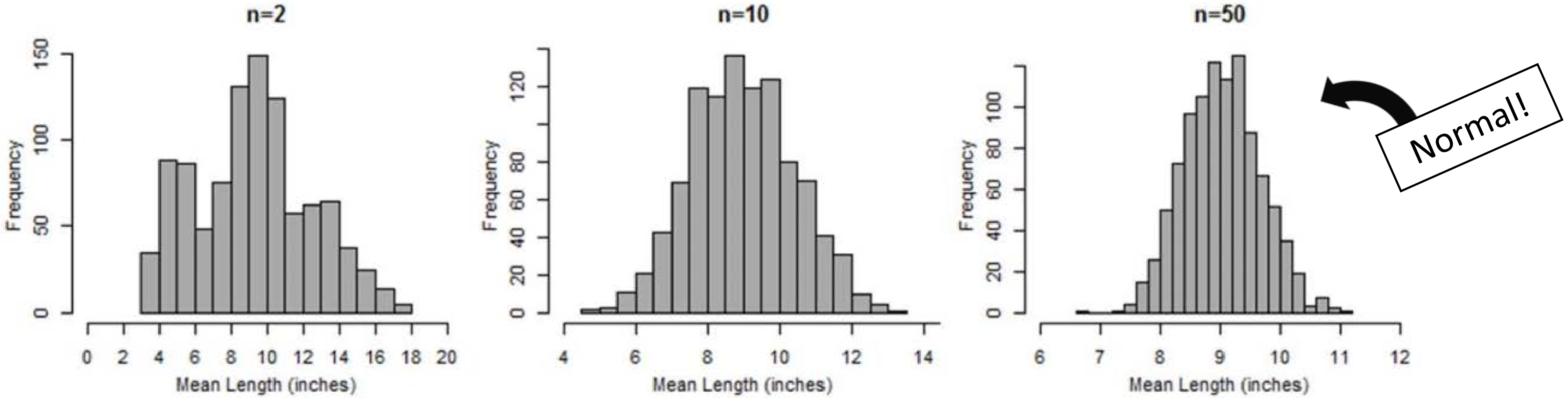
Example: Fishing

- However, we cannot measure the length of all 10,000 fish. Suppose instead that we catch 100 fish and measure their lengths, assuming that this is a random sample of the entire population. The distribution of the 100 fish lengths from this **sample** is shown below:



Example: Fishing

- Now, suppose that we went to the lake, caught a certain number of fish (n), and recorded the mean length of the fish. We do this for many days, assuming that the fish lengths don't change day-to-day. We can then plot this distribution of mean lengths from all of the days. This **sampling distribution of the mean** is shown below in three scenarios: one where we only catch 2 fish each day, one where we catch 10 fish each day, and one where we catch 50 fish each day.



Central Limit Theorem

If n is large enough, the sampling distribution of the mean will be approximately normal, regardless of the underlying population distribution that the samples were drawn from.

Central Limit Theorem

If n is large enough, the sampling distribution of the mean will be approximately normal, regardless of the underlying population distribution that the samples were drawn from.

How large is “large enough”?

- Depends on how normal the population distribution is
- If pop. distribution is normal, the sampling distribution of the mean will be normal regardless of the sample size
- The more abnormal the pop. distribution is, the larger the sample size has to be for the sampling distribution of the mean to be normal

CLT and Data Analysis

- Methods we'll talk about for the next 12 weeks rely on normality to be valid
- If the population distribution is normal, you're good
- If the population distribution is not normal, the CLT will kick in and help with your normality assumption as long as the sample size is large enough
- For small samples from a non-normal population distribution, stay tuned for **nonparametrics**

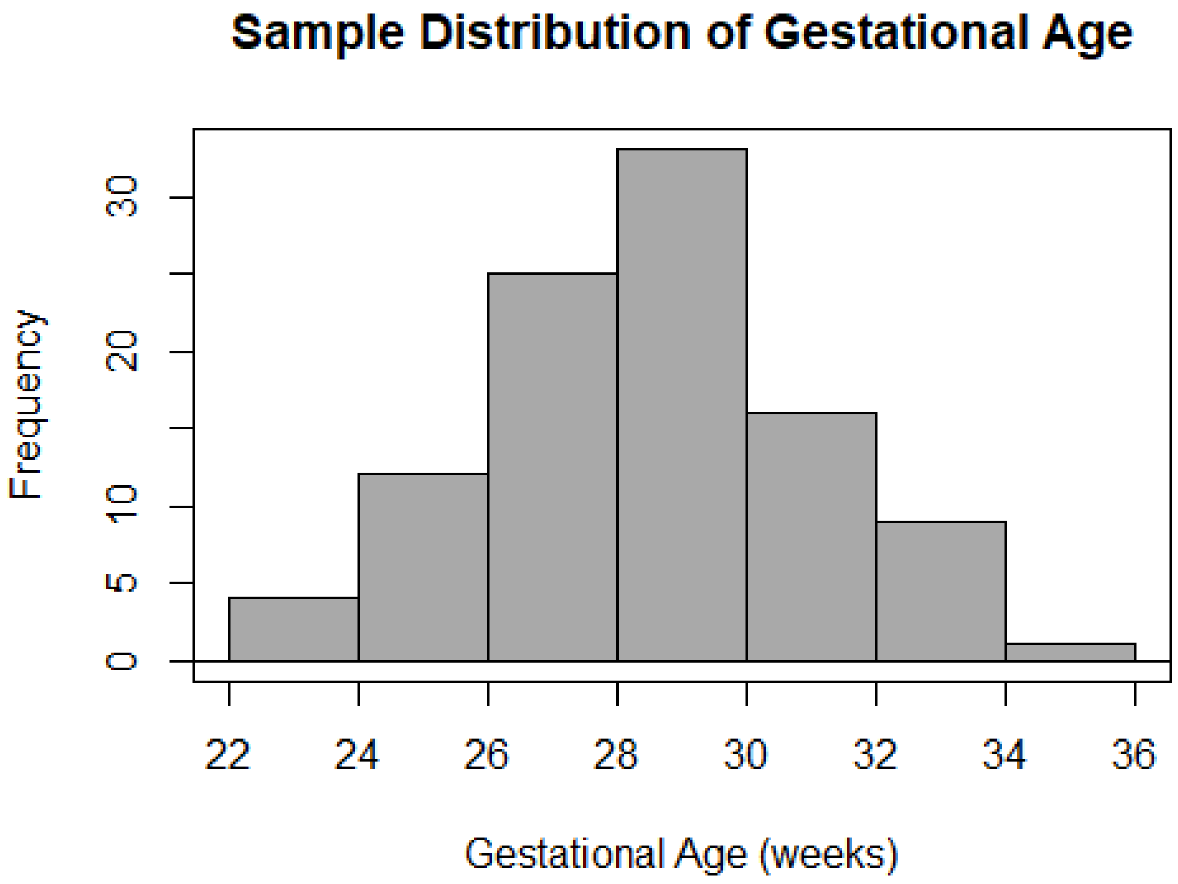
Low Birth Weight Data

- Information on 100 low birth weight infants born in two teaching hospitals in Boston, Massachusetts

Variable	Description
sex	Sex of the baby (Male, Female)
gestage	Gestational age at time of birth (weeks)
length	Length of the baby (cm)
birthwt	Birth weight of the baby (g)
headcirc	Baby's head circumference (cm)
apgar	Apgar score (integers, min=0, max=10). This is a scoring system used for assessing the clinical status of a newborn. 7 or higher is generally considered normal, 4-6 is low, and 3 or below is critically low.

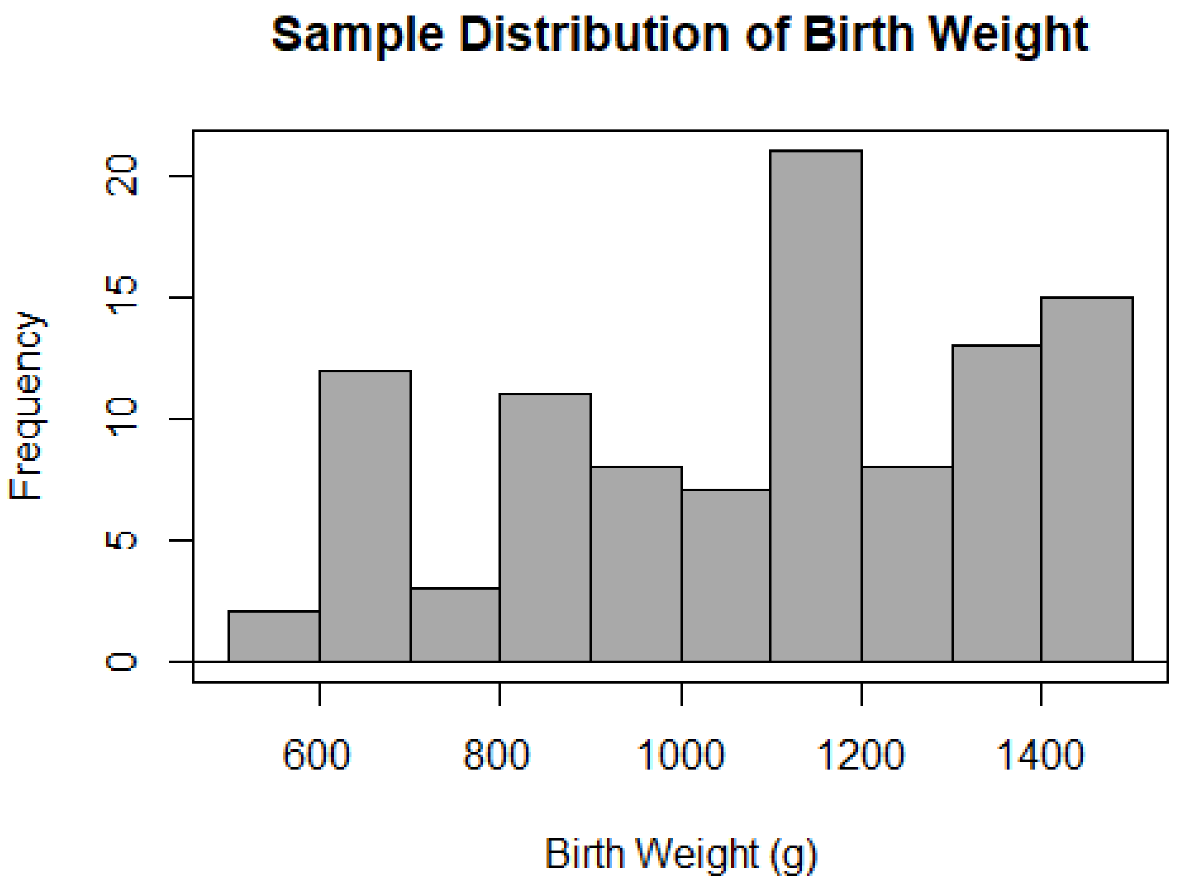
Find the dataset (lowbwt.xlsx) and the full data dictionary (lowbwt Data Dictionary.pdf) in the Data Module on the Canvas site

Example: Gestational Age



- Gestational age appears to be normally distributed in the sample
- Do you think the population distribution of gestational age is normal?
- Do you think the sampling distribution of the mean gestational age is normal?

Example: Birth Weight



- Birth weight does not appear to be normally distributed in the sample
- Do you think the population distribution of birth weight is normal?
- Do you think the sampling distribution of the mean birth weight is normal?

Important Points

- Difference between sample and population
- Normal distribution: bell-shaped, location and spread are determined by mean and standard deviation
- Concept of the sampling distribution of the mean
  - Relationship between population distribution, sample distribution, and sampling distribution of the mean
- How the Central Limit Theorem helps us assume normality for data analysis (and when does it not help)