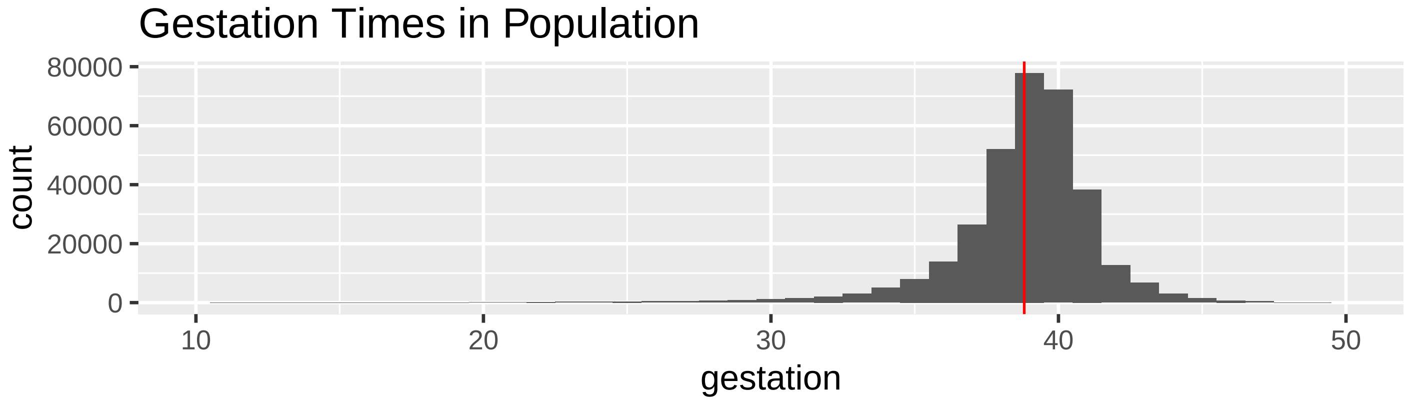
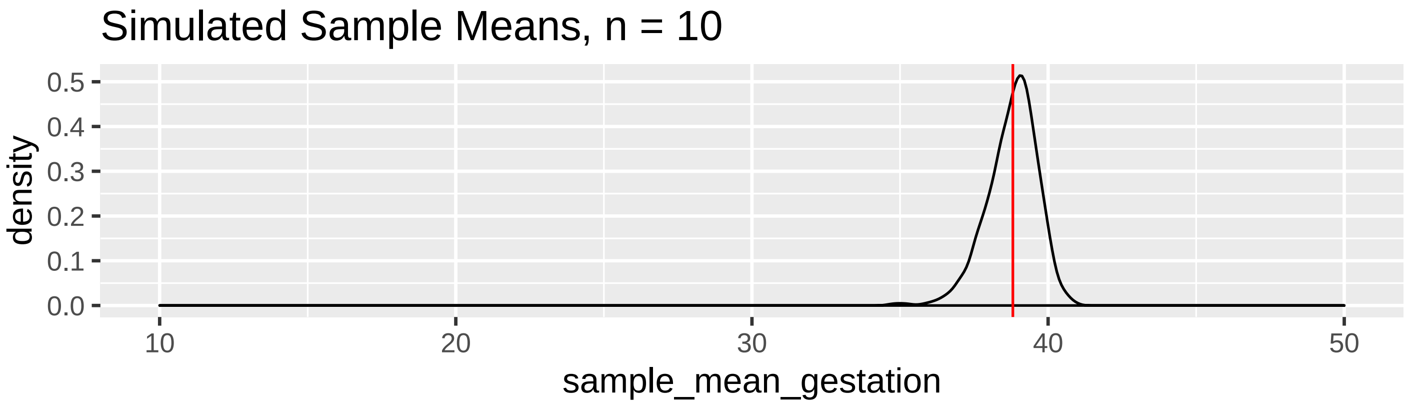
Sampling Distribution of the Sample Mean

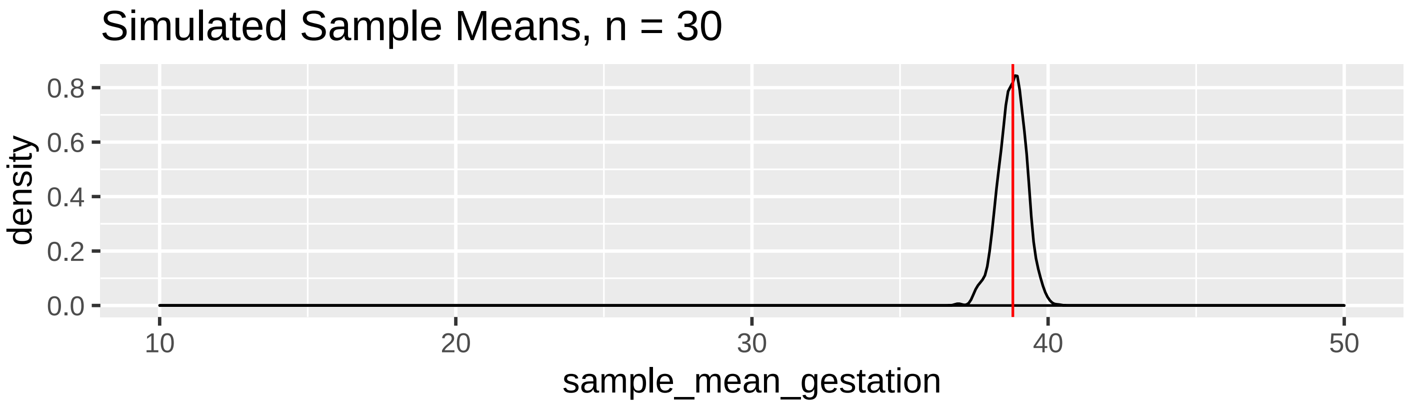
Our Goal: Estimate the mean of a quantitative variable, using data from a sample.

Sampling Distribution of the Sample Mean: The distribution of values of the sample mean that we would obtain from all possible samples of a certain size n.

Last class: Calculated the sample mean gestation time for many different samples from the population of all babies born in the US in December 1998:







Center: On average, the sample mean is close to the population mean

Spread: As n increases, there is less spread in the sample means from different samples

Shape: As n increases, the distribution of sample means becomes more symmetric.

Central Limit Theorem:

Suppose that the mean of a variable in the population is μ, and standard deviation is σ

The sampling distribution of $\bar{Y}$ based on n independent observations:

- has mean μ (on average, sample mean is equal to population mean)
- has standard deviation $\sigma/\sqrt{n}$ (more consistently near population mean for big n)
- is approximately normal, if n is large enough

Putting this together: If n is large enough, $\bar{Y} \sim \text{Normal}(\mu, \sigma/\sqrt{n})$ (approximately).

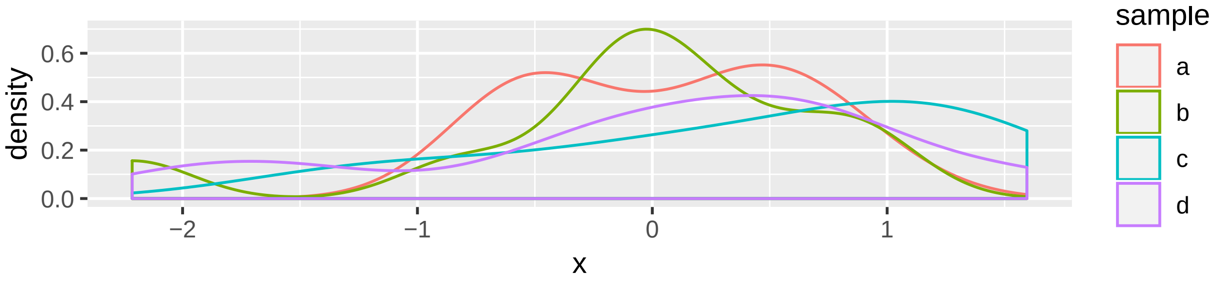
Conditions to Check:

1. There are no forms of **bias** in the sampling procedure.
2. We recorded a **quantitative** variable for each observational unit.
3. The **mean** is a reasonable summary of the center, and the **sample size** is large enough that the Central Limit Theorem applies.
4. The observational units in our sample are **independent** of each other.

Checking that using the mean is reasonable:

- Previously: unimodal, symmetric, no outliers.
- These conditions work for large sample sizes, but not for small sample sizes.

Here are density plots of 4 different samples of size 10 from a Normal(0, 1) distribution



For all 4 of these samples, using the mean is justified by theory. In practice, with small sample sizes the best we can do is to:

- calculate both the mean and the median and see if they are comparable (this will rule out serious outliers or serious skewness)
- think about the context of the data and whether we would expect there to be any outliers or serious skewness in the population.

Checking that the sample size is large enough:

- If the distribution looks good enough to calculate a mean, $n = 30$ is enough
- Smaller sample sizes are also OK if your distribution looks "good".

Example:

Suppose that in the population of all babies, the mean of all gestation times is 38.8 weeks and the standard deviation is 2.6 weeks. (This matches what we had for the babies born in 1998.)

(a) Suppose I plan to take a random sample of 100 babies and compute the sample mean gestation time, $\bar{Y}$ (of course I will avoid bias in my sampling). What is the sampling distribution of the sample mean? Check all conditions.

1. The sample is representative/there are no forms of **bias** in the sampling procedure.

2. We recorded a **quantitative** variable for each observational unit.

3. The **mean** is a reasonable summary of the center, and the **sample size** is large enough that the Central Limit Theorem applies.

4. The observational units in our sample are **independent** of each other.

(b) What is the probability that my sample mean will be more than 39.31 weeks? (For what proportion of samples would the sample mean be more than 39.31 weeks?)

Example:

(a) Suppose I plan to take a random sample of 64 houses in South Hadley, MA, and compute the sample mean square footage of the houses. According to town records, the mean square footage of all houses in the town is 1400 square feet, with a standard deviation of 100 square feet. The records also indicate that there is some skewness in the distribution of house square footage in the population, but it is not too severe. What is the sampling distribution of the sample mean? Check all conditions.

1. The sample is representative/there are no forms of **bias** in the sampling procedure.

2. We recorded a **quantitative** variable for each observational unit.

3. The **mean** is a reasonable summary of the center, and the **sample size** is large enough that the Central Limit Theorem applies.

4. The observational units in our sample are **independent** of each other.

(b) What is the probability that my sample mean will be more than 1425 square feet? (For what proportion of samples would the sample mean be more than 1425 square feet?)