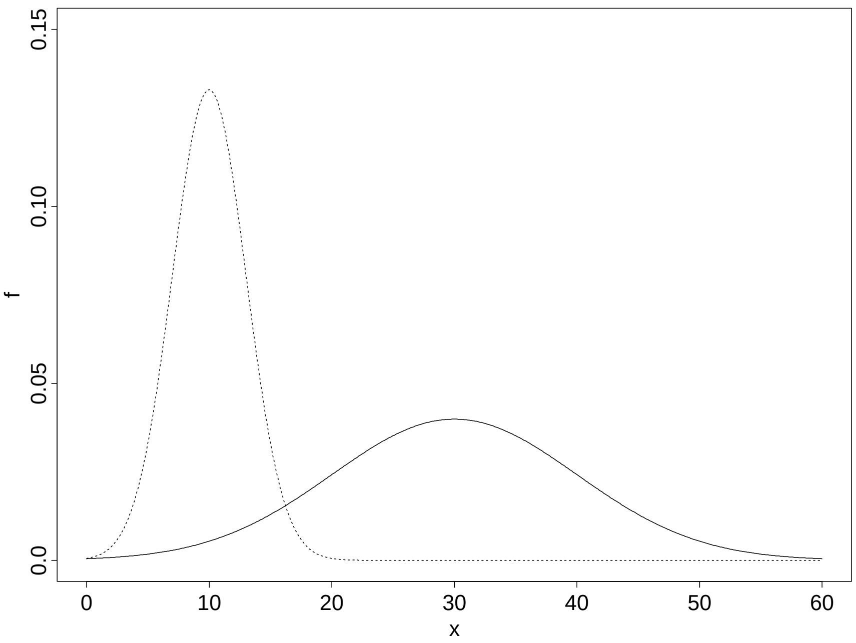
The Normal Distribution



The Normal Distribution is a common distribution in statistics. We've seen it before, it is a bell-shaped distribution. Many types of data follow this distribution—e.g., IQ test scores and physical measurement data like heights and weights.

- Let X denote a Normal random variable.
- μ denotes the mean of the distribution, and it determines the center of symmetry of the distribution.
- σ denotes the standard deviation, and it determines how wide the distribution is.

The Standard Normal Distribution

This distribution is simply a Normal Distribution with mean, $\mu = 0$, and standard deviation, $\sigma = 1$. We can convert Normal random variables to Standard Normal Random Variable as follows:

- Let $X \sim N(\mu, \sigma)$. Then

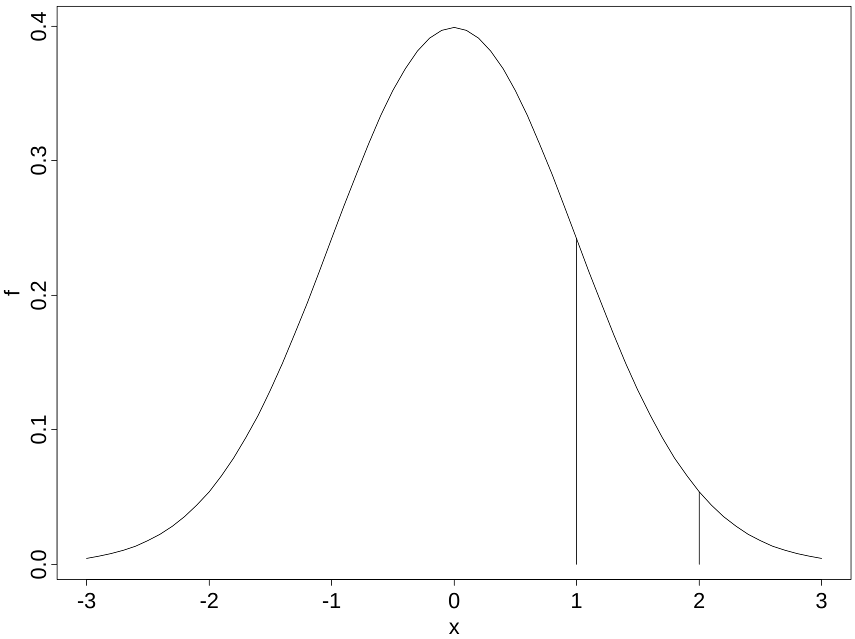
  $$Z = \frac{X - \mu}{\sigma}$$

  is a Standard Normal random variable.

- For example, If $X = 95$, and $X \sim N(100, 15)$, then

  $$\begin{aligned} Z &= \frac{95 - 100}{15} \\ &= \frac{-5}{15} \\ &= -0.33. \end{aligned}$$

Probability: Area Under the Curve



- $Z \sim N(0, 1)$
- $P(1 \leq Z \leq 2) = 0.136$, by calculating the area under the curve between the two vertical lines.

We can use the 68–95–99.7 percent rule to obtain an approximation to this value:

- Since 2 is two sd from the mean, $P(Z \leq 2) = 0.975$.
- Since 1 is one sd from the mean, $P(Z \leq 1) = 0.84$.
- We can combine these to find

$$\begin{aligned} P(1 \leq Z \leq 2) &= P(Z \leq 2) - P(Z \leq 1) \\ &= 0.975 - 0.84 \\ &= 0.135. \end{aligned}$$

Why Standardize?

We use the Standard Normal Distribution for the following reasons:

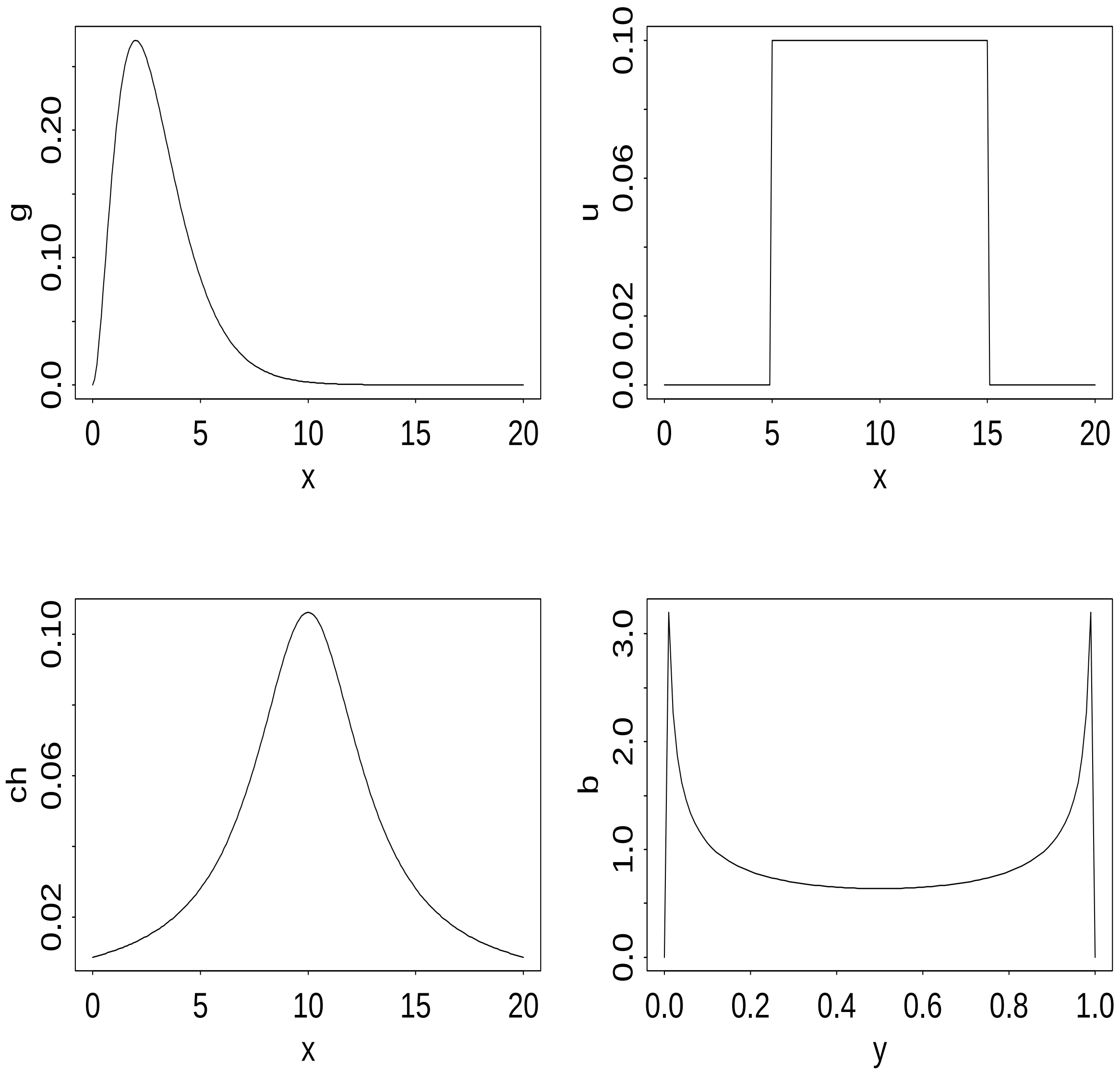
- So we can look up probabilities on a single table (note, for our purposes in this class, all we need to use is the 68–95–99.7 percent rule and linear interpolation).
- So we can compare Normal random variables on different scales (recall the homework question where you compared IQ test scores).

Why the Normal Distribution?

The Normal Distribution is important for a number of reasons:

- Lots of common data follow the Normal distribution.
- We can approximate other distributions with it.
- It has nice properties, like the 68–95–99.7 percent rule.
- The Central Limit Theorem.

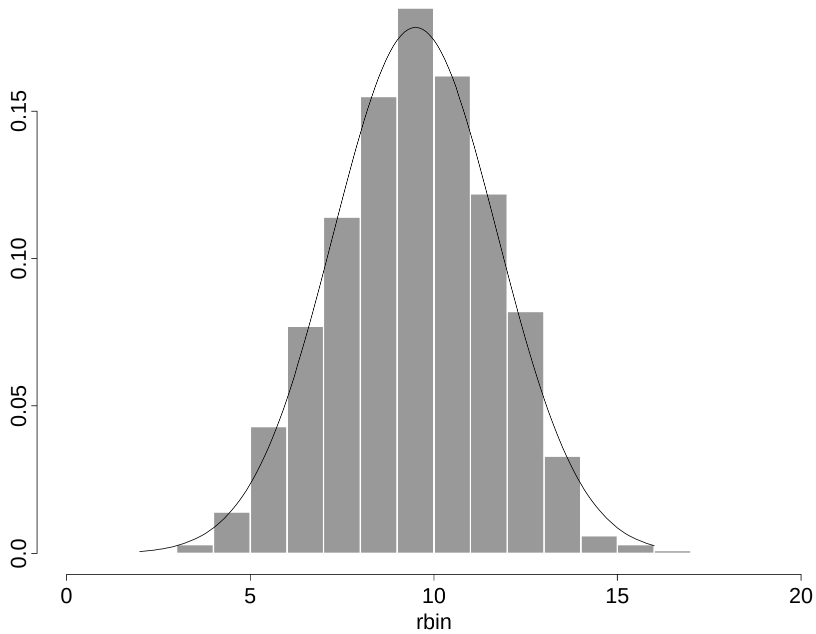
Non-normal Distributions



*

The binomial distribution is a Non-normal distribution. In some cases, however, it can be approximated by a Normal distribution. Here are some examples.

- Let X be a binomial random variable with $p = 0.5$ and $n = 20$.

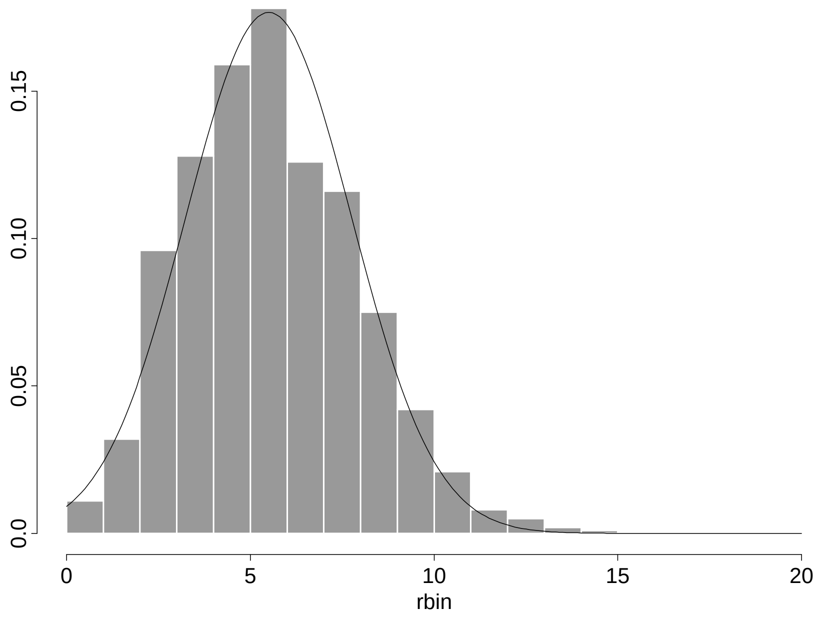


a Normal distribution with $\mu = np = 10$ and $\sigma = \sqrt{np(1-p)} = \sqrt{5}$ gives a fairly good approximation.

In general, if X is a binomial random variable, with success probability p and n trials, then the distribution of X can be approximated by a Normal distribution with

1. $\mu = np$
2. $\sigma = \sqrt{np(1-p)}$

- Let $X_1, X_2, \ldots, X_k$ be binomial random variables with $p = 0.15$ and $n = 40$.

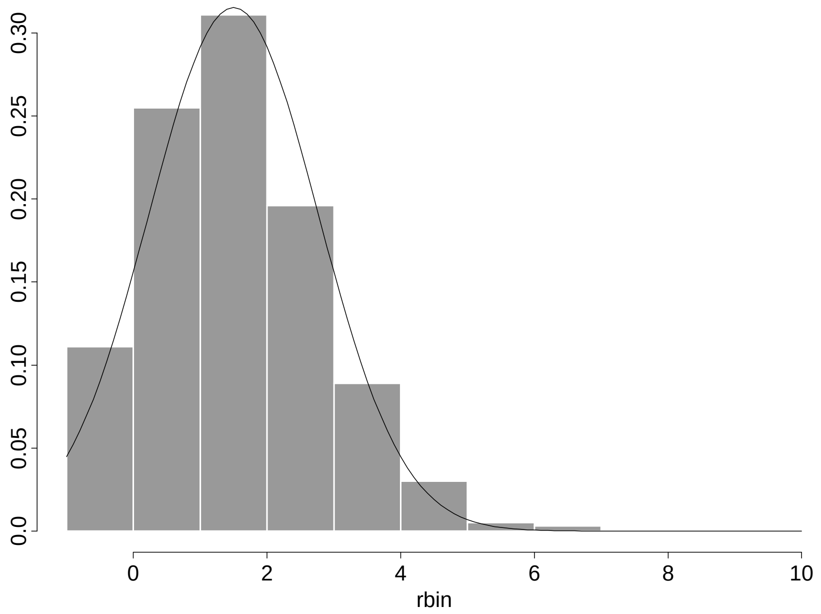


Here, a Normal distribution with $\mu = np = 6$ and $\sigma = \sqrt{np(1-p)} = \sqrt{5.1}$ approximates the binomial distribution fairly well.

The Normal Approximation to the Binomial

In general, this approximation works well if $np > 5$ *AND* $n(1-p) > 5$.

- In the first example, $np = n(1-p) = 10$.
- In the second example, $np = 6$ and $n(1-p) = 34$.
- Consider one more example: $X_1, X_2, \ldots, X_k$ are binomial random variables with $p = 0.2$ and $n = 10$.



Here the Normal approximation is not as good (though it's not bad!), since $np = 2$ in this case.

The Central Limit Theorem

This theorem tells us two important things about sample means of random variables. Suppose we take repeated samples of size n from a distribution that has some arbitrary shape. Then the Central Limit Theorem asserts that when n is large:

- The distribution of the sample means is Normal, with mean equal to the mean of the original distribution.
- The standard deviation of the sample means will equal the standard deviation of the original distribution divided by n, the size of the sample.

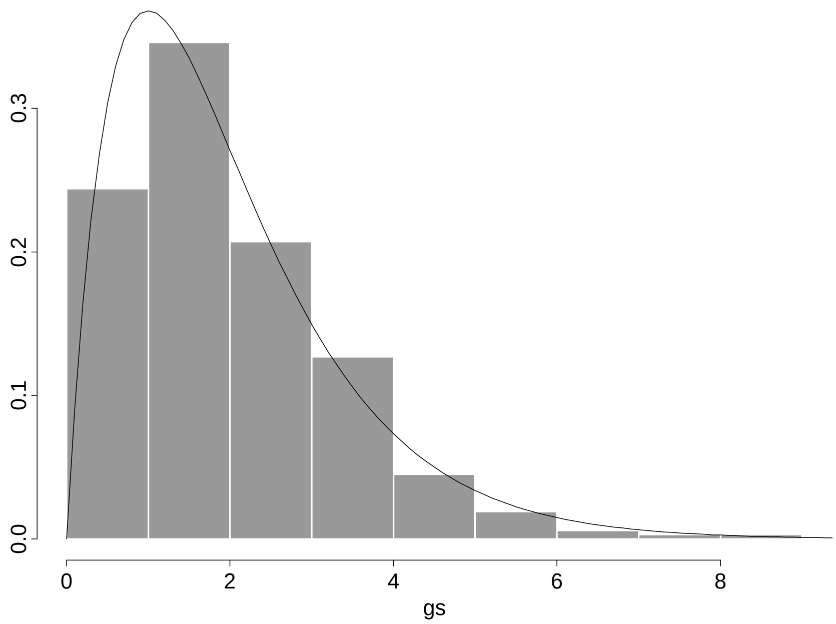
NOTE:

> The Central Limit Theorem applies, **regardless of the shape of the original distribution from which we sample**.

This means that if our data come from a skewed or multimodal distribution with mean μ and standard deviation σ, then the distribution of the mean of that distribution is $N(\mu, \frac{\sigma}{\sqrt{n}})$, where n is the size of the sample we used to calculate the mean.

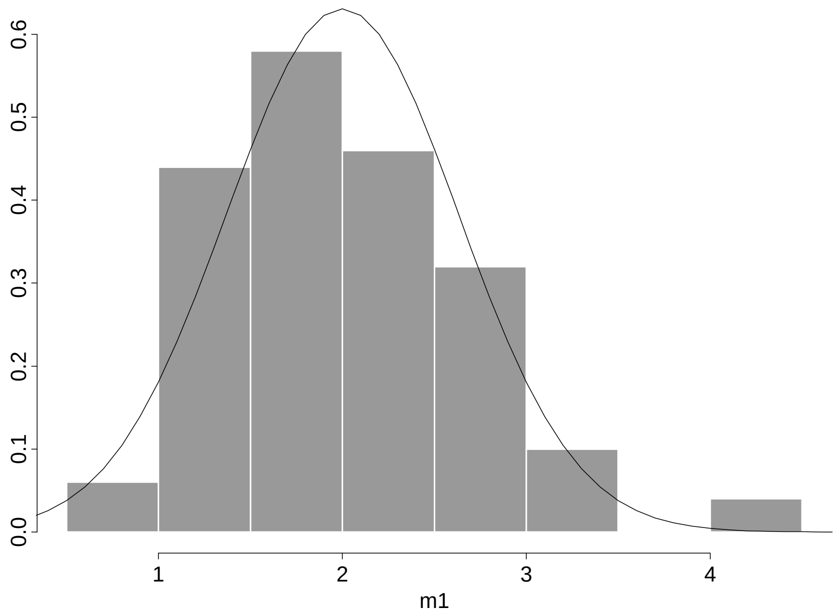
What Does "When n is large" Mean?

Usually, $n \geq 30$ is "large". Consider the following, highly skewed distribution. This distribution has mean $\mu = 2$ and standard deviation $\sigma = \sqrt{2}$.



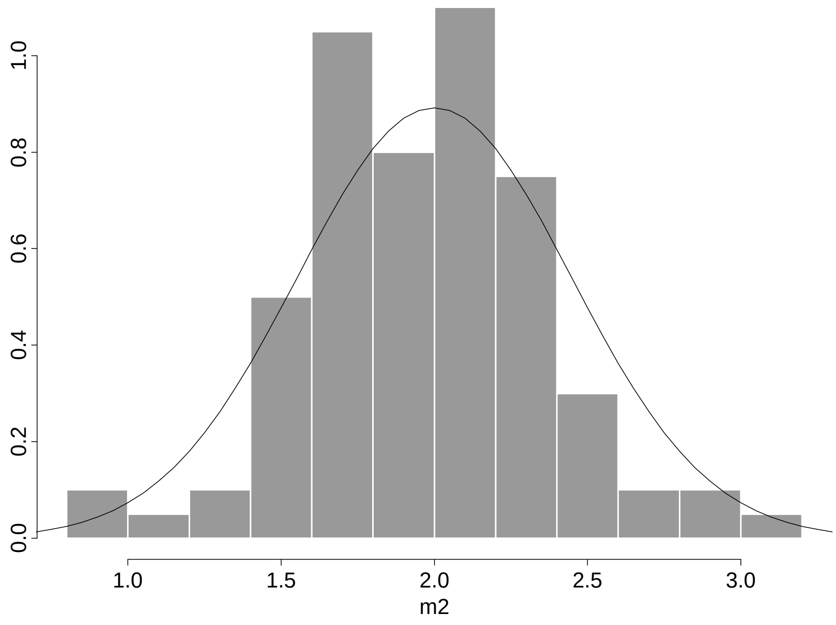
Now, take repeated samples from this distribution—first of size $n = 5$, then $n = 10$, and finally $n = 30$. For each sample, compute the mean, and then plot the histogram of the sample means for each sample size.

For $n = 5$:

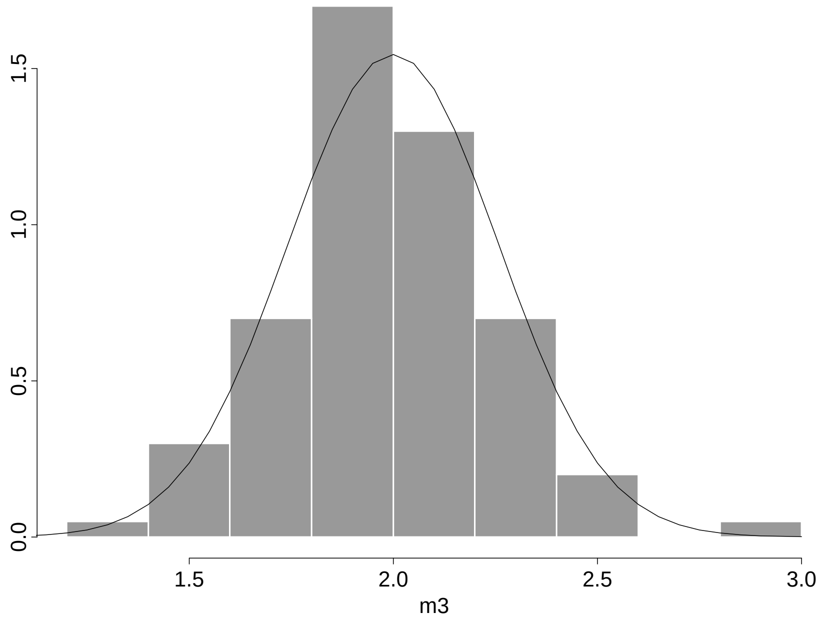


Already, the histogram looks fairly Normal!

For $n = 10$:



For $n = 30$:



Notice that the distribution of the sample mean is fairly normal even with samples of size 5.