M4S1 - Central Limit Theorem

Professor Jarad Niemi

STAT 226 - Iowa State University

September 18, 2018

Outline

- Sampling distribution
- Central Limit Theorem
- Standard error

Sampling distribution

Definition

A summary statistic is a numerical value calculated from the sample.

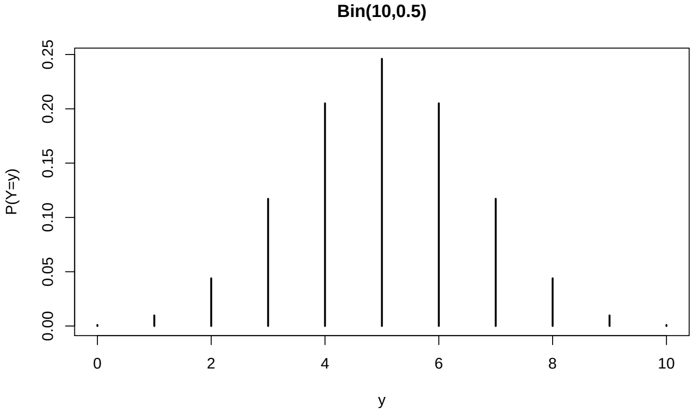
But this sample is only one of many possibilities. What could have happened if we had a different sample?

Definition

The sampling distribution of a statistic is the distribution of that statistic over different samples of a fixed size.

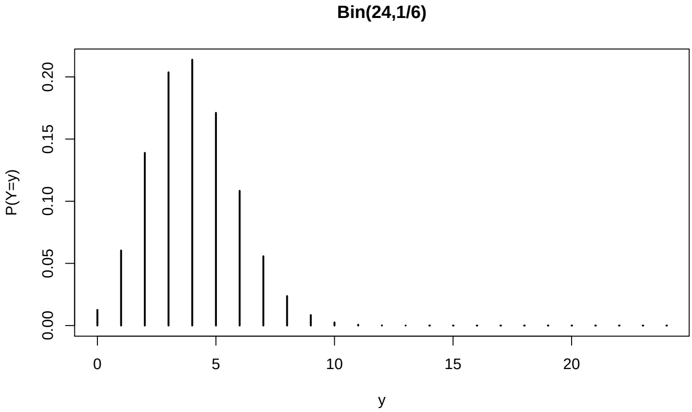
Flipping a coin

Suppose we repeatedly tossed a fair coin 10 times and recorded the number of heads. The sampling distribution is the binomial distribution with 10 attempts and probability of success 0.5.



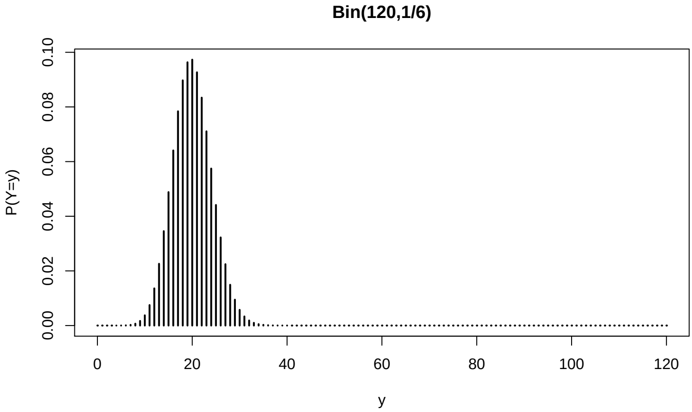
Rolling a die

Suppose we repeatedly rolled a fair 6-sided die 24 times and recorded the number of 1s. The sampling distribution is the binomial distribution with 24 attempts and probability of success $1/6$.



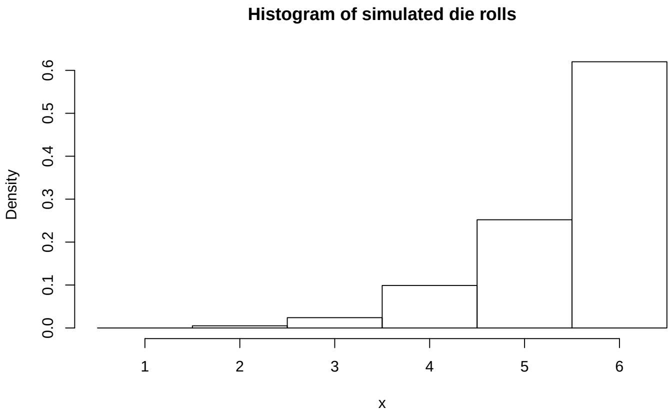
Rolling a die

Suppose we repeatedly rolled a fair 6-sided die 120 times and recorded the number of 1s. The sampling distribution is the binomial distribution with 120 attempts and probability of success $1/6$.



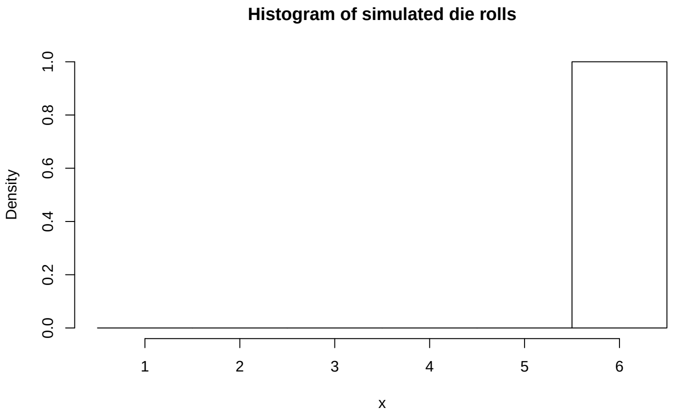
Rolling a die

Suppose we repeatedly rolled a fair 6-sided die 5 times and recorded the maximum. It's hard to analytically determine what happens, but we can use a computer to perform the experiment.



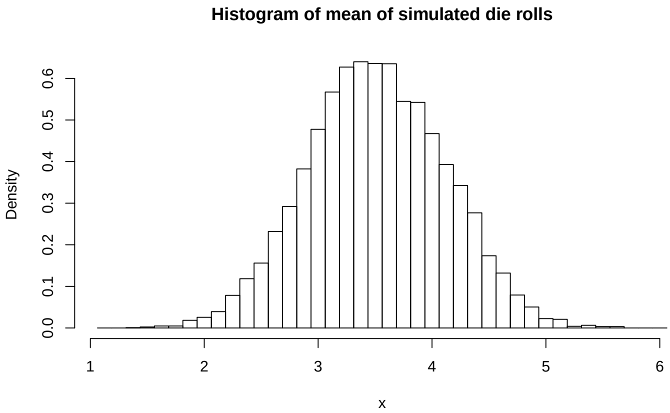
Rolling a die

Suppose we repeatedly rolled a fair 6-sided die 50 times and recorded the maximum. It's hard to analytically determine what happens, but we can use a computer to perform the experiment.



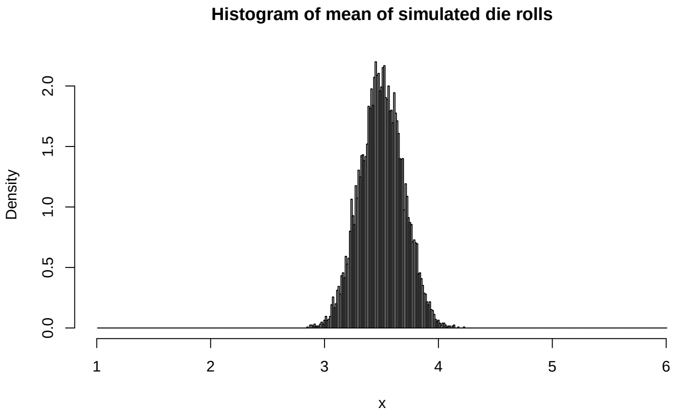
Sample mean

Suppose we repeatedly rolled a fair 6-sided die 8 times and recorded the mean. It's hard to analytically determine what happens, but we can use a computer to perform the experiment.



Sample mean

Suppose we repeatedly rolled a fair 6-sided die 80 times and recorded the mean. It's hard to analytically determine what happens, but we can use a computer to perform the experiment.



Central Limit Theorem

Theorem

Suppose you have a sequence of independent and identically distributed random variables $X_1, X_2, \ldots$ *with mean* $E[X_i] = \mu$ *and variance* $Var[X_i] = \sigma^2$*. The Central Limit Theorem (CLT) says the sampling distribution of the sample mean converges to a normal distribution. Specifically*

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \to N(0,1) \quad \text{as} \quad n \to \infty$$

where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$*. Thus, for large* n*, we can approximate the sample mean by a normal distribution, i.e.*

$$\overline{X} \mathrel{\dot\sim} N(\mu, \sigma^2/n)$$

where $\dot\sim$ *means "approximately distributed." The standard deviation of the sampling distribution of a statistic is known as the standard error, i.e.* $\sigma/\sqrt{n}$ *is the standard error from the CLT.*

Mean of the sample mean

Recall the following property:

$$E[aX + bY + c] = aE[X] + bE[Y] + c$$

If we have $E[X_i] = \mu$ for all i, then

$$\begin{aligned} E[\overline{X}] &= E\left[\frac{1}{n}\sum_{i=1}^n X_i\right] \\ &= \frac{1}{n}E\left[\sum_{i=1}^n X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^n E[X_i] \\ &= \frac{1}{n}\sum_{i=1}^n \mu \\ &= \frac{1}{n}n \cdot \mu \\ &= \mu \end{aligned}$$

Variance of the sample mean

Recall the following property for independent random variables X and Y:

$$Var[aX + bY + c] = a^2 Var[X] + b^2 Var[Y]$$

If we have $Var[X_i] = \sigma^2$ for all i, then

$$\begin{array}{rl} Var[\overline{X}] & = Var\left[\frac{1}{n}\sum_{i=1}^n X_i\right] \\ & = \frac{1}{n^2} Var\left[\sum_{i=1}^n X_i\right] \\ & = \frac{1}{n^2}\sum_{i=1}^n Var[X_i] \\ & = \frac{1}{n^2}\sum_{i=1}^n \sigma^2 \\ & = \frac{1}{n^2} n \cdot \sigma^2 \\ & = \sigma^2/n \end{array}$$

Sampling distribution of sample mean

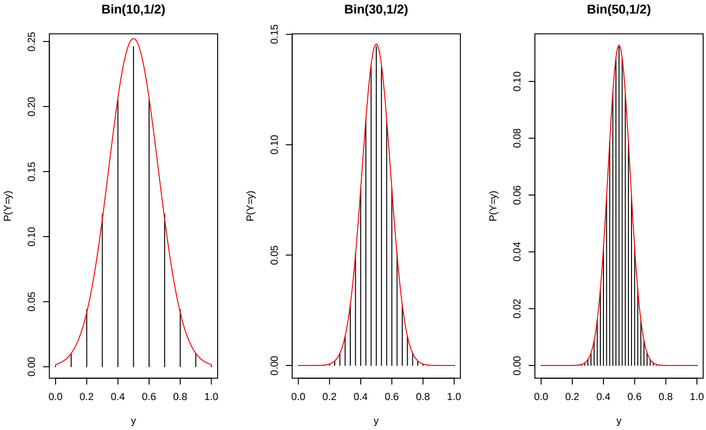
If $X_1, X_2, \ldots$ are a sequence of independent and identically distributed random variables with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2$, then

$$E[\overline{X}_n] = \mu \qquad Var[\overline{X}_n] = \sigma^2/n$$

for any n. The CLT says that, as n gets large, the sampling distribution of the sample mean converges to a normal distribution.

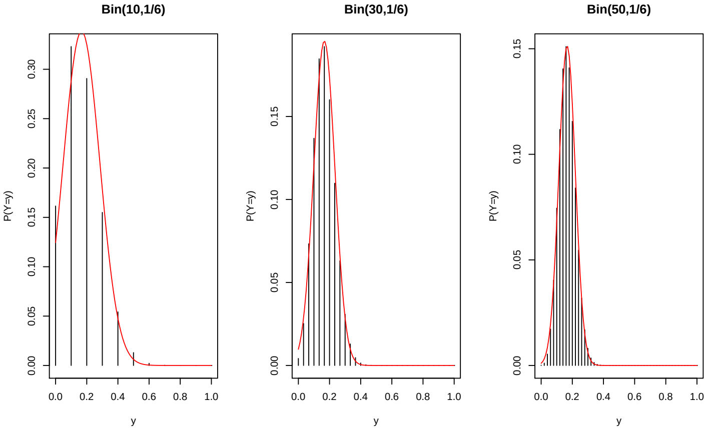
Coin flipping

Sampling distribution for the proportion of heads on an unbiased coin flip.



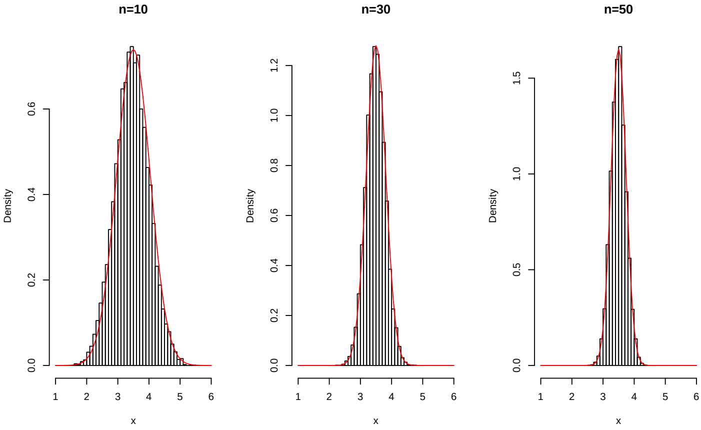
Die rolling

Sampling distribution for the proportion of 1s on an unbiased 6-sided die roll.



Die rolling

Sampling distribution for the sample mean of an unbiased 6-sided die roll.



Welfare

A certain group of welfare recipients receives SNAP benefits of \$110 per week with a standard deviation of \$20. A random sample of 30 people is taken and sample mean is calculated.

- What is the expected value of the sample mean?
  Let X_i be the SNAP benefit for individual i. We know $E[X_i] = \$110$ and $Var[X_i] = \$20^2$. Thus, $E[\overline{X}_{30}] = \$110$.
- What is the the standard error of the sample mean?
  The standard error is $\sigma/\sqrt{n} = \$20/\sqrt{30} \approx \$3.65$.
- What is the approximate probability the sample mean will be greater than \$120?
  We know $\overline{X}_{30} \dot{\sim} N(\$110, \$3.65^2)$.

$$
\begin{aligned}
P(\overline{X} > \$120) &= P\left(\frac{\overline{X}-\$110}{\$3.65} > \frac{\$120-\$110}{\$3.65}\right) \\
&\approx P(Z > 2.74) \\
&= 1 - P(Z < 2.74) \\
&= 1 - 0.9969 = 0.0031
\end{aligned}
$$