M4S1 - Central Limit Theorem

Professor Jarad Niemi

STAT 226 - Iowa State University

September 25, 2018

Outline

- Sampling distribution
  - Standard error
- Central Limit Theorem
- Estimation
  - Bias
  - Variability

Sampling distribution

Definition

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

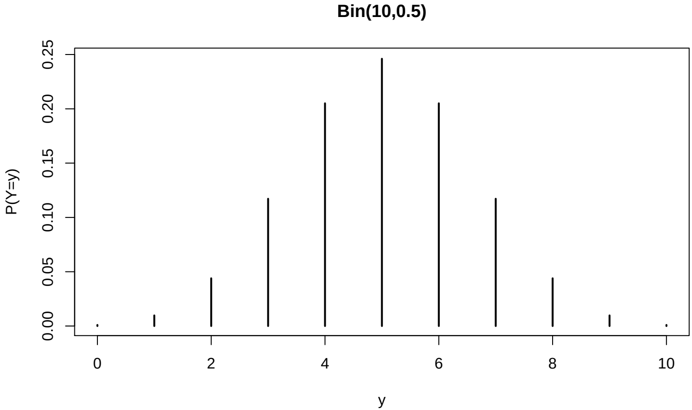
But this sample is only one of many possibilities. What could have happened if we had a different sample?

Definition

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

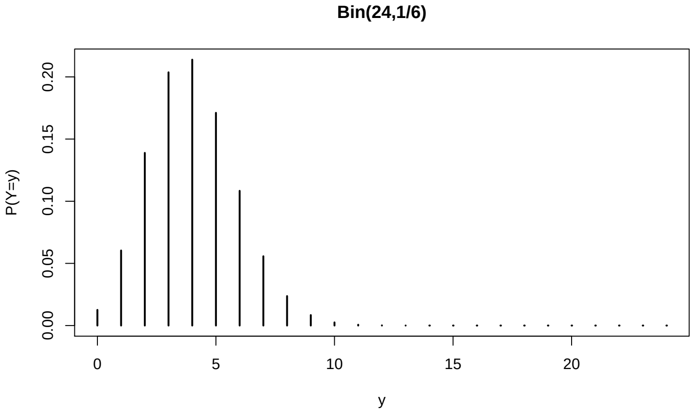
Flipping a coin

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.



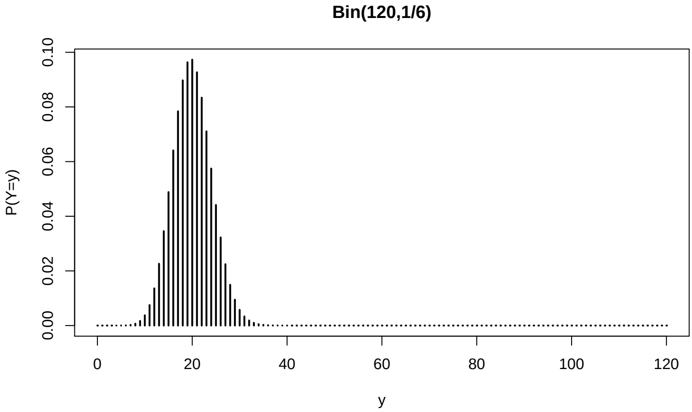
Rolling a die

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 $1/6$.



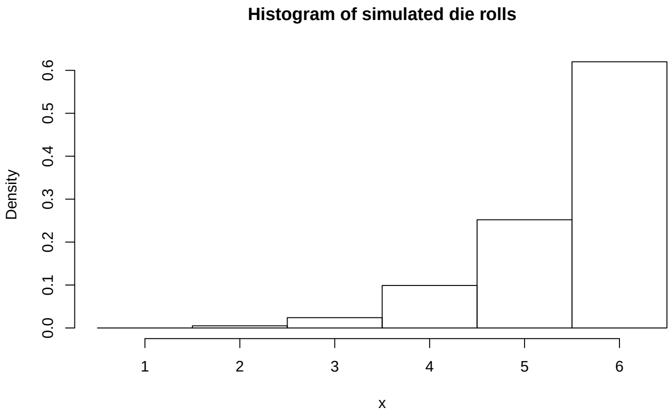
Rolling a die

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 $1/6$.



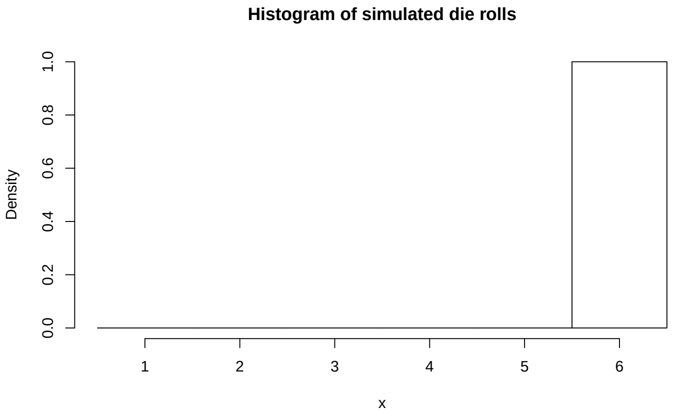
Rolling a die

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.



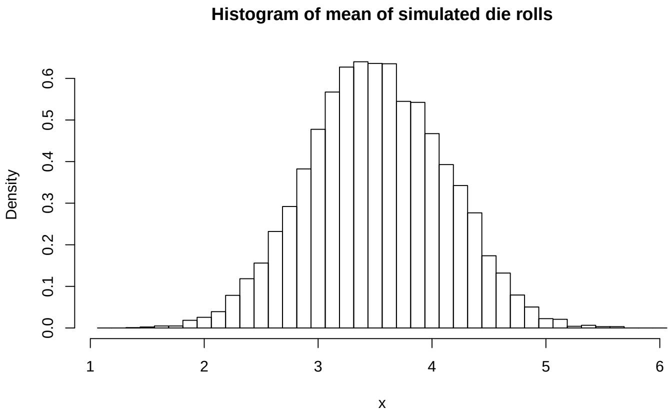
Rolling a die

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.



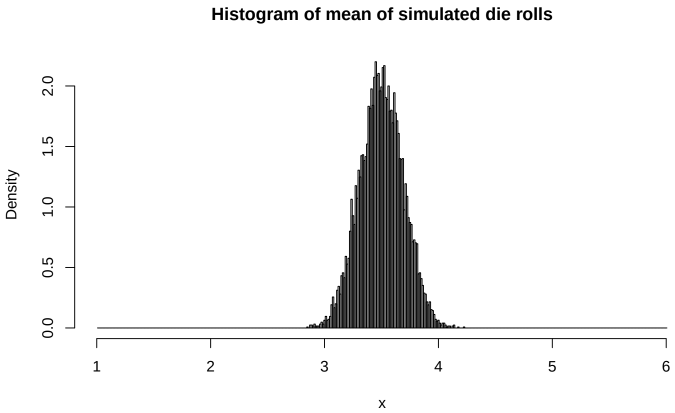
Sample mean

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 population mean* $E[X_i] = \mu$ *and population 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.*

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

where $\dot{\sim}$ *means "approximately distributed." The standard deviation of the sampling distribution of a statistic is known as the standard error (SE), 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

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

So the expectation/mean of the sample mean $(\overline{X})$ is the population mean μ.

Variance of the sample mean

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

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

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

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

So the variance of the sample mean ($\overline{X}$) is the population variance (σ^2) divided by the sample size (n). The standard error, which is the square root of the variance, is the population standard deviation (σ) divided by the square root of

Sampling distribution of sample mean

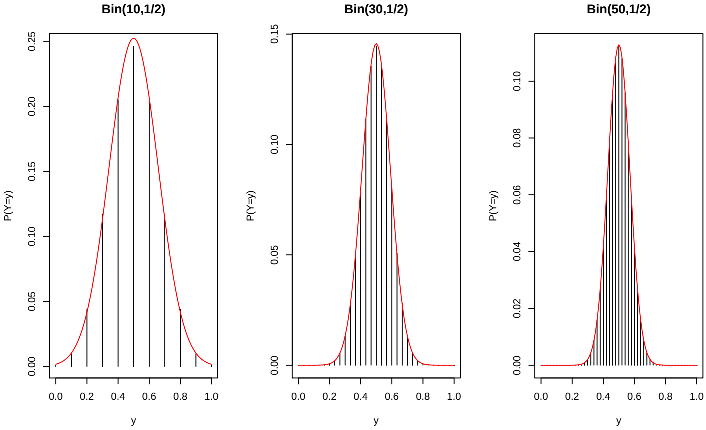
If $X_1, X_2, \ldots$ are a sequence of independent and identically distributed random variables with population mean $E[X_i] = \mu$ and population variance $Var[X_i] = \sigma^2$, then

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

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

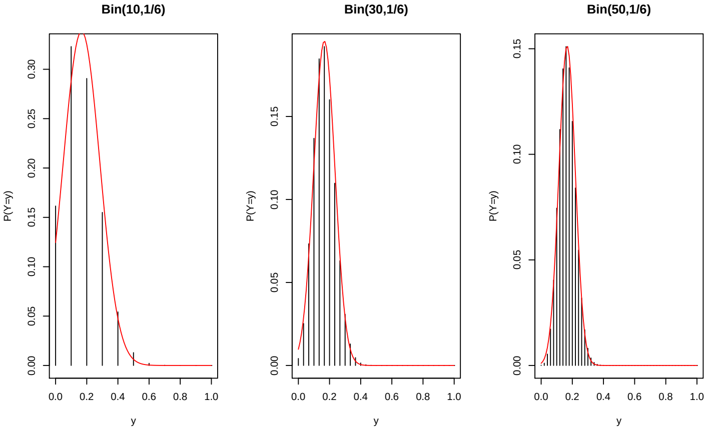
Coin flipping

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



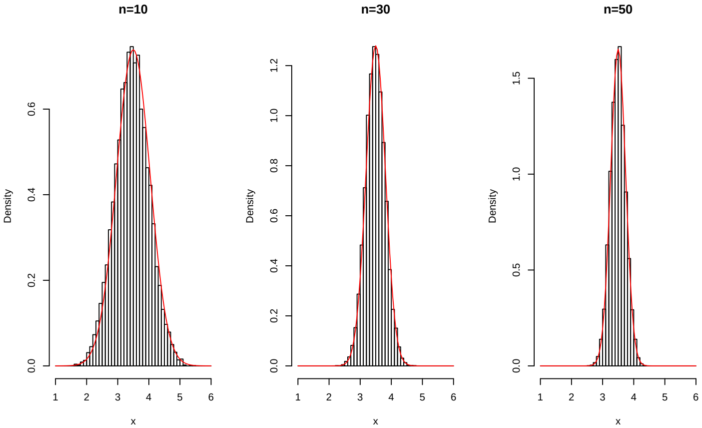
Die rolling

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 $\overline{X}_{30} \dot{\sim} N(\$110, \$3.65^2)$.

$$\begin{aligned} P(\overline{X}_{30} > \$120) &= P\left(\frac{\overline{X}_{30} - \$110}{\$3.65} > \frac{\$120 - \$110}{\$3.65}\right) \\ &\approx P(Z > 2.74) \\ &= 1 - P(Z < 2.74) \\ &= 1 - 0.9969 = 0.0031 \end{aligned}$$

Process to use CLT

Given a scientific question, do the following

1. Identify the random variables $X_1, X_2, \ldots$.
2. Verify these are independent and identically distributed.
3. Determine the expectation/mean and variance (or standard deviation) of the X_i.
4. Determine the sample size. Is the sample size large enough for the CLT to apply?
5. If yes, determine the approximate sampling distribution for the sample mean.
6. Write the scientific question in mathematical/probabilistic notation.
7. Calculate your answer.

Estimation

Definition

An estimator is a summary statistic that is used to estimate a population parameter.

Definition

An estimator is unbiased for a population parameter if the expectation/mean of the estimator is equal to the population parameter. Otherwise the estimator is biased.

The standard error of a statistic describes the variability of the statistic.

Sample mean

Let $X_1, X_2, \ldots$ be independent and identically distributed with population mean μ and population variance σ^2. Then the sample mean

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

has $E[\overline{X}] = \mu$ and standard error $SE[\overline{X}] = \sigma/\sqrt{n}$.

Thus, the sample mean is

- an unbiased estimator of the population mean and
- its variability (standard error) decreases by the square root of the sample size.

Bayesian estimator

In a Bayesian analysis, you specify your prior belief about μ before you observe the data. Suppose you are willing to specify that your prior belief about μ is normally distributed with mean m and variance v^2. Then you plan to collect data $X_1, X_2, \ldots$ that are independent and identically distributed with population mean μ and population variance σ^2.

A Bayesian estimator of the population mean μ is

$$\frac{1/v^2}{1/v^2 + n/\sigma^2} m + \frac{n/\sigma^2}{1/v^2 + n/\sigma^2} \overline{X},$$

and it has standard error

$$\frac{\sqrt{n/\sigma^2}}{1/v^2 + n/\sigma^2}.$$

Note that as $v^2 \to \infty$ (indicating a very uncertain prior belief), then this estimator becomes $\overline{X}$ which is unbiased and has standard error $\sigma/\sqrt{n}$.

Bayesian estimator (cont.)

The Bayesian estimator is biased because

$$\begin{aligned} E\left[\frac{1/v^2}{1/v^2+n/\sigma^2}m + \frac{1/v^2}{1/v^2+n/\sigma^2}\overline{X}\right] &= \frac{1/v^2}{1/v^2+n/\sigma^2}m + \frac{1/v^2}{1/v^2+n/\sigma^2}E[\overline{X}] \\ &= \frac{1/v^2}{1/v^2+n/\sigma^2}m + \frac{1/v^2}{1/v^2+n/\sigma^2}\mu \end{aligned}$$

but it has less variability because

$$\begin{aligned} \frac{\sqrt{n/\sigma^2}}{1/v^2+n/\sigma^2} &= \frac{1}{\frac{1/v^2}{n/\sigma^2}+\sqrt{n/\sigma^2}} \\ &< \frac{1}{\sqrt{n/\sigma^2}} \\ &= \frac{1}{\sqrt{n}/\sigma} \\ &= \sigma/\sqrt{n}. \end{aligned}$$

Thus the Bayesian estimator adds some bias to reduce variability. We call this the bias-variance tradeoff.

Bias and variability

Suppose you have the ability to take samples from one of two populations that both have the same mean. Population 1 has a standard deviation of 10 while population 2 has a standard deviation of 5. Due to the cost of sampling, you can either

1. take 100 samples of population 1 or
2. take 49 samples of population 2.

If your goal is to estimate the population mean using a sample mean, which of these two samples would you prefer to take?

The sample mean will have the same expectation/mean, so they are both unbiased. The standard error of population 1 is $10/\sqrt{100} = 10/10 = 1$ while the standard error of population 2 is $5/\sqrt{49} = 5/7 < 1$. Thus, on average, the sample mean from population 2 will be closer to the population mean than the sample mean from population 1. How few sample of population 2 would have the same standard error as the sample from population 1? 25