

Parameter and statistic

What is the average height of adult men in the US?

This is difficult to evaluate as there are 120 million adult men, but it can be *estimated* quite well with a relatively small sample.

The **population** consists of all adult men in the US (about 120 million).

A **parameter** is a quantity of interest about the population: the population average μ , or the population standard deviation σ .

A **statistic (estimate)** is the quantity of interest as measured in the sample: the **sample** average \bar{x} , or the sample standard deviation s .

The expected value

If we sample an adult male at random, then we expect his height to be around the population average μ , give or take about one standard deviation σ .

The **expected value** of one random draw is the population average μ .

How about \bar{x}_n , the average of n draws?

The **expected value of the sample average**, $E(\bar{x}_n)$, is the population average μ .

But remember that \bar{x}_n is a random variable because sampling is a random process.

So \bar{x}_n won't be exactly equal to $\mu = 69.3$ in: We might get, say, $\bar{x}_n = 70.1$ in. Taking another sample of size n might result in $\bar{x}_n = 69.1$ in.

How far off from μ will \bar{x}_n be?

The **standard error (SE)** of a statistic tells roughly how far off the statistic will be from its expected value.

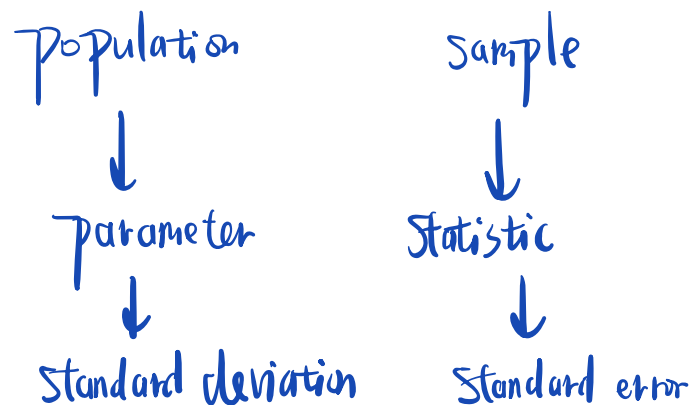
The standard error for the sample average

The **standard error (SE)** of a statistic tells roughly how far off the statistic will be from its expected value.

So the SE for a statistic plays the same role that the standard deviation σ plays for one observation drawn at random.

The **square root law** is key for statistical inference:

$$\text{SE}(\bar{x}_n) = \frac{\sigma}{\sqrt{n}}$$



The importance of the square root law is twofold:

- ▶ It shows that the SE becomes smaller if we use a larger sample size n . We can use the formula to determine what sample size is required for a desired accuracy.
- ▶ The formula for the standard error **does not depend on the size of the population**, only on the size of the sample.

Expected value and standard error for the sum

What if we are interested in the sum of the n draws, S_n , rather than the average \bar{x}_n ?

The sum and the average are related by $S_n = n\bar{x}_n$.

Both the expected value and the standard error can likewise be obtained by multiplying by n , therefore

$$E(S_n) = n\mu \quad SE(S_n) = \sqrt{n}\sigma$$






So the variability of the sum of n draws *increases* at the rate \sqrt{n} .

Expected value and standard error for percentages

What percentage of likely voters approve of the way the US President is handling his job?

The 'percentage of likely voters' is an average. This becomes clear by using the framework for counting and classifying:

- ▶ The population consists of all likely voters (about 140 million).
- ▶ Each likely voter falls into one of two categories: approve or not approve.
- ▶ Put the label '1' on each likely voter who approves, and '0' on each who doesn't.
- ▶ Then the number of likely voters who approve equals the sum of all 140 million labels.

▶     
1 0 0 1 0 = 2

- ▶ The percentage of likely voters who approve is the percentage of 1s among the labels.

Expected value and standard error for percentages

In a sample of n likely voters

- ▶ the number of voters in the sample who are approving is the sum S_n of the draws
- ▶ the percentage of voters approving is the percentage of 1s, which is $\frac{S_n}{n} \times 100\% = \bar{x}_n \times 100\%$

Therefore

$$E(\text{percentage of 1s}) = \mu \times 100\% \qquad SE(\text{percentage of 1s}) = \frac{\sigma}{\sqrt{n}} \times 100\%$$

where μ is the population average (=proportion of 1s) and σ is the standard deviation of the population of 0s and 1s.

All of the above formulas are for sampling with replacement. They are still approximately true when (sampling without replacement) if the sample size is much smaller than the size of the population.

Simple random sampling

Expected value and standard error when simulating

All of the above formulas are also true when the data are **simulated**, i.e. generated according to a probability histogram.

What are μ and σ in that case?

If the random variable X that is simulated has K possible outcomes x_1, \dots, x_K , then

$$\mu = \sum_{i=1}^K x_i \underbrace{\text{P}(X = x_i)}_{\frac{1}{K}} \quad \sigma^2 = \sum_{i=1}^K (x_i - \mu)^2 \underbrace{\text{P}(X = x_i)}_{\frac{1}{K}}$$

If the random variable X has a density f , such as when X follows the normal curve, then

$$\mu = \int_{-\infty}^{\infty} x f(x) dx \quad \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$



Applying the square root law

Toss a coin 100 times. How many 'tails' do you expect to see? Give or take how many?

1 = tails, 0 = heads, $p(1) = \frac{1}{2}$, $p(0) = \frac{1}{2}$

tails = sum of 100 draws

$$\begin{aligned} E(\text{sum}) &= n\mu, \quad \mu = 0 \cdot p(0) + 1 \cdot p(1) = \frac{1}{2} \\ &= 100 \times \frac{1}{2} \\ &= \underline{50} \end{aligned}$$

$$\text{give or take } SE(\text{sum}) = \sqrt{n}\sigma = \sqrt{100}\sigma = \sqrt{100} \cdot \sqrt{\frac{1}{4}} = \underline{5}$$

$$\sigma^2 = (0 - \frac{1}{2})^2 \cdot \frac{1}{2} + (1 - \frac{1}{2})^2 \cdot \frac{1}{2} = \frac{1}{4}$$

$$SE(\text{percentage}) = \frac{\sigma}{\sqrt{n}} = \frac{\frac{1}{2}}{\sqrt{100}} = 5\%$$

The sampling distribution

Toss a coin 100 times. The number of tails has the following possible outcomes:
 $0, 1, 2, \dots, 100$.

How likely is each outcome?

The number of tails has the binomial distribution with $n = 100$ and $p = 0.5$.
(‘success’ = coin lands tails)

So if the statistic of interest is S_n = ‘number of tails’, then S_n is a random variable whose probability histogram is given by the binomial distribution. This is called the **sampling distribution** of the statistic S_n .

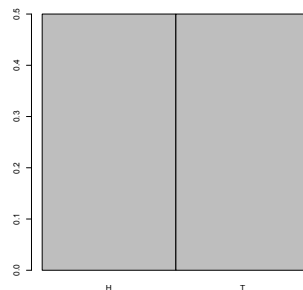
The sampling distribution of S_n provides more detailed information about the chance properties of S_n than the summary numbers given by the expected value and the standard error.

There are three histograms

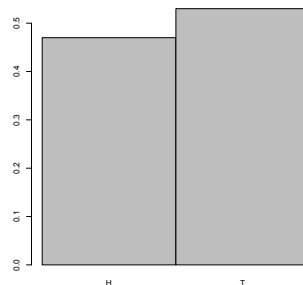
The chance process of tossing a coin 100 times comes with three different histograms:

1. The probability histogram for producing the data:

theoretical

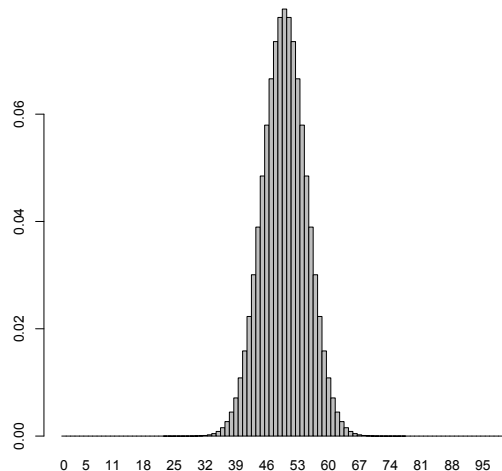


2. The histogram of the 100 observed tosses. This is an empirical histogram of real data:



There are three histograms

3. The probability histogram of the statistic $S_{100} = \text{'number of tails'}$, which shows the sampling distribution of S_{100} :



When doing statistical inference it is important to carefully distinguish these three histograms.

The law of large numbers

The square root law says that $SE(\bar{x}_n) = \frac{\sigma}{\sqrt{n}}$, the standard error of the sample mean, goes to zero as the sample size increases.

Therefore the sample mean \bar{x}_n will likely be close to its expected value μ if the sample size is large. This is the **law of large numbers**.

Keep in mind that the law of large numbers applies

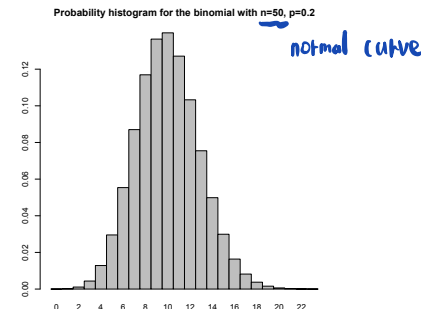
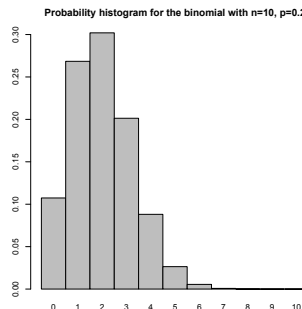
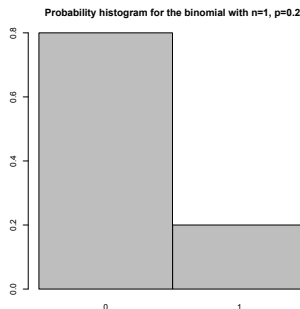
$$SE(S_n) = \sqrt{n} \sigma$$

- ▶ for averages and therefore also for percentages, but not for sums as their SE increases
- ▶ for sampling with replacement from a population, or for simulating data from a probability histogram

More advanced versions of the law of large numbers state that the empirical histogram of the data (the histogram in 2. in the previous section) will be close to the probability histogram in 1. if the sample size is large.

The central limit theorem

Recall the online game where you win a small prize with probability 0.2. We looked at the random variable X = 'number of small prizes' in n gambles and found that X has the binomial distribution with that n and $p = 0.2$.



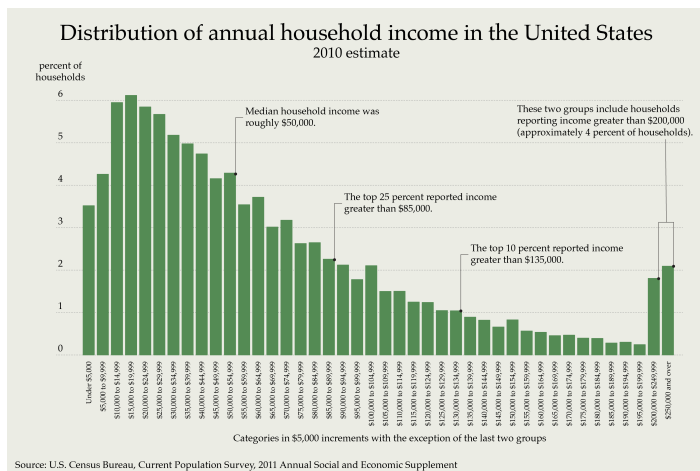
As n gets large, the probability histogram looks more and more similar to the normal curve. This is an example of the **central limit theorem**:

When sampling with replacement and n is large, then the sampling distribution of the sample average (or sum or percentage) approximately follows the normal curve. To standardize, subtract off the expected value of the statistic, then divide by its SE.

then use standard normal distribution to compute probability

The central limit theorem

The key point of the theorem is that we know that the sampling distribution of the statistic is normal *no matter what the population histogram is*:



$$\mu = \$67,000$$

$$\sigma = \$38,000$$

If we sample n incomes at random, then the sample average \bar{x}_n follows the normal curve centered at $E(\bar{x}_n) = \mu = \$67,000$ and with its spread given by

$$SE(\bar{x}_n) = \frac{\sigma}{\sqrt{n}} = \frac{\$38,000}{\sqrt{n}}.$$

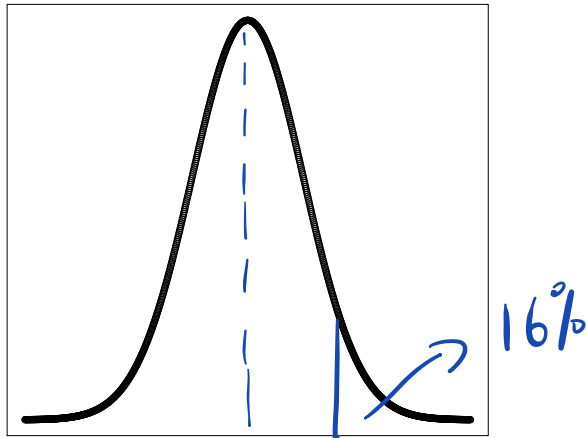
The central limit theorem

If we sample n incomes at random, then the sample average \bar{x}_n follows the normal curve centered at $E(\bar{x}_n) = \mu = \$67,000$ and with its spread given by

$$SE(\bar{x}_n) = \frac{\sigma}{\sqrt{n}} = \frac{\$38,000}{\sqrt{n}}.$$

$$\rightarrow SE(\bar{x}_n) = \frac{\sigma}{\sqrt{n}} = \frac{38000}{\sqrt{100}} = 3800$$

For example, if we sample 100 incomes, then by the empirical rule there is about a 16% chance that \bar{x}_n is larger than \$70,800:



$$\bar{\mu} \quad \bar{\mu} + SE(\bar{x}_n) = 67000 + 3800 = 70800$$

The central limit theorem

In the example about online gambling, we had X = 'number of small prizes' in n gambles.

Since we are counting the number of small prizes, we use a label for each gamble which shows '1' if a small prize is won and '0' otherwise.

Then X equals the sum of these labels and so the central limit theorem applies.

Using the formulas for μ and σ in the case of simulations we find $\mu = p$ and $\sigma = \sqrt{p(1-p)}$.
for binomial distribution, $\mu = p$, $\sigma = \sqrt{p(1-p)}$

Therefore we standardize the sum X with the expected value np and

$$\text{SE}(X) = \sqrt{np(1-p)}.$$

$= \sqrt{n} \sigma$

use to standardize

When does the central limit theorem apply?

For the normal approximation to work, the key requirements are:

- ▶ We sample with replacement, or we simulate independent random variables from the same distribution. *when sample size is way much smaller than population size, then sampling without replacement \approx sampling with replacement*
- ▶ The statistic of interest is a sum (averages and percentages are sums in disguise).
- ▶ The sample size is large enough: the more skewed the population histogram is, the larger the required sample size n .
 - { (if there is no strong skewness then $n \geq 15$ is sufficient)
 - strong skewness, then $n \geq 40$*

	expected value	standard error
average	μ	$\frac{\sigma}{\sqrt{n}}$
sum	$n\mu$	$\sqrt{n}\sigma$
percentage	μ	$\frac{\sigma}{\sqrt{n}}$

The percentage of Democratic voters in the sample is equal to the mean of the sample, and so the "around" value is the expected value of the sample mean, which in turn is equal to the population mean, 0.60 or 60%.



Quiz

1. A town has 10,000 registered voters, of whom 6,000 are voting for the Democratic party. A survey organization is taking a sample of 100 registered voters (assume sampling with replacement). The percentage of Democratic voters in the sample will be around ____, give or take _____. (You may use the fact that the standard deviation of 6,000 1s and 4,000 0s is about 0.5)

democratic: 1

- ☒ 60%, give or take 5%
☐ 40%, give or take 5%
☐ 60%, give or take 0.5%
☐ 40%, give or take 0.5%

$$\sigma = 0.5$$

$$IE(\text{percentage}) = \mu = 60\%$$

$$SE(\text{percentage}) = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{100}} = 5\%$$

$$\mu = 1 \cdot p(\text{democratic}) + 0 \cdot p(\text{non-democratic}) \\ = \frac{6000}{10000} = 60\%$$

2. You solicit 100 pledges for a charitable organization. Each pledge is equally likely to be \$10, \$50, or \$100. You may use the fact that the standard deviation of the three amounts \$10, \$50 and \$100 is \$37.

What is the expected value of the sum of the 100 pledges?

- ☒ \$5333
☐ \$533
☐ \$3700
☐ \$370

$$IE(\text{sum}) = n\mu = 100 \cdot \frac{160}{3} = 5333.33 \dots$$

$$\mu = 10 \cdot p(\$10) + 50 \cdot p(\$50) + 100 \cdot p(\$100) \\ = 10 \times \frac{1}{3} + 50 \times \frac{1}{3} + 100 \times \frac{1}{3} \\ = \frac{160}{3}$$

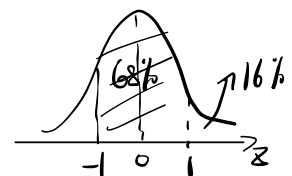
3. You solicit 100 pledges for a charitable organization. Each pledge is equally likely to be \$10, \$50, or \$100. You may use the fact that the standard deviation of the three amounts \$10, \$50 and \$100 is \$37.

What are the chances that the 100 pledges total more than \$5,700?

- ☒ 16%
☐ 32%
☐ 5%

$$SE(\text{sum}) = \sqrt{n}\sigma = \sqrt{100} \cdot 37 = 370$$

$$\text{Standardize: } \frac{5700 - 5333}{370} = 1 \quad \text{one standard deviation}$$



4. There are two candidates running for governor in CA and they are said to have roughly equal support from the voters. To get a better idea who is ahead, a company polls 400 of the 20 million registered voters in California. Likewise, there are two candidates running for mayor in Palo Alto who are said to have roughly equal support, and the company polls 400 out of the 20,000 registered voters in Palo Alto. Will the first poll be more accurate, equally accurate, or less accurate than the second poll?

- ☒ more accurate
☒ equally accurate
☐ less accurate

$$SE = \frac{\sigma}{\sqrt{n}}$$

5. The average taxable income reported on tax returns for the year 2016 is \$ 45,000, and the standard deviation of the taxable incomes is \$ 23,000.

Which of the following two statements are true? Both?

- ☒ The chances that the sum of 100 randomly selected taxable incomes exceeds \$ 4 million can be computed from the above information using normal approximation.
☒ The percentage of taxable incomes that fall below \$ 30,000 can be computed from the above information using normal approximation.

histogram of income is skewed and not normal.

6. Questions (a)-(d) below relate to the following situation: Someone tosses a fair coin 100 times.

Question (a): How many tails can she expect to get?

$$1: \text{tails}, 0: \text{heads}, p(1) = \frac{1}{2}, p(0) = \frac{1}{2}$$

$$IE(\text{sum}) = n\mu = 100 \times \frac{1}{2} = \underline{50}$$

$$\mu = 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} = \frac{1}{2}$$

7. Question (b): What is the "give and take" number for the result from Question (a)?

Enter answer here

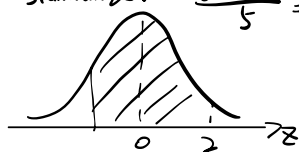
$$SE(\text{sum}) = \sqrt{n} \sigma = \sqrt{100} \cdot \sqrt{\frac{1}{4}} = 5$$

$$\sigma^2 = (1 - \frac{1}{2})^2 \cdot \frac{1}{2} + (0 - \frac{1}{2})^2 \cdot \frac{1}{2} = \frac{1}{4}$$

8. Question (c): What are the chances that she gets between 40 and 60 tails?

- ☐ 16%
☐ 68%
☒ 95%
☐ 99.7%

standardize: $\frac{60 - 50}{5} = 2$



9. A large group of people get together and everyone tosses a coin 100 times.

Question (d): About what percentage of people will get between 40 and 60 tails?

- ☐ 16%
☐ 68%
☒ 95%
☐ 99.7%