

## 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  $\mu$ , or the population standard deviation  $\sigma$ .

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  $\mu$ , give or take about one standard deviation  $\sigma$ .

The **expected value** of one random draw is the population average  $\mu$ .

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  $\mu$ .

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  $\mu$  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  $\sigma$  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


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

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.
- ▶  Five small black icons of people standing in a row, representing a sample of voters.
- ▶ 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  $\mu$  is the population average (=proportion of 1s) and  $\sigma$  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.

## 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  $\mu$  and  $\sigma$  in that case?

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

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

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?



## 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)

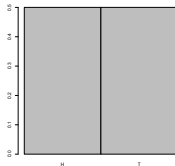
So if the statistic of interest is  $S_n$  = ‘number of tails’, then  $S_n$  is a random variable whose probability histogram is given by the binomial distribution. This is called the **sampling distribution** of the statistic  $S_n$ .

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.

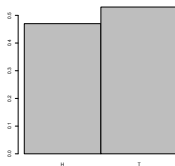
## There are three histograms

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

1. The probability histogram for producing the data:

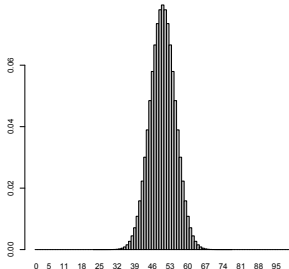


2. The histogram of the 100 observed tosses. This is an empirical histogram of real data:



## 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  $SE(\bar{x}_n)$ , 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  $\mu$  if the sample size is large. This is the **law of large numbers**.

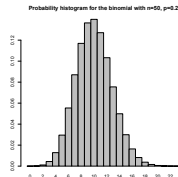
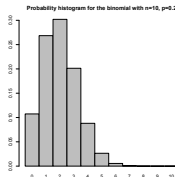
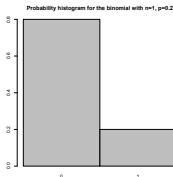
Keep in mind that the law of large numbers applies

- ▶ 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  $p = 0.2$ .

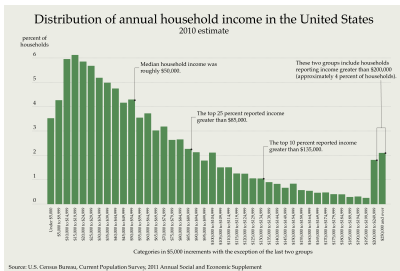


As  $n$  gets large, the probability histogram looks more and more similar to the normal curve. This is an example of the **central limit theorem**:

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.

# 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

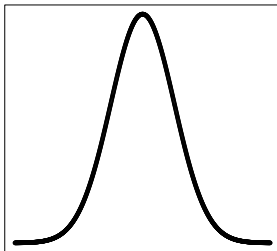
$$SE(\bar{x}_n) = \frac{\sigma}{\sqrt{n}} = \frac{\$38,000}{\sqrt{n}}.$$

## 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

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

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:



## 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  $\mu$  and  $\sigma$  in the case of simulations we find  $\mu = p$  and  $\sigma = \sqrt{p(1-p)}$ .

Therefore we standardize the sum  $X$  with the expected value  $np$  and  $SE(X) = \sqrt{np(1-p)}$ .



## 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.
- ▶ 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)