Lecture 1. What is econometrics?

Econometrics is about measuring economic relations

1. What is an economic relation?
2. How do you measure it?

1. Economic relations

An economic relation is a relation between economic variables

Examples

- The FED stabilizes economic activity by increasing or decreasing short term interest rates
- The recent tax cut was passed on the hypothesis that there is a relation between the level of taxes and economic growth
- A house buyer/seller is interested in the relation between price and amenities of a house
- A company wants to know what the effect is of its ad expenditures on sales or market share
- A student wants to know what the return is to (further) education

How do we represent economic relations?

In all examples there is/are

- Dependent variable that is the outcome of interest (economic activity, economic growth, house price, market share, income)
- Independent variable(s) the effect of which on the outcome we want to measure (interest rate, tax rate, surface of house, ad expenditures, level of education)

Methods in this course assume that there is exactly one dependent variable.

The dependent variable is denoted by: $Y$

There may be many independent variables

Independent variables are denoted by:

$$X_1, X_2, \ldots, X_K$$

We use mathematical expression to represent the relation between the dependent variable $Y$ and independent variables $X_1, X_2, \ldots, X_K$. 
For instance, if there is only one independent variable $X$ we could assume a linear relation

$$Y = \alpha + \beta X$$

The graph of this relation is a straight line with intercept $\alpha$ and slope $\beta$.

Measuring the relation is now measuring the coefficients $\alpha, \beta$.

Note that by redefining $X_1, X_2, \ldots, X_k$ we can also specify nonlinear relations

$$Y = \beta_1 + \beta_2 X + \beta_3 X^2 + \beta_4 X^3$$

is a nonlinear relation in $X$ (a cubic in $X$).

The key is that this relation is linear in the coefficients!

2. Measuring economic relations

In this course we only consider linear relations, i.e. linear in the coefficients.

In the case of one $X$ measuring the relation is obtaining numerical values for $\alpha, \beta$.

How do we do this?

Method 1: Use a graph. A linear relation/straight line is determined by two points. With two observations on $Y, X$ we can determine $\alpha, \beta$.

Example: for two houses we obtain price and square foot. Denote these numbers by $X_1, Y_1$ and $X_2, Y_2$.

If the linear relation is correct we have

$$Y_1 = \alpha + \beta X_1$$
$$Y_2 = \alpha + \beta X_2$$

We can solve this for $\alpha, \beta$ to find the numerical/computed values $\hat{\alpha}, \hat{\beta}$.
\[
\hat{\beta} = \frac{Y_2 - Y_1}{X_2 - X_1}
\]

\[
\hat{\alpha} = \frac{X_2 Y_1 - X_1 Y_2}{X_2 - X_1}
\]

Now consider the case that we have \( n \) observations denoted by the \( n \) pairs

\[Y_i, X_i, i = 1, \ldots, n\]

Will the method still work?

Only if all pairs \( Y_i, X_i \) are exactly on the straight line!

Consider an example: Data on selling price (in 1000\$) \( (Y) \) and living area (square foot) \( (X) \) of 14 houses in San Diego
The graph is called a scatterplot.

Note that the observations are not on a straight line!

Reason

1. Relation is nonlinear. For house prices we could fit a polynomial of degree 13. Is that a good idea?
2. Other variables are important for selling price, e.g. type of house, location, no. of bedrooms, no. of bathrooms etc.
Now assume that you have an unlimited budget to study the selling price of houses and also an unlimited capacity to measure variables. Then you may be able to measure all relevant variables that affect the selling price. Let these variables be

$$X_1, X_2, \ldots, X_K$$

Because these are all relevant variables we have the exact relation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$$

Hence if we have $n$ observations then for $i = 1, \ldots, n$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki}$$

In the scatterplot we had only $X_1$ as independent variable.

Because we do not have $X_2, \ldots, X_K$ we can rewrite the exact relation as

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i$$

with

$$u_i = \beta_2 X_{2i} + \cdots + \beta_K X_{Ki}$$

Hence $u$ captures the effect of all omitted variables in the exact relation.

In the graph $u_i$ is the deviation from the straight line for observation $i$.

We call $u_i$ the $i$-th residual. This is defined relative to some straight line.

$$u_i = Y_i - \alpha - \beta X_i$$

If we think that the deviations from the straight line are due to omitted variables, it makes sense to find the coefficients in the relation between $Y$ and $X$ (or $X_1$).

How do we do this?

The choice will be to compute the coefficients that best fit the data, i.e. that gives the smallest residuals.

Obvious that we can not make all residuals small at the same time. We need some criterion.
We choose

\[ S(\alpha, \beta) = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2 \]

Note

1. Squares treat negative/positive residuals symmetrically.
2. Large residuals get relatively more weight.
Lecture 2: Ordinary Least Squares and the Simple Linear Regression Model

Key observation:

If we plot $n$ observations

$$Y_i, X_i, i = 1, \ldots, n$$

on a dependent variable $Y$ and an independent variable $X$ in a scatterplot, then points $(X_i, Y_i)$ not on any straight line

$$Y = \alpha + \beta X$$

Reason: Omitted variables that affect $Y$ (beside $X$)

Contribution of omitted variables to $Y$ is the residual $u$

$$u = Y - \alpha - \beta X$$

See figure for illustration
\[ Y = \alpha + \beta X \]

\[ Y_i = Y_i - \alpha - \beta X_i \]

\[ Y = \tilde{\alpha} + \tilde{\beta} X \]
How do we find the hidden straight line if there are omitted variables?

Ordinary Least Squares (OLS) method

- Compute residuals for \( i = 1, \ldots, n \)

\[
 u_i = Y_i - \alpha - \beta X_i
\]

- Compute the sum of squared residuals

\[
 S(\alpha, \beta) = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2
\]

The sum of squared residuals depends on the coefficients \( \alpha, \beta \) (as is obvious from figure)

- Minimize \( S(\alpha, \beta) \) with respect to \( \alpha, \beta \)

The last step is not obvious and will be justified later.
OLS solution

\[ S(\alpha, \beta) = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2 = \]

\[ = \sum_{i=1}^{n} (Y_i^2 + \alpha^2 + \beta^2 X_i^2 - 2\alpha Y_i - 2\beta X_i Y_i + 2\alpha \beta X_i) = \]

\[ = \sum_{i=1}^{n} Y_i^2 + n\alpha^2 + \beta^2 \sum_{i=1}^{n} X_i^2 - 2\alpha \sum_{i=1}^{n} Y_i - 2\beta \sum_{i=1}^{n} X_i Y_i + 2\alpha \beta \sum_{i=1}^{n} X_i \]

Consider first case \( \alpha = 0 \)

\[ S(\beta) = \sum_{i=1}^{n} Y_i^2 + \beta^2 \sum_{i=1}^{n} X_i^2 - 2\beta \sum_{i=1}^{n} X_i Y_i \]

This is expression in \( \beta \) of form

\[ A + B\beta^2 + C\beta \]

with \( A, B, C \) constants.

The graph of such a function (if \( B > 0 \)) is a parabola with a minimum (see figure)
How do we find minimum?

At $\beta$ that minimizes $S(\beta)$ the slope of $S(\beta)$ is 0.

Slope of $S(\beta)$ in $\beta$ is derivative of $S(\beta)$

$$\text{Slope } S(\beta) \text{ in } \beta = \frac{dS(\beta)}{d\beta}$$
Now
\[
\frac{dS(\beta)}{d \beta} = 2B\beta + C
\]

Value of $\beta$ that minimizes $S(\beta)$ solution to
\[
2B\beta + C = 0 \quad \text{or} \quad \beta = -\frac{C}{2B}
\]

Now
\[
B = \sum_{i=1}^{n} X_i^2 \quad C = -2\sum_{i=1}^{n} X_i Y_i
\]

and
\[
\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}
\]

Next, case with $\alpha, \beta$

\[
S(\alpha, \beta) = \sum_{i=1}^{n} Y_i^2 + n\alpha^2 + \beta^2 \sum_{i=1}^{n} X_i^2 - 2\alpha \sum_{i=1}^{n} Y_i - 2\beta \sum_{i=1}^{n} X_i Y_i + 2\alpha \beta \sum_{i=1}^{n} X_i
\]

This a quadratic function in $\alpha, \beta$. Graph has same shape as parabola, but in 3D.

In minimum slope is again 0.

Now we have slope in $\alpha$-direction and slope in $\beta$-direction. These are the partial derivatives with respect to $\alpha$ and $\beta$:

\[
\frac{\partial S(\alpha, \beta)}{\partial \alpha} \quad \text{and} \quad \frac{\partial S(\alpha, \beta)}{\partial \beta}
\]

Computed like ordinary derivatives, except that other argument is kept constant: for slope/derivative in $\alpha$-direction we keep $\beta$ constant and for slope/derivative in $\beta$-direction we keep $\alpha$ constant.
With

\[
S(\alpha, \beta) = \sum_{i=1}^{n} Y_i^2 + n \alpha^2 + \beta^2 \sum_{i=1}^{n} X_i^2 - 2\alpha \sum_{i=1}^{n} Y_i - 2\beta \sum_{i=1}^{n} X_i Y_i + 2\alpha \beta \sum_{i=1}^{n} X_i
\]

we have

\[
\frac{\partial S(\alpha, \beta)}{\partial \alpha} = 2n\alpha - 2\sum_{i=1}^{n} Y_i + 2\beta \sum_{i=1}^{n} X_i
\]

\[
\frac{\partial S(\alpha, \beta)}{\partial \beta} = -2\sum_{i=1}^{n} X_i Y_i + 2\beta \sum_{i=1}^{n} X_i^2 + 2\alpha \sum_{i=1}^{n} X_i
\]

\[
S(\alpha, \beta) \text{ is minimal at the value of } (\alpha, \beta) \text{ where the slope in both directions is 0.}
\]

Hence

\[
\frac{\partial S(\alpha, \beta)}{\partial \alpha} = 2n\alpha - 2\sum_{i=1}^{n} Y_i + 2\beta \sum_{i=1}^{n} X_i = 0
\]

\[
\frac{\partial S(\alpha, \beta)}{\partial \beta} = -2\sum_{i=1}^{n} X_i Y_i + 2\beta \sum_{i=1}^{n} X_i^2 + 2\alpha \sum_{i=1}^{n} X_i = 0
\]

These are two (linear) equations in two unknowns \(\hat{\alpha}, \hat{\beta}\). These are called the normal equations.

Solution:

1. Solve first equation for \(\alpha\)

\[
\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} Y_i - \beta \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{Y} - \beta \bar{X}
\]

with \(\bar{Y}, \bar{X}\) the sample average of \(Y, X\)
2. Substitute this solution in the second equation

\[-\sum_{i=1}^{n} X_i Y_i + \hat{\beta} \sum_{i=1}^{n} X_i^2 + \left(\frac{1}{n} \sum_{i=1}^{n} Y_i - \hat{\beta} \frac{1}{n} \sum_{i=1}^{n} X_i\right) \sum_{i=1}^{n} X_i = 0\]

and

\[\hat{\beta} \left( \sum_{i=1}^{n} X_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} X_i \right)^2 \right) = \sum_{i=1}^{n} X_i Y_i - \frac{1}{n} \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i\]

Now

\[\sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} (X_i^2 - 2X_i\bar{X} - \bar{X}^2) = \sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 - n\bar{X}^2 = \]

\[= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = \sum_{i=1}^{n} X_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} X_i \right)^2 \]

and

\[\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} (X_i Y_i - \bar{Y}X_i - \bar{X}Y_i + \bar{X}\bar{Y}) = \]

\[= \sum_{i=1}^{n} X_i Y_i - n\bar{Y}\bar{X} - n\bar{X}\bar{Y} + n\bar{X}\bar{Y} = \sum_{i=1}^{n} X_i Y_i - \frac{1}{n} \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i\]

Using these results we find for \(\hat{\beta}\)

\[\hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}\]
Hence the values of $\alpha, \beta$ that minimize the sum of squared residuals are

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}$$

These are the Ordinary Least Squares (OLS) solutions to the problem of fitting a straight line to the points in a scatterplot. The least squares line

$$Y = \hat{\alpha} + \hat{\beta} X$$

is the straight line that ‘fits’ the scatterplot best (see figure).
\[ Y = \hat{\alpha} + \hat{\beta} X \]

\[ \hat{\alpha} = 52.35091 \]

\[ \hat{\beta} = 0.138750 \]

\[ R^2 = 0.820522 \]
Note

- If we divide numerator and denominator by \( n \) or \( n - 1 \), then

\[
\hat{\beta} = \frac{\text{Sample covariance of } X \text{ and } Y}{\text{Sample variance of } X}
\]

- From the OLS solution for \( \hat{\alpha} \)

\[
Y = \hat{\alpha} + \hat{\beta} X
\]

In words: The point \((\bar{Y}, \bar{X})\) is on the least squares line

- The residuals with respect to the least squares line are the OLS residuals

\[
e_i = Y_i - \hat{\alpha} - \hat{\beta} X_i
\]

From the normal equations

\[
0 = -n\hat{\alpha} + \sum_{i=1}^{n} Y_i - \hat{\beta} \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i) = \sum_{i=1}^{n} e_i
\]

\[
0 = \sum_{i=1}^{n} X_i Y_i - \hat{\beta} \sum_{i=1}^{n} X_i^2 - \hat{\alpha} \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} X_i (Y_i - \hat{\alpha} - \hat{\beta} X_i) = \sum_{i=1}^{n} X_i e_i
\]

Hence

\[
\sum_{i=1}^{n} e_i = \frac{1}{n} \sum_{i=1}^{n} e_i = 0
\]

\[
\sum_{i=1}^{n} X_i e_i = \frac{1}{n} \sum_{i=1}^{n} X_i e_i = 0
\]

In words: The sample average (and sum) of the OLS residuals is 0 and the sample covariance of these residuals and \( X \) is also 0

These are all consequences of the fact that we minimize the sum of squared residuals.
How good is the fit of the straight line to the scatterplot?

Define the fitted value

\[ \hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i \]

then by the definitions

\[ Y_i = \hat{Y}_i + e_i \]

Because the OLS residuals have average 0

\[ \bar{Y} = \bar{\hat{Y}} \]

In words: \( Y \) and \( \hat{Y} \) have the same sample average

Using this we have

\[ (Y_i - \bar{Y})^2 = (\hat{Y}_i - \bar{\hat{Y}})^2 + e_i^2 + 2e_i(\hat{Y}_i - \bar{\hat{Y}}) \]

If we take the sum over \( i \) we first observe

\[
\sum_{i=1}^{n} e_i(\hat{Y}_i - \bar{\hat{Y}}) = \sum_{i=1}^{n} e_i \hat{Y}_i - \bar{\hat{Y}} \sum_{i=1}^{n} e_i = \\
= \sum_{i=1}^{n} e_i(\hat{\alpha} + \hat{\beta} X_i) = \hat{\alpha} \sum_{i=1}^{n} e_i + \hat{\beta} \sum_{i=1}^{n} e_i X_i = 0
\]

Note that this implies that the sample covariance between the OLS residuals and the OLS fitted values is 0: OLS decomposes \( Y_i \) into two parts (residual and fitted value) that have covariance 0, i.e. are unrelated

Using this we find

\[
\frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - \bar{\hat{Y}})^2 + \frac{1}{n} \sum_{i=1}^{n} e_i^2
\]
The sample variance of $Y$ is equal to the sum of the sample variance of $\hat{Y}$ and the sample variance of $e$ or

Total Variance = Explained variance +
+ Unexplained variance

A measure of goodness of fit is

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

This is fraction of total variance that is explained by fitted straight line.

Note $R^2 = 1$ if and only if $Y_i = \hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$, i.e. if all observations are on the straight line.

Also $R^2 = 0$ if and only if

$$\bar{Y} - \hat{\beta}(\bar{X} - X_i) = \hat{\alpha} + \hat{\beta} X_i = \hat{Y}_i = \hat{\bar{Y}} = \bar{Y}$$
or

$$\hat{\beta}(\bar{X} - X_i) = 0$$

for all $i = 1, \ldots, n$. If the $X_i$ are not all equal, then this can only be the case if $\hat{\beta} = 0$. In that case $X$ does not help in explaining $Y$.

These are the extreme values for $R^2$. We have

$$0 \leq R^2 \leq 1$$

For the GPA data (see figure in lecture 1) we find

$$R^2 = 0.164933$$

$$\hat{\alpha} = 0.923692$$

$$\hat{\beta} = 0.523015$$
Lecture 3: The Simple Linear Regression Model

Goal in lectures 1&2 was measuring the relation between dependent variable $Y$ and independent variable(s) $X$

We consider simple relations, in particular the linear relation

$$Y = \alpha + \beta X$$

Problem was that if we have $n$ observations

$$Y_i, X_i, i = 1, \ldots, n$$

then these are usually not on the straight line.

The linear relation between $Y$ and $X$ is not exact. We have for $i = 1, \ldots, n$

$$Y_i = \alpha + \beta X_i + u_i$$

with $u_i, i = 1, \ldots n$ the residuals.

The residuals are present, because $X$ is not the only variable that determines $Y$. The combined effect of these omitted variables on $Y$ is the residual.

In lecture 2 we discussed the method of Ordinary Least Squares that computed $\hat{\alpha}, \hat{\beta}$ by minimizing the sum of squared residuals.

This produced the least squares line

$$Y = \hat{\alpha} + \hat{\beta} X$$

Although not all observations are on the least squares line, this line is as close as possible to the observations.
Is OLS a good idea?

We need a framework to answer this question.

The framework is provided by mathematical statistics.

Mathematical statistics: analysis of data generated by a random experiment in order to learn about that experiment.

Random experiment: Experiment/mechanism/phenomenon with outcome that cannot be (perfectly) predicted.

Example: Tossing a coin.

Two outcomes: head (H) or tail (T). If enough spin then one cannot predict the outcome with certainty.

Description of random experiment

- List of all possible outcomes
- Probabilities assigned to these outcomes

Tossing a coin

- Outcomes H,T
- Probability of H, notation, \( \text{Pr}(H) = p \)

Note

- Because \( \text{Pr}(H) + \text{Pr}(T) = 1 \), we have \( \text{Pr}(T) = 1 - p \)
- If coin is fair, then \( p = \frac{1}{2} \)
Assume we have a coin and want to determine whether the coin is fair.

Toss the coin 100 times and compute the fraction H, denoted by $\hat{p}$.

I find $\hat{p} = .51$. Can we conclude from this that the coin is fair? Why (not)?

I do another 100 tosses and find $\hat{p} = .48$.

Further experiments give $\hat{p} = .57$, $\hat{p} = .48$, $\hat{p} = .50$.

Note sampling variation in $\hat{p}$: for each 100 tosses (sample) we obtain a different $\hat{p}$.

If we repeat the 100 tosses many times, we obtain the probability/frequency distribution of $\hat{p}$. This distribution is called the sampling distribution of $\hat{p}$.

We get a good impression of this sampling distribution if we repeat the 100 tosses 1000 times. We can use the computer to do this.

I have plotted the sampling distribution of $\hat{p}$ for three different coins (see graphs) (why is this graph somewhat misleading? which values can $\hat{p}$ take?)
Frequency distribution fraction $H(N = 100, 1000$ repetitions$)$
Frequency distribution fraction $H \ (N = 100, \ 1000 \ repetitions)$
Frequency distribution fraction $H (N = 100, 1000$ repetitions$)$
Coin 1 seems to be fair, coin 2 seems to be biased, and for coin 3 the evidence is inconclusive.

How can we draw this conclusion?

Coin 1 is indeed fair and the graph is the sampling distribution of $\hat{p}$ for $p = \frac{1}{2}$.

Note the shape of this sampling distribution
- Close to symmetric around $p = \frac{1}{2}$
- Bell shape as for normal distribution

$\hat{p}$ is between .40 and .59 in 950 of the 1000 replications (95% of the replications).

If we toss a new coin 100 times and obtain $\hat{p}$, there are two possibilities
- $\hat{p}$ is between .40 and .59
- $\hat{p}$ is outside this interval

If outside then it is unlikely that the coin is fair; the probability of obtaining $\hat{p}$ if the coin were fair is less than 5%. If inside, then the coin is more likely to be fair.

Compare the other graphs that have $p = .6$ and $p = .55$. Note that our method is not full proof: if $p \neq \frac{1}{2}$, $\hat{p}$ can well be between .40 and .59, although probability is smaller for $p = .6$ than for $p = .55$.

How can we get more confidence in our conclusion?
Same graphs with 1000 tosses

**Sampling distribution** \( \hat{p} \) (N=1000, 1000 repetitions); \( p = .5 \)

Note that 95% interval is narrower for 1000 tosses
Note 95% interval does not overlap with that for fair coin
Sampling distribution $\hat{p}$ ($N=1000$, 1000 repetitions); $p = .55$
Next we formalize this argument

A random variable is a variable that assigns numerical values to outcomes of random experiments.

For coin tossing, e.g.

\[ X = 1 \quad \text{if outcome is H} \]
\[ X = 0 \quad \text{if outcome is T} \]

Here simple relabeling. Often outcome of random experiment is high-dimensional and then a single random variable highlights an aspect of the outcome, e.g. US economy in 2002 with random variables Consumption, Business investment, GDP, etc.

The values of \( X \) occur with a certain probability. In coin tossing

\[ \Pr(X = 1) = p \]
\[ \Pr(X = 0) = 1 - p \]

Before we toss the coin 100 times, we do not know the result of these tosses.

We represent these tosses by a set of random variables \( X_1, X_2, \ldots, X_{100} \).

Note

\[ \Pr(X_i = 1) = p \]
\[ \Pr(X_i = 0) = 1 - p \]

for \( i = 1, \ldots, 100 \), because the same coin is tossed repeatedly. Also the outcomes of any two tosses are independent.

A collection of random variables that have the same distribution and are independent is called a random sample.
The fraction $H$ is

$$\hat{p} = \frac{1}{100} \sum_{i=1}^{100} X_i$$

This is just the sample average of $X_1, X_2, \ldots, X_{100}$. These are random variables and hence $\hat{p}$ is a random variable too.

$\hat{p}$ only depends on the random sample. It is called an estimator of $p$ (the numerical value that is obtained after the 100 tosses is called an estimate of $p$).

An estimator has a sampling distribution (see above) and the quality of the estimator depends on its sampling distribution.

Important
- Sampling distribution symmetric around $p$
- Sampling distribution has small variance

Before we derived the sampling distribution experimentally. There is another method: study the sampling distribution of $\hat{p}$ using our knowledge of distribution of $X_1, X_2, \ldots, X_{100}$.

Remember expectation/expected value of a random variable $Y$ that takes values $y_1, \ldots, y_k$ and has probability distribution

$$\Pr(Y = y_k) = p_k, \quad k = 1, \ldots, K$$

i.e.

$$E(Y) = \sum_{k=1}^{K} p_k y_k$$

Properties of expectation ($a_1, a_2$ are constants)

$$E(a_i) = a_i$$

$$E(a_1 Y_1 + a_2 Y_2) = a_1 E(Y_1) + a_2 E(Y_2)$$

If $Y_1, Y_2$ are independent random variables, then

$$E(Y_1 Y_2) = E(Y_1) E(Y_2)$$

This is enough to derive the mean and variance of the sampling distribution of $\hat{p}$.
\[ E(\hat{p}) = E\left( \frac{1}{100} \sum_{i=1}^{100} X_i \right) = \frac{1}{100} \sum_{i=1}^{100} E(X_i) = \frac{1}{100} \cdot 100 \cdot p = p \]

Conclusion: The mean of the sampling distribution of \( \hat{p} \) is \( p \). If this is true for an estimator we call this estimator unbiased.

Next, the sampling variance

\[ Var(\hat{p}) = E\left( (\hat{p} - p)^2 \right) = E\left( \left( \frac{1}{100} \sum_{i=1}^{100} X_i - p \right)^2 \right) = \]

\[ = \frac{1}{10000} E\left( \sum_{i=1}^{100} (X_i - p)^2 \right) \]

Separating squares and cross-products we find that the last expression on the rhs is

\[ E\left[ \sum_{i=1}^{100} (X_i - p)^2 \right] = E\left[ \sum_{i=1}^{100} (X_i - p)^2 \right] + \]

\[ + 2E\left[ \sum_{i<j}^{100} (X_i - p)(X_j - p) \right] \]

Now

\[ E\left[ (X_i - p)^2 \right] = p(1 - p) \]

and

\[ E\left[ (X_i - p)(X_j - p) \right] = 0 \]

Hence

\[ Var(\hat{p}) = \frac{1}{100} \cdot p(1 - p) \]

Note that we can estimate this variance by substituting \( \hat{p} \) for \( p \).