Introduction

Welcome to EEP 118 Section!

- We will be holding in-person sections each week. We will post the notes before each section and share the video, slides, and note solutions after.
- Attendance is not mandatory. But please try and attend the section at the *time* you signed up for (it's fine to swap here and there).
- Schedule (all Mulford 240, except W 4-5pm in Dwinelle 243)
  - Wednesday 9AM-10AM; 4PM-5PM (Jed)
  - Friday 9AM-10AM; 12PM-1PM (Pierre)
- Office hours (Giannini 203)
  - Jed: Wednesdays 10AM-12PM.
  - Pierre: Tuesdays from 2-4PM
- Email policy: We will try our best to respond within 48 hours during the week. However, don't expect responses after 6PM and on the weekend.
- Please reach out if you are having any issues with technology or in general. We will try our best to be accomodating, but we need to know in order to help!
- Class websites
  - bCourses: course announcements, files, videos, assignments
  - Piazza: forum for questions and discussions; GSIs monitor but student interactions encouraged!
  - DataHub server: host for assignment Jupyter notebooks; do your work here and then save as PDF to submit
  - If you haven't checked out videos on Jupyter notebook and R yet, please do so!
  - Gradescope: website for submitting completed assignments

Let's dive into the material, then!

1. Regressions and Regression Models (w/Solutions)[1]

Simply put, economists use regression models to study the relationship between two variables. If Y and X are two variables, representing some population, we are interested in "explaining Y in terms of X", or in determining "how Y varies with changes in X".

The classic example, common in labor economics, is to try and understand the relationship between income (Y) and education (X). When we talk about adding other X's (covariates or regressors) into our estimation, this means that we believe that other variables aside from education are also important in explaining variation in income such as work experience or parents' education.

We will go into this in much more detail as the course progresses, but for now, we can think of regression models as an estimated relationship between X and Y variables found in actual data.

The linear regression model assumes that the relationship between Y and X is linear - and as economists we then try to find the line that most closely approximates the true relationship. The appropriate picture to have in mind is the following:[2]

[1] The section notes for this course are based off of those originally created by Erin Kelley. Many thanks!
[2] Figure taken from: Lay, David C. Linear Algebra and Its Applications. 4th ed. Boston: Addison-Wesley, 2012. Print.

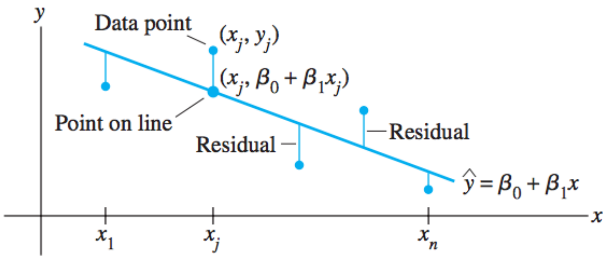


FIGURE 1 Fitting a line to experimental data.

Any data set you work with will have some outcomes you are interested in (the Y term) and some explanatory variables (the X term). Plotting these data points will often produce something that resembles a line. The economist's role is to estimate the equation of that line - notice how the equation in the graph $\hat{y} = \beta_0 + \beta_1 x$ is the equation of a line, as you learned in calc 1.

2. Model Example

Watching TV one evening you come across a news program talking about Berkeley residents' health. The video shows an image of an emergency room in a hospital packed with people. The conditions of the hospital look to be very poor.

The news anchor says: "As you can see, health services here are so bad that going to a hospital is actually worse than staying at home. The following statistics demonstrate that you are better off staying away from hospitals." The following table is then shown on the screen:

Percent of sick patients who fully recover	
Stayed at home	Went to hospital
68%	25%

What is the implied research question in this news story?

What is the effect of hospitals on fully recovering from illness?

Do you agree with the news anchor's conclusion? What other factors might contribute to whether or not someone recovers from illness? How could additional data or information improve your confidence in the anchor's conclusion?

I do not agree with the news-achors conclusions because the sample of people that go to the hospital is very different from the sample that doesn't go to the hospital. Those that go to the hospital may have more severe illnesses for example.

What are the following components of the regression model that would analyze this question if you had the data?

1. Dependent Variable (Y): Fully Recover
2. Explanatory/ independent of primary interest (X_1): Went to Hospital
3. Additional Explanatory/ independent variables or covariates ($X_2, X_3, ...$): Age, Medical History, Healthy Eating...

3. Data Types

Knowing and understanding your data is *critical* to being a good economist. The form of your data will dictate which methods of data analysis we can choose from, which will, in turn, determine which different types of questions we can answer with it.

1. **Cross-Sectional Data:** Contains observations of different people, countries, firms, farmer etc. *at a single point in time*.
   - Example: a 2014 survey of Berkeley seniors on their academic record and extra-curricular participation.
2. **Time Series Data:** Contains a single person, country, firm, farmer etc. *over multiple points in time*.
   - Example: Data containing the weekly number of violent crimes committed in Los Angeles from 2010-2014.
3. **Pooled/Repeated Cross Section:** Contains multiple cross sections of people, countries, firms, farmers, etc *over multiple points in time where the observations are not necessarily repeated across rounds*.
   - Example: Current Population Survey in the United States. Each month a different set of households is surveyed about employment, unemployment etc.
   - This is also often referred to as a *repeated* cross section.
4. **Panel or Longitudinal Data:** You observe *the same* set of people, countries, firms, farmers, etc. *over multiple points in time*.
   - Example: The Indonesian Family Life Survey has been conducted across five rounds of data collection between 1993 and 2007 returning to the same households throughout the course of the study.
   - Panel data can either be shaped in long or wide format: below is an example of each:

Table 1: Data in Wide Format

	ID	Income_97	Income_98
1.	1	1000	2000
2.	2	4320	5000

Table 2: Data in Long Format

	ID	Year	Income
1.	1	97	1000
2.	1	98	2000
3.	2	97	4320
4.	2	98	5000

4. Functional Forms Review

Preliminaries: Proportions, Percentages, and Elasticities

Below we provide formulas and/or definitions for some key concepts:

- *Proportional change*: $\frac{x_1 - x_0}{x_0} = \frac{\Delta x}{x_0}$
- *Percentage change*: $\frac{x_1 - x_0}{x_0} \times 100 = \frac{\Delta x}{x_0} \times 100$
- *Elasticity*: $\frac{\Delta z / z}{\Delta x / x} = \frac{\partial z}{\partial x} \frac{x}{z}$
  - The percent change in one variable in response to a given percent change in another variable, holding all other relevant variables constant. In other words it summarizes the responsiveness of one variable to a change in another variable.
  - If an elasticity $\eta = 0$, we say the relationship is perfectly inelastic. If $0 < \eta < 1$, the relationship is inelastic (but not perfectly inelastic). If $\eta > 1$, the relationship is elastic.

Functional Forms and Marginal Effects

Choosing an appropriate functional form is a critical choice in econometric modeling. Your choice of model and selection of variables will greatly influence the fit of your model when mapping independent variables to your dependent variable.

1. Linear functions and Unit-Unit changes

If we assume a linear functional form, the model is: $y = \beta_0 + \beta_1 x$

Interpretation: First, take the derivative of the expression to get: $\frac{dy}{dx} = \beta_1$. Now, we can rewrite as:

$$\frac{\Delta y}{\Delta x} = \frac{dy}{dx} = \beta_1$$

Then we can rearrange to see that

$$\Delta y = \beta_1 \Delta x$$

Suppose $\Delta x = 1$, so that x changes by 1 **unit**. Then we can plug this into the above expression to see that y will change by β_1 **units**.

2. Logarithmic functions and Percent-Unit changes

If we assume a logarithmic functional form, the model is: $y = \beta_0 + \beta_1 \log(x)$

Interpretation: First, take the derivative of our model, $\frac{dy}{dx} = \frac{\beta_1}{x}$ and again notice that we can rewrite this (in this case it's an approximation, for small changes in x):

$$\frac{\Delta y}{\Delta x} \approx \frac{dy}{dx} = \frac{\beta_1}{x}$$

Then we can rearrange to see that

$$\Delta y = \beta_1 \frac{\Delta x}{x}$$

Suppose we know that x changes by 10 **percent**, so that the proportional change in x is 0.1: $\frac{\Delta x}{x} = 0.1$. Plug this value into the expression we derived, and we see that y will change by $\beta_1 * 0.1$ **units**.

3. Exponential functions and Unit-Percent changes

If we assume an exponential functional form, the model is: $y = e^{\beta_0 + \beta_1 x}$ or $\log(y) = \beta_0 + \beta_1 x$

Interpretation: Once again, we take a derivative of our model with respect to x to find $\frac{d\log(y)}{dx} = \beta_1$, and we rewrite it in terms of small changes in $\log(y)$ and x:

$$\frac{\Delta \log(y)}{\Delta x} \approx \frac{d\log(y)}{dx} = \beta_1$$

Use the fact that $\Delta \log(y) = \frac{\Delta y}{y}$:

$$\frac{\left(\frac{\Delta y}{y}\right)}{\Delta x} \approx \frac{d\log(y)}{dx} = \beta_1$$

Then we can rearrange to see that

$$\frac{\Delta y}{y} = \beta_1 \Delta x$$
$$\Rightarrow \%\text{ change} = \frac{\Delta y}{y} \times 100 = (\beta_1 \Delta x) \times 100$$

Suppose x changes by 1 **unit** and plug this into the expression we just derived. We see that the proportional change in y is β_1, and the percent change in y is $100 \times \beta_1$ **percent**.

4. Log-Log functions and Percent-Percent changes

If we assume a log-log functional form, the model is: $\log(y) = \beta_0 + \beta_1 \log(x)$

Interpretation: As usual, start by taking a derivative of our model, $\frac{d\log(y)}{dx} = \beta_1 \frac{1}{x}$ and re-writing it in terms of small changes:

$$\frac{\Delta \log(y)}{\Delta x} \approx \frac{d\log(y)}{dx} = \beta_1 \left(\frac{1}{x}\right) \Rightarrow \beta_1 \left(\frac{\Delta x}{x}\right) = \Delta \log(y) = \frac{\Delta y}{y}$$
$$\Rightarrow \frac{\Delta y}{y} = \beta_1 \left(\frac{\Delta x}{x}\right)$$
$$\Rightarrow \frac{\Delta y}{y} \times 100 = \beta_1 \left(\frac{\Delta x}{x}\right) \times 100$$

Suppose we know that x changes by 10 **percent**. Plug this value into the expression we derived, and we see that y will change by $\beta_1 * 10$ **percent**.

Practice

This Table (Table 2.3 in Wooldridge) is meant to practice and continue familiarizing ourselves with these functional forms.

Model	DepVar	IndepVar	How does Δy relate to Δx?	Interpretation
Linear	y	x	$\Delta y = \beta_1 \Delta x$	$\Delta y = \beta_1 \Delta x$
Logarithmic	y	$\log(x)$	$\Delta y = \beta_1 \frac{\Delta x}{x}$	$\Delta y = (\beta_1/100)\%\Delta x$
Exponential	$\log(y)$	x	$\frac{\Delta y}{y} = \beta_1 \Delta x$	$\%\Delta y = (100\beta_1)\Delta x$
Log-Log	$\log(y)$	$\log(x)$	$\frac{\Delta y}{y} = \beta_1 \frac{\Delta x}{x}$	$\%\Delta y = \beta_1 \%\Delta x$

Examples

Example 1. Suppose you've collected data on household gasoline consumption (gallons) in the Bay Area and gas prices ($ per gallon), and you estimate the following model:

$$\log(gasoline) = 12 - 0.21price$$

According to the model, how does gas consumption change when *price* increases by $1?

Sol. If *price* increases by $1, the model predicts that *gasoline* will decrease by 21%. This is an exponential model, our y variable is gasoline consumption and our x variable is the price of gas. From the table we see that in the exponential functional form.

$$\frac{\Delta y}{y} = \beta_1 \Delta x$$

This means that if x changes by 1 unit (i.e. $\Delta x = 1$), then the *proportional* change in y is $\beta_1 * 1$. If we plug in -0.21 for β_1, we know that $\frac{\Delta y}{y} = -0.21 * 1$. We can multiply the proportional change by 100 to get the percentage change in y: -21%.

Example 2. Professor Magruder uses firm data from Kenya to investigate how basket sales were affected by straw prices. In this example, he looks at the share of basket purchases that were made while baskets were on sale. The following model can be estimated:

$$\log(basketshare) = 0.83 + 0.491 \log(strawprice)$$

How does *basketshare* change if straw prices rise by 2%? Does this relationship make sense?

Sol. If the price of straw increases by 2%, the model predicts that the share of baskets sold on sale increases by 0.98%. This is a log-log model, our y variable is the share of baskets sold on sale, and our x variable is the price of straw. From the table we see that

$$\frac{\Delta y}{y} = \beta_1 \frac{\Delta x}{x}$$

If straw prices increase by 2% this is a proportional change of 0.02, so we know that $\frac{\Delta x}{x} = 0.02$. Plugging in this value and the estimated value of β_1, we can see that $\frac{\Delta y}{y} = 0.491 * 0.02 = 0.00982$. Because the *proportional* change in y is 0.00982, the percentage change in y is 0.982%. Another way to see this is that $\%\Delta y = \beta_1 \%\Delta x$ (multiply both sides of the formula from the table by 100), so that the percentage change in y, $\%\Delta y = 0.491 * 2\% = 0.98\%$.

Example 3. Suppose you've collected data on CEO salaries (hundred thousand $) and annual firm sales (million $), and you estimate the following model:

$$salary = 2.23 + 1.1 \log(sales)$$

According to the model, how does *salary* change if annual firm sales increase by 10%?

Sol. If annual firm sales increase by 10%, the model predicts that CEO salary increases by $11,000. This is a log model, our y variable is CEO salary and our x variable is annual firm sales. From the table we see that,

$$\Delta y = \beta_1 \frac{\Delta x}{x}$$

If annual firm sales increase by 10%, then we know that the proportional change in sales (x) is $\frac{\Delta x}{x} = 0.10$. We can plug this and our estimate of β_1 into the formula from the table to see that $\Delta y = 1.1(0.1) = 0.11$. Since the units of CEO salaries is $100,000, an increase of 0.11 units is an increase of $11,000.

Example 4. Wooldridge exercise 3.4: Here is the result of a regression of median salary for new law school graduates on their LSAT score, median undergraduate GPA of the class, number of volumes in the law library, cost of attendance, and rank of the law school (1 being the best). The unit of observation is a law school:

$$\widehat{\log(salary)} = 8.34 + .0047lsat + .248gpa + .095 \log(libvol) + .038 \log(cost) - .0033rank$$

How does the median salary change when *libvol* (volumes in law library) change?

Sol. If the number of volumes in a law school's law library increases by 10%, the predicted median salary of its graduates increases by 0.95%, controlling for LSAT scores, GPA, cost and rank.

The coefficient on *rank* is pretty small. Does this mean the rank of a law school doesn't matter a lot for graduates' salaries?

Sol. What does the rank measure? The quality of the school (and it might even be somewhat determined by the salaries of its graduates, I don't know). The rank of the school is probably partly determined by $lsat$, gpa, $\log(libvol)$, which all measure the quality of the school and its students. Once we control for these measures of school quality, rank is not very informative, because it itself is a proxy of school quality.

5. Random Variables and Distributions

a. Definitions

Let's briefly review the definitions of random variables and distributions

Random variables and their probability distributions: In essence, a random variable is a number that is taken from some distribution of possible outcomes. It can be **discrete** where there are a finite number of possible values (number of completed years of school) or **continuous** where there are infinite possible values (a person's height). Once a random variable is drawn from the distribution, it becomes the **realization** of a random number.

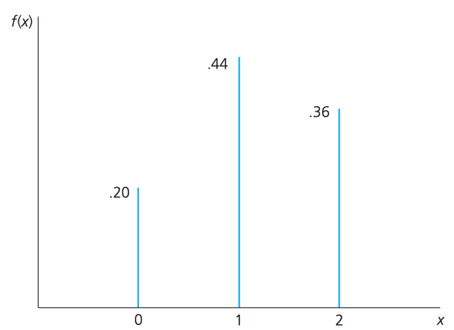
Why do we care about random variables? Let's start with an example from physics $\Delta x = v_{avg} \Delta t$. This is an example of a deterministic relationship, if we know the average velocity (v_{avg}) and the time that has an object has traveled (t), we know the change in it's position (Δx) with certainty. There are few (if any!) relationships like this in economics. If we know someone's education level and gender, we may have a good sense of their *expected* wages, but we don't have a formula for their exact wages. Thus, we treat wages, education, and gender as random variables, and explain their relationships using statistical techniques.

Any discrete random variable can be completely described by detailing the possible values it takes, as well as the associated probability that it takes each value. The **probability density function** (**pdf**) of X summarizes the information concerning the possible outcomes of X and the associated probabilities.

$$f(x_j) = P(X = x_j), \;\; j = \{1, 2, 3, 4, 5, ...k\}$$

$$f(0) = 0.20 \,; f(1) = 0.44 \,; f(2) = 0.36$$

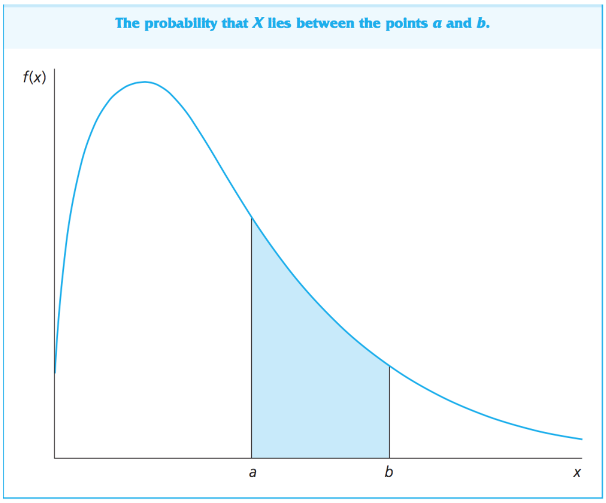
The pdf of the number of free throws made out of two attempts.



We can define a probability density function for continuous variables as well. However it doesn't make sense to talk about the probability that a continuous random variable takes on a particular value,[3] rather the pdf computes the probability of events involving a certain range. The probability that X takes on a value within the interval $[a, b]$ is given by

$$Pr(a < X < b) = \int_a^b f(x)dx$$

[3] "The idea is that a continuous random variable X can take on so many possible values that we cannot count them or match them up with the positive integers, so logical consistency dictates that X can take on each value with probability zero" (Woolridge p.717)



Cumulative Distribution Function: When computing probabilities for continuous random variables, it is easiest to work with the cumulative distribution function (**cdc**). The CDF of a random variable is defined as:

$$F(x) = P(X \leq x)$$

For discrete random variables, this is obtained by summing the pdf over all values x_j such that $x_j \leq x$. For a continuous random variable, F(x) is the area under the pdf, f, to the left of the point x. For a continuous random variable, F(x) is the area under the pdf to the left of the point x. Two important properties of cdf's that we will use later in the course:

$$P(X > c) = 1 - F(c)$$
$$P(a < X \leq b) = F(b) - F(a)$$

Joint Distributions, Conditional Distributions, and Independence: Let X, and Y be discrete random variables. Then (X,Y) have a joint distribution, which can be described by the **joint probability density function** of (X,Y):

$$f_{X,Y} = P(X = x, Y = y)$$

Two variables are **independent** if the joint PDF is equal to the product of the individual variables' pdf.

$$f_{X,Y} = f_X(x)f_Y(y)$$

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

We might also be interested in establishing how X varies with different values of Y: this is the conditional distribution of Y given X, which is described by the **conditional probability density function** :

$$f_{(Y|X)}(y|x) = P(Y = y|X = x)$$

b. Example

Take the following example of a survey of 652 women applying for a job at a factory. Two pieces of information that were collected include whether a woman was the head of her household and how much education she had completed. Look below at the following charts:

	Head of household	
	Yes	No
Incomplete primary	30	124
Primary only	44	192
Secondary	123	139

	Head of household	
	Yes	No
Incomplete primary	0.05	0.19
Primary only	0.07	0.29
Secondary	0.19	0.21

Note that the chart on the left gives the total number of women who fit in each *cell* of the chart. The sum of these cells is 652. From this chart, we could then calculate the chart on the right which tells us what proportion of women fall into each category. Each cell of the chart on the right provides us with the joint probability of two events happening.

- What is the joint probability that a randomly drawn person from the sample is a secondary school graduate and not a head of household? $f(secondary, no) = 0.21$
- What is the conditional probability that a randomly drawn head of household has not completed primary school? $f(Noprimary|yes) = 30/197 = 0.15$
- Is this the same as the (unconditional) probability of someone randomly drawn from the full sample not having completed primary school? *No*: $0.05 + 0.19 = 0.24 \neq 0.15$
- Are head of household status and education independent variables?
  Two variable are independent if

$$f_{X,Y} = P(X = x_j) \times P(Y = y_j)$$

or

$$f(x|y) = f(x)$$

Now using the first definition

$$\begin{aligned} f(secondary, headofhousehold) &= 0.19 \\ f(secondary) * f(headofhousehold) &= (0.19 + 0.21) * (0.05 + 0.07 + 0.19) = 0.40 * 0.31 = 0.12 \end{aligned}$$

Now using the second definition

$$\begin{aligned} f(secondary|headofhousehold) &= 123/197 = 0.62 \\ f(secondary) &= (123 + 139)/652 = 0.40 \end{aligned}$$

It looks like secondary education and head of household status are not independent, as women who have completed secondary education are more likely to be heads of household.

6. Features of Probability Distributions

a. The Expected Value

If X is a random variable, the **expected value** (or expectation) of X, denoted E(X), is the weighted average of all possible values of X. The weights are determined by the probability density function. The expected value is also called the population mean. Formally

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \cdots + x_k f(x_k) = \sum_{j=1}^{k} x_j f(x_j)$$

Note if X is continuous

$$E(X) = \int_{-\infty}^{+\infty} x f(x) d(x)$$

Now for a quick example:

x_j	$p(X = x_j)$
-1	1/8
0	1/2
2	3/8

$$E(X) = (-1)(1/8) + 0(1/2) + 2(3/8) = 5/8$$

b. The Variance and Standard Deviation

The **variance** tells us the expected distance from X to its mean:

$$\begin{aligned} Var(X) &= E[(X - E(X))^2] \\ &= \sum_{j=1}^{k} f(x_j)(x_j - E(X))^2 \text{ for discrete case} \\ &= \int_{-\infty}^{+\infty} f(x)(x - E(X))^2 dx \end{aligned}$$

Going back to our example:

$$Var(X) = \frac{1}{8}(-1 - \frac{5}{8})^2 + \frac{1}{2}(0 - \frac{5}{8})^2 + \frac{3}{8}(2 - \frac{5}{8})^2$$

Note the squaring eliminates the sign from the distance measure; the resulting positive value corresponds to the notion of distance, and treats values above and below symmetrically.

The **standard deviation** of a random variable, denoted sd(X) is the positive square root of the variance: $sd(X) = +\sqrt{Var(X)}$

Note

$$Var(aX + b) = a^2 Var(X)$$
$$sd(aX + b) = a \cdot sd(X)$$

This last property makes the standard deviation more natural to work with than the variance. As an example, take a random variable X measured in dollars. Next define Y=1000X. Suppose E(X)=20 and sd(X)=6. Then:

$$E(Y) = 1000E(X) = 20,000$$
$$sd(Y) = 1000sd(X) = 6,000$$
$$Var(Y) = (1000)^2 sd(X) = 6,000,000$$

The expected value and the standard deviation both increase by the same factor (1,000), whereas the variance of Y scales by 1,000,000.

c. Sample Mean and Law of large numbers

In this class we talk a lot about samples versus populations. In an ideal world we would have data on the universe of individuals in order to estimate a relationship between X and Y. In other words we would have data on the population at large. However, it's almost always the case that we do not have all this data at our disposal, rather we have a sample of individuals (Berkeley students rather than the universe of students at American universities). We can calculate statistical properties of these smaller samples: which we call sample mean and sample variance. The sample mean is given below:

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

The law of large numbers says that if we draw a sample consisting of n realizations of our random variable, and take the average, this sample mean will approach the population mean as n approaches infinity.

As an example, let X be the roll of a die, which can take on values 1,2,3,4,5,6. The population average (or the mathematical expectation) is the average if we were to throw the die infinitely many times:

$$E(X) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = 3.5$$

Next, lets calculate the sample mean, which is the value we get after only 100 throws of the die:

n	x_j	$\bar{X}_n$
2	6,6	12/2=6
3	1,2,2	5/3= 1.67
5	1,1,6,3	11/4=2.75
⋮		

As n→ ∞ any irregularities that occur due to the small sample size are muted, and the sample mean will converge to the population mean.

d. Sample Variance

Let X denote a random variable, with Variance Var(X). The sample variance is given by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Note: it seems like we should divide by n, but instead we divide by $n - 1$. We do this to ensure that the sample variance estimator is an unbiased estimator of population variance (lots of terminology there, which we will explore later)