LLN, CLT, NORMAL MODELS AND LINEAR REGRESSION

Contents

1. Reading

- Read chapter 4 of [1]

2. Learning objectives

- The Central limit theorem (what it does and does not tell us).
- Working with Normal random variables (linear transformations of them, calculating probabilities).
- Correlation Normal random variables (Multivariate normal distribution).
- Basic linear regression model with a single predictor.

3. The central limit theorem and sample distribution

- We now have the formalism in place to state the Central Limit Theorem (CLT) in more precise terms.

Theorem 1. *Let X_i be a sequence of iid random variables and let*

$$E[X_i] = \mu, \quad \text{var}(X_i) = \sigma^2$$

and set

$$S_N = \sum_{i=1}^{N} X_i.$$

Then

$$P\left(\frac{S_N - N\mu}{\sqrt{N\sigma^2}} < z\right) \to P(Z < z) \tag{1}$$

Where

$$Z \sim \text{Normal}(0, 1)$$

- The normal variable Z with zero mean and variance one is called a **standard normal** random variable. Since evaluations of the CDF of a standard normal random variable appear so often, we use the shorthand

$$\Phi(z) = P(Z < z).$$

Example 1 (Binomial)**.** *Let*

$$Y \sim \text{Binomial}(N, q)$$

Question: Assume N is even and use the central limit theorem to approximate $P(Y < N/2)$ with a Normal distribution. How does the accuracy depend on N and q?

Solution: Using that $\mu = E[X_i] = q$ and $\sigma^2 = \text{var}(X_i) = q(1-q)$, we find that the normal approximation to Y is

$$P\left(\frac{Y - Nq}{\sqrt{N\sigma^2}} < z\right) \to P(Z < z)$$

for

$$Z \sim \text{Normal}(0, 1).$$

Now we write

$$\begin{aligned} P(Y < N/2) &= P(Y - Nq < N/2 - Nq) \\ &= P\left(\frac{Y - Nq}{\sqrt{Nq(1-q)}} < \frac{N/2 - Nq}{\sqrt{Nq(1-q)}}\right) \\ &= P\left(\frac{Y - Nq}{\sqrt{Nq(1-q)}} < \sqrt{N}\frac{1-2q}{2\sqrt{q(1-q)}}\right) \\ &\to P\left(Z < \sqrt{N}\frac{1-2q}{2\sqrt{q(1-q)}}\right) = \Phi\left(\sqrt{N}\frac{1-2q}{2\sqrt{q(1-q)}}\right) \end{aligned}$$

We can use colab to plot both sides of this equation.

Do Exercise 3

- **Note on iid assumption:** One of the most important things to recognize about the CLT when it comes to application is that the assumptions that the X_i are independent are not that important, so long as they are not too correlated. Even though the precise quantitive statement of the CLT won't when there are correlations, the sum will still be well approximated by a Normal distribution. We can state this in what I call the qualitative CLT. This is a point we will return to later!

3.1. Properties of Normal random variables.

- **Linear transformations of Normal random variables:** Suppose

$$Z \sim \text{Normal}(0, 1)$$

and define

$$X = \sigma Z + \mu$$

Then

$$\begin{aligned} P(X < x) &= P(\mu + \sigma Z < x) \\ &= P\left(Z < \frac{x - \mu}{\sigma}\right) \end{aligned}$$

Hence

$$X \sim \text{Normal}(\mu, \sigma^2).$$

- With this understanding of how to linearly transform a Normal random variable, we can see that the CLT can be informally stated as

$$S_N \approx S_{\text{CLT}} \sim \text{Normal}(N\mu, N\sigma^2)$$

- More generally,

$$X \sim \text{Normal}(\mu_x, \sigma_x)$$

Now consider

$$Y = aX + b$$

At this point it should make sense that Y is also normal. but what are the mean and variance? Taking the average of both sides,

$$E[Y] = a\mu + b$$

and

$$\text{var}(Y) = \text{var}(aX) + \text{var}(b)$$

Form the formula for variance, we know $\text{var}(aX) = a^2\text{var}(X)$. Also, $\text{var}(b) = 0$ So

$$Y \sim \text{Normal}(a\mu_x + b, a^2\sigma_x^2).$$

Note that in going from Z to X and X to Y, we are just multiplying and shifting everything. Think about what this does to the histogram.

- The process of going from X to Z is called standardizing. For any variable X the **standardized** variable is defined as

$$Z = \frac{X - \mu_x}{\sigma_x}$$

Transforming X to a standard Normal is equivalent to measuring X in units of standard deviations. For example, if we make a histogram of X, all this transformation does is change the X axis to units of standard deviations from the mean.

- **Combing normal random variables:** Suppose

$$\begin{aligned} Y_1 &\sim \text{Normal}(\mu_1, \sigma_1^2) \\ Y_2 &\sim \text{Normal}(\mu_2, \sigma_2^2) \end{aligned}$$

What is the distribution of $X_1 + X_2$? One way to see that this should be Normal is to note that Normal random variables emerge form the CLT as sums over many (roughy) iid variables. Let's say Y_1 and Y_2 are respectively approximations to sums S_1 and S_2 over N_1 and N_2 terms. Let f_X be the distribution of terms in the first sum and $f_{X'}$ the second. Then, if we sample a random term from the $N_1 + N_2$ terms that make up the sum $S_1 + S_2$, the chance that is was a sample from f_X is $N_1/(N_1 + N_2)$. The chance it come from $f_{X'}$ is $N_2/(N_1 + N_2)$. Therefore, it has a distribution which is

$$f_X \frac{N_1}{N_1 + N_2} + f_{X'} \frac{N_2}{N_1 + N_2}.$$

If we draw another sample from these $N_1 + N_2$ terms, its distribution will depend on the value of the first sample we obtained. For example, if samples from f_X are very likely to take on very large values and $f_{X'}$, then a large value of the first sample we drew will tell us there are probably now less of the terms from the first sum in our list of $N_1 + N_2$ numbers. However, the distribution of $S_1 + S_2$ will still be approximately normal, since as we already mentioned, these correlations are not that important. This is illustrated in Figure 1.

- We can summarize everything we learned above in the following result.

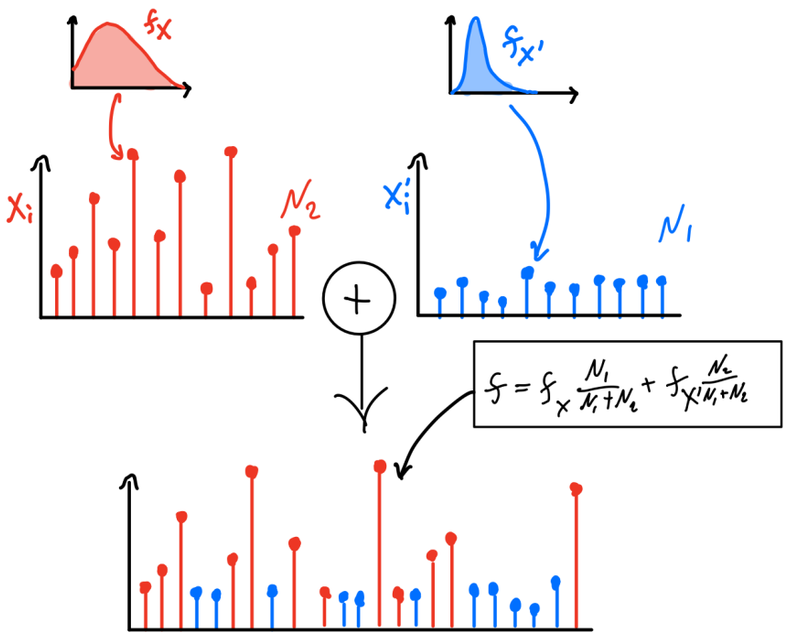

Figure 1. If we add two sums containing N_1 and N_2 terms respectively then the resulting sum behaves similar to the sum over $N_1 + N_2$ variables whose distribution is a mixture of the two. In particular, if we pick a random term from the sum, there is $N_1/(N_1 + N_2)$ chance for it to follow the distribution of the first N_1 terms (denoted f_X in the figure) and an $N_2/(N_1 + N_2)$ chance for it to follow the second distribution.

Theorem 2 (Special case of Theorem 4.6.1 in [1])**.** *Let*

$$X_1 \sim \text{Normal}(\mu_1, \sigma_1^2)$$
$$X_2 \sim \text{Normal}(\mu_2, \sigma_2^2)$$

be independent, then

$$aX_1 + bX_2 + d \sim \text{Normal}\left(a\mu_1 + b\mu_2 + d, a^2\sigma_1^2 + b^2\sigma_2^2\right)$$

4. Regression modeling

- Now we want to understand relationships between Normal random variables.

Example 2 (A taste of linear regression)**.** *Consider the following model:*

$$X \sim \text{Normal}(\mu_x, \sigma_x^2)$$
$$Y|X \sim \text{Normal}(\beta_1 X + \beta_0, \sigma^2)$$

Question: What is the marginal distribution of Y? What is $E[XY]$? How does this compare to $E[X]E[Y]$?

Solution: We know that

$$Y|X = \beta_1 X + \beta_0 + Z, \quad Z \sim \text{Normal}(0, \sigma^2)$$

Thus, the marginal distribution of Y is the sum of two Normal random variables with mean and variance $(\beta_1\mu_x + \beta_0, a\sigma_x^2)$ and $(0, \sigma^2)$ respectively. By Theorem 2,

$$Y \sim \text{Normal}(\beta_1\mu_x + \beta_0, \beta_1^2\sigma_x^2 + \sigma^2)$$

To compute $E[XY]$, we note that

$$E[XY|X = x] = E[xY|X = x] = xE[Y|X = x]$$

therefore

$$E[XY] = E[XE[Y|X]] = E[X(\beta_1 X + \beta_0)] = \beta_1 E[X^2] + \beta_0 E[X]$$

Using

$$E[X^2] = \text{var}(X) + E[X]^2 = \sigma_x^2 + \mu_x^2$$

Therefore

$$E[XY] = \beta_1\sigma_x^2 + \beta_1\mu_x^2 + \beta_0\mu_x$$

On the other hand,

$$E[X]E[Y] = \mu_x(\beta_1\mu_x + \beta_0) = \beta_1\mu_x^2 + \beta_0\mu_x$$

The difference between the two is the additional term $\beta_1\sigma_x^2$, which we picked up from the variance of x.

- The example above motivates the definition of **covariance**

$$\operatorname{cov}(X,Y) = E[XY] - E[X]E[Y] \tag{2}$$

Note that another way to write this is

$$E[(Y - E[Y])(X - E[X])] = E[XY] - 2E[X]E[Y] + E[X]E[Y] = \operatorname{cov}(X,Y)$$

so if we replaced X with Y, this becomes the variance.

- We can generalize the relationship between the covariance, slope (a) and variance of X (σ_x^2) in Example 2 to any model of the form

$$\begin{aligned} X &\sim \text{some distribution with mean } \mu_x \text{ and variance } \sigma_x^2 \\ Y|X &\sim \text{Normal}(\beta_1 X + \beta_0, \sigma^2). \end{aligned} \tag{3}$$

Such a model is a linear **linear regression model**. The variable X is called the **predictor** and Y the **response** variable. Recall that we can write $E[X^2]$ as

$$E[X^2] = \operatorname{var}(X) + E[X]^2 = \sigma_x^2 + \mu_x^2.$$

Now for any linear regression model

$$\begin{aligned} E[XY] &= E[XE[Y|X]] = E[X(\beta_1 X + \beta_0)] = \beta_1 E[X^2] + \beta_0 E[X] \\ &= \beta_1 \sigma_x^2 + \beta_1 \mu_x^2 + \beta_0 \mu_x \\ E[Y] &= \beta_1 \mu_x + \beta_0 \end{aligned}$$

so

$$\operatorname{cov}(X,Y) = \beta_1 \sigma_x^2$$

regardless of the distribution of X (assuming $\sigma_x^2 < \infty$). The intuition for this formula is given in figure 2.

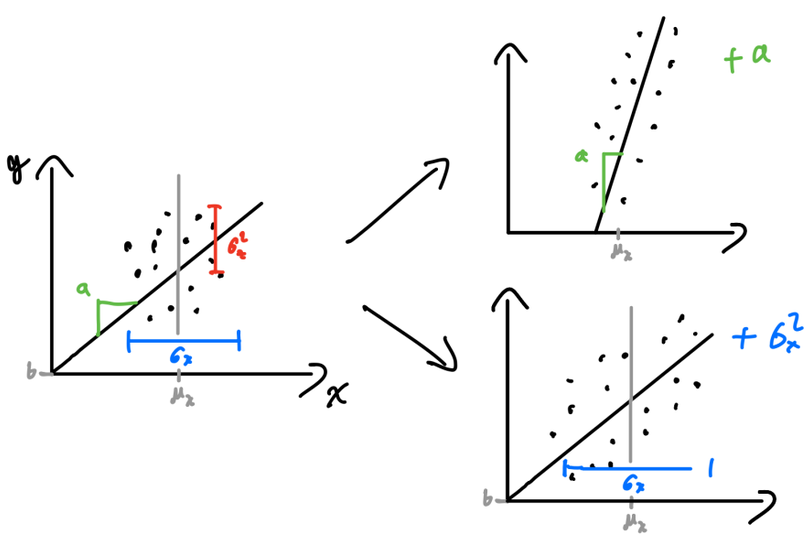

Figure 2.

Do Exercises 4 an 5

- A crucial observation is that the covariance allows us to relate the parameter β_1 (the slope) in the model above to averages over X and Y. In other words, it provides us with a means to estimate the slope from samples $(x_1, y_2), \ldots, (x_n, y_n)$.

$$\beta_1 \approx \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{X})^2}$$

The is what the function

```
> np.cov(x,y)[0,1]
```

computes in Python. The reason for the $[0,1]$ is that the covariance function in numpy actually computes a 2D array (a Matrix), where the off diagonal entries are the covariance. The diagonal entries are the variances.

We can also estimate β_0. Using $E[Y] = \beta_1 \mu_x + \beta_0$ we have

$$\beta_0 = E[Y] - \beta_1 \mu_x \approx \hat{\beta}_0 = \overline{Y} - \left(\frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{X})^2} \right) \overline{X}$$

4.1. **Least square interpretation.**

- Suppose we plot X and Y points in a place. Regardless of where these X and Y points come from (Normal model or not), we can compute $\hat{\beta}_1$ and $\hat{\beta}_0$. These estimators are known as **least squares** estimators (note that we haven't formally defined what an estimator is) because it happens that these values minimize the sum of the squared difference between our data points and the line $\hat{a}x + \hat{b}$. That is, they are the values that make the **residual sum of squares**(RSS) smallest

$$RSS = \sum_{i=1}^{n} r_i^2, \quad r_i = Y_i - (\hat{\beta}_1 X_i + \hat{\beta}_0)$$

smallest. The R

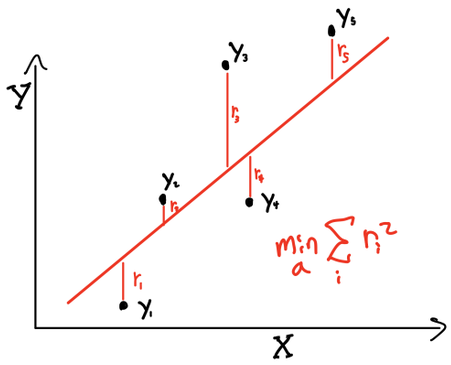

Figure 3.

There are many other ways we could draw a line through a set of (x, y) points. This particular way of estimating the slope – by minimizing RSS – happen to make sense under the assumption that the data is sampled from a Linear regression model (Equation 3).

Example 3 (Marketing data)**.** *Here we consider the some on advertising budgets and sales for a company. We will explore whether the budget for TV advertisements is associated with higher sales.*

Questions: Fit the data to a linear regression model with the TV budget at the predictor and sales as the response variable.

(a) **Fit linear regression model:** *What are the estimates of β_1 and β_0?*
(b) **Visualize the data:** *Plot the regression long along with a scatter plot of the data.*
(c) **Accessing model assumptions:** *Using the fitted values of β_1 and β_0, simulate 10 "fake" data sets which have the same number of points as the real data set and the same x values. Make plots of these and compare to the real data.*

Solution: see colab notebook.

References

[1] John Tabak. *Probability and statistics: The science of uncertainty*. Infobase Publishing, 2014.