Notation

First a bit of notation. The symbol to the left of the semicolon is the variable which we are measuring or using as an input to the function. In Eq. 1 below this is x. The symbols to the right of the semicolon are known. Below, these are μ and σ. This if often read as "f as a function of x parameterized by μ and σ" or "f as a function of x given μ and σ"

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{1}$$

To reiterate, we are measuring x and we already know μ and σ.

For many problems, however, we have a collection of data and wish to obtain μ and σ. Notationally, we can express this as,

$$f(\mu,\sigma;x_1 \ldots x_N) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}. \tag{2}$$

This can be read as f as a function of μ, and σ given x_i through x_N. In other words, μ and σ are our inputs, and the x-values are fixed.

Why Maximize the Likelihood?

This is by no means rigorous, but I want to motivate why we might choose our values of μ and σ that maximize the likelihood given our collection of date. In Fig. 1, I sample 10 points from the normal distribution with $\mu = 2$ and $\sigma = 2$. These are the blue circles on the x-axis. I also plot the likelihood function using those parameters. For clarity, I added dotted lines showing the corresponding likelihood of each sampled x value.

Notice where most of these points cluster. There are a couple of outliers, but most cluster around $x = 2$. Notice also the value of the distribution function for these points. The likelihood is high relative to the outliers. In fact, the likelihood is close to its maximum value.

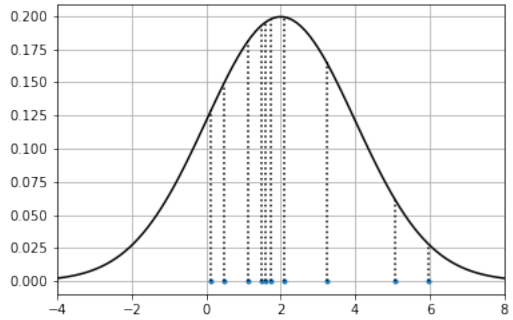

Figure 1: 10 points sampled from the normal distribution with $\mu = \sigma = 2$. I also plot the density function for these parameter values.

So it makes sense to assume the likelihood of a point is near the maximum. Therefore, when we try to estimate μ and σ we choose values that maximize the likelihood.

Likelihood of Multiple Observations

While likelihood is not exactly the same as probability, think of it in a similar way. If we have, for example a fair die, we can ask how to calculate the probability of getting three sixes in a row. We can do that by multiplying the probability of each event together. Since the probability of rolling a six is 1/6, the probability of three sixes in a row is,

$$\frac{1}{6} \times \frac{1}{6} \times \frac{1}{6}.$$

It is the same for the likelihood. The likelihood of N points is the product of the likelihood of each individual event. In mathematical form,

$$f = f(x_1)f(x_2)...f(x_{N-1})f(x_N) = \prod_{i=1}^{N} f(x_i)$$

so for the normal distribution,

$$f(\mu, \sigma; x_1...x_N) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} \tag{3}$$

Finding the Maximum of a Function

In calculus, we learn one can find the maximum (or minimum or inflection point) of a function by taking its derivative, setting it equal to zero, then solving for the variable. In Fig. 2, I plot the normal distribution with $\mu = 2$ along with the tangent line at three different points (x = -1, 2, and 5). Recall the the derivative gives us the slope of the tangent line. Notice that for values other than $x = 2$, the tangent line has a finite slope. However, at $x = 2$, it is flat. In other words has zero slope. This corresponds to our maximum.

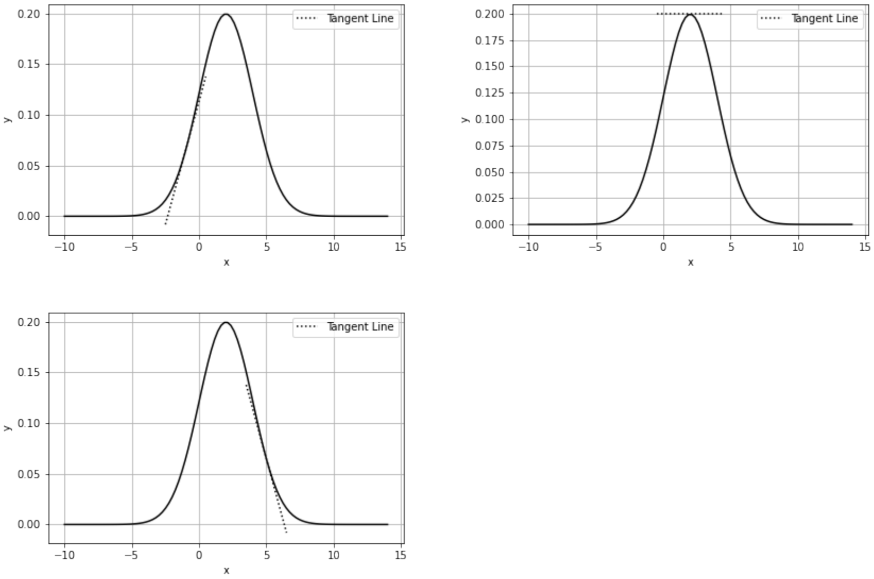

Figure 2: Plots of three different tangent lines.

Our problem is only slightly more difficult than one might find in an introductory calculus course. We have a function that depends on two values, namely μ and σ. Therefore we are going to have to take derivatives with respect to each of these variables resulting in us having to solve two equations,

$$\frac{\partial f}{\partial \mu} = 0$$

and,

$$\frac{\partial f}{\partial \sigma} = 0$$

.

Why Use the Log of the Likelihood?

In practice, we maximize the log of the likelihood. This will simplify the calculation of the derivative. Taking the log does not affect the position of the maximum.

I am going to make use of the fact

$$\ln(AB) = \ln(A) + \ln(B). \tag{4}$$

Thus, we will transform Eq. 3 from a product into a summation. We can then differentiate term by term.

Estimation of μ

First we take the log of the likelihood,

$$\ln(f(\mu,\sigma;x_1...x_N)) = \ln\left[\prod_{i=1}^{N}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right].$$

This becomes

$$\ln(f(\mu,\sigma;x_1...x_N)) = \sum_{i=1}^{N}\left[\ln\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right].$$

Now, we need to take the derivative and set it equal to zero,

$$\frac{d}{d\mu}\sum_{i=1}^{N}\left[\ln\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right] = 0.$$

Now, we use the property of logs to simplify things a bit,

$$\frac{d}{d\mu}\sum_{i=1}^{N}\left[\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \ln\left(e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right)\right] = 0.$$

The log of an exponential just gives us the exponent so we can further simplify,

$$\frac{d}{d\mu}\sum_{i=1}^{N}\left[\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \left(-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)\right] = 0.$$

The first term is constant with respect to μ so its derivative is zero,

$$\frac{d}{d\mu}\sum_{i=1}^{N}\left(-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)=0.$$

This gives us,

$$\sum_{i=1}^{N}-\left(\frac{x_i-\mu}{\sigma}\right)\left(\frac{-1}{\sigma}\right)=0.$$

After multiplying through by σ we have,

$$\sum_{i=1}^{N}(x_i-\mu)=0.$$

We can split this into two summations,

$$\sum_{i=1}^{N}x_i-\sum_{i=1}^{N}\mu=0$$

The sum of μ N times is just $N\mu$ so we get,

$$\sum_{i=1}^{N}x_i-N\mu=0.$$

Finally,

$$\mu=\frac{1}{N}\sum_{i=1}^{N}x_i.$$

It turns out our optimal μ is just the mean of our data points.

Estimation of σ^2

As with the estimation of μ, we differentiate the log likelihood with respect to σ and set it equal to zero,

$$\frac{\partial}{\partial\sigma}\left[\ln\prod_{i=1}^{N}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right]=0. \tag{5}$$

Making use of Eq. 4, we can rewrite this as,

$$\frac{\partial}{\partial\sigma}\sum_{i=1}^{N}\left[\ln\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right)\right]=0. \tag{6}$$

Again using Eq. 4

$$\frac{\partial}{\partial\sigma}\sum_{i=1}^{N}\left[\left(\ln\left[\frac{1}{\sigma\sqrt{2\pi}}\right]\right)+\ln\left[e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right]\right]=0. \tag{7}$$

Above in the right-hand term involving he log, we are taking the log of an exponential leaving us just with exponent,

$$\frac{\partial}{\partial\sigma}\sum_{i=1}^{N}\left[\left(\ln\left[\frac{1}{\sigma\sqrt{2\pi}}\right]\right)-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right]=0. \tag{8}$$

Now, we differentiate term by term to get,

$$\sum_{i=1}^{N}\left[\frac{-1}{\sigma}+\frac{(x_i-\mu)^2}{\sigma^3}\right]=0. \tag{9}$$

We then multiply every term by σ getting,

$$\sum_{i=1}^{N}\left[-1+\frac{(x_i-\mu)^2}{\sigma^2}\right]=0. \tag{10}$$

Singe we are summing -1 N times, and we can pull σ out of the summation, this further simplifies to,

$$-N+\frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2=0. \tag{11}$$

Solving for σ^2 we arrive at,

$$\sigma^2=\frac{1}{N}\sum_{i=1}N(x_i-\mu)^2,$$

which is just the formula for the variance.