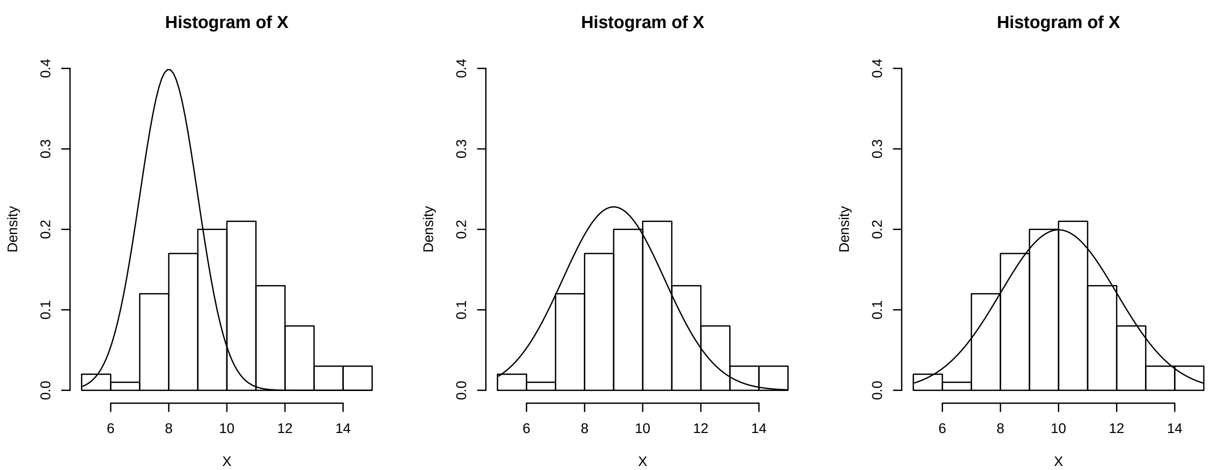
MOTIVATION: MULTIVARIATE VS UNIVARIATE

Before we discuss random vectors, here is some statistical motivation for why we care about multivariate distributions. These emphasize two things: First, linear algebra is a fundamental part of applied statistics. Don't avoid it, embrace it! Second, a common use of *density* and *probability mass functions* for parameter estimation are to define *likelihoods*, which are joint mass (or density) functions but where we flip-flop our notions about which quantities in these equations are constants vs variables.

PROBABILITY DENSITY VS LIKELIHOOD

Here's a crude, graphical way of fitting a normal distribution to a large number of iid data: Plot a histogram, choose an initial mean μ and variance σ^2 then overlay the corresponding normal density curve. Adjust your guesstimates until it looks like a good fit. In **R**...

```r
set.seed(661); ## See ?set.seed or ask me :-)
X=rnorm(100,10,2); ## 100 replicates drawn from Normal(mean=10,sd=2)
par(mfrow=c(1,3));
xvals = seq(min(X),max(X),length=100); # for plotting...
hist(X,freq=FALSE,ylim=c(0,.4)); points(xvals,dnorm(xvals,8,1),type="l")
hist(X,freq=FALSE,ylim=c(0,.4)); points(xvals,dnorm(xvals,9,1.75),type="l")
hist(X,freq=FALSE,ylim=c(0,.4)); points(xvals,dnorm(xvals,10,2),type="l")
```



Formally, we'd like to compute some "goodness of fit" measure instead of just trusting our intuition with what "looks like a good fit". This might be the *SSD* (sometimes called the *sum of squared errors* [*SSE*]) from the OLS example above, but another options comes from some theoretical results in mathematical statistics: the *likelihood* of parameters μ and σ given the data X. Here, our estimates are the values of μ and σ that maximize the likelihood.

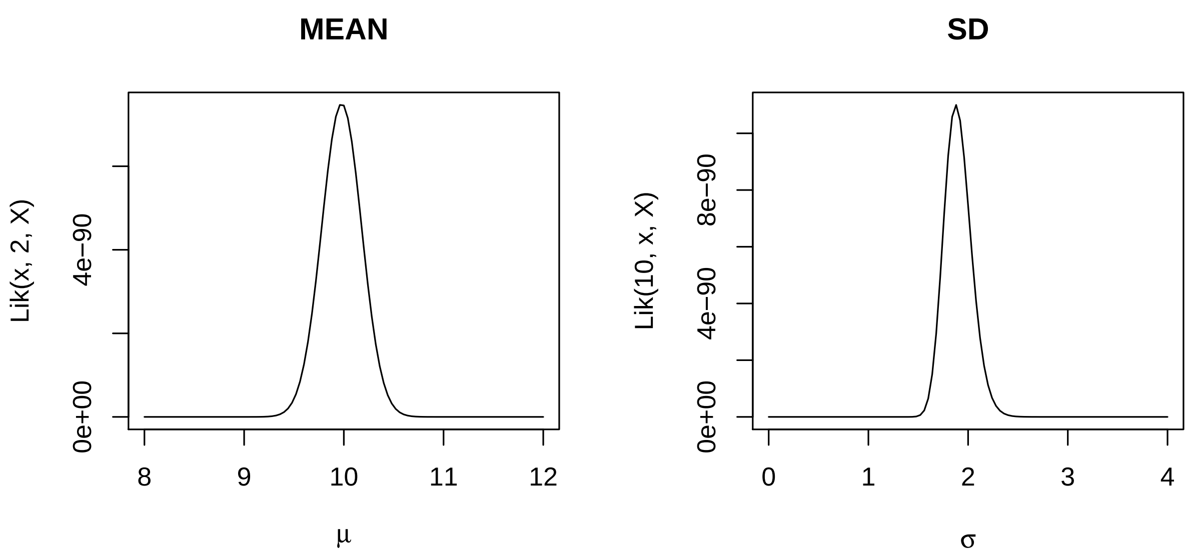
What is this likelihood? This is defined by the distribution for random vector X, but where we flip around our notion of what's fixed and what varies. That is, we treat the x values (our data) as fixed and our candidate parameter estimates μ and σ are treated as variable quantities. Lets look at a specific example to see how we define and use a likelihood function in practice.

Likelihood Example: Assume all X_i are iid with normal density $f(x_i; \mu, \sigma)$. This implies the joint density $f_X(x_1, ..., x_n; \mu, \sigma) = \prod_{i=1}^n f(x_i, \mu, \sigma)$. Here we can write it as a simple product, thanks to the independence of the individual random variables. This density function defines the likelihood function for parmeters μ and σ

$$\mathcal{L}(\mu, \sigma; \mathbf{x}) = \prod_{i=1}^{n} f(x_i, \mu, \sigma)$$

Note that we've gone from a function of n variables (number of data points) down to a function of 2 variables (number of parameters), and our domain is no longer the sample space but is instead the range of possible parameters ($\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$). Again, in **R**...

```r
par(mfrow = c(1, 2))
Lik = Vectorize(function(mu, sd, xs) prod(dnorm(xs, mean = mu, sd = sd)), "mu")
# fix sd=2, vary mu
curve(Lik(x, 2, X), from = 8, to = 12, main = "MEAN", xlab = expression(mu))
# fix mu, vary sd
Lik = Vectorize(function(mu, sd, xs) prod(dnorm(xs, mean = mu, sd = sd)), "sd")
curve(Lik(10, x, X), from = 0, to = 4, main = "SD", xlab = expression(sigma))
# Optimization algorithms can then be used to refine estimates.
```



Concluding Remark: We typically do statistics by treating all of our data as a single outcome from a joint distribution, even when estimating values from a simple univariate distribution!

ORDINARY LEAST SQUARES (OLS)

Suppose you have data y_i that are assumed to be observations of normally distributed random variables Y_i with standard deviation σ and a mean μ_i that depends on different factors X_i that can be manipulated (or that can otherwise vary) for each experiment. For example, heights of individuals (Y_i) as a function of age, gender, etc. might look like

$$Y_i = Normal(\mu_i = \beta_0 + \beta_i X_{i1} + \cdots + \beta_k X_{ik}, \sigma)$$

Since a normal r.v. with mean μ and standard deviation σ can be written as μ plus a normal r.v. with mean 0 (i.e., $\mu + N(0, \sigma)$) it follows that

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i$$

where each ϵ_i are independent normals with mean 0 and standard deviation σ. Writing these n equations in matrix form yields

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 + X_{11} + \cdots + X_{1k} \\ 1 + X_{21} + \cdots + X_{2k} \\ \vdots \\ 1 + X_{n1} + \cdots + X_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

or written in more compact matrix and vector notation,

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

Note that $E(\mathbf{Y}) = \mathbf{X}\beta$. Assuming the observed outcomes (data) $y = (y_1, \cdots, y_n)^T$ and inputs $\mathbf{X}$ are known, and the goal is to estimate best-fit parameters β (call this estimate $\hat{\beta}$). A good way to compute that estimate is to take the *sum of squared differences* (SSD) between the observed data and the expected model output for a given set of parameters β (i.e., $SSD = r^T r$ where $r = y - E(\mathbf{Y})$; a measure of "distance" between model and data) then use the β that minimizes that distance as our estimate for $\hat{\beta}$. It can be shown with a little linear algebra that

$$\hat{\beta} = (\mathbf{X^T X})^{-1} \mathbf{X^T y}$$

Therefore we've used linear algebra and a little multivariate calculus to turn an optimization problem into a relatively simple matrix computation!

Concluding Remark: In this course, we will tend to use univariate and simple multivariate examples to facilitate learning important concepts in probability. *However*, in practice, *statistics is a multivariate endeavor* and therefore it pays to be familiar with these basic probability concepts in a multivariate setting, and also the basic linear algebra tools used in both theoretical and applied statistics.