VectorBiTE Training 2018
Methods Workshop

Basics of Probability and Statistics

www.vectorbite.org

Assumed Background

I assume that you've seen most of this before. This material is covered in more depth in the assigned reading. In particular, I expect that you know:

- axioms of probability and their consequences.
- conditional probability and Bayes theorem
- definition of a random variable (discrete and continuous)
- the idea of a probability distribution and likelihood

We'll do a VERY fast review of these and do some practice exercises, including with R.
Note – We won't review confidence intervals, p-values, the central limit theorem, etc.

Some probability notation

We have a set, S of all possible events. Let $\Pr(A)$ be the probability of event A. Then:

- A^c is the complement to A (all events that are not A).
- $A \cup B$ is the union of events A and B ("A or B").
- $A \cap B$ is the intersection of events A and B ("A and B").
- $\Pr(A|B)$ is the conditional probability of A given that B occurs.

Axioms of Probability

- $\sum_{i \in S} \Pr(A_i) = 1$, where $0 \leq \Pr(A_i) \leq 1$
- $\Pr(A) = 1 - \Pr(A^c)$
- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$
- $\Pr(A \cap B) = \Pr(A|B)\Pr(B)$
- If A and B are independent, then $\Pr(A|B) = \Pr(A)$

Probability Practice

Supposed you have two fair 6 sided die.

1. What is the probability that you roll dice and get two of the same number?
2. What is the probability that you roll the dice and get snake eyes?
3. What is the probability that your dice the sum will equal 7? What is the probability that the sum will be 7 if the first die is a 2?

Bayes Theorem

Bayes Theorem allows us to related the conditional probabilities of two events A and B:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

$$= \frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)\Pr(A) + \Pr(B|A^c)\Pr(A^c)}$$

Example: Drug Testing

All athletes at the Olympics are subject to random drug screening. The probability that an athlete uses drugs is 0.1%. The drug test has a sensitivity of 99% (the proportion of drug users who test positive) and a specificity of 95% (the proportion of those who are drug free that test negative). An athlete tests positive for the drug. What is the probability that they actually use the drug? That is, find:

$$\Pr(\text{They are a drug user}|\text{positive test}) \rightarrow \Pr(D|+)$$

Example: Drug Testing

$$\Pr(D|+) = \frac{\Pr(+|D)\Pr(D)}{\Pr(+)}$$

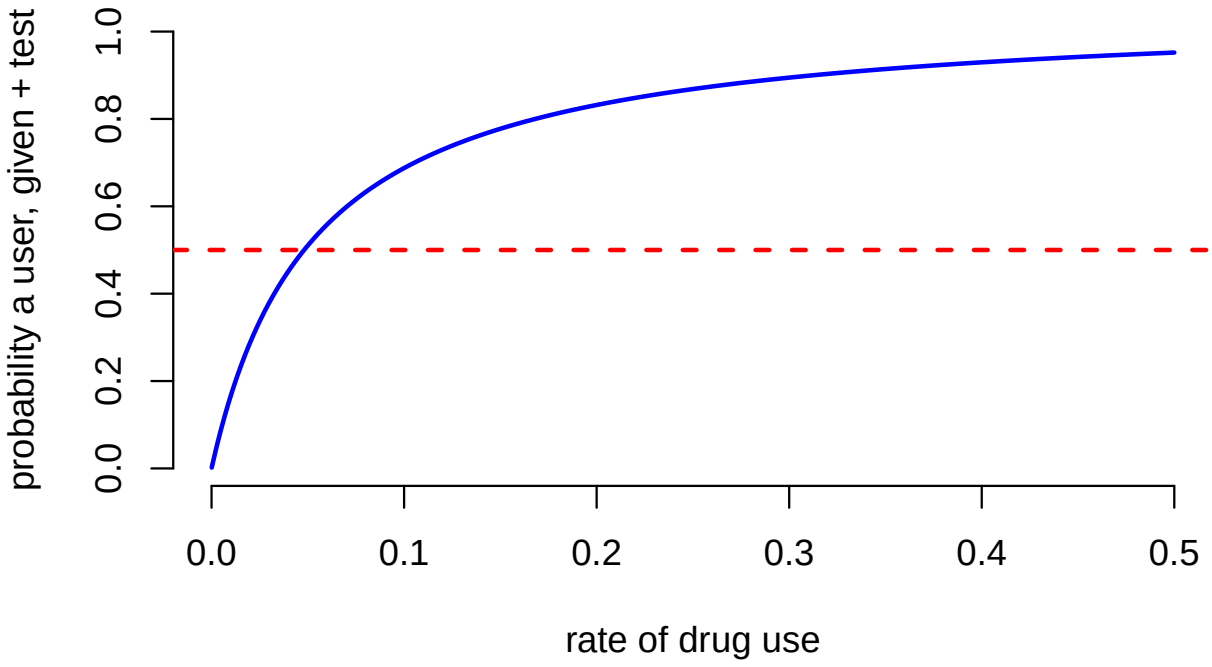
$$= \frac{\Pr(+|D)\Pr(D)}{\Pr(+|D)\Pr(D) + \Pr(+|ND)\Pr(ND)}$$

$$= \frac{\Pr(+|D)\Pr(D)}{\Pr(+|D)\Pr(D) + (1 - \Pr(-|ND))\Pr(ND)}$$

$$= \frac{0.99 \times 0.001}{0.99 \times 0.001 + (1 - 0.95) \times 0.999} \approx 0.02$$

Example: Drug Testing

The underlying rate of the event is very important.



Folks often forget this. It's known as the Base Rate Fallacy, and is related to the Prosecutor's Fallacy.

Random Variables (RVs)

A random variable is a variable whose value is subject to randomness. Its possible values and their probabilities are described by a probability distribution. They come in two varieties

- discrete (numbers of items or successes)
- continuous (heights, times, weights)

Some Notation

- X, Y - capital, sometimes bold or with subscripts - RVs
- x, y, k - lower case - the values that the RV takes
- $f(x)$, f_k - functions that return a value for an input x or k
- $\Pr(\cdot)$ - "Probability of $\cdot$"
- $\mathbb{E}[\cdot]$ - "the expected value of $\cdot$"

Probability Distributions

A Probability Distribution is a "is a mathematical description of a random phenomenon in terms of the probabilities of events." (Wikipedia)

That is, it's a mathematical function that gives the probabilities of an RV taking on various alternative values.

Discrete RVs

For discrete RVs, the distribution of probabilities is described by the probability mass function (pmf), f_k such that:

$$f_k \equiv \Pr(X = k)$$

$$\text{where } 0 \leq f_k \leq 1 \text{ and } \sum_k f_k = 1$$

For example, for a fair 6-sided die:
$f_k = 1/6$ for $k = \{1, 2, 3, 4, 5, 6\}$

Discrete RVs

Related to the PMF is the cumulative distribution function (cdf), $F(x)$.

$$F(x) \equiv \Pr(X \leq x)$$

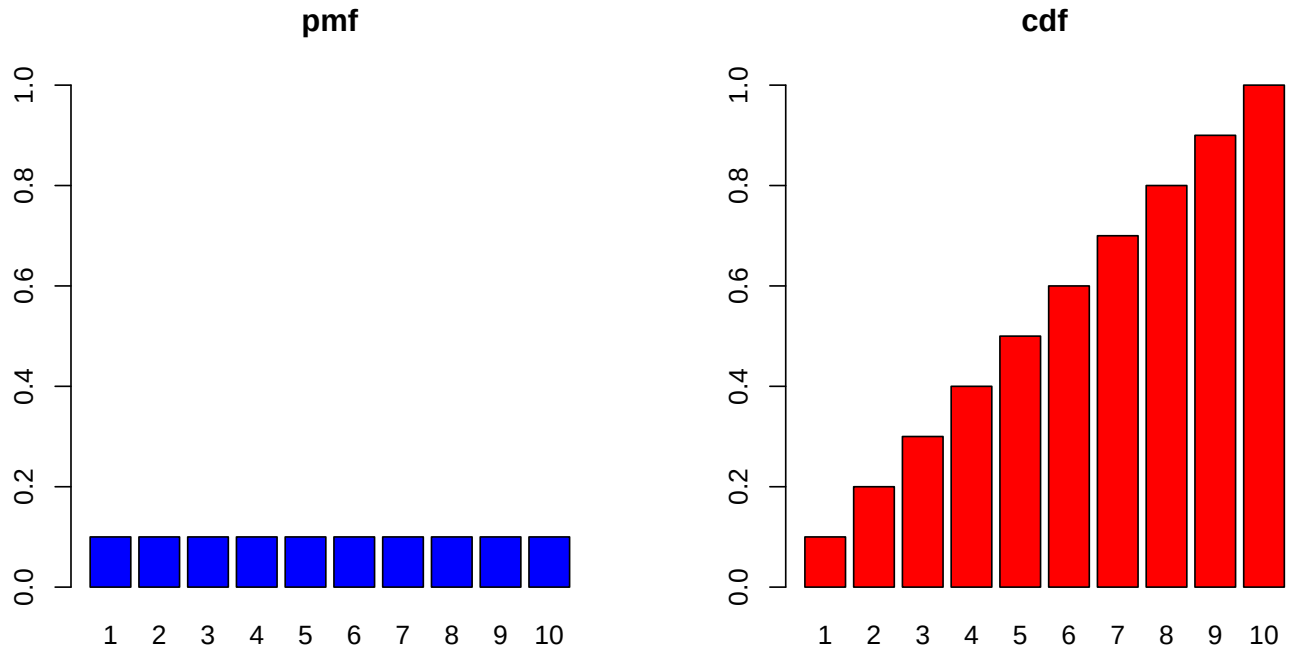
For the 6-sided die

$$F(x) = \sum_{k=1}^{x} f_k$$

where $x \in 1 \ldots 6$.

Discrete RVs

Example: RV can take values 1 through 10, each with probability 0.1:



Discrete RVs: Practice

For a fair 6-sided die:

1. What is f_k if $k = 7$? $k = 1.5$?
2. What is $F(3)$? $F(7)$? $F(1.5)$?

Continuous RVs

Things are just a little different for continuous RVs. Instead we use the probability density function (pdf) of the RV, and denoted by $f(x)$. It still describes how relatively likely are alternative values of an RV, but does not return a probability.

Continuous RVs

An analogy:

Probabilities are like weights. The PMF tells you how much weight each possible value or outcome contributes to a whole. The PDF just tells you how dense it is around a value. To calculate the weight, you need to also know the size of the area that you're interested in.

Continous RVs

Related to the pdf is the cumulative distribution function (cdf), $F(x)$.

$$F(x) \equiv \Pr(X \leq x)$$

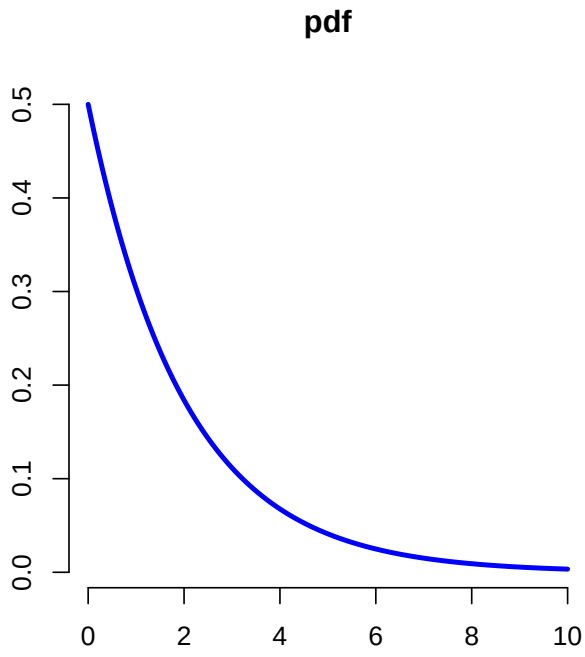
For a continuous distribution

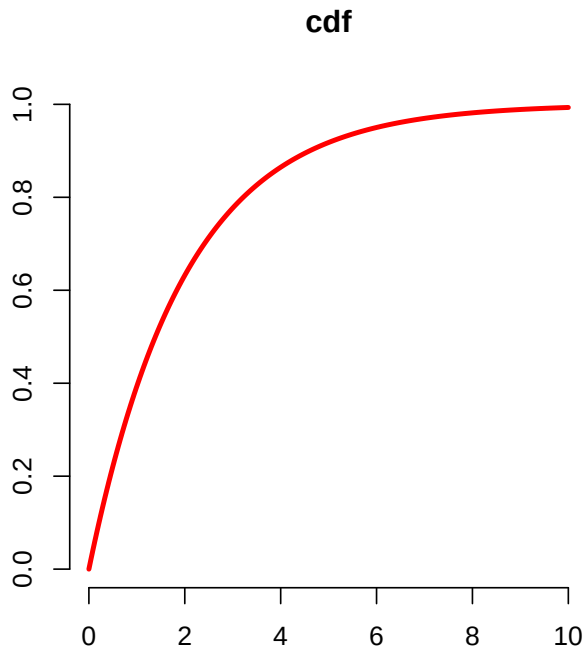
$$F(x) = \int_{-\infty}^{x} f(x')dx'$$

For a normal distribution with mean 0, what is $F(0)$?

Continuous RVs

Example: exponential RV, where $f(x) = re^{-rx}$:





Moments of a RV

The probability distribution of a random variable can usually be characterized by a small number of parameters, called "moments". These are usually defined as:

$$\mu'_n = \mathbb{E}[(X - c)^n] = \int_{-\infty}^{\infty} (x - c)^n f(x) dx$$

Here c is some value, and $\mathbb{E}[\cdot]$ is read as "the expected value of $\cdot$". These moments give you lots of information about the distribution – and you have worked some of them before...

First moment: mean

The expected value (mean), denoted μ or $\mathbb{E}[X]$ of an R.V. is given by the first moment:

$$\mathbb{E}[X] = \int xf(x)dx \qquad \text{(continuous RVs)}$$

$$\mathbb{E}[X] = \sum_x xf_x \qquad \text{(discrete RVs)}$$

This moment is the main descriptor of the central tendency or location of your RV.

The summation notation/integral is a fancy way of writing down an arithmetic mean of the distribution.

Discrete RVs: mean

Recall, for a fair 6-sided die:
$f_k = 1/6$ for $k = \{1, 2, 3, 4, 5, 6\}$

What is $\mathbb{E}[k]$?

$$\mathbb{E}[k] = \sum_k k f_k = \sum_{k=1}^{6} k \frac{1}{6}$$
$$= \frac{1}{6} \sum_{k=1}^{6} k = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6)$$

It's like we rolled 6 die and got one of each number - a normal average where the data are perfectly proportioned according to the relative probabilities of outcomes.

Second moment: variance

The second central moment (i.e., around the mean) is the variance, denoted σ^2 or var(X)

$$\text{var}(X) = \mathbb{E}[(X-\mu)^2] = \int (x-\mu)^2 f(x) dx \quad \text{(continuous RVs)}$$

$$\text{var}(X) = \mathbb{E}[(X-\mu)^2] = \sum_x (x-\mu)^2 f_x \quad \text{(discrete RVs)}$$

This moment is a descriptor of variability that is independent of translation (i.e., if you move the mean).

Second moment: variance

Variances are convenient to work with because if X and Y are uncorrelated then $\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y)$. However, the variance has units of the mean^2, which makes it hard to compare to the mean. There are two common solutions:

$$\sigma = \mathrm{sd}(X) = \sqrt{\mathrm{var}(X)} \qquad \text{standard deviation}$$

$$\mathrm{CV}(X) = \frac{\sigma}{\mu} \qquad \text{coefficient of variation}$$

Probability distributions and their moments are descriptors of what we think a Population is like.

If you know the parameters of the distribution you know everything about the population – moments, how likely are you to see some values than others, etc. You can calculate a lot of quantities yourself, or you can look things up, or use R.

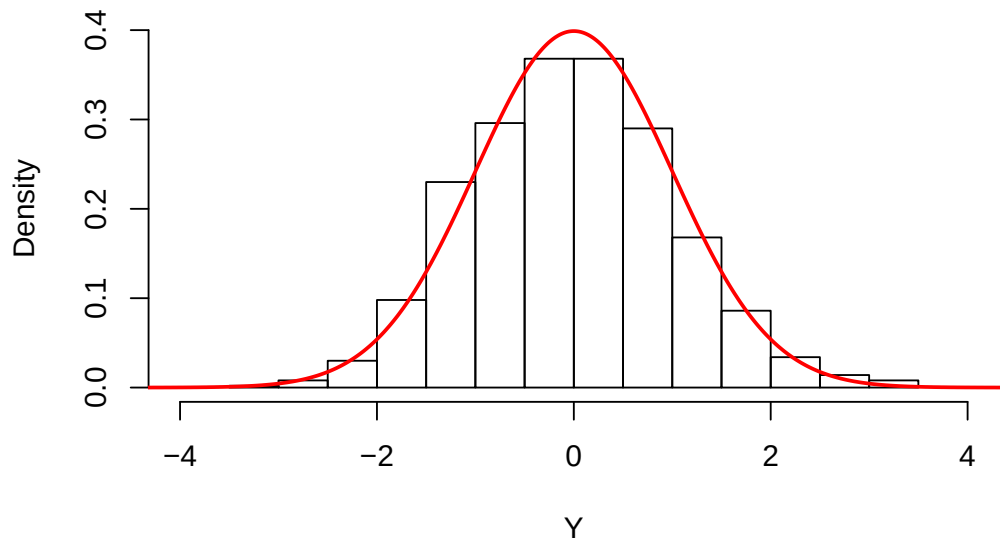
Probability Distributions in R

For standard probability distributions R can take random draws, calculate the cumulative distribution, probability density, and quantiles. For the normal distribution:

```
> rnorm(3, mean=0, sd=1) ## random draws
[1]  0.010666837 -1.518043592 -0.106643773
> qnorm(p=c(0.025, 0.975), mean=0, sd=1) ## quantiles
[1] -1.959964  1.959964
> pnorm(q=0, mean=0, sd=1) ## cdf, Pr(x<0.5)
[1] 0.5
```

Probability Distributions in R

```
> Y <- rnorm(1000, mean=0, sd=1)
> hist(Y, freq=F, ylim=c(0, 0.4),
    xlim=c(-4, 4), main="")
> x<-seq(-10, 10, length=1000)
> lines(x, dnorm(x, mean=0, sd=1), col="red", lwd=2)
```



Normal Distribution

For the purposes of this course, the Normal (or Gaussian) distribution is the most important. Its pdf is:

$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

with $\mathbb{E}[x] = \mu$ and $\text{var}(x) = \sigma^2$. Thus the mean and variance are independent. Consider two independent, normally distributed RVs, $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$.

- $cX \sim \mathcal{N}(c\mu_x, c^2\sigma_x^2)$
- $X + b \sim \mathcal{N}(\mu_x + b, \sigma_x^2)$
- $X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$

Distribution of the sample mean

Consider the mean for an *iid* sample of n observations of a random variable $\{X_1, \ldots, X_n\}$.

Suppose that $\mathbb{E}(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$, then

- $\mathbb{E}(\bar{X}) = \frac{1}{n} \sum \mathbb{E}(X_i) = \mu$
- $\text{var}(\bar{X}) = \text{var}\left(\frac{1}{n} \sum X_i\right) = \frac{1}{n^2} \sum \text{var}\left(X_i\right) = \dfrac{\sigma^2}{n}$.

If X is normal, then $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

Method of Moments

Consider an *iid* sample of n observations of a random variable $\{x_1, \ldots, x_n\}$. You can calculate sample values of the moments of the RV from these, i.e.:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n} \sum (x - \bar{x})^2$$

Then you can estimate the parameters of a probability distribution by "matching" these up with the theoretical values of the moments calculated above. This approach can be biased, so it's good to follow up with a maximum likelihood estimate.

Exercise 1: Gamma distribution

This is the distribution of waiting times until a certain number of events take place. For instance a Gamma(shape=3, scale=2) tells you the distribution of the length of time, in days, you would wait for 3 deaths if the average survival time is 2 days. The pdf is

$$\frac{1}{s^a \Gamma(a)} x^{a-1} e^{-x/s}$$

where a is the shape and s is the scale, $\mathbb{E}[x] = as$ and $\text{var}(x) = (as^2)$.

1. Take 20 draws from a gamma (`rgamma`) with $a = 3$ and $s = 2$, and plot the histogram of samples and overlay the pdf.
2. What are the MoM estimates for a and s?
3. Again take 20 draws from the gamma, and calculate the MoM estimates. Repeat 100 times. (hint: `replicate()` or `lapply` may be useful.) Plot histograms of your MoM for each parameter individually as well as jointly on a scatter plot. Indicate the true parameter values in your plots. How does the MoM perform?

Likelihoods

Recall that $f(Y_i)$ is the pmf (pdf), and it tells us the probability (density) of some yet to be observed datum Y_i given a probability distribution and its parameters. If we make many observations, $\mathbf{Y} = y_1, y_2, \ldots, y_n$, we are interested how probable it was that we obtained these data, jointly. We call this the "likelihood" of the data, and denote it as

$$\mathcal{L}(\theta; Y) = f_\theta(Y)$$

where $f_\theta(Y)$ is the pdf (or pmt) of the data interpreted as a function of θ.

Likelihoods

For instance, for binomial data:

$$\Pr(Y_i = k|\theta = p_k) = \binom{N}{k} p_k^k (1 - p_k)^{N-k}.$$

If we have data $\mathbf{Y} = y_1, y_2, \ldots, y_n$ that are i.i.d. as binomial RVs, the probabilities multiply, and the likelihood is:

$$\mathcal{L}(\theta; Y) = \prod_{i=1}^{n} \binom{N}{y_i} p_i^y (1 - p)^{N-y_i}.$$

Likelihoods vs. probability

"Likelihood is the hypothetical probability [density] that an event that has already occurred would yield a specific outcome. The concept differs from that of a probability in that a probability refers to the occurrence of future events, while a likelihood refers to past events with known outcomes." (1)

Further, the likelihood is a function of θ (the parameters), assuming fixed data.

---

1. Weisstein, Eric W. "Likelihood." From MathWorld–A Wolfram Web Resource.

http://mathworld.wolfram.com/Likelihood.html

Likelihoods

We are usually interested in relative likelihoods – e.g., is it more likely that the data we observed came from a distribution with parameters θ_1 or θ_2? Thus we only worry about the likelihood up to a constant. Further, it is often easier to work with the log-likelihood:

$$L(\theta; Y) = \ell(\theta; Y) = \log(\mathcal{L}(\theta; Y))$$

where $\log(\cdot)$ is the natural log.

Maximum Likelihood Estimators (MLEs)

We can find the parameters that are most likely to have generated our data – the maximum likelihood estimate (mle) of the parameters. To do this we maximize the likelihood (or equivalently minimizing the negative log-likelihood) by taking its derivative and setting it equally to zero:

$$\frac{\partial \mathcal{L}}{\partial \theta_j} = 0 \quad \text{or} \quad -\frac{\partial L}{\partial \theta_j} = 0$$

where j denotes the j^{th} parameter. We usually denote the MLE as $\hat{\theta}_j$.

Likelihoods

The likelihood DOES NOT tell you the probability that parameters have a certain value, given the data. To obtain that quantity, usually called the "posterior probability of the parameters" in Bayesian statistics, you have to use Bayes Theorem.

Exercise 2:

Imagine that you flip a coin N times, and then repeat the experiment n times. Thus, you have data $k = k_1, k_2, \ldots k_n$ that are the number of times you observed a head in each trial. p is the probability of obtaining a head.

1. Write down the likelihood and log-likelihood for the data.
2. Take the derivative of the negative log-likelihood, set this equal to zero and find $\hat{p}$.
3. Simulate some data in R with $p = 0.25$.
4. Calculate the log-likelihood of your simulated data across a range of p (from 0 to 1), and plot them. This is called a "likelihood profile". If you couldn't fine the MLE analytically, how might you use the likelihood profile to estimate it?