COS 424: Interacting with Data

Lecturer: David Blei
Scribe: Rebecca Fiebrink

Lecture #2
9 February 2007

Administrative remarks:

- *Chenwei Zhu has been added as a second teaching assistant for the course. The course is now open for those who were previously unable to register.*
- *Students should begin to familiarize themselves with R (see link from course website). Both the GUI and emacs interfaces are useful. There will be an R tutorial on Monday, 12 February, at 7pm (location TBA).*
- *Students should do "Homework 0" if they haven't already.*

1 Overview: Some Probability and Statistics

This lecture covers the basics of core concepts in probability and statistics to be used in the course. These include random variables, continuous and discrete distributions, joint and conditional distributions, the chain rule, marginalization, Bayes Rule, independence and conditional independence, and expectation. Probability models are discussed along with the concepts of independently and identically distributed (IID) random variables, model parameters, likelihood, and the maximum likelihood estimator (MLE).

2 Random Variables

Probability is the study of random variables. A random variable is a "probabilistic" outcome, such as a coin flip or the height of a person chosen from a population. Here, the observed value of heads or tails, or the observed height, depend on some known or unkown probability distribution. It may also sometimes be useful to model probabilistically quantities that are not "probabilistic" in this sense, for example the temperature on some future date, the temperature on some past date, or the number of times the word "streetlight" appears in a document.

Random values take on values in a *sample space*. This space may be discrete or continuous, and the space may be defined differently for different scenarios. For example, the sample space for a coin flip is $\{H, T\}$; the sample space for height might be defined as the positive real values in $(0, \infty)$; for temperature, it might be defined as real values in $(-\infty, \infty)$; for the number of occurrences of a word in a document, it might be the positive integers $\{1, 2, \ldots\}$.

There is not necessarily one uniquely "correct" sample space for a particular concept. For example, one may argue that using a sample space of $(0, \infty)$ for a person's height allows for impossibly tall people, and one might define an alternative sample space that puts a finite upper limit of the height one would measure. However, it typically doesn't matter if the sample space is "too big," for example infinite height or infinitely negative temperature (as long as the distribution on the sample space places no or very low probability on the impossible events).

2.1 Terminology

The values of a random variable are called *atoms*. Random variables are written using capital letters, and realizations of the random variables are written using lowercase letters. For example, X is a coin flip, and x is the value (H or T) of the coin flip.

3 Discrete Distributions

A discrete distribution assigns a probability to every atom in the sample space of a random variable. For example, if X is an (unfair) coin, then the sample space consists of the atomic events $X = H$ and $X = T$, and the discrete distribution might look like:

$$\begin{aligned} P(X = H) &= 0.7 \\ P(X = T) &= 0.3 \end{aligned}$$

For any valid discrete distribution, the probabilities over the atomic events must sum to one; that is,

$$\sum_x P(X = x) = 1$$

The probability of a disjunction is a sum over part of the probabilities of a set of atoms in the sample space. For example, the probability that the outcome of a single die roll (D) is bigger than 3 is equivalent to the probability that the outcome is 4, or the outcome is 5, or the outcome is 6. The probabilities that the die is 4, 5, or 6 are added together:

$$P(D > 3) = P(D = 4) + P(D = 5) + P(D = 6)$$

An *event* is a subset of atoms (one or more). In the example above, $D > 3$ is an event that consists of three of the six possible atoms for a die roll. The probability of an event is the sum of the probabilities of its constituent atoms.

A distribution may be visualized with the following picture, where an atom is any point inside the box. Those points inside the circle correspond to outcomes for which $X = x$; those outside the circle correspond to outcomes for which $X \neq x$.

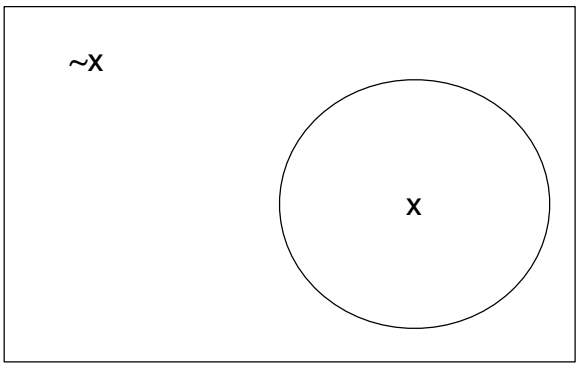

Figure 1: Venn diagram representation of a probability distribution for a single random variable

3.1 Joint Distributions

Typically, one considers *collections* of random variables. For example, the flipping of four coins involves four random variables, one for each coin. The *joint distribution* is a probability distribution over the configuration of all the random variables in the ensemble. For example, the joint distribution for a flip of each of four coins assigns a probability to every outcome in the space of all possible outcomes of the four flips. If all coins are fair, this would look like:

$$\begin{aligned} P(HHHH) &= 0.0625 \\ P(HHHT) &= 0.0625 \\ P(HHTH) &= 0.0625 \\ &\ldots \end{aligned}$$

Note that one can consider the outcome of the flip of four coins as a single random variable with 16 possible outcomes ($HHHH$, $HHHT$, $HHTH$, ...).

A joint distribution can be visualized using the same approach as a distribution of a single random variable. The figure below represents the joint distribution of X and Y.

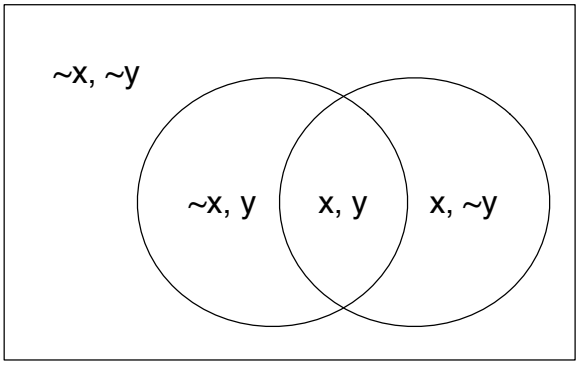

Figure 2: Venn diagram representation of a probability distribution for a single random variable

3.2 Conditional Distributions

A *conditional distribution* is the distribution of some random variable given some evidence, such as the value of another random variable. For example, $P(X = x|Y = y)$ is the probability that $X = x$ when $Y = y$.

A conditional distribution gives more information about X than the distribution of $P(X)$ alone. For example, if Dave really likes Steely Dan, there might be an appreciable probability that Dave is listening to Steely Dan at any given time:

$$P(\text{Dave listens to Steely Dan}) = 0.5$$

However, if Dave's wife does not like Steely Dan, the probability that he listens to the music might be influenced by whether or not she is home. For example:

$$\begin{aligned} P(\text{Dave listens to Steely Dan}|\text{Toni is home}) &= 0.1 \\ P(\text{Dave listens to Steely Dan}|\text{Toni is not home}) &= 0.7 \end{aligned}$$

The above expressions give a more precise definition of the probability of Dave listening to Steely Dan.

The conditional distribution $P(X = x|Y = y)$ is a *different distribution* for each value of y. So, we have

$$\sum_x P(X = x|Y = y) = 1$$

just as we saw before with

$$\sum_x P(X = x) = 1$$

This is like saying "P(Dave listens to Steely Dan | Toni is home)+P(Dave doesn't listen to Steely Dan | Toni is home) = 1", which makes sense.

However, remember that

$$\sum_y P(X = x|Y = y) \neq 1(necessarily)$$

If the above did necessarily sum to 1, it would be like saying that "P(Dave listens to Steely Dan | Toni is home) +P(Dave listens to Steely Dan | Toni is not home) = 1", which doesn't necessarily hold.

4 Conditional Probability and Related Concepts

Conditional probability can be defined in terms of the joint and single probability distributions:

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

(which holds when $P(Y) > 0$). This can be thought of as scaling the probability that $X = x$ and $Y = y$ by $P(Y = y)$, or by comparing the intersection of $X = x$ and $Y = y$ to the whole area of $Y = y$ in the Venn diagram representation. That is, the conditional probability $P(X = x|Y = y)$ is the relative probability of $X = x$ in the space where $Y = y$.

4.1 The Chain Rule

The definition of conditional probability leads to the *chain rule*, which lets us define the joint distribution of two (or more) random variables as a product of conditionals:

$$\begin{aligned} P(X, Y) &= P(X, Y)\frac{P(Y)}{P(Y)} \\ &= P(X|Y)P(Y) \end{aligned}$$

The chain rule can be used to derive the $P(X, Y)$ when it is not known. For example, let Y be a disease (e.g., a cold) and X be a symptom (e.g., a sneeze). We may know $P(X|Y)$ and $P(Y)$ from data. The chain rule can be used to obtain the probability of having the disease and the symptom (e.g., both sneezing and having a cold).

The chain rule can be used for any set of N variables:

$$P(X_1, \ldots, X_N) = \prod_{n=1}^{N} P(X_n|X_1, \ldots, X_{n-1})$$

This holds for *any* ordering of the random variables.

4.2 Marginalization

Given a collection of random variables, we are often interested in only a subset of them. For example, we might want to compute $P(X)$ from a joint distribution $P(X, Y, Z)$. *Marginalization* allows us to compute $P(X)$ by summing over all possible realizations of the other variables:

$$P(X) = \sum_y \sum_z P(X, y, z)$$

This beautiful property actually derives from the chain rule:

$$\begin{aligned}
\sum_y \sum_z P(X, y, z) &= \sum_y \sum_z P(X)P(y, z|X) \quad \text{(by the chain rule)} \\
&= P(X) \sum_y \sum_z P(y, z|X) \quad \text{(because } P(X) \text{ doesn't depend on } y \text{ or } z\text{)} \\
&= P(X) \quad \text{(because } \sum_y \sum_z P(y, z|X) = 1\text{)}
\end{aligned}$$

4.3 Bayes Rule

By the chain rule,

$$\begin{aligned}
P(X, Y) &= P(X|Y)P(Y) \\
&= P(Y|X)P(X)
\end{aligned}$$

This is equivilently expressed as *Bayes rule*:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

(provided $P(Y)$ is nonzero; the expression $P(X|Y)$ is irrelevant if $P(Y) = 0$).

Again, suppose Y is a disease and X is a symptom. From $P(X|Y)$ and $P(Y)$, we can compute the (useful) quantity $P(Y|X)$. For example, we can compute the probability that a sneezing person has a cold, given the (easier to estimate) probability that one sneezes if one has a cold, and the probability of one having a cold in the first place.

Bayes rule is important in *Bayesian statistics*, where Y is a parameter that controls the distribution of X.

5 Independence

Random variables are *independent* if knowing about X tells us nothing about Y. That is,

$$P(Y|X) = P(Y)$$

This means that their joint distribution factorizes:

$$X \perp\!\!\!\perp Y \iff P(X,Y) = P(X)P(Y).$$

This factorization is possible because of the chain rule:

$$\begin{aligned} P(X,Y) &= P(X)P(Y|X) \\ &= P(X)P(Y) \end{aligned}$$

Examples of random variables that are independent include flipping a coin once and flipping the same coin a second time, or using an electric toothbrush and having blue as a favorite color. Using a *blue* electric toothbrush and having blue as a favorite color would *not* be independent. Other examples of random variables that are not independent are being registered as a Republican and voting for Bush in the last election, or the color of the sky and the time of day.

Notice that just because variables are not independent, it does not mean that there is a causal relationship between them.

Other examples of independence relationships include

- The rolls of two twenty-sided dice: *independent*
- Rolling three dice and computing $(D_1 + D_2, D_2 + D_3)$: *not independent*
- The number of enrolled students and the temperature outside today: *independent? Maybe, unless you consider the fact that temperature changes with time, and so do the number of enrolled students.*
- The number of attending students and the temperature outside today: *clearly not independent!*

5.1 Conditional Independence

Suppose we have two coins, one biased and one fair, with

$$P(C_1 = H) = 0.5 \quad P(C_2 = H) = 0.7.$$

We choose one of the coins at random: choose $Z \in \{1, 2\}$ (suppose we may keep this choice secret). We flip this coin C_Z twice and record the outcome (X, Y) (i.e., X is the outcome of the first flip, and Y is the outcome of the second flip). Are X and Y independent? Well, no. For example, if the first flip of the coin, X, is heads, it is more likely that we are flipping Coin 2, which means that it is more likely than not that the second flip, Y, will also be heads. But what if we know Z, the choice of the coin that was flipped? If we know that Z is 2, for example, we know precisely the probability distribution for the second flip. Knowing the result of the first flip gives us no additional information.

In such a scenario, we can say that X and Y are *conditionally independent* given Z. Here,

$$P(Y|X, Z = z) = P(Y|Z = z)$$

for all possible values of z. Again, this implies a factorization:

$$X \perp\!\!\!\perp Y|Z \iff P(X,Y|Z = z) = P(X|Z = z)P(Y|Z = z),$$

for all possible values of z. Note the difference between conditional and marginal independence: conditional independence implies a factorization regardless of the value of Z; marginal independence implies a factorization only when we do not know what Z is.

6 Continuous Random Variables

So far, we've only considered random variables that take on discrete values, such as dice rolls or coin flips. However, random variables may also be continuous. For discrete variables, recall that we use a probability distribution whose values sum to 1; for continuous variables, we use a *probability density*, $p(x)$, which *integrates* to 1. For example, if the sample space for a random variable X is the set of all real numbers (i.e., $x \in \mathbb{R}$), then

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

The probability density $p(x)$ is a function defined over the sample space of X. Note that $p(x)$ is *not* interpretable as the the probability that $X = x$; it is the *density* of x. Instead, probabilities themselves are integrals over intervals in the sample space. For example, the probability that X takes a value in $(-2.4, 6.5)$ is

$$P(X \in (-2.4, 6.5)) = \int_{-2.4}^{6.5} p(x)dx$$

6.1 The Gaussian Distribution

The *Gaussian* (or *Normal*) *distribution* is a commonly used continuous distribution. Its distribution is

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

where μ is the mean and σ^2 is the variance. Technically speaking, the Gaussian distribution specifies that the probability density associated with a point x is proportional to the negative exponentiated half-distance to μ scaled by σ^2. A more compelling explanation is supported by the graphical representation of the distribution we know and love, shown below.

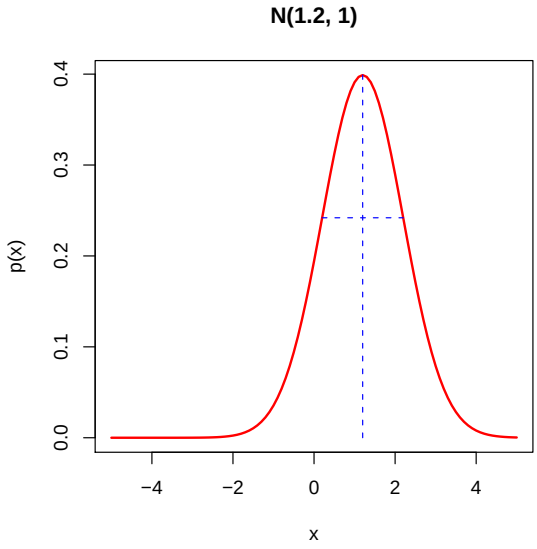

Figure 3: An example Gaussian distribution

This distribution is parameterized with $\mu = 1.2, \sigma^2 = 1$. μ controls the location of the "bump", and σ^2 controls the spread of the bump.

6.2 Notation

For discrete random variables, p denotes the *probability mass function*, which is the same as the distribution on atoms. We can therefore use P and p interchangeably for atoms, treating the probability density function as a set of density "spikes" for values in the sample space and 0 values everywhere else; the issue of integrating this is swept under the rug.

For continuous random variables, however, p is the *density*, which is not interchangeable with probability. This is an unpleasant detail, and you should ask questions if you are confused.

7 Expectation

If $f(X)$ is a function of a random variable X, then $f(X)$ is itself also a random variable. In the case that X is discrete, the *expectation* of $f(X)$ is a weighted average of f, where the weighting is determined by $p(x)$:

$$\mathrm{E}[f(X)] = \sum_x p(x)f(x)$$

(Note that $f(X)$ might not be discrete, even if X is.)

In the continuous case, the expectation is an integral

$$\mathrm{E}[f(X)] = \int p(x)f(x)dx$$

In both of these cases, the expectation $E[f(X)]$ is a scalar value (possible infinite).

The conditional expectation is defined similarly:

$$\mathrm{E}[f(X)|Y = y] = \sum_x p(x|y)f(x)$$

This might raise the questions, what is $\mathrm{E}[f(X)|Y = y]$? What is $\mathrm{E}[f(X)|Y]$? $\mathrm{E}[f(X)|Y = y]$ is a scalar, as in the previously discussed, non-conditional expectation. However, $\mathrm{E}[f(X)|Y]$ is a function of a random variable (and therefore a random variable itself)! This is because it depends on the distribution of Y, which we may or may not know.

7.1 Iterated Expectation

What happens when you take the expectation of the conditional expectation? You arrive at a nice property of *iterated expectation*, illustrated here:

$$\begin{aligned}
\mathrm{E}[\mathrm{E}[f(X)]|Y]] &= \sum_y p(y)\mathrm{E}[f(X)|Y=y] \quad \text{by the definition of expectation} \\
&= \sum_y p(y)\sum_x p(x|y)f(x) \quad \text{again, by the definition of expectation} \\
&= \sum_y \sum_x p(x,y)f(x) \quad \text{by using the chain rule in reverse} \\
&= \sum_y \sum_x p(x)p(y|x)f(x) \quad \text{by using the chain rule forward} \\
&= \sum_x p(x)f(x)\sum_y p(y|x) \quad \text{by pushing out the terms that don't depend on } y \\
&= \sum_x p(x)f(x) \quad \text{by noting that } \sum_y p(y|x) = 1 \\
&= \mathrm{E}[f(X)] \quad \text{by the definition of expectation!}
\end{aligned}$$

The following example illustrates that iterated expectation can be a very useful tool. Suppose we flip a coin with probability π of heads until we see a heads. What is the expected waiting time for a heads?

A naive solution might go like this: Call N the number of flips we have to perform until we see a heads. N is a random variable, and we are interested in its expected value. The probability that $N = 1$ is π. The probability that $N = 2$ is $(1-\pi)\cdot\pi$, or the probability that we *didn't* see a heads on the first flip, but we did on the second flip. The probability that $N = 3$ is $(1-\pi)^2\cdot\pi$, and so on. Therefore, we can use the definition of expectation to write:

$$\begin{aligned}
\mathrm{E}[N] &= 1\pi + 2(1-\pi)\pi + 3(1-\pi)^2\pi + \ldots \\
&= \sum_{n=1}^{\infty} n(1-\pi)^{(n-1)}\pi
\end{aligned}$$

This gives us an infinite sum that is not immediately obvious how to compute.

Let's try another approach, this time using iterated expectation. Call X_1 the first coin flip. We can define the expected value of N as

$$\begin{aligned}
\mathrm{E}[N] &= \mathrm{E}[\mathrm{E}[N|X_1]] \quad \text{by the definition of iterated expectation} \\
&= \pi\cdot\mathrm{E}[N|X_1 = H] + (1-\pi)\mathrm{E}[N|X_1 = T] \quad \text{by the definition of expectation}
\end{aligned}$$

Here, $\mathrm{E}[N|X_1 = H]$ is just 1; we've observed a heads on the first try, so our count is 1 and we are done. To compute $\mathrm{E}[N|X_1 = T]$, we can just add 1 for the coin we've already flipped and "start over," pretending the next flip is the first and computing the expectation for N. This allows us to write

$$\begin{aligned}
\mathrm{E}[N] &= \pi\cdot 1 + (1-\pi)(\mathrm{E}[N]+1)] \\
&= \pi + 1 - \pi + (1-\pi)\mathrm{E}[N] \\
&= 1/\pi
\end{aligned}$$

This is quite a nice method of solution compared to the infinite sum approach.

8 Probability Models

Probability distributions are used as *models* of data we observe. We can pretend that our data is drawn from some unknown distribution, then infer the properties of that distribution from the data. Examples of properties we may wish to infer include the bias of a coin, the average height of a student, the chance that someone will vote for Hillary Clinton, the chance that someone in Vermont will vote for Hillary Clinton, the proportion of gold in a mountain, the number of bacteria in our body, and the evolutionary rate at which genes mutate. In this class, we will often deal with models as tools for learning about and working with data.

8.1 Independent and Identically Distributed Random Variables

Independent and identically distributed (or IID) random variables are mutually independent of each other, and are identically distributed in the sense that they are drawn from the same probability distribution. For example, if we flip the same coin N times and record the outcome, then $X_1, \ldots, X_N$ are IID. The IID assumption can be useful in data analysis, even when the data is not strictly IID.

8.2 Parameters

Parameters are values that *index* a particular family of distributions. For example, a coin flip is distributed according to a Bernoulli distribution, whose parameter π is the probability of occurrence of some event (here, a heads). For a coin flip, the Bernoulli distribution can be expressed as

$$p(x|\pi) = \pi^{1[x=H]}(1-\pi)^{1[x=T]},$$

where $1[\cdot]$ is called an *indicator function*. It is 1 when its argument is true and 0 otherwise. The value π indexes the Bernoulli distribution; changing π leads to different Bernoulli distributions. We now have a convenient and general representation for the distribution.

A Gaussian distribution has two parameters, the mean and the variance. Recall that a Gaussian distribution is expressed as

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

8.3 The Likelihood Function

Again, suppose we flip a coin N times and record the outcomes. Further suppose that we *think* the probability of heads is some value π (this is distinct from whatever the true probability of heads "really" is). Given our π, whatever it is, the probability of an observed sequence of coin flips is

$$p(x_1, \ldots, x_N|\pi) = \prod_{n=1}^{N} \pi^{1[x_n=H]}(1-\pi)^{1[x_n=T]}$$

In fact, we can think of this expression $p(x_1, \ldots, x_N|\pi)$, the probability of a set of observations, as a function of π. This is called the *likelihood function*. Taking the log of this

expression allows for more elegant reasoning later on, so we introduce the log of this, the *log likelihood* function (also called "loglihood," if you're being funny):

$$\begin{aligned}\mathcal{L}(\pi) &= \log(p(x_1, \ldots, x_N|\pi)) \\ &= \log\left(\prod_{n=1}^{N} \pi^{1[x_n=H]}(1-\pi)^{1[x_n=T]}\right) \\ &= \sum_{n=1}^{N} 1[x_n = H]\log\pi + 1[x_n = T]\log(1-\pi)\end{aligned}$$

Consider an observed sequence of coin flips: $HHTHTHHTHHTHHTH$. The diagram below plots the log likelihood for values of π in (0,1) for this observed sequence. The value of π that maximizes the log likelihood is marked with a black dot: it is 2/3.

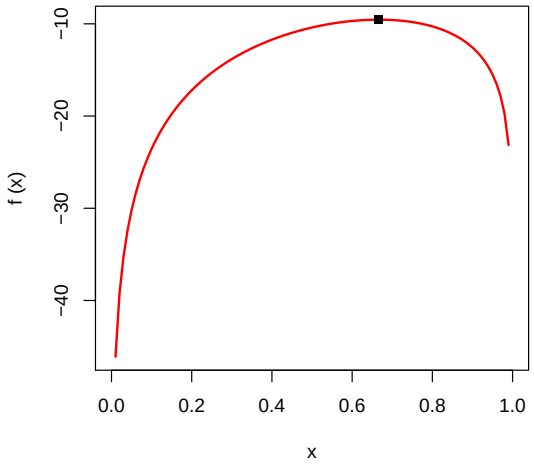

Figure 4: Maximizing the Bernoulli log likelihood

8.4 The Maximum Likelihood Estimate

The *maximum likelihood estimate*, or MLE, is the value of the parameter that maximizes the log likelihood (notice that this is equivalent to maximizing the likelihood itself, but the math in the log maximization might be friendlier). In the Bernoulli example, the MLE $\hat{\pi}$ is the observed proportion of heads:

$$\hat{\pi} = \frac{1}{N}\sum_{n=1}^{N} 1[x_n = H]$$

Intuitively, and in some sense mathematically, this is the value that "best explains" our observations. In fact, this "maximum likelihood principle" is an important idea in statistics.

The MLE is "good" for several reasons, one of which is consistency. For example, suppose you flip a coin N times, and its true bias is π^*. Perform your own estimate of the bias from the data $x_1, \ldots x_N$ with the MLE $\hat{\pi}$. Then,

$$\lim_{N\to\infty} \hat{\pi} = \pi^*$$

This property holds for IID random variables in general. This is a good thing. It lets us sleep at night.

The plot below shows in blue the MLE $\hat{\pi}$ computed after each of 5000 coin flips. The true bias π^* is shown in red. It is apparent that the MLE does converge toward the true bias as the number of trials increases. Frequentist statistics is generally concerned with long-term properties of estimators, such as this convergence. For this reason, consistency has historically been the most important property when choosing an estimator. However, Bayesian statistics addresses the fact that, in practice, there will always be a finite amount of data. In light of this view, other estimators are also used.

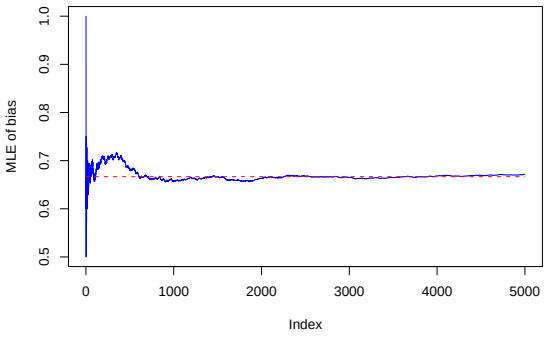

Figure 5: Consistency of MLE

8.5 Gaussian MLEs

Suppose we make observations $x_1, \ldots, x_N$ on a continuous random variable, and we choose to model this data with a Gaussian distribution:

$$p(x_1, \ldots, x_N | \mu, \sigma^2) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{-(x_n - \mu)^2}{2\sigma^2} \right\}$$

The log likelihood is therefore

$$\mathcal{L}(\mu, \sigma) = -\frac{1}{2} N \log(2\pi\sigma^2) - \sum_{n=1}^{N} \frac{(x_n - \mu)^2}{2\sigma^2}$$

The MLE of the mean, μ, is the *sample mean*

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

and the MLE of the variance, σ^2, is the *sample variance*

$$\hat{\sigma^2} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \hat{\mu})^2$$

As an example, consider $x_1, \ldots, x_N$ to be the approval ratings of US presidents from 1945 to 1975, recorded four times per year. The data points are plotted below as blue 'X's, and the Gaussian parameterized with the maximum likelihood estimators computed from the data is shown in red.

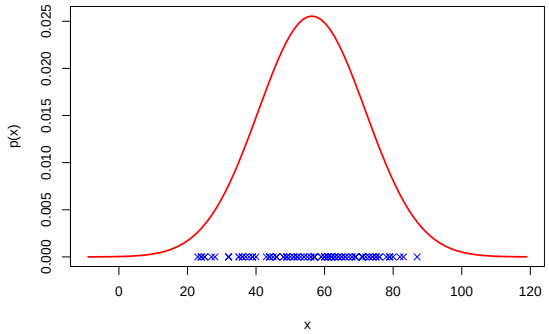

Figure 6: Gaussian model of approval ratings

There are several valid criticisms of this model. For one thing, the Gaussian assigns non-zero probabilities to approval ratings below 0 and above 100. It ignores the sequential nature of the data, and it assumes that the approval ratings are IID, which they certainly aren't! It also ignores historically available information, such as the party of the president, whether a war was going on, etc.

The famous statistician George Box once said, "All models are wrong. Some are useful." In dealing with data, one has to decide whether a particular model is suitable for a given task, or whether improving a model to make it "less wrong" is worthwhile.

9 Future Probability Concepts

Future probability concepts to be dealt with in class include Naive Bayes classification; linear regression and logistic regression; hidden variables, mixture models, and the EM algorithm; graphical models; factor analysis; sequential models; and perhaps generalized linear models and Bayesian models if time allows.