CO902
Probabilistic and statistical inference

Lecture 2

Tom Nichols
Department of Statistics &
Warwick Manufacturing Group

t.e.nichols@warwick.ac.uk

Last Time

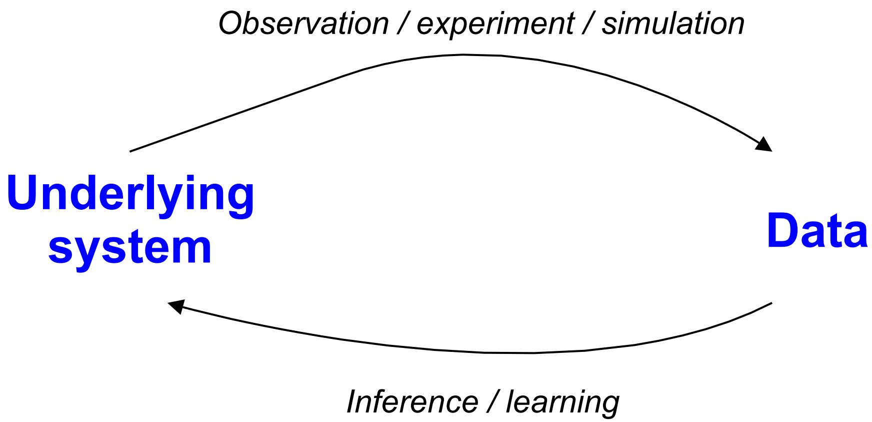
- Law of total probability, aka “sum rule”
- Random Variable
- Probability Mass Function (PMF)
- Expectation, Variance
- Joint distribution of 2 or more random variables
- Conditional probability
- Product rule
- Bayes theorem
- Independence
- Parameterized distributions

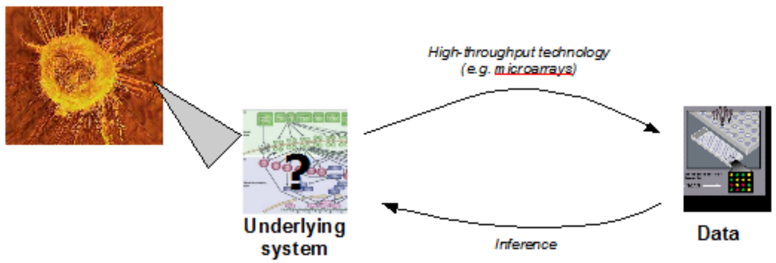
Outline

- **Estimation**
  - Parameterized families
  - Data, estimators
  - Likelihood function, Maximum likelihood
- **(In)dependence**
  - The role of *structure* in probabilistic models
  - Dependent RVs, Markov assumptions
  - Markov chains as structural models
- **Properties of estimators**
  - Bias
  - Consistency
  - Law of large numbers

Inference: from data to prediction and understanding

- Today we'll talk about problem of **inferring parameters from data**
- First, what's a parameter?





Parameterized distributions

- We saw in L1 that a function *P* called the pmf gives the probability of every possible value of an RV
- And we introduced the idea of **parameterized families of pmfs**

$$P(X = x \mid \theta) \quad = \quad f(x; \theta)$$

- This is a distribution for *x*, which depends on a (fixed) theta.
- That is, *P* is a function which
  - maps all possible values of *x* to [0,1]
  - And sums to one

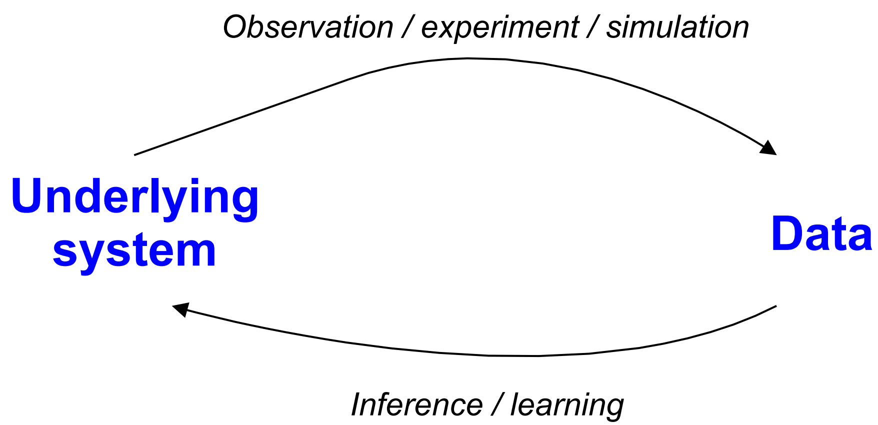
Parameterized distributions

$$P(X = x \mid \theta) \quad = \quad f(x; \theta)$$

- **Parametrized pmfs**
  - Simple parameterized distributions, when combined in various ways can lead to interesting, powerful models

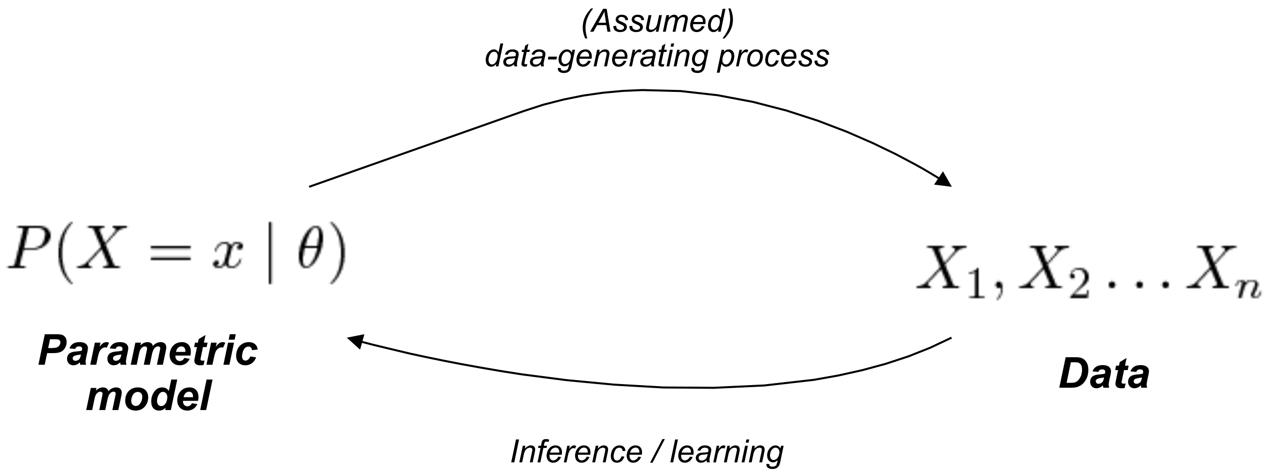
- So we start by looking at the problem of learning parameters from data generated by a parametric pmf

Inference...


Observation / experiment / simulation
Underlying system
Data
Inference / learning

...with a model

- In the simplest case, we *assume* a parameterized distribution is a reasonable description of the data-generating process
- We use the data to say something about the unobserved parameters



- Often, we combine simple parametric models together in various ways, to build up powerful models for tough, real-world problems
- E.g. BNs or MCs are built up from simple elements
- But the basic theory and concepts of estimation we'll learn today are *very* relevant, no matter how complicated the model

Bernoulli distribution

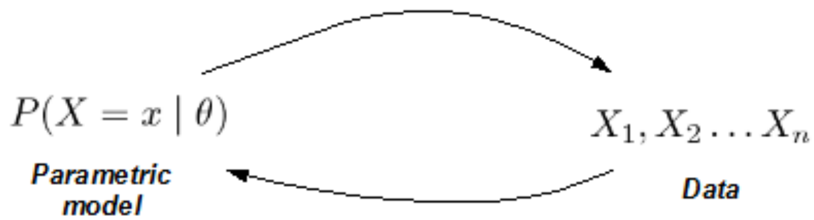
- X has two possible outcomes, one is "success" (X=1) other "failure" (X=0). PMF (one free parameter):

$$P(X = x \mid \theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \end{cases}$$

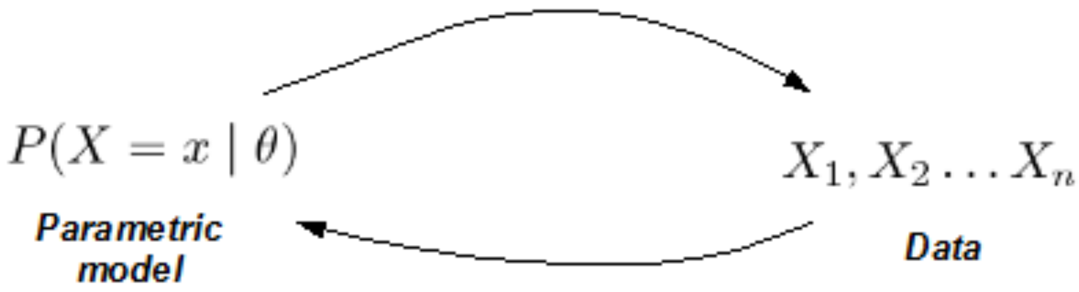
$$X \in \{0, 1\}$$

$$\theta \in [0, 1]$$

- Q: what does data ***generated* from a Bernoulli** look like?



PMF as a data-generating model



- Using a computer, how would you generate or simulate data from the Bernoulli?
- Notice we're assuming the RVs Xi are independent, and all have the exact same Bernoulli PMF
- In a certain sense, there are **two aspects** to the overall model: the pmf(s) involved, and some assumptions about *how* RVs are related

i.i.d. data

- **Data:** the results of *N* completed tosses

H, H, T, H, T, H

$$X_1, X_2 \dots X_n$$

- **Model:** "i.i.d" Bernoulli

$$\begin{aligned} X_i &\stackrel{iid}{\sim} Bernoulli(\theta) \\ P(X_1, X_2 \dots X_n \mid \theta) &= \prod_{i=1}^{n} P(X_i \mid \theta) \\ &= \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)} \end{aligned}$$

- That is, we assume: *(i) Each toss has the same probability of success, (ii) the tosses are independent*
- This means the probability of the next toss coming up heads is simply θ
- Prediction is related to estimation, here very closely...

Estimators

- An **estimator** is a function of random data (“a statistic”) which provides an estimate of a parameter:

$$\hat{\theta} = \hat{\theta}(X_1, X_2 \ldots X_n)$$

- Note terminology/notation: **parameter**, **estimate** and **estimator**

- Several ways of estimating parameters, we will look at:
  - **Maximum likelihood estimator or MLE**
  - **Bayesian inference**
  - **Maximum A Posteriori (MAP) estimator**

Likelihood function

- When we think of "fitting" a model to data (curve-fitting, say), we're thinking of adjusting free parameters to make the model and data match as closely as possible
- Let's take this approach to our **probabilistic models**
- Joint probability of all of the data given the parameter(s):

$$P(X_1, X_2 \dots X_n \mid \theta)$$

- Now, write this as a *function of the unknown parameter(s)*:

$$\mathcal{L}(\theta) \;=\; P(X_1, X_2 \dots X_n \mid \theta)$$

- This is the **likelihood function**

Likelihood function

- **Likelihood function:**

$$\mathcal{L}(\theta) \;=\; P(X_1, X_2 \ldots X_n \mid \theta)$$

- **NOT** a probability distribution over possible values of parameter
- Rather, simply a function which for any value of parameter gives a measure of how well the model specified by that value fits the data
- The key link between a probability model and data

- For N Bernoulli trials…

Probability Mass Function
Domain: $\{0,1\}^N$ Range: R^+
For a particular θ, probability of the data

$$P(X_1, X_2, ..., X_N; \theta) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$$

Likelihood function
Domain: [0,1] Range: R^+
For this particular data, how "likely" are different θ's

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$$

Maximum likelihood estimator (MLE)

- Loosely speaking, the likelihood function tells us how well models specified by various values of the parameter fit the data
- A natural idea then is to construct an estimator in the following way:

$$\begin{aligned}\hat{\theta} &= \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\theta) \\ &= \underset{\theta}{\operatorname{argmax}}\, P(X_1, X_2 \ldots X_n \mid \theta)\end{aligned}$$

- This would then be a sort of "best fit" estimate
- This estimator is called the **Maximum likelihood estimator** or **MLE**

Example: coin tosses

- Let's go back to the coin tossing example
- This will be very simple, but will illustrate the steps involved in getting a MLE, which are essentially the same in more complicated situations

H, H, T, H, T, H ... **?**

Example: coin tosses

- **Data:** the results of *N* completed tosses

$$X_1, X_2 \dots X_n$$

- **Model:** i.i.d Bernoulli

$$\begin{aligned} X_i &\overset{iid}{\sim} Bernoulli(\theta) \\ P(X_1, X_2 \dots X_n \mid \theta) &= \prod_{i=1}^{n} P(X_i \mid \theta) \\ &= \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)} \end{aligned}$$

- **Q: Write down the likelihood function for this model. Write down the *log-likelihood*. Using differential calculus, maximise the likelihood function to obtain the MLE.**

Example: coin tosses

- Likelihood function for our i.i.d. Bernoulli model:

$$\begin{aligned} P(X_1, X_2 \dots X_n \mid \theta) &= \prod_{i=1}^{n} P(X_i \mid \theta) \\ &= \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)} \end{aligned}$$

- Often easier to deal with the log-likelihood
- Log-likelihood:

$$\begin{aligned} \log(P(X_1, X_2 \dots X_n \mid \theta)) &= \sum_{i=1}^{n} x_i \log(\theta) + (1-x_i)\log(1-\theta) \\ &= \mathcal{L}(\theta) \end{aligned}$$

($\mathcal{L}(\theta)$ will denote likelihood or log-likelihood, will be obvious from context, though some authors use $\mathcal{L}(\theta)$ only for likelihood, $l(\theta)$ for log-likelihood)

MLE

- Log-likelihood function for i.i.d. Bernoulli model:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} x_i \log(\theta) + (1 - x_i) \log(1 - \theta)$$

- Set derivative wrt θ to zero and simplifying:

$$\hat{\theta}_{MLE} = \frac{n_1}{n}$$

$$n_1 = \sum_{i=1}^{n} x_i$$

Note the "hat"

θ True, unknown parameter
(Fixed. Influences data)

$\hat{\theta}$ Estimated parameter
(Random. A function of the data)

- That is, the estimate is simply the proportion of successes, which accords with intuition

Dependent RVs

- Introduce a new, graphical notation
  - Vertices represent RVs
  - Edges represent dependencies
- i.i.d. structure...

H, H, T, H, T, H

$$X_1, X_2 \dots X_n$$

Dependent RVs

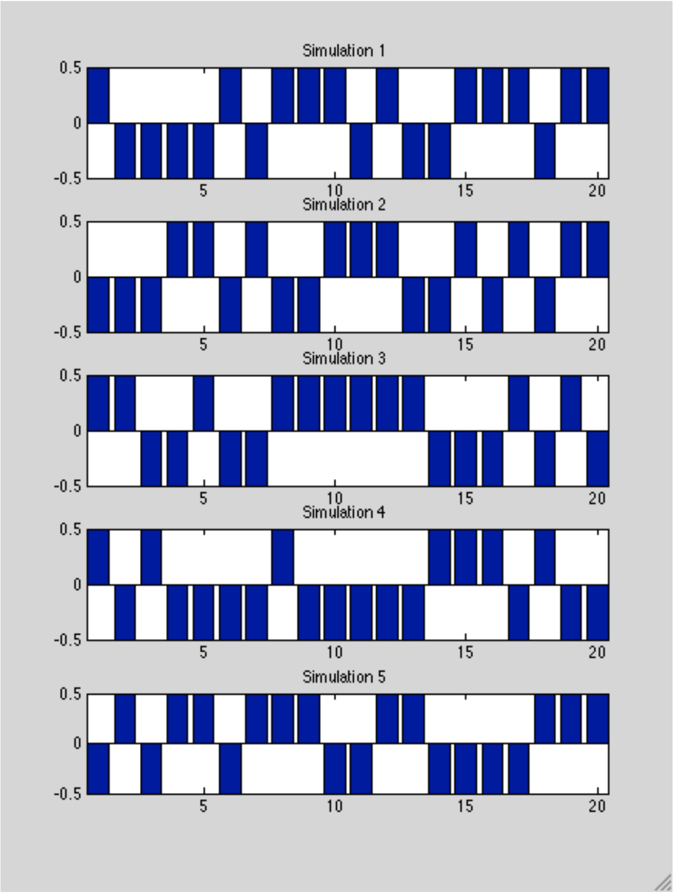
- Let's stick to binary RVs for now
- Binary RVs don't have to be i.i.d. - even though so far we've assumed this.
- Independence has pros and cons...
- Cons: Independence *not* a good model for, say:
  - Sequence of results (win/lose) of football matches
  - Status of proteins in a pathway
  - Time series
- Pros: simplicity! Allowed us to write down the joint distribution and likelihood function as a very simple product - the full joint is a big thing, with many parameters
- Compromise: permit a restricted departure from complete independence...

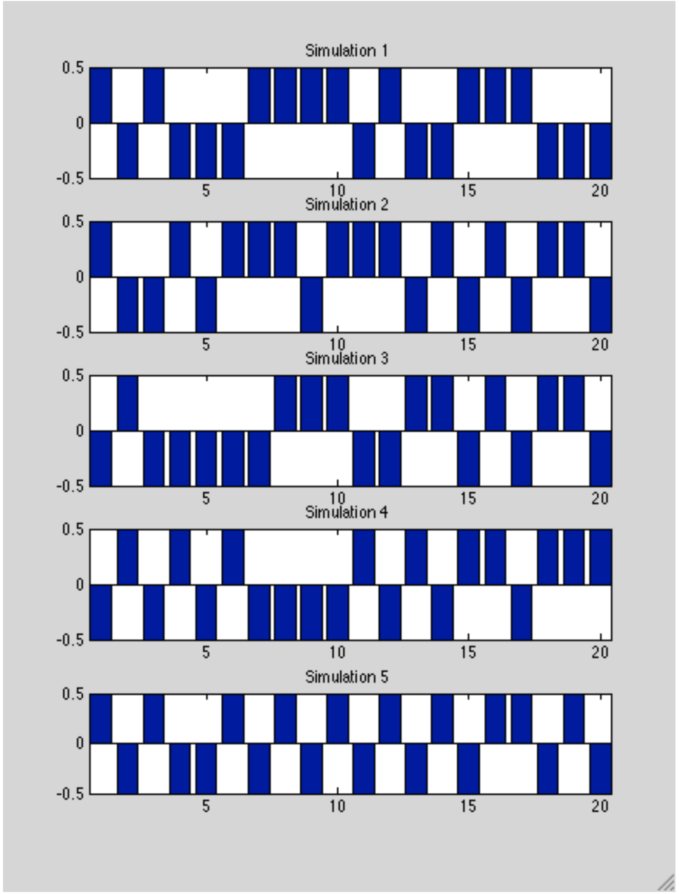
Football results

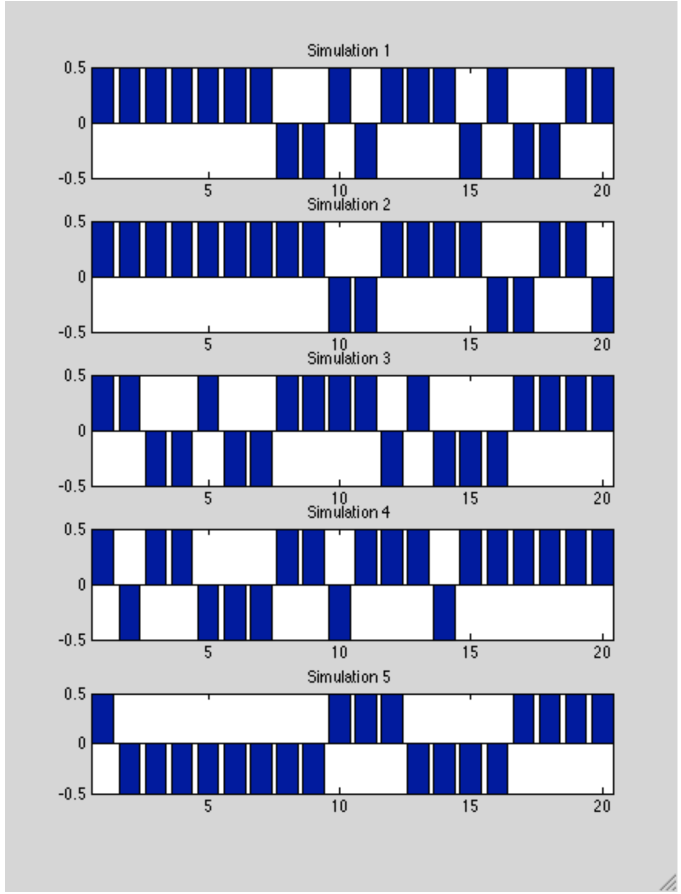
- Sequence of results
- Let each result depend on the one before, but not directly on the previous ones
- We can draw this using the graphical notation...

- **Q: Suppose we wanted to generate data from this model – what would we need to do, what do we need to specify? How many free parameters do we end up with?**

Samples from Football Markov Chain







Three parameter settings (not in order; 0.5 for initial state)…

$$P(X_i|X_{i-1}=0)=0.4 \quad P(X_i|X_{i-1}=0)=0.6 \quad P(X_i|X_{i-1}=0)=0.5$$
$$P(X_i|X_{i-1}=1)=0.6 \quad P(X_i|X_{i-1}=1)=0.4 \quad P(X_i|X_{i-1}=1)=0.5$$

Q: Which is which!?

Markov chains

- We've built a (discrete-index, time-invariant) **Markov chain** and you've generated data or sampled from it using ***ancestral sampling***
- More formally, the elements are:
  - An *initial distribution* P_0
  - *A transition matrix* $\boldsymbol{T}$
- MCs are interesting mathematical objects, with many fun properties, you'll encounter them in that form during stochastic processes
- But they can also be viewed as special case of something called a **probabilistic graphical model**, which is a model with a graph which allows some dependence structure, but is still **parsimonious**
- Applications abound: DNA sequences, speech, language, protein pathways etc. etc.
- We'll encounter probabilistic graphical models later

Conditional distribution

- The RVs in our MC are all binary, and the transition matrix ***T*** is fixed

- The (1st order) Markov assumption underlying our chain is

$$P(X_i \mid past) = P(X_i \mid X_{i-1})$$

- In our case these conditionals are simply **Bernoulli**
- In other words, the MC we've constructed is built from a one-step conditional probability idea and a humble Bernoulli distribution

- Finally, what's the joint distribution over X_1 ... X_T?
- That is, *global joint* can be expressed in terms of *local conditionals*

Likelihood

- Finally, what's the joint distribution of *n* datapoints sampled from the chain?
- That is, *global joint* can be expressed in terms of *local conditionals*

$$
\begin{aligned}
P(X_1, X_2, ..., X_N) &= P(X_N | X_1, X_2, ..., X_{N-1}) \times \\
&\quad P(X_{N-1} | X_1, X_2, ..., X_{N-2}) \times \\
&\quad \cdots \\
&\quad P(X_2 | X_1) \times \\
&\quad P(X_1)
\end{aligned}
$$

always true, for any ordering

$$
\begin{aligned}
&= P(X_N | X_{N-1}) \times \\
&\quad P(X_{N-1} | X_{N-2}) \times \\
&\quad \cdots \\
&\quad P(X_2 | X_1) \times \\
&\quad P(X_1)
\end{aligned}
$$

Based on 1st order Markov property

$$
= P(X_1) \prod_{i=2}^{N} P(X_i | X_{i-1})
$$

- This is the joint distribution of the data given the parameters, leading to a very compact likelihood function
- Let's find the MLE's of our binary Markov chain…

Estimators

- **Estimator** is function of random data (“a statistic”) which provides an estimate of a parameter:

$$\hat{\theta} \quad = \quad \hat{\theta}(X_1, X_2 \ldots X_n)$$

- Estimation is how we go from real-world data to saying something about underlying parameters
- We've seen a simple example of building up a more complicated model using a simple pmf, so even in complex settings, the ability to estimate properly is crucial
- This is why it's worth looking at **properties of estimators**

Properties of estimators

- The estimator is a **function of RVs**, so is itself a RV:

$$\hat{\theta} = \hat{\theta}(X_1, X_2 \dots X_n)$$

- Two key properties:
  - **Bias**
  - **Consistency**

Estimators

- Estimator is an RV.
- Let's use subscript *n* to indicate the number of datapoints ("sample size"):

$$\hat{\theta}_n = \hat{\theta}(X_1, X_2 \ldots X_n)$$

- Then $\hat{\theta}_n$ is a RV whose distribution is the distribution of values you'd obtain if you
  - repeatedly sampled *n* datapoints
  - applied the estimator
  - and noted down the estimate

Random variation in estimators

- Estimator is a RV, itself subject to **random variation**
  - Easy to forget that when dealing with randomness, even the “answer” is subject to variation
  - Have to be careful to see that what we think are “good” methods are consistently useful, and that a good result isn't just a fluke

Bias

- Estimator is a RV, itself subject to **random variation**
- A natural question then is this: *how different is the average of the estimator from the true value of the parameter*?
- The quantity

$$\mathbb{E}[\hat{\theta}_n] - \theta$$

captures this idea and is called the **bias** of the estimator

- An estimator with zero bias for all possible values of the parameter, i.e.:

$$\forall \theta \cdot \mathbb{E}[\hat{\theta}_n] = \theta$$

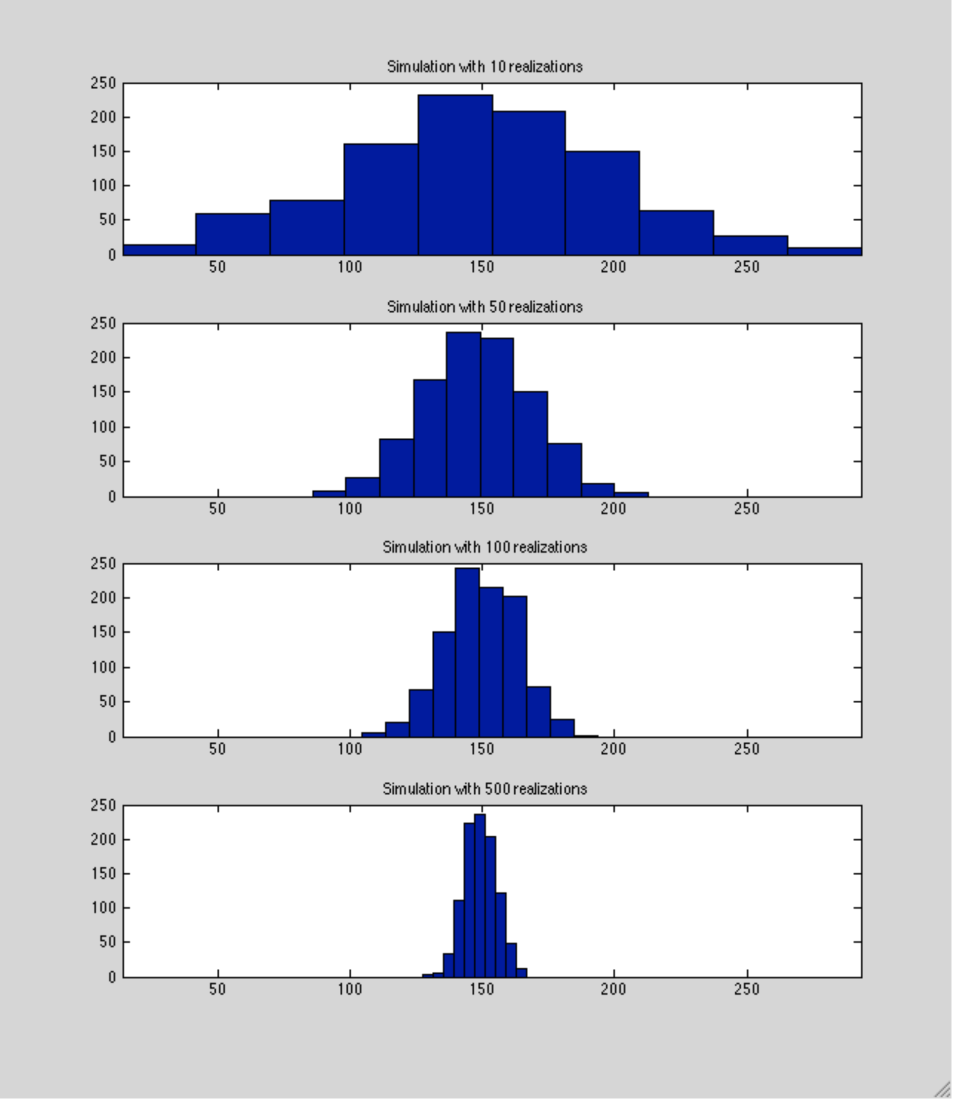
is said to be **unbiased**

Consistency

- Notion of bias is tied to sample size *n*
- What if we had **lots** of data?
- You'd hope that with enough data you'd pretty much definitely get the right answer…
  - Remember the lab?
  - More simulations allowed us to accurate estimate the variance of X^2 (X was roll of a die)

– We we don't get the "right" answer with lots of data, we should worry

– So, we're interested in the behaviour of the estimator as *n* grows large



Convergence in probability

- RVs don't converge deterministically: there's always *some* chance, even for large *n*, that we don't get the right answer
- Instead we will use a probabilistic notion of convergence

- We say that a sequence $X_1, X_2 \ldots$ of RVs **converges in probability** to a constant *k*, if

$$\forall \epsilon > 0, \quad \lim_{n \to \infty} P(|X_n - k| \geq \epsilon) \quad = \quad 0$$

Consistency

- We can now say something about how an estimator behaves as *n* grows large

- We say that an estimator is **consistent** if it converges in probability to the true value of the parameter. That is, if:

$$\forall \epsilon > 0, \quad \lim_{n \to \infty} P(|\hat{\theta}_n - \theta| \geq \epsilon) \quad = \quad 0$$

- Sufficient conditions for consistency:

$$\lim_{n \to \infty} \mathbb{E}[\hat{\theta}_n - \theta] = 0$$

$$\lim_{n \to \infty} \mathbb{V}[\hat{\theta}_n] = 0$$

  - … asymptotically unbiased, zero variance

Example: Bernoulli MLE

- The estimator:

$$\hat{\theta}_{MLE} = \frac{n_1}{n}$$

$$n_1 = \sum_{i=1}^{n} x_i$$

- **Q: can you write down the expectation of the estimator? (Just apply the E operator...)**

Example: Bernoulli MLE

- The estimator:

$$\hat{\theta}_{MLE} = \frac{n_1}{n}$$

$$n_1 = \sum_{i=1}^{n} x_i$$

- Expectation of estimator:

$$\begin{aligned}\mathbb{E}[\hat{\theta}_n] &= \mathbb{E}[n_1/n] \\ &= \frac{n\theta}{n} = \theta\end{aligned}$$

- That is, unbiased

Example: Bernoulli MLE

- Consistency: we've shown the estimator is unbiased, so all we need is to show that

$$\lim_{n\to\infty} VAR(\hat{\theta}_n) = 0$$

- Variance of estimator:

$$VAR(\hat{\theta}_n) = \frac{VAR(n_1)}{n^2}$$

- Result follows
- Of course, we can **verify these properties computationally**

Example: Bernoulli MLE

- Consistency: we've shown the estimator is unbiased, so all we need is to show that

$$\lim_{n \to \infty} VAR(\hat{\theta}_n) = 0$$

- Variance of estimator:

$$\begin{aligned} VAR(\hat{\theta}_n) &= \frac{VAR(n_1)}{n^2} \\ &= \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n} \end{aligned}$$

- Result follows
- Of course, we can **verify these properties computationally**

Weak Law of Large Numbers

- A very general and intuitive result
- If $X_1, X_2 \ldots X_n$ are i.i.d. RVs with:

$$\begin{aligned} \mathbb{E}[X_i] &= \mu_X \\ VAR(X_i) &= \sigma_X^2 < \infty \end{aligned}$$

Then the **sample mean**:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

converges in probability to the true mean:

$$\forall \epsilon > 0, \quad \lim_{n\to\infty} P(|\bar{X}_n - \mu_X| \geq \epsilon) = 0$$

Properties of estimators

- Theory is interesting, but what is really important are the concepts
  - The estimator *itself* is subject to variation
  - How much of a difference this makes depends on interplay between how many parameters, how much data etc.
  - Sometimes theory can tell us what problems to expect, but failing neat closed-form expressions, theory at least guides us towards what we should simulate to understand what's going on