Basic Statistics

Machine Learning

Statistical Analyses

- Statistical analyses usually involve one of 3 things: (1) The study of populations using samples; (2) The study of variation in the population and amongs samples; and (3) Techniques for data abstraction and data reduction
- Statistical analysis is more than calculating using recipies:
  1. What is the question to be answered?
  2. Can it be quantitative (i.e. can we make measurements about it)?
  3. How do we collect data?
  4. What can the data tell us?

Some things you may want to do with data

1. Visualise
2. Summarise
3. Determine the distribution of values
4. Compare groups of instances
5. Identify or describe relationships
6. Identify groups of similar instances
7. Identify groups of variables

Probability Theory and Statistics I

- Probability theory is concerned with mathematical descriptions of randomness, using the idea of a random variable (along with *events* and *stochastic processes*)
- Probability theory provides the mathematical foundations for statistics:
  - One practical example of the use of probability theory in statistics is the use of probability models that use theoretical distributions to model relative frequencies
  - Another example is the derivation of results that are relevant to sampling. Two prominent examples are the ‘ ‘Law of Large Numbers” and the “Central Limit Theorem”
  - Probability theory also provides ways of calculating bounds on the relative frequency of outcomes, irrespective of the underlying distribution. An example of such a bound is given by ”Chebyshev’s inequality”

Probability Theory and Statistics II

- What about events that do not occur repeatedly? *Subjective probabilities* are used to deal with once-off events.

  > *The Miracle of the Sun was an event on 13 October 1917 in which 30,000 to 100,000 people, who were gathered near Fatima, Portugal, claimed to have an witnessed extraordinary solar event as a sign from The Virgin Mary. According to many witnesses, after a period of rain, the dark clouds broke and the sun appeared as an opaque, spinning disc in the sky. It was significantly duller than normal, and cast multicolored lights across the landscape. The sun then careened towards the earth in a zigzag pattern.*

  What is the probability of such an event? Clearly a sample-based estimate is not possible. ayesian probability theory gives us the machinery to estimate probabilities like these

The Law of Large Numbers I

- A probability distribution over the values of a random variable describes how likely each value is. That is, one view of a probability distribution is as the relative frequency—in the long run—with which a random variable will take its different values.
- The difference between the probability of some event and the relative frequency with which it is observed in a sample necessarily approaches 0. This is the essence of the *Law of Large Numbers*
- This is what underlies the justification for repeated collection of observational data to measure some quantity
- Conversely, the law provides the theoretical support for using probability dstributions to deal with randomness in the real-world

The Law of Large Numbers II

- But it does not justify the gambler's fallacy: a win (or a head) is due after a long sequence of losses (or tails)
- Here is what the law implies when estimating the (true) mean of a distribution using a sample:
  1. Take a sample of size n
  2. See if the sample mean is "close" to the real mean. That is, for some ϵ, check if the following occurs:

$$|\text{Mean} - \mu| < \epsilon$$

  3. If you perform this experiment many times, then you will find that this occurs nearly all the time. That is:

$$P\left(|\text{Mean} - \mu| < \epsilon\right) \rightarrow 1$$

- More generally:

$$\lim_{n \rightarrow \infty} P\left(\left|\hat{\theta} - \theta\right| < \epsilon\right) \rightarrow 1$$

Monte Carlo Estimation I

- What is the probability that 10 dice throws add up exactly to 34?
- The exact probability that 10 dice throws to add up to 34 is $4325310/(6^{10}) \approx 0.072$. Here is an example of how relative frequencies approach this value in samples of size n:

n	Approx P
10	0.000
100	0.090
250	0.048
500	0.056
5000	0.062
10000	0.065

- The approximate probabilities are called *Monte Carlo estimates*. There are 2 ways to obtain probabilities:

Monte Carlo Estimation II

Exact. Calculate this exactly by counting all possible ways of making 34 from 10 dice. This is the exact answer.

Approximate. Simulate throwing the dice (say 500 times), count the number of times the results add up to 34, and divide this by 500. This is one estimate of the exact answer.

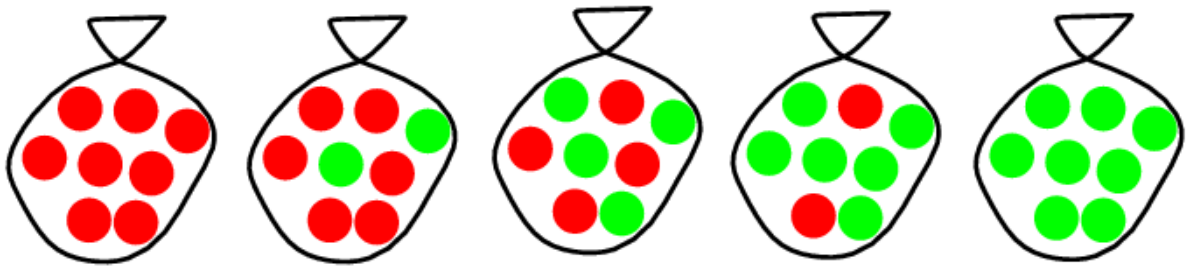
- In fact, the estimate from the approximte way can get quite close to the correct answer quite quickly (as you can see from the table)
- Use Monte Carlo estimation when:
  1. We have all the probabilities we need for the computation, but
  2. . . . actually doing the computation is too hard.

Fitting Models to Data

- Theoretical distributions are *parametrised*. That is, they are defined as functions of some parameters. For example, the Gaussian distribution is $N(\mu, \sigma)$
- If it is known that a theoretical distribution is an appropriate model for data, we may still not know the exact values of the parameters of the distribution
- It is sometimes possible to find the value for parameters of the distribution that maximises the probability $P(D|\theta)$ of obtaining the data observed. This probability is called the *likelihood*
  - We are asking: what value of θ for the known theoretical distribution maximises the probability of observing the data D
- That is, we find values for parameters that *maximises likelihood*

Example: The Maximum Likelihood Principle I

- Suppose there are five kinds of bags of lollies from Russell and Norvig):
  1. 10% are h_1: 100% cherry lollies
  2. 20% are h_2: 75% cherry lollies + 25% lime lollies
  3. 40% are h_3: 50% cherry lollies + 50% lime lollies
  4. 20% are h_4: 25% cherry lollies + 75% lime lollie
  5. 10% are h_5: 100% lime lollies



- Then we observe lollies drawn from some bag:

Example: The Maximum Likelihood Principle II

- What kind of bag is it? What flavour will the next lolly be?
- To answer these questions, we will first have to fit a model to the data

The Maximum Likelihood Principle (contd.) I

- Bags have a fraction θ of cherry lollies
- We are therefore dealing with binomial models (cherry vs lime lollies) in which we do not know θ. We will take this set of models to be characterised by the *parameter* θ
- Now we unwrap N lollies, and find c and $N - c$ limes. We will have to assume that these are i.i.d. (independent, identically distributed) observations
- What can we say about the probability of observed data, using the binomial distribution as our theoretical model. This is:

$$\text{Prob}(c \text{ cherries and } (N-c) \text{ limes } \propto \theta^c(1-\theta)^{(N-c)}$$

- Question: For what value of θ will this probability be highest?

The Maximum Likelihood Principle (contd.) II

- Ans: Find maximum by differentiating and setting first differential to 0. Actually easier to differentiate $\log(P)$ and set that to 0:

$$\log(P) \;=\; L(P) \;=\; c\log\theta \;+\; (N-c)\log(1-\theta)$$

Differentiating w.r.t. θ and setting this to zero:

$$\frac{dL(P)}{d\theta} \;=\; \frac{c}{\theta} - \frac{N-c}{1-\theta} \;=\; 0$$

which gives $\theta = c/N$

- This is the "Maximum Likelihood Estimate" for θ ($L(P)$ is called the likelihood function)
(Seems sensible, but causes problems with 0 counts! But more on that later.)

Bayes Rule Revisited

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

Or

$$P(Hypothesis|Data) = \frac{P(Data|Hypothesis)P(Hypothesis)}{P(Data)}$$

Bayes Rule can be used as a rule of inference for solving ML problems. Given data and prior information, derive:

1. Parameter value (hypotheses are paratemeter values)
2. Predictions (hypotheses are values of a random variable)
3. Preference over models (hypotheses are models)

In all cases, we have to understand Bayes rule as an update of priors to posteriors (which become priors for the next update *etc*.)

Discrete Hypotheses, Discrete Data I

- Assume we have a model with a parameter Θ that takes values from a discrete set $\{\theta_1, \theta_2, \ldots, \theta_n\}$. We want to work out which value of $\Theta = \theta$ is best for the data d we have have observed
  - d is one of many data sets $d_1, d_2, \ldots, d_k$ that could have been drawn. So, d is the value of some random variable D
  - θ is one of many values $\theta_1, \theta_2, \ldots, \theta_n$ that could be assigned to the random variable Θ
  - For each combination d_i, θ_i, the probability $p(d_i, \theta_i)$ is a value from the joint distribution $p(D, \Theta)$
- Here, we are considering the case where D and Θ are discrete r.v.'s
  - Hypotheses are statements like $\Theta = \theta$. So $P(Hypothesis) = P(\Theta = \theta_i)$ where P is the p.m.f. of Θ

Discrete Hypotheses, Discrete Data II

- Bayes:

$$P(\Theta = \theta | D = d) = \frac{P(D = d | \Theta = \theta) P(\Theta = \theta)}{P(D = d)}$$

- Example
  - Suppose we have a coin, whose probability of *heads* could be one of $0.5, 0.6$ or 0.9, with prior probabilities $0.4, 0.4$ and 0.2 respectively.
  - The coin is flipped once, and it lands *heads*. That is the data is $D = h$
  - Then:

H	Prior	Likelihood	Numerator	Denom	Posterior
$\Theta = 0.5$	0.4	0.5	0.2	0.62	0.3226
$\Theta = 0.6$	0.4	0.6	0.24	0.62	0.3871
$\Theta = 0.9$	0.2	0.9	0.18	0.62	0.2903

Discrete Hypotheses, Discrete Data III

- NOTE: The hypothesis with the highest likelihood is different to the hypothesis with the highest posterior probability

Continuous Hypotheses, Discrete Data I

- Assume we have a model structure with some parameter Θ, and we want to work out which value of $\Theta = \theta$ is best for the data d we have observed
- We now consider the case where D is a discrete r.v. and Θ is a continuous r.v
  - Hypotheses are statements like $\Theta = \theta$
  - Correctly: "Θ lies in an interval $d\theta$ around θ". So $P(Hypothesis) = f(\Theta = \theta)d\theta$ where f is the p.d.f
- Bayes:

$$f(\Theta = \theta | D = d)d\theta = \frac{P(D = d|\Theta = \theta)f(\Theta = \theta)d\theta}{P(D = d)}$$

Continuous Hypotheses, Discrete Data II

- This is usually simplified to:

$$f(\theta|d) = \frac{P(d|\theta)f(\theta)}{P(d)}$$

(BUT: remember how we got here, with the $d\theta$'s)

- Example
  - Suppose we have a biased coin, whose probability of *heads* is some unknown value θ. The prior p.d.f is $f(\theta) = 2\theta$
  - The coin is flipped once, and it lands *heads*
  - Then:

Data.	This is $D = \textit{heads}$ (or $D = h$ for short)
Hypothesis.	This is $\Theta = \theta$ (θ is not known)
Likelihood.	This is $P(D = h\|\Theta = \theta) = \theta$
Prior.	$f(\theta)\ d\theta = 2\theta\ d\theta$
Numerator.	$2\theta^2\ d\theta$
Denominator.	$P(D = h) = \int_0^1 2\theta^2\ d\theta = 2/3$

Continuous Hypotheses, Discrete Data III

Posterior. $(2\theta^2\ d\theta)/(2/3) = 3\theta^2\ d\theta$

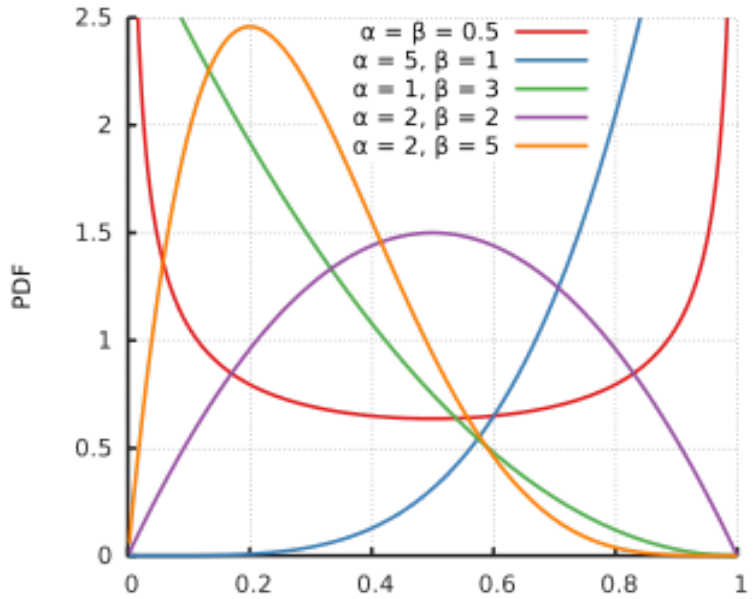
Posterior p.d.f $3\theta^2$

- A very flexible function for specifying the prior p.d.f. of a continuous r.v. uses the *Beta* distribution
- The Beta distribution is a 2-parameter distribution, specifying the p.d.f.

$$Beta(\alpha, \beta) = f(\theta; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

We omit the proof here that f is a density function (that is, $\int_0^1 f(\theta; \alpha, \beta)\ d\theta = 1$. Varying α and β gives us many different kinds of p.d.f's:

Continuous Hypotheses, Discrete Data IV



- In general, any p.d.f. of the form:

$$f(r) \;=\; c\; r^{\alpha-1}(1-r)^{\beta-1}$$

for $0 \leq r \leq 1$ is necessarily $Beta(\alpha, \beta)$ with $c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$

Continuous Hypotheses, Discrete Data V

- Let us take the case of data drawn from a Binomial distribution. That is, the number of successes y_N (*heads*) is the value of a r.v. $Y_N \sim Binom(N, \theta)$
- Example
  - Suppose we observe 8 *heads* in 12 tosses of a coin with $P(heads) = \theta$. Assume the prior p.d.f is $f(\theta) = 1$. What is the posterior p.d.f for θ?

$$
\begin{aligned}
\text{Data.} \quad & Y_{12} = y_{12} = 8 \text{ (8 } h\text{'s in 12 trials)} \\
\text{Hypothesis.} \quad & \Theta = \theta \\
\text{Likelihood.} \quad & \binom{12}{8}\theta^8(1-\theta)^4 \\
\text{Prior.} \quad & 1d\theta \\
\text{Numerator.} \quad & \binom{12}{8}\theta^8(1-\theta)^4 d\theta \\
\text{Denominator.} \quad & \binom{12}{8}\int_0^1 \theta^8(1-\theta)^4 d\theta \\
\text{Posterior.} \quad & c\theta^8(1-\theta)^4 d\theta = Beta(9,5)d\theta
\end{aligned}
$$

  - So, the posterior p.d.f. is $Beta(9, 5)$

Continuous Hypotheses, Discrete Data VI

- BUT: the prior p.d.f. ($f(\theta) = 1$) is just $Beta(1, 1)$. So, after observing 8 *heads* and 4 *tails*, a $Beta(1, 1)$ prior p.d.f has been updated to a $Beta(1 + 8, 1 + 4)$ posterior p.d.f
- GENERALISE: if we started with a $Beta(a, b)$ prior p.d.f and we observe s *heads* and t *tails*, the posterior p.d.f will be $Beta(a + s, b + t)$

- This property (prior and posterior both having a *Beta* distribution) happens because the similarity of the functional form of the likelihood (Binomial) and the prior p.d.f (Beta):

$$\begin{aligned} \text{Likelihood.} \quad & c_1\theta^s(1-\theta)^t \\ \text{Prior.} \quad & c_2\theta^{a-1}(1-\theta)^{b-1}d\theta \\ \text{Posterior} \quad & c_3\theta^{s+a-1}(1-\theta)^{b+t-1}d\theta \end{aligned}$$

- The *Beta* distribution is said to be a *conjugate* prior p.d.f. for a Binomial likelihood function.
  - For every such conjugage prior-likelihood pair, the posterior will have the same functional form as the prior

Continuous Hypotheses, Discrete Data VII

- There are other pairs with this property (Dirichlet-Multinomial; Gaussian-Gaussian *etc.*)

- Conjugate prior-likelihood pairs make Bayesian posterior probability calculation easy (easier)

Discrete Hypotheses, Continuous Data I

- Assume we have a model with a parameter Θ that takes values from a discrete set $\{\theta_1, \theta_2, \ldots, \theta_n\}$. We want to work out which value of $\Theta = \theta$ is best for the data x (previously we used d) we have have observed
  - x is the value of some r.v. X
  - θ is one of many values $\theta_1, \theta_2, \ldots, \theta_n$ that could be
- Here we want to consider the case where X is a continuous r.v. and Θ is a discrete r.v.
- The only thing that changes is that likelihoods are calculated using p.d.f's. That is:

Hypothesis.	$\Theta = \theta$
Data.	X in an interval dx around x
Prior.	$P(\Theta = \theta)$
Likelihood.	$f(x\|\theta)dx$ (correctly, $f_{X\|\Theta}(x\|\theta)dx$)
Numerator.	$f(x\|\theta)dxP(\Theta = \theta)$

Discrete Hypotheses, Continuous Data II

Denominator. $f(x)dx = (\int_0^1 f(\theta)d\theta)\ dx$

Posterior. $P(\Theta = \theta|x)$

- Bayes:

$$P(\Theta = \theta|x) = \frac{f(x|\Theta = \theta)dxP(\Theta = \theta)}{f(x)dx}$$

(sometimes, the dx's are not shown)

- Example
  - Suppose data are drawn from one of 2 clusters whose centres are at $\mu_1 = 2$ and $\mu_2 = 5$, both with prior probability 0.5. Data from a cluster μ_i are drawn according $N(\mu_i, 1)$
  - An instance $x = 3$ is observed. What is the posterior probability of the clusters given this data?
  - Then:

Discrete Hypotheses, Continuous Data III

H	Prior	Likelihood	Numerator	Denom	Posterio
$\Theta = 2$	0.5	$0.4e^{-(3-2)^2/2}\ dx$	$0.12\ dx$	$0.19\ dx$	0.63
$\Theta = 5$	0.5	$0.4e^{-(3-5)^2/2}\ dx$	$0.07\ dx$	$0.19\ dx$	0.37

Continuous Hypotheses, Continuous Data I

- Assume we have a model structure with some parameter Θ, and we want to work out which value of $\Theta = \theta$ is best for the data x (previously we used d) we have observed.
  - As before, θ and x will be values of random variables Θ and X (used to be D)
  - Both Θ and X are now continuous r.v.'s
- The only thing that changes is that both likelihood and priors are calculated using p.d.f's. That is:

Hypothesis.	$\Theta = \theta$
Data.	X in an interval dx around x
Prior.	$f(\theta)d\theta$
Likelihood.	$f(x\|\theta)dx$ (correctly, $f_{X\|\Theta}(x\|\theta)dx$)
Numerator.	$f(x\|\theta)dxf(\theta)d\theta$
Denominator.	$f(x)dx = (\int_0^1 f(\theta)d\theta)\ dx$
Posterior.	$f(\theta\|x)d\theta$

Continuous Hypotheses, Continuous Data II

- Example
  - Suppose $X \sim N(\theta, 1)$ and $\Theta \sim N(2, 1)$
  - Then:

$$\begin{aligned}
\text{Prior.} \quad & c_1 e^{-(\theta-2)^2/2} d\theta \\
\text{Likelihood.} \quad & c_2 e^{-(5-\theta)^2/2} dx \\
\text{Numerator.} \quad & c_3 e^{-(2\theta^2-14\theta+29)/2} dx d\theta = \\
& c_3 e^{-((\theta^2-7/2)^2+9/4)} dx d\theta = c_4 e^{-(\theta^2-7/2)^2} dx d\theta \\
\text{Denominator.} \quad & \left(\int c_4 e^{-(\theta^2-7/2)^2} d\theta\right) dx = c_5 dx \\
\text{Posterior.} \quad & c_6 e^{-(\theta^2-7/2)^2} d\theta
\end{aligned}$$

  - The posterior p.d.f. is $N(7/2, \sigma)$ where $2\sigma^2 = 1$ and $c_6 = 1/(\sigma\sqrt{2\pi})$
- So a Gaussian is conjugate prior for Gaussian likelihood.

Beyond Parameter Estimation I

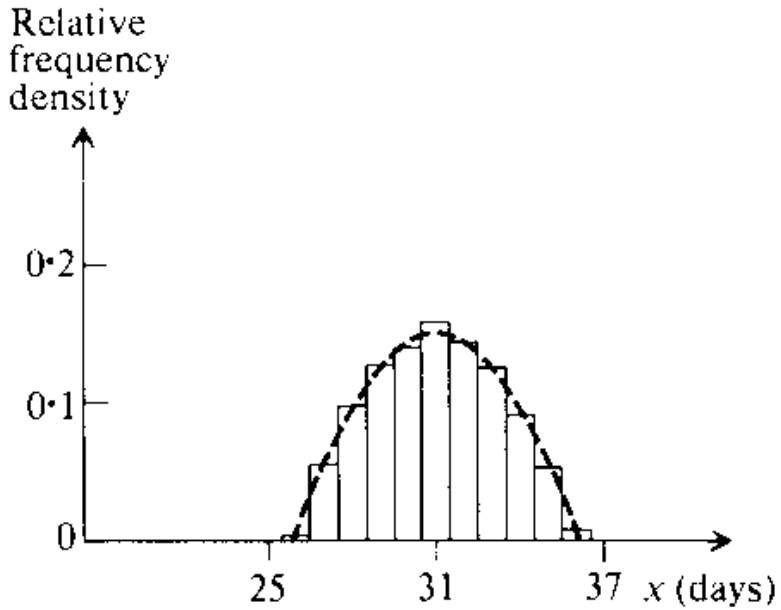
- Theoretical distributions (Gaussian, Poisson, Binomial...) have a well-defined structure: parameter estimation is all that is needed
- BUT: what if there is no known probability model? That is, there is no known theoretical distribution, and therefore no known structure for the function? This is an example of a second kind of estimation problem (*structure estimation*)
- In principle, any mathematical function will do, provided it satisfies some properties
- Here is a table of frequency values observed for some event:

Beyond Parameter Estimation II

x days	Frequency
$25.5 \leq x < 26.5$	4
$26.5 \leq x < 27.5$	55
$27.5 \leq x < 28.5$	99
$28.5 \leq x < 29.5$	127
$29.5 \leq x < 30.5$	140
$30.5 \leq x < 31.5$	158
$31.5 \leq x < 32.5$	142
$32.5 \leq x < 33.5$	125
$33.5 \leq x < 34.5$	90
$34.5 \leq x < 35.5$	52
$35.5 \leq x < 36.5$	8

Beyond Parameter Estimation III

- Here is a histogram of relative frequency density and possible theoretical probability density function



- How do you fit the correct model? (The mean is estimated to be $m = 39.98$ and spread is estimated to be $s^2 = 4.94$.)
- Let us fit a parabolic density function $\phi(x) = b^2 - a^2x^2$ with $\mu = 31$ and $\sigma^2 = 5$

Beyond Parameter Estimation IV

- It is easier to work with a shifted function that has the origin at $x = 31$. The new density function $\phi(x')$ will therefore have a mean of 0
- We want the total area under the function to be 1. That is

$$\int_{-b/a}^{b/a} (b^2 - a^2 x'^2 dx' = \frac{4b^3}{3a} = 1$$

- We also want $\sigma^2 = 5$. So:

$$\sigma^2 = \int_{-b/a}^{b/a} x' 2\phi(x') dx' - \mu^2 = \frac{4b^5}{15a^3} = 5$$

- Solving simulataneously, we get $b^2 = 0.15$ and $a^2 = 0.006$, giving $\phi(x') = 0.15 - 0.006x'^2$

Beyond Parameter Estimation V

- In general, we will need to solve 2 kinds of estimation problems:
  1. Parameter estimation
  2. Structure estimation
- Sometimes, we will decompose these into 2 separate tasks. At other times we will try to formulate it as an optimisation problem that finds the best combination of structure and parameters
- From now on, we will distinguish between:
  - Probability models, which specify parametrised theoretical distributions; and
  - Models for data, which are pairs (π, θ), where π denotes a structure from a set of structures Π and θ denotes parameters from a set of parameters Θ

(The structures in Π can be mathematical functions, trees, graphs, relations, programs . . .)

Beyond Parameter Estimation VI

- *Hypothesis selection* in Machine Learning means selecting a model $h = (\pi^*, \theta^*)$ for data d that satisfies some properties (for example, $P(h|d)$ is a maximum)

Estimation from Samples I

- Estimating some aspect of the population using a sample is a common task. Along with the estimate, we also want to have some idea of the accuracy of the estimat (usually expressed in terms of *confidence limits*: we will come to this later)
- Some measures calculated from the sample are very good estimates of corresponding population values. For example, the sample mean m is a very good estimate of the population mean μ. But this is not always the case. For example, the range of a sample usually under-estimates the range of the population
- We will have to clarify what is meant by a "good estimate". One meaning is that an estimator is correct on average. For example, on average, the mean of a sample is a good estimator of the mean of the population

Estimation from Samples II

- For example, when a number of samples are drawn and the mean of each is found, then average of these means is equal to the population mean
- Such an estimator is said to be *statistically unbiased*. More on this later

- Let us first look at what happens when we take lots of samples, and calculate an estimate from each sample
- For example:

Sample	Observations	Mean
1	3, 0, 0, 4, 2, 0, 2, 0, 12	2.3
2	4, 1, 1, 0, 17, 1, 0, 3, 1, 2	3.1
3	...	2.0
...	...	0.6
...	...	...
...	...	...

Estimation from Samples III

- Sample means vary: so there is a distribution of means in the last column. We will see in a moment that this distribution is approximately Normal (for almost all distributions that the original data may come from)
- The mean of this distribution is the same as the mean of the population from which the samples are drawn
- That is, the means are scattered approximately symmetrically about the mean of the population of the population. This scatter or standard deviation (called the *standard error of the mean*) is a scaled-down version of the population s.d.
- These are consequences of the *Central Limit Theorem*

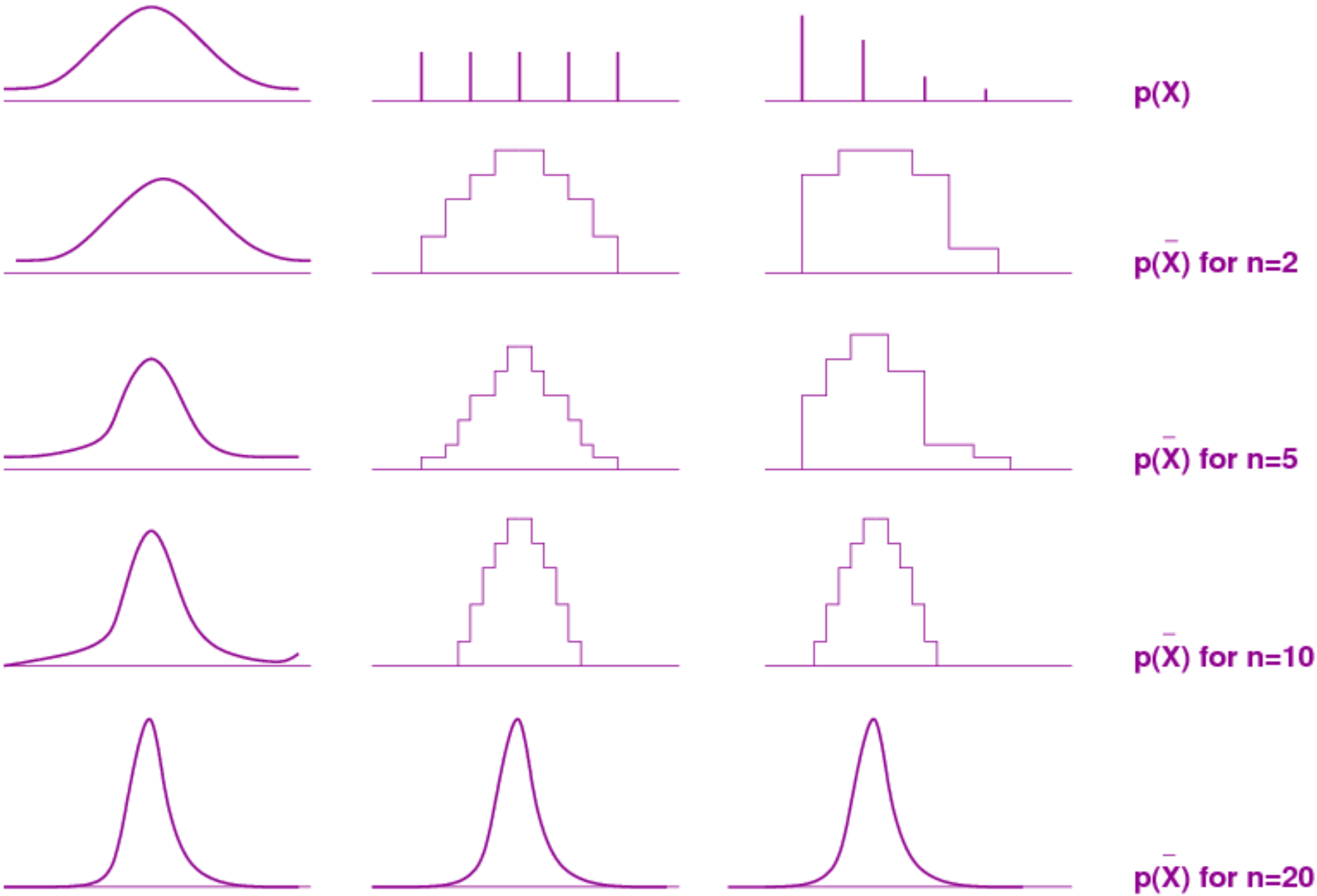
The Central Limit Theorem I

- Suppose we had n independent samples drawn from some population, which has some well-defined mean μ and variance σ^2
- Then, in the form we have seen this before, we said that, for large n, the (arithmetic) mean of the n data points is approximately symmetric and bell-shaped centred around μ and with a spread of σ^2/n. That is, we can model it with a normal (or Gaussian) distribution with mean μ and variance σ^2/n
- In general, the theorem actually applies not just to the mean, but to the sum of the data points. That is, for large n, the sum is well-modelled by a normal distribution with mean $n\mu$ and variance σ^2. The result on the mean is therefore a corollary of this more general statement of the theorem.
  - We will not look at the proof of the theorem ay this point

The Central Limit Theorem II

- The theorem even holds if the n observations are from different populations, provided no one observation dominates the sum
- It will not hold for power-law distributions which have infinite means and variances (see "Expectation Calculations" for these)

The Central Limit Theorem III



Approximations with a Single Sample

- The results about the distribution of the means is not very helpful, since we do not know either the population mean or its variance
- We also do not have many samples: we usually have a few, or often only one sample. What can we do in such cases?
- We can only *estimate* the parameters. For example:

$$\text{estimated std. error of the mean} = \frac{s}{\sqrt{n}}$$

- The CLT tells us that distribution of means is approximately Normal provided:
  - The sample size is about 10 or more, if frequencies in the original population are distributed Normally
  - The sample size is about 100 or more, if frequencies in the original population are distributed in a skewed manner

Approximations with a (Small) Single Sample

- The results about approximations to the Normal distribution of the means is not very close when sample sizes are small
- When sample sizes is small, the "t-distribution" gives the proportion of times different values of a specific ratio occurs in samples of that size. The ratio is like a standardised variable:

$$t = \frac{(\text{sample mean} - \text{pop. mean})}{\text{estimated std. error of mean}}$$

- Now, for each sample, we can calculate a t value. If the means of the samples is a Normal distribution then the resulting distribution of the t-values is a t-distribution
- For small samples, the t-values follow a t-distribution when the population is approximately Normal

Accuracy of an Estimate (Confidence Intervals) I

- If sample sizes are large enough, then the sampling distribution of a statistic like the mean will be approximately Normal with mean μ and s.d. equal to the standard error

Accuracy of an Estimate (Confidence Intervals) II

- The frequency-based interpretation of a confidence limit is a bit complicated. But in practice the means of most random samples will be somewhat similar, and it is usually good enough to act as though there is a very high chance that the population mean is between $2 \times \mathrm{std.err}$ of the sample mean

Small Sample Confidence Intervals

- The 95% interval for the population mean:

$$\mu = \text{sample mean} \pm 2 \times \text{std.err.}$$

Small Sample Confidence Intervals

- The 95% interval for the population mean:

$$\mu \;=\; \text{sample mean} \pm 2 \times \text{std.err.}$$

has two difficulties: (1) *std.err.* is equal to $\sigma/\sqrt{n}$. So, to calculate this, we need to know the population's standard deviation σ; and (2) The Normal approximation is not very good for small samples

Small Sample Confidence Intervals

- The 95% interval for the population mean:

$$\mu = \text{sample mean} \pm 2 \times \text{std.err.}$$

has two difficulties: (1) *std.err.* is equal to $\sigma/\sqrt{n}$. So, to calculate this, we need to know the population's standard deviation σ; and (2) The Normal approximation is not very good for small samples

- Both these problems are "solved" using the t distribution. So, the the 95% confidence interval for small samples with unknown σ is:

$$\mu = \text{sample mean} \pm t_{95\%,n-1} \times \text{sample std.err.}$$

where $t_{95\%,n-1}$ is a value from the t distribution. For small n this will be a bit larger than 2. The *sample std.err.* is equal to $s/\sqrt{n}$.

What Does a "Confidence Interval" Mean?

- The general approach for constructing a confidence interval for a parameter μ using a sample estimate m requires us to do something line $m \pm k \times s.e.$
- A 95% c.i. $m \pm a$ for some μ does not mean: we are 95% sure that μ lies between $m + a$ and $m - a$. It means in 5 times out of 100, the interval centred on m will not include μ
- To understand this, we must understand that estimates of the mean vary from one sample to the other. If we knew the distribution of how mean-estimates varied, then we could use that to construct the confidence interval

Significance I

- Sometimes, we have a prior hypothesis about the population. For example, female literacy is higher in smaller households. Is this really true?

Significance II

- If the difference is probably a fluke, then we say that our prior hypothesis cannot be ruled out. The prior hypothesis is usually called the *null hypothesis*
- If the probability of a fluke is very low, then we have reason to think that the null hypothesis is, in fact, not true. Put another way, *if* the null hypothesis was in fact true, then the probability of getting this sample of values is very small
- To get this probability, we will need to know the distribution of the sample statistic
  - From the CLT, we know that the sampling distribution of the sample mean is approximately Normal

Significance III

- There is a big difference between:
  - *Hypothesis testing*, in Statistics; and
  - *Hypothesis selection*, in Machine Learnng

Efficient Estimators

- As well as being correct on average, we would also like the distribution of sample values to have a low scatter
- Estimators can therefore be compared on the basis of the variance of their sample distributions

$$\text{Efficiency of V vs. W} = \frac{\textit{variance of W}}{\textit{variance of V}}$$

- If this value is greater than 1 then V is *more efficient* than W. For example:
  - When samples are drawn from a population that is approximately Normal, the distribution of sample medians has a variance of about 1.6 times the variance of the distribution of the sample means.
  - When samples are drawn from a population that has a specific power law distribution called the Laplace distribution, the distribution of sample medians has a variance of 0.5 times the variance of the distribution of sample means

The sample mean is a less efficient estimator than the median if power law distributions are involved

The Bias-Variance Tradeoff I

- An estimate $\hat{\theta}$ of a parameter θ is said to be an *unbiased estimate* of θ if:

$$\mathrm{E}\left(\hat{\theta} - \theta\right) = 0$$

  the one with minimum variance (that is, the most efficient estimator)
- In general, we would be comparing estimators that have some bias and some variance
- For example, we can combine the bias and variance of an estimator by obtaining the *mean square error* of the estimator, or MSE. This is the average value of squared deviations of an estimated value V from the true value of the parameter θ. That is:

$$\mathrm{MSE} \;=\; \text{Avg. value of} (V - \theta)^2$$

The Bias-Variance Tradeoff II

- Now, it can be shown that:

$$\text{MSE} = (\text{variance}) + (\text{bias})^2$$

- So, we can re-define the efficiency of unbiased estimators:

$$\text{Efficiency of V vs. W} = \frac{\textit{MSE of W}}{\textit{MSE of V}}$$

- Since

$$\text{MSE} = (\text{variance}) + (\text{bias})^2$$

  The lowest possible value of MSE is 0

The Bias-Variance Tradeoff III

- In general, we may not be able to get to the ideal MSE of 0. Sampling theory tells us the minimum value of the variance of an estimator. This value is known as the *Cramer-Rao* bound. So, given an estimator with bias b, we can calculate the minimum value of the variance of the estimator using the CR bound (say, v_{min}). Then:

$$\text{MSE} \geq v_{min} + b^2$$

  The value of v_{min} depends on whether the estimator is biased or unbiased (that is $b = 0$ or $b \neq 0$)
- It is not the case that v_{min} for an unbiased ($b = 0$) estimator is less than v_{min} for a biased estimator. So, the MSE of a biased estimator can end up being lower than the MSE of an unbiased estimator

Correlation I

- The *correlation coefficient* is a number between -1 and +1 that indicates whether a pair of variables x and y are associated or not, and whether the scatter in the association is high or low
  - High values of x are associated with high values of y *and* low values of x are associated with low values of y, and scatter is low
  - A value near 0 indicates that there is no particular association and that there is a large scatter associated with the values
  - A value close to -1 suggests an inverse association between x and y
- Only appropriate when x and y are roughly linearly associated (doesn't work well when the association is curved)

What Does Correlation Mean? I

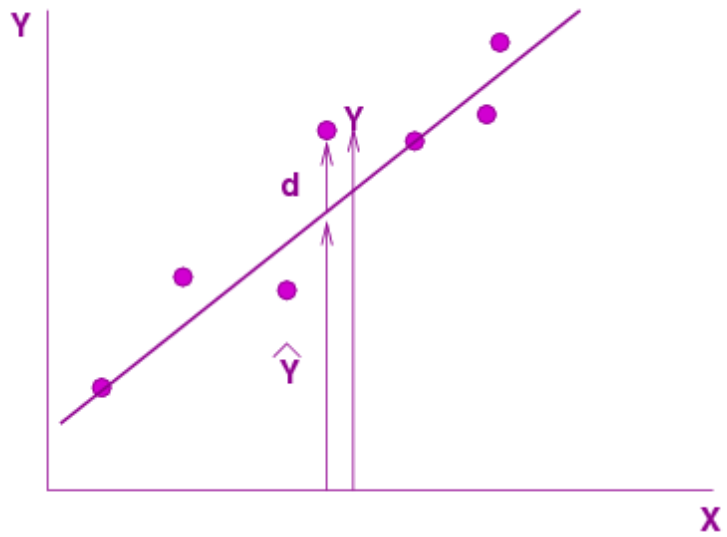
- r is a quick way of checking whether there is some linear association between x and y
- The sign of the value tells you the direction of the association
- All that the numerical value tells you is about the scatter in the data
- The correlation coefficient does not model any relationship. That is, given a particular x you cannot use the r value to calculate a y value
  - It is possible for two datasets to have the same correlation, but different relationships
  - It is possible for two datasets to have the different correlations but the same relationship
- Cannot use correlations to compare datasets. All you can derive is whether there is a positive or negative relationship between x and y

MAXIM #4: Correlation isn't causation

Regression

- Given a set of data points x_i, y_i, what is the relationship between them?
- One kind of question is to ask: are these linearly related in some manner? That is, can we draw a straight line that describes reasonably well the relationship between X and Y
- Remember, the correlation coefficient can tell us if there is a case for such a relationship
- In real life, even if such a relationship held, it will be unreasonable to expect all pairs x_i, y_i to lie precisely on a straight line. Instead, we can probably draw some reasonably well-fitting line. But which one?

Linear Relationship Between 2 Variables I



- GOAL: fit a line whose equation is of the form $\hat{Y} = a + bX$
- HOW: minimise $\sum_i d_i^2 = \sum_i (Y_i - \hat{Y})^2$ (the "least squares estimator")

Linear Relationship Between 2 Variables II

- The calculation for b is given by:

$$b = \sum(xy) / \sum x^2$$

where $x = (X_i - \overline{X})$ and $y = (Y_i - \overline{Y})$

- $a = \overline{Y} - b\overline{X}$

Linear Relationship Between 2 Variables III

- Under aome assumptions (called *Gauss-Markov* assumptions, these are unbiased estimates of a "true" line between X and Y.
  - Since $MSE = variance + (bias)^2$, these values of a and b will give the most efficient estimate of the true line
  - Under soem further assumption (about the distribution of errors), this line is the line with maximum likelihood

From Relationships to Generalisations I

- An empirical relationship—like a regression line— obtained under some conditions, cannot constitute the basis of a generalisation that holds under substantially different conditions
- For example, a a linear relationship between log(*weight*) and *height* of children, obtained from data can only become generally acceptable if it is shown to hold (with approximately the same coefficients) across a range of conditions
  - The result holds irrespective of race, gender, geography, socio-economic status, time *etc.*
- When different sets of data are modelled by the same relationship, then there is a case for the relationship being *lawlike*

Using Relationships I

- Summarisation. The obvious use of a relationship is that it summarises the data from which it was obtained
- Prediction. For data drawn under the same conditions, we would expect to be able to use the relationship for prediction
  - Distinguish here between *prediction* and *extrapolation*
  - We will use the first to mean using the relationship within the operating range; and the second to mean using the relationship outside the operating range
  - When we predict, we expect to do so with high accuracy
  - When we extrapolate, a successful outcome is unexpected and suggests we might have found a lawlike relationship
  - *MAXIM: Empirical relationships do not hold universally*
- Understanding. Although empirical relationships do not tell us why something happens, they can form the low-level building blocks for developing a better understanding.