Were Germany “robbed” of the 1966 World Cup?

3D reconstruction
two video cameras
football laws decree that the football ground is rectangular

Was it a goal?
Should it have been awarded?
Historical fact

97,000 people crammed inside London's Wembley Stadium to watch.

After 12 minutes Helmut Haller had put West Germany ahead, but the score was levelled by Geoff Hurst four minutes later. Martin Peters put England in the lead in the 78th minute; England looked set to claim the title when the referee awarded a free kick to West Germany. The ball was launched goalward and Wolfgang Weber managed to poke it across the line, with England appealing in vain for handball as the ball came through the crowded penalty area.

With the score level at 2-2 at the end of 90 minutes, the game went to extra-time.

In the 98th minute Hurst found himself on the score sheet again, when his shot hit the crossbar and was controversially deemed to have crossed the line by the referee. It didn't matter however as Hurst netted his third in the 120th minute, just as the gathered crowd invaded the pitch to celebrate with the team. Geoff Hurst is the only player ever to have scored three times in a World Cup Final.

BBC commentator Kenneth Wolstenholme’s description of the match's closing moments has gone down in history: "Some people are on the pitch. They think it's all over." *(Hurst scores)* "It is now!"
Does a second view resolve the issue?
Does freeze frame help?

officials can’t do freeze frame, but they do need to be sure
Referee’s judgement & reasonable level of confidence

The whole of the ball must cross the line for a goal to be awarded

In case of doubt, the defending side gets the benefit of the doubt … usually
All measurements, and all perceptual judgements, are subject to uncertainty

A goal should be awarded if, and only if, the most trailing edge of the confidence interval is inside the goal area
So, should the goal have been awarded?

• Two video feeds
  – Different syncs; but known shape of ground enables transformation between them to be estimated precisely
  – The ground is defined to be rectangular, and the posts vertical
• Tracking the ball
• Infer best estimate of the error in the region of impact of the ball on the goal line
Reconstructed 3D scene

Bird’s eye view of the goal

This relies on projective geometry & stereo vision. So far as we are concerned, the ball moves on a trajectory in 3D. Uncertainty is explicit
Kalman filtering
maintaining & updating a model of uncertainty (in the ball’s position)

An environment; but it’s foggy and our range sensor has limited range. Blue lines mark known walls.

A planned trajectory

An actual trajectory

The ellipses represent the uncertainty in position
How much mileage does your car do on 10 litres of unleaded petrol?

- **Obvious variations:**
  - city vs motorway; flat vs mountains; calm day vs strong headwind; using air conditioning vs not; accelerating slowly vs not; …

- Does BP fuel give better mileage than Shell?
- How would you determine superiority for a given driving condition?
- How many “trials” would it take to convince you that such a claim is “substantially true”?
Heights and weights

• “Generally” males are taller and heavier than females
• Are the people in this lecture “representative” of the normal population?
• Are there “significant differences” between first year Engineering students and PPE?
3G phones
from engineering considerations to business decisions

- The financial viability rests on the cost of providing sufficient bandwidth, and this increases sharply as the bandwidth increases.
- We’d like to keep costs low, yet provide a “sufficient quality of service.”
- QoS:
  - Guaranteed connection; speed of response
- Design parameters
  - How many simultaneous users? 5 yrs from now?
  - How long will each call be? How large a “typical” msg? (note that 3G phones aimed primarily at www and images)
  - What carrying capacity meets demands: (a) all of the time; (b) most of the time; “on the average”
- Business considerations
  - Can we create a market, eg among business users, or schoolkids?
  - What additional features can we add – eg transmission of images?
Microcalcifications: a breast cancer screening target

In this image, all the localised white spots are microcalcifications, the earliest sign of breast cancer detectable in images.

We have developed a program that detects microcalcifications based on knowledge of the formation of a x-ray mammogram.
Mammogram fragment

Anisotropic diffusion

Wiener filtering hint

Adaptive foveal filter fusion
Truth and Judgement
This applies to all decision making, not just in medicine

<table>
<thead>
<tr>
<th></th>
<th><em>Is a microcalc</em></th>
<th><em>Is not a microcalc</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Judged to be a</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>microcalc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judged <em>not</em> to</td>
<td>FN</td>
<td>TN</td>
</tr>
<tr>
<td>be a microcalc</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note that the judgement depends on parameters, for example local image contrast

Aim is to maximise TPs while minimising FPs
Sensitivity and Specificity

- Sensitivity = TP/(TP+FN)
- Specificity = TN/(FP+TN) = 1 - FP/(FP+TN)
- Parameter value $c$ determines $(\text{Sensit}(c), \text{Specif}(c))$
Receiver Operator Characteristics curves

98.5% TP rate for a FP rate of 0.03 per image
Variability and uncertainty are endemic

• All sensory data is subject to noise
  – Signal/image enhancement, pattern classification, object tracking, estimating effects of therapy, detecting need for preventative maintenance, …

• All measurements are prone to variations between “observers”
  – Segmenting brain structures, “ground truth”
Statistical inference

• Often, vast number of examples to test
  – Voting intention for UK election
  – Time precludes sampling the whole set
• Select a manageable subset
  – 1,000 voters
• From response of sample set, infer the likely behaviour of the whole population

A “poorly selected” sample would give a misleading result
Statistical sample set

• What does it mean for a sample set to be “representative”?
• How can we design a set to be free from bias?
• Given that there are 30 million voters in the UK, does 1,000 suffice for a representative sample for voting intentions? 10,000? 100?
Probability vs Statistics

• Probability
  – **Assume** that variation follows a particular “distribution”
  – Example: average transistor gain is 2.64, variance is 0.11
  – **Assume** that the distribution is known in advance

• Statistics
  – **estimate** the distribution from observed samples
Who is this man?

Answer: “probably” the most famous person in the history of Probability theory
<table>
<thead>
<tr>
<th>concept</th>
<th>explanation</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial or (random)</td>
<td>An action, the outcome of which is not predictable in advance</td>
<td>1. Roll of a dice</td>
</tr>
<tr>
<td>experiment</td>
<td></td>
<td>2. Woman having a mammogram</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. Breakdown time of an insulating fluid</td>
</tr>
<tr>
<td>outcome</td>
<td>Simplest result of an experiment</td>
<td>1. 5 is thrown</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. All clear</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. 0.78 minutes</td>
</tr>
<tr>
<td>Sample space</td>
<td>All the possible outcomes</td>
<td>1. {1, …, 6}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Many …</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. All positive reals, subject to measurement accuracy</td>
</tr>
<tr>
<td>event</td>
<td>A set of one or more outcomes</td>
<td>1. A number less than 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. A benign mass</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. Less than 5 mins</td>
</tr>
<tr>
<td>Event space</td>
<td>All possible events</td>
<td>1. Prime, 3, &lt; 5, …</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Calcs, cancer, clear, …</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. 0.25 mins, &lt; 1 hr, …</td>
</tr>
</tbody>
</table>
Dice throwing events: Venn diagrams

- Outcome is a prime number
- Outcome is an even number
- Outcome is both prime and even, i.e., 2
<table>
<thead>
<tr>
<th>concept</th>
<th>explanation</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A \cap B$</td>
<td>Intersection of the events $A$ and $B$ (those outcomes that are in common)</td>
<td>$P \cap E = {2}$</td>
</tr>
<tr>
<td>$A \cup B$</td>
<td>Union: those outcomes that are in $A$ or $B$ or both</td>
<td>$P \cup E = {2,3,4,5,6}$</td>
</tr>
<tr>
<td>$A \setminus B$</td>
<td>Difference: those outcomes that are in $A$ but not in $B$</td>
<td>$P \setminus E = {3,5}$</td>
</tr>
<tr>
<td>$A \subseteq B$</td>
<td>All the outcomes in $A$ are also in $B$</td>
<td>$P \subseteq \Omega$</td>
</tr>
<tr>
<td>$</td>
<td>A</td>
<td>$</td>
</tr>
<tr>
<td>$A^c$</td>
<td>Complement of the event $A$</td>
<td>$P^c = {1,4,6}$</td>
</tr>
</tbody>
</table>
Combinations of events

\[(A \cup B) \cap C = (A \cap C) \cup (B \cap C)\]
Axioms of probability, $P$

- Aim to assign to each event $E$ a number $P(E)$ so that $0 \leq P(E) \leq 1$
- If $S$ is the sample space, then $P(S) = 1$
- For any two events $E_1$ and $E_2$ for which $E_1 \cap E_2 = \phi$

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Amazingly, any set of events $\{E_1, E_2, \ldots\}$ and any function $P$ that satisfies these three axioms, satisfies the laws of probability
Derived relationships

\[ P(\emptyset) = 0 \]
\[ P(A^c) = 1 - P(A) \]
\[ P(A \setminus B) = P(A) - P(A \cap B) \]
\[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]

These intuitively reasonable relationships can be *proved* rigorously by applying the three rules set out on the previous slide.
Example

For all events $A$, we have $A \cup A^c = S$ and $A \cap A^c = \emptyset$

From these, we have $P(A) + P(A^c) = P(A \cup A^c) = P(S) = 1$

So $P(A^c) = 1 - P(A)$
An urn at a fairground contains 1000 lottery tickets, numbered 1 through 1000

A fairground performer offers to pay £3 to anyone who selects a ticket at random whose number is not divisible by 2, 3, or 5

It costs £2 to have a go, and you lose the stake if the number is not divisible by 2, 3, or 5.

Is it a good idea to have a go?
Denote by $A, B, C$ the events of being divisible by 2, 3, 5

$P(A) = 0.5, P(B) = 0.333, P(C) = 0.2$

$P(A \cap B) = P(\text{divisible by 6}) = 0.166$

Similarly for $P(A \cap C), P(B \cap C), P(A \cap B \cap C)$

Putting it all together:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

$$- P(A \cap B) - P(A \cap C) - P(B \cap C)$$

$$+ P(A \cap B \cap C)$$

$$= 0.734$$

How would the odds change if the lottery ticket was not put back in to the urn? Fairground performers are rarely charitable! Occasionally, they are not good at maths though.
Conditional probability

- Conditional probability
- Combinations & permutations
- Multiplication rule & total probability rule
- Bayes’ Theorem
- Random variables
Conditional probability

In a manufacturing process, 10% parts contain visible surface defects. It turns out that 25% parts with visible surface defect don’t work; however, only 5% of parts without a surface defect don’t work.
Surface flaw $\Rightarrow$ probability of not working = 25%
No surface flaw $\Rightarrow$ probability of not working = 5%

Denote by $D$ event that a part is defective; and by $F$ the event that the part has a surface flaw.

Probability of a part being defective given that it has a surface flaw is denoted $P(D|F)$ and called the *conditional probability of $D$ given $F$*

$$P(D | F) = 0.25$$
$$P(D | F^c) = 0.05$$
Conditional probability $P(B|A)$ of an event $B$ given an event $A$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

Whenever $P(A)>0$
Samples of cast aluminium are classified on the basis of surface finish (microns) & on basis of length measurements. Denote by $A$ the event that a sample has excellent finish; and by $B$ the event that the sample has excellent length. A typical (unbiased) sample of 100 parts:

<table>
<thead>
<tr>
<th>$B$ Excellent length</th>
<th>$B^c$ Less than excellent length</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$ Excellent surface finish</td>
<td>75</td>
</tr>
<tr>
<td>$A^c$ Less than excellent surface finish</td>
<td>10</td>
</tr>
</tbody>
</table>

Determine: $P(A)$, $P(B)$, $P(A|B)$, $P(B|A)$.

Note that $P(A | B) \neq P(B | A)$.
Combinations & permutations

- We often need to calculate (or estimate) the number of outcomes that constitute an event.
- Often, a key discriminant is whether or not a choice made at random is made with/without replacement.

Example: a PC board has but 8 different locations on which a component can be placed. Four components are to be placed. How many designs are possible?

\[ 8 \times 7 \times 6 \times 5 = \frac{8!}{4!} \]
In a room of 25 people, what is the probability that at least two share a birthday (ignoring leap year birthdays)?

Sample space = 365 days.

Evidently, repetition is allowed, for a room with \(N\) people the sample space has size \(365^N\)

Let \(E\) be the event that all the people in the room have different birthdays. The number of these is:

\[
\frac{365 \cdot 364 \cdot 363}{365 \cdot 365 \cdot 365} \ldots = \frac{365!}{(365 - N)!} \cdot \left(\frac{1}{365^N}\right)
\]

\(P(E)\) for \(N = 25\) is

\[
P(E) = \frac{365!}{(365-25)! \cdot 365^{25}} = 0.4313
\]

So probability at least two share a birthday is \(1.0 - 0.4313 = 0.5687\)
What is the probability of being dealt a royal flush (10, J, Q, K, A) in poker?

\[ P(R) = 4 \times \left( \frac{5}{52} \times \frac{4}{51} \times \frac{3}{50} \times \frac{2}{49} \times \frac{1}{48} \right) = \frac{1}{649740} \]

If you have been dealt the Ace of spades as first card, how does the probability change?

\[ P(R \mid \text{Ace of Spades}) = \frac{4}{51} \times \frac{3}{50} \times \frac{2}{49} \times \frac{1}{48} \]
\[ = \frac{4 \times 13}{20} P(R) \]
Multiplication rule

The conditional probability equation \( P(B \mid A) = \frac{P(A \cap B)}{P(A)} \)

Can be rewritten in a form called the multiplication rule:

\[
P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)
\]

The multiplication rule is useful for determining the probability of an event that depends on other events.
$B = (B \cap A) \cup (B \cap A^c)$
Total probability rule

\[ P(B) = P(B \cap A) + P(B \cap A^c) \]

\[ P(B) = P(B \mid A)P(A) + P(B \mid A^c)P(A^c) \]
Surface defect example

In a manufacturing process, 10% parts contain visible surface defects. It turns out that 25% parts with visible surface defect don’t work; however, only 5% of parts without a surface defect don’t work.

\[ P(F') = P(F \mid D)P(D) + P(F \mid D^c)P(D^c) \]

\[ P(F') = 0.25 \times 0.1 + 0.05 \times 0.9 = 0.07 \]
Total probability rule: $n$ mutually exclusive & exhaustive events

$$P(B) = \sum_{i=1}^{n} P(B \cap E_i) P(E_i)$$
Independence

Conditional probability $P(B|A)$ gives the probability of B conditional on prior knowledge of the occurrence of A. Perhaps this prior knowledge doesn’t change the probability of B, so that

$$P(B | A) = P(B)$$

In this case we say that the events A and B are independent
Equivalent statements of independence

\[ P(A \mid B) = P(A) \]
\[ P(B \mid A) = P(B) \]
\[ P(A \cap B) = P(A) * P(B) \]

How would you prove that these were equivalent statements?
Bayes’ Theorem

\[ P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \]
The reverend Thomas Bayes
Born, London 1702
Died, Tunbridge Wells, 1761

Thomas Bayes was ordained, a Nonconformist minister like his father. It seems likely he was educated in maths by De Moivre.

Bayes set out his theory of probability in Essay towards solving a problem in the doctrine of chances published in the Philosophical Transactions of the Royal Society of London in 1764.

Bayes's conclusions were accepted by Laplace in a 1781 memoir, rediscovered by Condorcet (as Laplace mentions), and remained unchallenged until Boole questioned them in the Laws of Thought. Since then Bayes' techniques have been subject to controversy.
Why is Bayes’ Theorem so important?

• **Example from pattern classification**
  – Trying to recognise hand-printed alphabetic characters
  – Establish a set of *classes* corresponding to instances of A, B, C, …
  – Presented with an image $I$
  – Our task is to assign it to a pattern class
JMB
THE CAT
Given the image $I$ we want to compute the class $k$ for which $P(C_k | I)$ is highest. Bayes’ Theorem enables us to calculate this by:

$$P(C_k | I) = \frac{P(I | C_k)P(C_k)}{P(I)}$$

Given a model for each character, this term gives the probability of this image given that we know the identity of the character. The normalisation factor $P(I)$ ensures that the probabilities sum to 1.
$$P(C_k \mid I) = \frac{P(I \mid C_k)P(C_k)}{P(I)}$$

*Posterior probability* since it determines the class *given* the image

*Class conditional probability* of the image assuming that the class is known

*Prior probability* for the class
Application of Bayes’ Theorem: segmenting brain structures from MRI images

Original image

Cerebrospinal fluid  Gray matter  White matter
Original noisy MRI image of the brain – two slices

Classification of brain tissue by an algorithm developed by Chen Xiao Hua and Michael Brady, 2005

“ground truth” = the classifications laboriously done by a clinical specialist
Proton density MRI image of the brain

T₂ weighted image of the same brain

Clinician’s classification of the brain tissue

Classification by the algorithm of Chen and Brady, 2005
Random variables

- Measure the output from a device every day:
  \[ x_1 = 2.45, x_2 = 2.43, x_3 = 2.39, \ldots \]
- A *random variable* \( X \) has instances \( x_i \)
- The *range* of \( X \) is the set of possible values that the random variable may assume
Discrete random variable

• Finite or countably infinite range

• Examples
  – number of bits transmitted
  – Number of scratches on a surface
  – Proportion of defective parts in a batch of 1000
  – Number of women recalled for further tests after breast screening
Continuous random variable

- Range comprises real numbers, perhaps an interval $[0.72, 1.36]$
- Measurements are always finite resolution; but it is often convenient to pretend that we have infinite precision
- Examples
  - Physical parameters: time, pressure, voltage, …
Continuous to discrete

• May be convenient to replace a continuous random variable by a discrete one

• Example
  – A shoe company may reasonably decide to offer only a finite set of shoe sizes: 7, 7 ½, 8, ..
Discrete random variables

- Probability distribution, probability mass function, cumulative distribution
- Mean (or expected) value
- Variance
- Examples
  - Uniform
  - Binomial
  - Poisson
  - Geometric
Probability distribution

There is a chance that a bit transmitted through a digital transmission line is received in error.

The discrete random variable $X$ is the number of bits in error in the next four bits transmitted. $\text{Range of } X = \{0, 1, 2, 3, 4\}$

The probability distribution of a random variable $X$ is a representation of the probabilities associated with the possible values of $X$.

\[
P(X = 0) = 0.6561 \\
P(X = 1) = 0.2916 \\
P(X = 2) = 0.0486 \\
P(X = 3) = 0.0036 \\
P(X = 4) = 0.0001
\]
Graphical representation of the probability distribution
For a discrete random variable $X$, the probability mass function $f$ is defined by

$$f(x_i) = P(X = x_i)$$

Evidently:

$$f(x_i) \geq 0$$

$$\sum_i f(x_i) = 1$$
Example

X denotes the number of semiconductor wafers that need to be analysed in order to detect a large particle, indicative of contamination.

Contamination is relatively rare: prob of occurrence is 0.01

Wafers are generated separately, so can be considered independent events.

What is the probability distribution of X?

Let p denote a wafer that has a large particle, a the absence.

Sample space \{p, ap, aap, aaap, \ldots\}

\[
P(X = 1) = P(p) = 0.01
\]

\[
P(X = 2) = P(ap) = 0.99 \times 0.01
\]

\[
P(X = 3) = P(aap) = 0.99^2 \times 0.01
\]
Shape of probability distribution

A single peak, **monomodal**

Distribution is approximately symmetric about the peak

The rate at which it does so is called the spread of the distribution

There are numerous mathematical formalisations of this generic shape: Binomial, Poisson, and normal (Gaussian).

The law of large numbers says why this shape predominates
Cumulative distribution

Given a discrete random variable $X$, like the contaminated wafers, we may be interested in $P(X \leq 3)$. Of course, this is the union of the events $X=0$, $X=1$, …

More generally, for a discrete random variable with ordered values $\{x_1, x_2, \ldots\}$ we are often interested in

$$P(X \leq x) = \sum_{x_i \leq x} f(x_i)$$

This is called the cumulative distribution $F(X)$.
Simple properties of the cumulative distribution

\[ 0 \leq F(x) \leq 1 \]

\[ x \leq y \Rightarrow F(x) \leq F(y) \]
Mean of a discrete random variable

The discrete random variable $X$ denotes the number of messages sent over a computer network per second.

<table>
<thead>
<tr>
<th>$x$ = number of messages per second</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(x)$</td>
<td>0.08</td>
<td>0.15</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.07</td>
</tr>
</tbody>
</table>

How many messages “on the average” do we expect to send?

$10 \times 0.08 + 11 \times 0.15 + 12 \times 0.3 + 13 \times 0.2 + 14 \times 0.2 + 15 \times 0.07$
Mean of a discrete random variable

if \( X \) is a discrete random variable with values \( \{ x_1, x_2, \ldots \} \)

The mean or expected value of \( X \) is given by

\[
E(X) = \sum_{x} x f(x)
\]
Variance of a discrete random variable

$$\sigma^2 = V(X) = \sum_x (x - \mu)^2 f(x)$$

It is easy to show that

$$V(X) = \sum_x x^2 f(x) - \mu^2$$

$$V(X) = E((x - \mu)^2)$$
Uniform distribution

The simplest discrete random variable, typified by:

- fair dice
- deck of cards
- number of 48 voice lines that is in use at an exchange at any particular time

all values in the range of $X$ are equally likely. If the range is $\{x_1, x_2, \ldots\}$ then the probability of each possible value is given by $f(x_i) = \frac{1}{n}$

If the range of $X$ is consecutive integers $\{a, \ldots, b\}$

The mean and variance are:

$$E(X) = \frac{(b + a)}{2} \quad \quad V(X) = \frac{(b - a + 1)^2 - 1}{12}$$
In practice, distributions are often *approximated* as sums of uniform distributions.

The red distribution is the one observed for a particular brain. The three uniform distributions shown are those for which the sum best approximates the real red distribution.
Binomial distribution

The Binomial distribution arises from a block of trials each of which has one of two outcomes success and failure (or, up and down, heads and tails, foo and baz, whatever).

Such a trial is labelled a Binomial (or Bernoulli) trial.

It is assumed that any pair of trials in the experiment are independent of each other.

It is reasonable to assume that the probability of success in each trial is constant throughout the sequence of trials.

Suppose that there are $n$ trials (this, together with $p$, is the characteristic parameter for the Binomial distribution). For an $n$-trial experiment, the sample space $x$ of $X$ may assume values $x = 0, 1, \ldots, n$
Binomial distribution

Suppose that there are $x$ successes in $n$ trials. The number of ways this can happen is $\binom{n}{x}$.

Each of these has probability $p^x (1 - p)^{n-x}$.

The binomial distribution is defined by

$$f_{\text{binomial}}(x) = \binom{n}{x} p^x (1 - p)^{n-x}, x = 0, 1, \ldots, n$$
Shape of the Binomial Distribution

Bin(n,p) depends only on the two parameters n, p.

For fixed n, the Binomial distribution becomes more symmetric as p increases from 0 to 0.5, or decreases from 1 to 0.5;

For fixed p, the Binomial distribution becomes more symmetric as n increases

The mean of Bin(n,p) is equal to np

V(Bin(n,p)) = np(1-p)
Binomial distribution, $n = 100, p = 0.5$
Binomial distribution, n = 100, p = 0.8
Geometric distribution

concerns the case where Bernoulli trials are continued until the first success.

The mass distribution function $f(x)$ for success on the $x$ trial after $(x-1)$ failures is

$$f(x) = (1 - p)^{(x-1)} p, \quad x \geq 1$$

$$E(X) = \frac{1}{p}$$

$$V(X) = \frac{(1 - p)}{p^2}$$
**Poisson distribution**

Consider the transmission of $n$ bits over a digital communication channel. Suppose that the discrete random variable $X$ equals the number of bits transmitted incorrectly.

If we assume that the probability of a bit being in error remains constant over time and that the outcomes of each being transmitted (correctly or in error) are independent of each other, then $X$ has a Binomial distribution, with mean $= np$.

Now suppose instead that improvements in technology mean that we can transmit far more bits $n^*$ because the probability $p^*$ of an error in transmission decreases.

We assume that the improvement in technology is such that $np = n^*p^*$ - that is $n$ increases and $p$ decreases but in such a way that the mean value $np$ remains constant. Call this constant $\lambda$. We seek the probability mass function in this case.
Poisson distribution

\[ f_{\text{binomial}}(x) = \binom{n}{x} p^x (1 - p)^{(n-x)} = \binom{n}{x} \left( \frac{\lambda}{n} \right)^x (1 - \left( \frac{\lambda}{n} \right))^{(n-x)} \]

It can be shown that

\[ \lim_{x \to \infty} f(x) = \frac{e^{-\lambda} \lambda^x}{x!} \]

And this defines the distribution \textit{Poisson}(\lambda)

\[ E(X) = V(X) = \lambda \]
Optical character recognition

$X$ is the number of incorrectly identified characters.

Typical number: 2.3 per 10000 characters.

The probability is very low and the number of characters to be scanned very large, so it is better to use a Poisson distribution to model the probability distribution.

$p = \frac{2.3}{10000}, n = 10000$ so $\lambda = 2.3$

Probability that there are 2 incorrectly scanned characters:

$$P(X = 2) = \frac{e^{-2.3} \cdot 2.3^2}{2!} = 0.265$$
Probability that there is at least one flaw in 20000 characters scanned

\[ P(X \geq 1) = 1 - P(X = 0) \]

In this case

\[ \lambda = \frac{2.3}{10000} \times 20000 = 4.6 \]

so

\[ P(X \geq 1) = 1 - \frac{e^{-4.6} \times 4.6^0}{0!} = 0.9899 \]
Continuous random variables

The continuous analogue of a probability mass function is the **probability density function (pdf)** $f(x)$ where $x$ is a value of a continuous random variable $X$.

The figure shows a typical pdf, for which the value of $f(x)$ is non-zero only on an interval. This is not always the case.
Probability density function

Any continuous function $f(x)$ that satisfies:

$$f(x) \geq 0, \forall x$$

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

The area under the curve is:

$$P(a \leq x \leq b) = \int_{a}^{b} f(x) \, dx$$
Introduce discrete random variable $Y$, where $y_n = [x_n, x_{n+1}]$ and let the integral of the corresponding rectangle be $g_n$.

Divide the continuous range of $x$ into contiguous intervals:

$\{[x_1, x_2], [x_2, x_3], \ldots, [x_n, x_{n+1}], \ldots\}$
Cumulative distribution function

This is defined by

\[ F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\,du, -\infty < x < \infty \]
Example: Cumulative dist. fn.

The time until a chemical reaction is complete (msec) is approximated by the cumulative distribution function:

\[ F(x) = \begin{cases} 
0, & x < 0 \\
1 - e^{-0.01x}, & x \geq 0 
\end{cases} \]

The pdf for X is obtained by differentiating the expression for F:

\[ f(x) = \begin{cases} 
0, & x < 0 \\
0.01e^{-0.01x}, & 0 \leq x 
\end{cases} \]

The probability that a reaction completes within 200msec is:

\[ P(X < 200) = F(200) = 1 - e^{-2} = 0.8647 \]
Mean and variance of a continuous random variable

Mean

\[ E(X) = \int_{-\infty}^{\infty} x f(x) \, dx \]

Variance

\[ V(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x) \, dx \]

\[ V(X) = \int_{-\infty}^{\infty} x^2 f(x) \, dx - E(X)^2 \]
Uniform distribution

Suppose that the range of a continuous random variable $X$ is the finite interval $[a,b]$. The pdf is given by:

$$f(x) = \frac{1}{b-a}$$

**mean** $E(X) = \frac{(a + b)}{2}$

**variance** $V(X) = \frac{(b-a)^2}{12}$

Example: $X$ denotes the current measured in a thin copper wire (mA), with the range of $X$ $[0,20\text{mA}]$
Normal or *Gaussian* distribution

The most widely used (and abused) mathematical model for the pdf of a continuous random variable is the *normal* distribution, the familiar bell-shaped curve:

1. This distribution occurs approximately very often in practice
2. Many other distributions are well-approximated as weighted sums of normal distributions
Equation for the normal distribution

\[ f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \]

- Mean: \( \mu \)
- Variance: \( \sigma^2 \)

Note that \( f(x) > 0 \) for all finite values of \( x \). The fact that \( f(x) \) never reaches zero, causes some difficulties in signal processing. There are a variety of practical solutions in signal filtering.
Percentiles of the normal distribution

\[ P(\mu - \sigma < X < \mu + \sigma) = 0.6827 \]
\[ P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9545 \]
\[ P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973 \]

Note that more than 95% of values lie within 2 s.d.s of the mean, corresponding to < 2.5% in each tail.

Manufacturing aims at 6\( \sigma \)
Suppose that we have a general continuous random variable X whose mean is µ and whose variance is σ². Perform a long sequence of experiments, each of which is independent of the others, and each of which has the same random variable X. That is we have a sequence of rvs: \( X_1, X_2, \ldots, X_n, \ldots \)

After \( n \) experiments, the sample mean is given by: 
\[
\overline{X}_n = \frac{X_1 + \ldots + X_n}{n}
\]

This can also be thought of as a random variable and it is easy to show that 
\[
E(\overline{X}_n) = \mu \quad \text{and} \quad V(\overline{X}_n) = \frac{\sigma^2}{n}
\]

DeMoivre (1733) showed that as \( n \) increases, the pdf of \( \overline{X}_n^n \) rapidly approximates a normal distribution.
Standardising the normal distribution: the $z$ statistic

There are tables for the normal distribution (e.g. HLT).

If $X$ is normally distributed, we may want to know $P(X<27.34)$.

Evidently this depends upon the *particular* normal distribution, that is to say on its parameters $(\mu, \sigma)$.

Instead of printing an infinite number of tables, one per $(\mu, \sigma)$, we convert $X(\mu, \sigma)$ to a *standard form* $Z(0,1)$, whose mean is 0 and whose variance is 1. This is easily done:

$$Z = \frac{(x - \mu)}{\sigma}$$
Example

The current measurements in a strip of wire follow a normal distribution with a mean of 10mA and variance of 4mA$^2$. What is the probability that a measurement will exceed 13mA.

Let $X$ denote the current, so we seek $P(X>13)$.

Standardising: $Z = (X-10)/2$.

$X > 13$ corresponds to $Z>1.5$

$$P(X > 13) = P(Z > 1.5)$$
$$= 1 - P(Z \leq 1.5)$$
$$= 1 - 0.993319$$
$$= 0.06681$$
Example: meeting specifications

The diameter of a shaft of an optical disk drive is normally distributed, with mean 0.2508cm, standard deviation 0.0005cm.

The specifications on the shaft are 0.2500±0.0015cm.

What proportion of shafts conform to the specifications? That is, find the proportion of shafts no larger than 0.2515cm, no smaller than 0.2485cm.

\[
P(0.2485 < X < 0.2515) = P\left(\frac{0.2485 - 0.2508}{0.0005} < Z < \frac{0.2515 - 0.2508}{0.0005}\right)
\]

\[
= P(-4.6 < Z < 1.4)
\]

\[
= P(Z < 1.4) - P(Z < -4.6)
\]

\[
= 0.91924 - 0.0000
\]

\[
= 0.91924
\]

Most nonconforming shafts are at the upper end of the range. Reset the mfr process so that the mean is equal to 0.25cm!
Binomial and Normal distributions

Sometimes the Binomial distribution looks quite like a normal distribution. Not always, though.

Suppose $X(n,p)$ is a binomial distribution, so that the mean is $np$ and variance is $np(1-p)$. Then

$$Y = \frac{X - np}{\sqrt{np(1-p)}}$$

Is approximately normally distributed. The approximation is particularly good for $np > 5$ and $np(1-p) > 5$.

Similarly for a Poisson distribution $X(\lambda)$:

$$Y = \frac{X - \lambda}{\sqrt{\lambda}}$$
Two* random variables

Suppose that we have two random variables X and Y, for example a person’s height and the person’s weight.

We don’t expect them to be independent, rather that there be some kind of relation between them, or that the one depends on the other, for example: taller ⇒ heavier (mostly).

For some X,Y (for example: X = number of cousins, Y = £/$ exchange rate, you might expect no/very weak dependence

Note that, for a given height (say), we would expect a distribution of weights, and that this distribution would depend on the height. So, people who are 195cm (6’5’’) would, on the average, be expected to be heavier than people who are 155cm (5’2’’). Of course, some people who are 195cm are tall and skinny, others are just huge! *or more
Different weight distributions for different heights

For people height 155cm, we observe a mean weight of 50Kg, and a standard deviation of 5Kg.

People of height 195cm have a mean weight of 90Kg, and a larger standard deviation of 10Kg.
Two random variables and marginal distributions

If we have two variables $X$ and $Y$, then we can define the joint probability

$$p(x, y) = P(X = x, Y = y)$$

The probability mass function $p_X$ for the variable $X$ is defined by

$$p_X(x) = \sum_y p(x, y)$$

Similarly, the probability mass function $p_Y$ for the variable $Y$ is defined by

$$p_Y(x) = \sum_x p(x, y)$$
Selecting balls from a bag

Suppose that 3 balls are randomly selected from a bag containing 3 red, 4 white, and 5 blue balls. Denote by $R$, $W$ the number of red and white balls picked out. Then the joint probability mass function of $R$, $W$ is given by:

$$
p(0,0) = \frac{\binom{5}{3}\binom{3}{3}}{\binom{12}{3}} = \frac{10}{220}, \text{ since all three balls chosen are blue}
$$

$$
p(1,0) = \frac{\binom{3}{1}\binom{5}{2}}{\binom{12}{3}} = \frac{30}{220}, \quad p(0,1) = \frac{\binom{4}{1}\binom{5}{2}}{\binom{12}{3}} = \frac{40}{220}, \text{ etc...}
$$
\[ P(R=r, W=w) \]

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Row sum = ( P(R=r) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td>40</td>
<td>30</td>
<td>4</td>
<td>84</td>
</tr>
<tr>
<td>220</td>
<td>220</td>
<td>220</td>
<td>220</td>
<td>220</td>
<td>220</td>
</tr>
<tr>
<td>220</td>
<td>30</td>
<td>18</td>
<td>0</td>
<td>60</td>
<td>108</td>
</tr>
<tr>
<td>220</td>
<td>220</td>
<td>220</td>
<td>220</td>
<td>220</td>
<td>220</td>
</tr>
<tr>
<td>220</td>
<td>15</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>220</td>
<td>220</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>220</td>
</tr>
<tr>
<td>220</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ \text{Column sum} \]

\[
\begin{array}{cccc}
56 & 112 & 48 & 4 \\
220 & 220 & 220 & 220 \\
\end{array}
\]
Independence of $X, Y$

$X, Y$ are independent when

$$p(x, y) = p_X(x)p_Y(y)$$

Expectation of the product $XY$:

$$E(XY) = \sum_{x, y} xy p(x, y)$$

$X, Y$ are independent $\iff E(XY) = E(X)E(Y)$
Covariance of $X, Y$

Recall that the Variance of a variable $X$ is defined by:

$$Var(X) = E[(X - \mu_X)^2]$$
$$= E[(X - \mu_X)(X - \mu_X)]$$

In like manner, the Covariance of $X, Y$, denoted $Cov(X,Y)$ is defined by:

$$Cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

$$Cov(X,Y) = E[XY - \mu_X Y - X \mu_Y + \mu_X \mu_Y]$$
$$= E[XY] - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y$$
$$= E[XY] - \mu_X \mu_Y$$

It is easy to show that:

$$Cov(X,Y) = Cov(Y,X)$$
Covariance matrix

\[
\begin{pmatrix}
\text{Cov}(X, X) & \text{Cov}(Y, X) \\
\text{Cov}(X, Y) & \text{Cov}(Y, Y)
\end{pmatrix}
\]

The diagonal terms are the usual variances.

Since the off-diagonal terms are equal, the matrix is symmetric, hence has real eigenvalues.
The ellipse is a pictorial representation of the 95% probability of acceptance of a point allegedly in the distribution.
Kalman filtering

An environment; but it’s foggy and our range sensor has limited range

A planned trajectory

An actual trajectory

The ellipses represent the covariance matrix