1) Basic Probability Theory

Sample Space – Set of possible outcomes of a random experiment. Usually denoted with set notation, can be finite, countably or uncountably infinite. e.g. Coin Toss, S={H, T}, 2 Coin Tosses, S={(H, H),(H, T),(T, H),(T, T)}
Choice of Odd number S = {x ∈ N|∃y ∈ N.[2y + 1 = x]}
Event – a subset of the sample space. It is the collection of **some** of outcomes e.g. Coin Tossing E = {H}, Even Dice Roll E = {2, 4, 6}.
Extreme events (∅) never happen) - empty set
Event S always happens as it is the entire sample space.

1.1) Probability – if sample space S is Finite or Countable, we can assign probabilities. If uncountably infinite, we cannot have the probabilities reasonably sum to 1. Thus when defining a probability function on S, we define the collection of subsets we'll measure as **F**.
F has the following properties:
1. Nonempty 2. Closed under complement
3. Closed under countable union.
A collection of sets is known as a σ-algebra
Probability Measure: A function P : F → [0, 1] on the pair (S, F) such that:
1. ∀E ∈ F.[0 ≤ P(E) ≤ 1] 2. P(S) = 1
3. C$P(\bigcup_i E_i) = \sum_i P(E_i)$ sets $E_1, E_2, \dots \in F$:

From this we derive on a probability measure:
1. P(E) = 1 – P(E) 2. P(∅) = 0
3. Cou$P(\bigcup_i E_i) = \sum_i P(E_i)$ sets $E_1, E_2, \dots \in F$:

1.2) Probability Interpretations

1) Classical: P(E) = |E| / |S|
2) Frequentist: Through repeated observations of identical random experiments in which E can occur, the proportion of experiments where E occurs tends towards the probability of E. At an infinite number experiments, the proportion of occurrences of E is equal to P(E).
3) Subjective: Probability is the degree of belief held by the individual.

1.3) Joint Events

Joint Events: events E and F that occur at the same time. AKA the **and event E ∩ F.**
Independent:
Two events are independent if P(E ∩ F)=P(E)P(F).
This can be extended to n events. The events are **dependent** if this doesn't hold.
Propositions:
1) If events E and F are independent, then !E and F are also independent. Easily provable with set algebra.
2) P(E ∪ F) = P(E) + P(F) – P(E ∩ F).
We can solve these problems using tables quite easily.

1.4) Conditio

Definition of Conditional Probability: $P(E|F) = \frac{P(E \cap F)}{P(F)}$
P(E|F) = P(E) if E and F are independent which makes sense and is easily proved with set algebra manipulation.
Conditional Independence: P(•|F) defines probability measure obeying the axioms of probability on set F (When have just reduced S to F)
Three events E1, E2, F are **conditionally independent** if and only if: P(E1 ∩ E2|F) = P(E1|F) × P(E2|F)
I
$$P(E|F) = \frac{P(E) \times P(F|E)}{P(F)}$$

Partition Rule: (The Law of Total Probability):
Consider a set of events $\{F_1, F_2, \dots\}$ which form a partition of S. Then for any event E ⊆ S, the **Law of**
$$P(E) = \sum_i P(E|F_i)P(F_i).$$
This makes complete sense! The probability of E is of course the sum of the probability of each event occurring, and then E occurring given them.
Remember: Probabilities of the form P(E | F) are conditional probabilities
Probabilities of the form P(E ∩ F) are joint probabilities.
Probabilities of the form P(E) as marginal probabilities

2) Random Variables

Probability Space: (S, F, P) Models a random experiment where probability measure P(E) is defined on subsets E ⊆ S belonging to sigma algebra F.
Random Variable: Random variable is a mapping from the sample space to the reals, e.g: X: X : S → R
Each element in the sample space s ∈ S is assigned to a numerical value by X(s). When referring to the value of a random variable we use its name, e.g X in P(5 < X ≤ 30)
Simple RV: Finite set of possible outcomes. (dice faces)
Discrete RV: Countable outcomes. (distance (m))
Continuous RV: Can be a continuous range (temp)

M1) Example Discrete Random Variable

S = {1,2,3,4,5,6}, for any s ∈ S.P({s}) = 1/6. We can define an RV st: X(1) = 1, X(2) = 2 ... X(6) = 6. Then we can use X: $P_x(1 < X <= 5)$ = P({2,3,4,5}) = 2/3. We can also define a random variable Y, Y(e) = 0 if e is odd,

2) Induced Probability

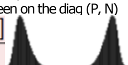
Induced Probability: The probability measure P defined on a sample space S induces (creates) a probability distribution on the rand var in R (distribution of its outcomes): S_x = {s ∈ S|X(s) ≤ x}.
$P_x(X ≥ x) = P(S_x)$. $P_x(\dots)$ is often written as P(...).
Example: We define random variable X : {H, T} → R over the continuum R such that: X(T) = 0 and X(H) = 1.
$$S_X = \begin{cases}\emptyset & \text{if } x<0\\ \{T\} & \text{if } 0\le x<1 \\ \{H,T\} & \text{if } x \ge 1\end{cases}$$
X represents the number of heads flipped.
$$P_X(X \le x) = P(S_X) = \begin{cases} P(\emptyset)=0 & \text{if } x<0 \\ P(\{T\}) = 1/2 & \text{if } 0 \le x < 1 \\ P(\{H,T\}) = 1 & \text{if } x \ge 1\end{cases}$$
Now we can use x to compactly show some probabilities: $P_x(X=1)$ = ½. Overall, Induced Probability just refers to the creation of a Probability Dsitribution on a sample space for some probability of some event we want to measure.
Support (Range): The set of all possible values of a random variable X.
supp(X) = X(S) = {x ∈ R|∃s ∈ S.X(s) = x}
$P_x(X ≤ x)$ is defined for all x ∈ supp(X)

2.1) Cumulative Distribution Functions

The CDF of a random variable X is the probability that X takes some value less than or equal to some x:
F_x : R → [0, 1] such that $F_x(x) = P_x(X ≤ x)$
To be a valid CDF:
1) **Probability must be between 0 and 1:** ∀x ∈ R.0 ≤ FX(x) ≤ 1
2) **Monotonicity:** $\forall x_1, x_2 \in R x_1 < x_2 \Rightarrow F_x(x_1) \le F_x(x_2)$
3) **Infinite Bounds:** $F_x(-\infty) = 0$, $F_x(\infty) = 1$
Thus CDFs are right continuous. We can determine the probability over finite intervals using the cumulative distribution: for (a, b] ⊆ R $P_x(a < X ≤ b) = F_x(b) – F_x(a)$

2.2) Probability Mass Functions:

Gives probability that a DRV is exactly equal to its value. The sample space of S is mapped onto elements in the support of X. We can then partition the sample space into a countable, disjoint collection of event subsets:
$s \in E_i \Leftrightarrow X(s) = x_i$, i = 1, 2...
A PMF is valid only if:
1) No Negative Probabilities: ∀x ∈ supp(X).$p_x(x) ≥ 0$
2) Probabilities sum to 1: $\sum_{x\in supp(x)} p_X(x) = 1$

Expectation: The mean of the distribution X.
E(X) = Σ_x x p(x)
E(g(X)) = Σ_x g(x)p(x)
E(aX+b) = aE(X) + b
E(g(X) + h(X)) = E(g(X)) + E(h(X))
Variance: Measure of spreadness of values X can take
Var(X) = E[(X – E(X))²] ⇒ Var(X) = E(X²) – E(X)².
Var(aX + b) = a²Var(X).
Standard Deviation = root Var.
Skewness: Measure of asymmetry of a distribution:
Can be positive or negative as seen on the diag (P, N)
$$\gamma_1 = \frac{E[(X - E(X))^3]}{sd(X)^3}$$


2.3) Discrete Random Variables

For a DRV, we define the PMF as: $p_x(x_i) = P(X=x_i) = P(E_i)$ where x_i ∈ supp(X), x_i is the outcome of event E_i.
We can define it in terms of PMFs or CDFs:
$p_x(x_i) = P_x(X=x_i) = P(X≤x_i) - P(X≤x_{i−1}) = F_x(x_i) - F_x(x_{i−1})$
Discrete CDFs have the follow properties:
1) Limiting Cases: $\lim_{x\to-\infty} F_x(x) = 0$ $\lim_{x\to\infty} F_x(x) = 1$
2) Continuous from right:
For x ∈ R $\lim_{h\to 0+} F_x(x + h) = F_x(x)$
3) Non-Decreasing: a < b ⇒ $F_x(a) ≤ F_x(b)$
4) Covers a range: For a < b. P(a<X≤ b)=FX(b)-FX(a)

2.4) Combining Random Variables

Let $X_1, X_2, \dots, X_n$ be n random variables with diff distribution and not necessarily independent: Let $S_n = \Sigma^n_{i=1}X_i$, and S_n / n be their average.
$E(S_n) = \Sigma^n_{i=1}E(X_i)$, $E(S_n/n) = E(S_n) / n$.
$Var(S_n) = \Sigma^n_{i=1}Var(X_i)$, $Var(S_n / n) = Var(S_n) / n^2$.
Combining IID Distributions: If $X_1, X_2, \dots, X_n$ are IID with $E(X_i) = u_x$, $Var(X_i) = \sigma^2_x$
$E(S_n / n) = u_x$, $Var(S_n / n) = \sigma^2_x / n$

2.5) Probability Distributions

1) Bernoulli Distribution: Basically a binomial but only 1 trial. It models an experiment with two outcomes, a random variable X takes values 1 with p or 0 with (1-p).
X ~ Bernouilli(p), pmf is $p(x) = p^x(1-p)^{1-x}$, x = 0, 1
μ = p, σ² = Var(X) = p(1 – p)
2) Binomial Distribution: Given n trials with two options, binomial models the number of outcomes. (e.g 3 tosses, num of ways for 2 heads from total outcomes).
X ~ Binomial(n, p) where X takes values 0, 1, 2,...,n and 0 ≤ p ≤ 1:
PMF: $P_x(x) = \binom{n}{x}p^x(1-p)^{n-x}$
E(X) = np — Note that choice is: $\binom{n}{x} = \frac{n!}{x!(n-x)!}$
Var(X) = np(1-p)
Skewness = Skewness $\gamma_1 = \frac{1-2p}{\sqrt{np(1-p)}}$

2.5.3) Poisson Distribution

Given a constant mean number of events per fixed item interval, provides probabilities of different numbers of events occurring. (e.g we find avg 6p an hour, what is probability that we find 10p in a given hour)
PMF: $P_x(x) = \frac{e^{-\lambda}\lambda^x}{x!}$
E(X) = λ
Var(x) = E(X) = λ
Skewness = 1 / (λ)^{1/2} (always positive).

4) Geometric Distribution

A potentially infinite number of trials to get an outcome (attempts required to shoot a target, given probability of hit). We can consider it infinite Bernoulli trials $X_1, X_2, \dots$, where X = {i|X_i = 1} (X is number of attempts to get outcome 1).
For X ~ Geometric(p) where X takes all values in Z⁺ = {1, 2, ...} and 0 ≤ p ≤ 1:
PMF = $p_x(x) = p(1-p)^{x-1}$
u = E(x) = 1/p
Var(x) = 1-p / p²
Skewness = 2 – p / (1-p)^{1/2}
We can also consider the number of trials **before** we get an outcome too:
Y = X – 1 takes values N ={0,1,2,...}:
PMF = $p(1-p)^Y$
U = E(Y) = 1-p / p
Variance and Skewness are unchanged.

5) The Discrete Uniform Distribution

Where a discrete number of outcomes are equally likely (e.g fair dice).
For X ~ U({1, 2, ... , n}):
PMF = $p_x(x)$ = 1/n
u = E(x) = n+1 / 2
Var(x) = n²-1 / 12
Skewness = 0

2.6) Poisson Limit Theorem:

We can use the **Binomial Distribution to approximate the Poisson Distribution**.
Poisson(λ) ≈ Binomial(n, p) when λ = np and n is very large, p is very small
Explanation:
This is for a Poisson distribution mean and variance are equal and for binomial, mean is np, variance np(1−p) so as p gets smaller (and n larger) np ≈ np(1 − p).

3) Continuous Random Variables

For continuous random variables we want to track quantities in R (e.g temperature, volume).

3.1) Probability Density Function:

For a random variable X : S → R the induced probability is defined as: $P_X((-\infty, x]) = P(S_X) = F_X(x)$
A variable x is absolutely continuous if
$\exists f_X$: R → R such that: $F_x(x) =$
$$\int_{u=-\infty}^{x} f_X(u)du$$
$f_x(x) = F'(x) = d/dx\, F_x(x)$. $f_x(x)$ is a probability density function, and F_x is the CDF.
To find the probability X ∈ (a, b]:
$P_x(a<X≤b) = P_x(X≤b) – P_x(X≤a) = F_x(b) – F_x(a)$ = the integral of $f_x(x)$ wrt x from a to b.
Notes:
1) We can use < and <= interchangeably as the probability of a specific event P(X=x) = 0 ⇔ P(X <= x) = P(X < x). 2) The sum over a range != 0.
3) Hence the range of a CRV is uncountable.
The integral from infinity to minus infinity is 1.

3.2) Mean, Variance and Quantiles

The **mean** of a CRV X:
$U_x = E_x(X) = \int_{-\infty}^{\infty} x f_X(x)dx$
$E_x(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)dx$
E(aX + b) = aE(X) + b
E(g(X) + h(X)) = E(g(X)) + E(h(X))
The **Variance** of a CRV X:
$Var_x(X) = E(X^2) – (E(X))^2$
Var(aX+b) = a_2Var(X)
Quartiles of a CRV X:
To find LQ, median, UQ, or the nth percentile just integrate the pdf for 0.25, 0.5, 0.75 or n/100 respectively (or use the CDF).

3.3) Notable Continuous Distributions

1) The Uniform Distribution: X ~ U(a, b)
PDF: $f_x(x)$ =1 / (b-a)
CDF: 0 for x <= a, x-a/b-a for a < x < b, 1 for x >= b
E(x) = u = a + b / 2
Variance = (b-a)² / 12
The standard uniform distribution is X~U(0,1).

2) Exponential Distribution

Given a rate of events λ, what is the probability of waiting X time for the event to occur.
PDF: $f_x(x) = \lambda e^{-\lambda x}$
CDF: $F_x(x) = 1 - e^{-\lambda x}$, where x >= 0
E(x) = u = 1 / λ
Variance = 1 / λ²
This distribution is **memoryless** – the time waited already does not affect the future behaviour of the distribution.

Given X ~ Poisson(λ) the time between events is modelled by X ~ Exp(λ) (interval time for one event).

There is a variant with Exp(θ), θ = 1/λ.
3) Normal Distribution: A symmetric distribution with a mean value u and variance σ². For X ~ Normal(μ, σ²) or X ~ N(μ, σ²) where σ > 0:
PDF:
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$
CDF:
$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\}dt$$
The standard normal distribution is X ~ N(0, 1). It has PDF and CDF:
$$\phi(x) = \frac{1}{\sqrt{2\pi}}exp\left\{-\frac{1}{2}x^2\right\} \quad \Big| \quad \Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt$$

M2) Convert to Standard Normal

X ~ N(u, σ²), aX + b ~ N(aμ + b, a²σ²)
So, X~N(u, σ²) ➔ X – u / σ ~ N(0,1)
and so P(X <= x) = Φ(x-u / σ)
So we can use P(X <= x) = (x-mean) / standard deviation.
Example: P(X <= 7), in X~N(4, 2²) = 0.9332 = P(X <= x) = (7-4) / 2 = Φ(1.5) = 0.9332

4) Lognormal Distribution

X ~ N(μ, σ²) and Y = eˣ we
$f_y(y) = \frac{1}{\sigma y\sqrt{2\pi}}exp\left[-\frac{(\log y-\mu)^2}{2\sigma^2}\right]$

3.4) Central Limit Theorem

A Moment Generating Function M_x for a Continuous Random Variable X is:
$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx}f_X(x)dx$$
The nth moment is given by the following, assuming our integral is valid:
$$E[X^n] = \left.\frac{d^nM_x(t)}{dt^n}\right|_{t=0}$$
MGFs provide a way of generating properties about our distributions such as the Expectation, Variance (these properties are known as ***moments***). Moments are important because they provide a way to summarize the properties of a probability distribution and compare different distributions.

Example: Deriving E(x) and Var

$$E[X] = \left.\frac{dM_x(t)}{dt}\right|_{t=0} = \left.\frac{dE[e^{tX}]}{dt}\right|_{t=0} = \left.\frac{d\int_{-\infty}^{\infty}e^{tx}f_X(x)dx}{dt}\right|_{t=0} = \left.\int_{-\infty}^{\infty}xe^{tx}f_X(x)dx\right|_{t=0} = \int_{-\infty}^{\infty}xe^{0x}f_X(x)dx = \int_{-\infty}^{\infty}xf_X(x)dx$$
For E[X²] we start with $\left.\frac{d^2M_x(t)}{dt^2}\right|_{t=0}$
We end up applying the differentiation step twice and end with $x^2f_x(x)$ dx in the integral instead.
Var[X] = E[X²] – (E[X])²

3.5) Product of Random Variables

Given independent random variables $Z_1, Z_2 \dots, Z_n$:
$$E[\prod_{i=1}^{n} Z_i] = \prod_{i=1}^{n} E[Z_i]$$
The sum of the random variables is the products of their Moment Generating Functions.
$M_{z1+z2}(t) = E[e^{t(Z1+Z2)}] = E[e^{tZ1}e^{tZ2}] = E[e^{tZ1}]E[e^{tZ2}] = M_{Z1}(t)M_{Z2}(t)$
$$S_n = \sum_{i=1}^{n} Z_i \Rightarrow M_{S_n}(t) = \prod_{i=1}^{n} MX_j(t)$$

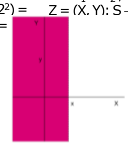
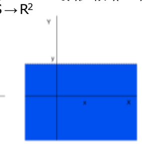
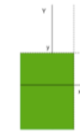
3.6) Central Limit Theorem

Given $X_1, X_2, \dots, X_n$ are independent and identically distributed random variables from any distribution with mean μ and finite variance σ², S_n = sum from to n of all x_i.
Hence, we have a distribution with known expectation (μ) and variance so we can form a Normal Distribution:
$$\begin{array}{lll} Y = S_n & E(Y) = n\mu & Var(Y) = n\sigma^2 \\ Y = S_n - n\mu & E(Y) = 0 & Var(Y) = n\sigma^2 \\ Y = \frac{S_n - n\mu}{\sqrt{n}\sigma} & E(X) = 0 & Var(X) = 1 \end{array}$$
Y can be used to approximate the standard normal:
$$\lim_{n\to\infty}\frac{S_n - n\mu}{\sqrt{n}\sigma} \sim N(0,1)$$
Thus, for large finite n,
$$\overline{X} \approx N(\mu, \frac{\sigma^2}{n}) \text{ and } \sum_{i=1}^{n} X_i \approx N(n\mu, n\sigma^2)$$
Where X bar is the average value of our random variables X_i from 1 to n.

4) Joint Random Variables

4.1) Joint Cumulative Density Functions

Suppose we have random variables X and Y s.t.:
X : S_X → R and Y : S_Y → R
We can define Z operating on sample space S s.t:
S = $S_1 \times S_2$, with S = {(s_X, s_Y)|$s_X \in S_X \wedge s_Y \in S_Y$}
Z = (X, Y): S → R²


^Marginal of X

^Marginal of Y

^Marginal of X and Y

Hence the induced probability function for Z will be:
$F(x, y) = P_Z(X ≤ x, Y ≤ y) = P_Z((−∞, x],(−∞, y]) = P(S_{XY})$. We can use the marginals of the joint distribution to get the distribution of the two random variables:
$F_X(x) = F(x, ∞)$
$F_Y(y) = F(∞, y)$
For valid Joint CDFs we require Monotonicity:
$\forall x_1, x_2, y_1, y_2 \in R.[x_1 < x_2 \Rightarrow F(x_1, y_1) \le F(x_2, y_1) \wedge y_1<y_2 \Rightarrow F(x_1, y_1) \le F(x_1, y_2)]$
∀x, y ∈ R. F(x, −∞) = F(−∞, y) = 0
F(∞, ∞) = 1

To compute the interval of a CDF:
$P_Z(x_1 < X ≤ x_2, y_1 < Y ≤ y_2) = F(x_2, y_2) – F(x_1, y_2) – F(x_2, y_1) + F(x_1, y_1)$

4.2) Joint Probability Mass Function

$p(x, y) = P_Z(X = x, Y = y)$ where x, y ∈ R
We can get the original pmfs of the two variables as: $p_X(x) = \sum_y p(x,y)$ and $p_Y(y) = \sum_x p(x,y)$
To be a valid pmf:
- $\forall x, y \in \mathbb{R}.\ 0 \le p(x,y) \le 1$
- $\sum_y \sum_x p(x,y) = 1$

4.3) Joint Probability Density Function

When the variables being joined are continuous we have R × R → R, in this case:
$$F(x,y) = \int_{a=-\infty}^{y}\int_{b=-\infty}^{x} f(b,a)\ db\ da$$
$$f(x,y) = \frac{\sigma^2}{\sigma x \sigma y}F(x,y)$$
We differentiate to get the PMF from the PDF, for CRVs remember.
To be valid:
1. ∀x, y ∈ R.f(x, y) ≥ 0
2. $\int_{y=-\infty}^{\infty}\int_{x=-\infty}^{\infty} f(x,y)\ dx\ dy = 1$

4.4) Marginal Density Functions

$$f_X(x) = \int_{y=-\infty}^{\infty} f(x,y)\ dy$$
$$f_Y(y) = \int_{x=-\infty}^{\infty} f(x,y)\ dx$$
AKA: We integrate over all of y to get marginal pdf for x, and integrate over all of x to get marginal pdf for y.
Example: Given continuous variables (X, Y) ∈ R²:
$$f(x,y) = \begin{cases} 1 & |x| + |y| < \frac{1}{\sqrt{2}} \\ 0 & otherwise \end{cases}$$
To determine the marginal PDFs for x and y, we use the formulae from above.
First notice that: |x|+|y|< 1/√2 ⇔ |y| < 1/√2 - |x|.
So, given an x, y must be between:
-1/√2 + |x| < y < 1/√2 - |x|
$F_x(x) = \int^{\infty}_{y=-\infty} f(x, y)\, dy = \int^a_b 1\, dy$ = 2/√2 - 2|x|.
Where a = 1/√2 - |x|, b = -1/√2 + |x|.
X follows a similar process to yield 2/√2 - 2|y|.

4.5) Multinomial Distribution

Given a sequence of n independent, identical experiments (same distribution and parameters), r possible outcomes for each experiment, each probability q_i corresponds to the outcome i, and the sum of all q_is is 1, we can model this with the **Multinomial Distribution.** We have a set of random variables where each X_i represents the number of experiments resulting in outcome i:
$P(X_1=n_1, X_2=n_2, \dots, X_r= n_r) = n!/n_1!*n_2!*\cdots*n_r! * q_1^{n1} \times q_2n^2 \times \cdots \times q_r^{nr}$
Example:
Given 4 different political parties with popularities: Ingsoc 40%, Techno Union 20%, Norsefire 15% Birthday Party 25%
We ask 10 people of what party they prefer, what is the probability that: 2 support Ingsoc, 4 support the Techno Union, 1 supports Norsefire, 3 support the Birthday Party?
$P(X_{ingsoc} = 2, X_{techno-union}=4, X_{norsefire}=1, X_{birthday} = 3)$ = 10! / (2! * 4! * 3! * 1!) x (0.4)² x (0.2)⁴ x (0.15)³ x (0.25)¹ = 0.756%.

4.6) Joint Conditional Random Variables

Given random variables X and Y: if they're independent $F(x, y) = F_x(x)F_y(y)$
Specifically:
For Discrete Variables $p(x, y) = p_x(x)p_y(y)$ (pmf)
For Continuous Variables $f(x, y) = f_x(x)f_y(y)$ (pdf)
Consider the previous example:
$f_x(x)$ = 2/√2 - 2|y|.
$f_y(y)$ = 2/√2 - 2|x|.
f(x, y) =/= $f_x(x)f_y(y)$; X and Y are dependent.

4.7) Conditional PMFs

$p_{X|Y}(x|y) = p(x, y) / p_Y(y)$ where ∀y.pY (y) > 0
Bayes' Theorem: states that on some partition of the sample space S, $P_1, \dots P_k$:
$P(X) = \Sigma^k_{i=1} P(X|E_i) P(E_i)$
Given each partition the probability of some X occurring sums to the total probability of X occurring.
Using the conditional joint pmf we can also express this theorem (over a single partition) as:
$p_{X|Y}(x|y) \times p_{Y|X}(y) = p_{Y|X}(y|x) \times p_X(x)$
Conditional Margin PDF properties:
$$p(x) = \sum\nolimits_y P_{X|Y}(x|y)p_y(y)$$
(To find the probability of x, go through every y, summing the probability of x occurring with that y, multiplied by the probability of that y)

4.8) Conditional PDFs

We class the joint PDF as $f_{X|Y}(x|y) = f(x, y) / f_Y(y)$
X and Y independent ⇔ ∀x, y ∈ R. $f_{X|Y}(x, y) = f_X(x)$
So Bayes' Theorem is: $f_{X|Y}(x|y) = f_{Y|X}f_X(x) / f_Y(y)$

Conditional PDF Marginal Joint Probabilities:
$$f_X(x) = \int_{y=-\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)\ dy$$
With the culmulative distribution:
$$F_X(x) = \int_{y=-\infty}^{\infty} F_{X|Y}(x|y)f_Y(y)\ d_y$$

Example:
Given X ~ Exp(λ) and Y ~ Exp(μ) what is P(X < Y).
$P(X<Y) = \int_{x<y} f(x,y)dx\, dy$
$= \int^{\infty}_{y=-\infty} \int^{\infty}_{c=-\infty} f(x, y)\, dx\, dy$
$= \int^{\infty}_{y=-\infty} \int^{\infty}_{c=-\infty} f_X(x) f_Y(y)\, dx\, dy$ (independent X, Y)
$= \int^{\infty}_{y=-\infty} F_X(y) * \mu e^{-\mu y}\, dx\, dy$
From here we continue substituting definitions and integrating and end with λ / λ + μ

4.9) Expectation and Variance for Joint Random Variables

Joint Expectation: g is a function of two variables on random variables X, Y:
For discrete variables: $E(g(X, Y)) = \Sigma_y\Sigma_x g(x,y)p(x,y)$
For continuous variables: $E(g(X, Y)) = \int_{y=-\infty}^{\infty}\int_{x=-\infty}^{\infty} g(x,y)f(x,y)\ dx\ dy$
Hence:
For all $g(X, Y) = g_1(X) + g_2(Y) \Rightarrow E(g_1(X) + g_2(Y)) = E_X(g_1(X)) + E_Y(g_2(Y))$
If X and Y are independent $E(g_1(X) \times g_2(Y)) = E_X(g_1(X))) \times E_Y(g_2(Y))$
If g(X, Y) = X × Y we have $E(XY) = E_X(X) \times E_Y(Y)$.

4.10) Covariance – Similar to variance but for two variables. Calculated the same way – $\sigma_{XY} = Cov(X, Y) = E[(X – \mu_X)(Y – \mu_Y)] = E[XY] – \mu_X\mu_Y$. When X and Y are independent, it is 0.
4.11) Correlation – a version of covariance that is ignorant to the scale of X and Y: $\rho XY = Cor(X, Y) = \sigma_{XY} / \sigma_X * \sigma_Y$

4.12) Multivariate Normal Distribution

Given a random vector $X=(X_1,\dots,X_n)$ with means $\mu=(\mu_1,\dots, \mu_n)$ has joint pdf:
$$f_X = \frac{1}{\sqrt{(2\pi)^n det \sum}}exp(-1/2(x-\mu)^T \sum^{-1}(x-\mu))$$
Where Σ is the covariance matrix, with Σ(i, j) = Cov(X_i, X_j). The covariance matrix is positive definite for a pdf to exist.

4.13) Conditional Expectation

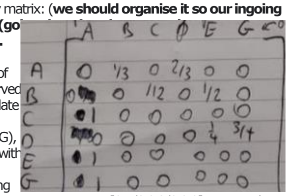
In general E(XY) =/= $E_X(X) E_Y(Y)$. For discrete random the conditional expectation of Y given that X = x is: $E_{Y|X}(Y|x) = \sum_y y p_{y|X}(y|x)$
For continuous random variables: $E_{Y|X}(Y|x) = \int_{y=-\infty}^{\infty} y f_{Y|X}(y|x)\ dy$
The conditional expectation is a function of x and not y in both cases.
Intuition: We get the weighted sum over all Ys, for each value (x) of X.
Expectation of a Conditional Expectation: $E_Y(Y) = E_X(E_{Y|X}(Y|X))$
Intuitively, the expected value of Y can be calculated in two step:1) calculate the conditional expected value of Y given X, which is denoted by EY|X 2) Take the expected value of EY|X over the entire range of X.
This can be generalised into the **Tower Rule** which calculates chained conditional expectations**:**
$E(Y) = E_{X1}(E_Y(Y|X_1)) =$
$E_{X2}(E_{X1}(E_Y(Y|X_1, X_2)|X_2)) = \dots$
$= E_{Xn} (E_{Xn-1}(\dots E_{X1}(E_Y(Y|X_1, \dots , X_n)|X_2, \dots , X_n)\dots|X_n))$

5) Discrete Time Markov Chains:

- A series of random variables modelling the state at each time step: $X_0, X_1\dots$
- The state at step n is denoted as J_n
- A stochastic process is a **Markov Chain** if the probability of from the current state to the next state depends **only on the current state**.

Important:
- The state space J contains all the states that we can be in at any step
- A sequence is just a path through the states $X_0, X_1 \dots$
- The initial probability vector π determines the starting state.
- The transition probability matrix r determines our probability of moving into each of the next states from our current state.

Example:
Consider an urn that contains initially two red balls and one green ball. At every time step you draw a ball at random. This is then put back into the urn and another ball of the same color is also added to the urn before the next draw. Assume that the experiment is stopped after n draws, when the urn contains n + 3 balls in total. Upon stopping, the urn is emptied and the experiment is restarted placing again two red balls and one green ball in the urn. Assume throughout that n = 2:
The state space for the urn is:
A B C D E G
J = {<2R, 1G>, <2R, 2G>, <2R, 3G>, <3R, 1G>, <3R, 2G>, <4R, 1G>}
Since we start in state 1 always, the initial probability vector π is [1 0 0 0 0 0]
The transition probability matrix: (**we should organise it so our ingoing states are on the left (go** **the top (going right)).**



	A	B	C	D	E	G
A	0	1/3	0	2/3	0	0
B	0	0	1/2	0	1/2	0
C	0	0	0	0	0	0
D	0	0	0	0	1/4	3/4
E	0	0	0	0	0	0
G	0	0	0	0	0	0

a) Let X be the number of red balls in the urn observed emptying the urn. Calculate the probability P(X < 4).
This probability = P(Not G), since G is the only state with 4 red balls.
We compute this by doing intial probability vector * transition matrix = [0 1/3 0 2/3 0 0]. Now we have 3 balls. Repeat, [0 1/3 0 2/3 0 0] * transition matrix = [0 0 1/6 0 1/3 ½].
We have done it twice now, as needed, as n = 2. So P(x<4) = P(Not G After 2 Draws) = 1 – ½ = 1/2 .

6) Statistics and Estimation

Statistics and Probability are kind of opposite - in probability we used distributions to predict the likelihood of events. In statistics, we use events/empirical data to determine or validate the probability distribution that models these results.

Sample – A subset of the population. Statistical methods use it to make inferences about the population.

Statistical Models – a structure (often a distribution) developed from a sample that can be used to make inferences about a population. They're parametric, ie: can be described entirely by their parameters, they have a finite set of parameters. If the probability of each outcome only depends on their parameters, then we can assume these points are IID.

$X_1, X_2, \dots, X_n \sim \text{Model}(\theta_1, \theta_2, \dots, \theta_k)$ given IID.

Examples: Normal, Poisson.

6.1) Central Limit Theorem for Statistics

Given a random variable X belonging to a distribution, the mean value of the sample size from X is:

$Y \sim N(\mu, \sigma^2/n)$. As the sample size increases, the variance in mean between different samples increases. At infinity we can use standard normal. AKA, the CLT lets us form a distribution without needing to know it.

6.2) Estimators

Statistic – a function operating on random variables of a sample. $T = T(X_1, X_2, \dots, X_n) = T(\underline{X})$. It is a function of random variables, and so it is a random variable itself. Hence if distribution X's parameters are known, we can use it, if T is the sum of ages of a class of 10, and we know the mean age, variance we can calculate probabilities for various T. When given some sample $x = (x_1, x_2, \dots, x_n)$ we have: $t = t(\underline{x}) = t(x_1, x_2, \dots, x_n)$.

Examples: mean, stdev, median.

Estimator: *A statistic used to approximate the parameter of the distribution of its arguments.* Given a sample $\underline{x}$ the estimator $t = t(\underline{x})$ is called an estimate. If we can approximately identify the sampling distribution of the statistic ($P_{T|\theta}$) we can find the expectation, variance (and more) related to our statistic. The CLT holds still

Examples of Estimators (some are better):

1) Using the first / any X_i as the estimator: $T[X_1, X_2, \dots, X_n] = X_1 \sim P_{X|\theta}$

2) Median: $T_{median}[X_1, X_2, \dots, X_n] = X\,|(n+1)/2| \sim P_{X|\theta}$

3) Mean $T_{\overline{X}}[X_1, X_2, \dots, X_n] = \frac{\sum_{i=1}^n X_i}{n} \sim N(\mu, \frac{\sigma^2}{n})$

Estimators can be biased thanks to being based on a sample rather than the population. $\text{bias}(T) = E[T|\theta] - \theta$

Unbiased estimator: bias = 0.

For any distribution the sample mean $\underline{x}$ is an unbiased estimate for the population mean μ.

For the variance: If we know the **population mean μ** we can use the unbiased estimator: $S_\mu^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$

The sample variance **is a biased estimator** and is defined as:

$$S^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$$

We apply **Bessel's Correction** to get the **unbiased sample variance**

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$$

6.3) Efficient Consistent Estimator

We can quantify how exactly good estimators are. We use a metric called **Estimator Efficiency:**

Given two unbiased estimators $E_1(\underline{X})$ and $E_2(\underline{X})$ where $X = (X_1, \dots, X_n)$ (a sample containing n observations X...)

We can compare the mean, variance etc to see which estimator is more efficient. We want a low variance.

E_1 is more efficient than E_2 if:

$\forall\theta \text{Var}\, E_1(E_1|\theta) \le \forall\theta \text{Var}\, E_2(E_2|\theta)$

or $\exists\theta\, \forall\theta \text{Var}\, E_1(E_1|\theta) < \forall\theta \text{Var}\, E_2(E_2|\theta)$

More efficient means less variance in estimates. If an estimator is more efficient than any other possible estimator, it is called efficient.

Example:

Given a population with mean μ and variance σ².

We have a sample: $X = (X_1, \dots, X_n)$

Consider two estimators:

1. $E_1 = \underline{X}$ (sample mean) 2. $E_2 = X_1$

We can compute the bias for both:

1. The expected value of the sample mean is the population mean μ, hence E_1 is unbiased.

2. The expected value of any observation is μ, so the first observation in the sample is also unbiased.

Now, we compute the variance: For a single sample the variance is σ² hence: $\text{Var}E_2(E_2|\mu \text{ and } \sigma^2) = \text{Var}(X_1) = \sigma^2$

For the sample mean, we use the CLT – so the variance is the mean of the sample divided by the size = σ²/n

So the variance of $E_2 \le E_1$ variance, so E_1 is the more efficient estimator.

The consistency of an estimator grows as the sample size grows. Note that the sample mean is a consistent estimator always.

7) Confidence Intervals

Case 1) We know the true variance of a population. We work out the sample mean, and it is distributed as:

$X \sim N(\underline{x}, \sigma^2/n)$

If μ (population mean) = $\underline{x}$ then (using the standard normal distribution) we can say that there is a 95% probability that the observed statistic is in the range $[\underline{x} - 1.96\,\sigma / n, \underline{x} + 1.96\,\sigma / n]$. This is using a two tailed standard normal value at the 95% confidence level.

So the formula is $[x - z\,\sigma / n, x + z\,\sigma / n]$ where z is the two tailed standard normal value at the right confidence interval.

Case 2) The true variance is unknown. We have to obtain the bias corrected variance:

$$S_{n-1} = \sqrt{\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}}$$

We must use the student's t distribution to calculate our t score:

We set degrees of freedom: v = n – 1.

For a double ended confidence (100 – α)%, we compute $t_{v=n-1,\,1-\alpha/2}$ to find the critical values.

So the formula is:

$[\underline{x} - t_{v=n-1,\,1-\alpha/2}\, S_{n-1} / \sqrt{n}, \underline{x} + t_{v=n-1,\,1-\alpha/2}\, S_{n-1} / \sqrt{n}]$

When using the tables for t values, we use the size we want (e.g 0.975 for 95% double-ended confidence interval), and then use the degrees of freedom (n – 1).

To do these questions we simply just find each parameter and slot it into the appropriate formula.

8) Hypothesis Testing:

Given two samples we determine whether the difference is significant enough to suggest the parameters of the distribution are different for the two of them. I know **Null Hypothesis H_0, Alternative Hypothesis H_1.**

We can have a 1) "has changed" two sided test (H_0: θ=θ₀ versus H_1: θ=/= θ₀) or a "is less than" or "is more than" one sided test (H_0: θ > θ₀ versus H_1: θ < θ₀)

Steps:

1. Choose a test statistic $T(\underline{X})$ to use on the data.
2. Find a distribution P_T under H_0 from the test statistic.
3. Determine the rejection region (the region in which a result would invalidate H0).
4. Calculate the observed test statistics $t(\underline{x})$.
5. If $t(\underline{x})$ is in the rejection region, reject H0 and accept H_1, else retain H_0.

8.1) Test Errors

The significance level α ∈ (0, 1) of a hypothesis test determines the size of the rejection regions.

α → 0 Less and less likely to reject H_0, rejection region smaller, confidence in our result is lower - easier test. (remember we use 1-α for the p value).

α → 1 More and more likely to reject H_0, rejection region larger, confidence higher - stricter test / easier to fail.

(5% significance is standard).

The p-value of a test is the significance level threshold between rejection/acceptance of H0 for a given test.

Type 1: Reject H_0 when it is actually true.

α = P(T ∈ R|H_0)

Type 2: Accepting H_0 when H_1 is true. β = P(T∉ R|H_1)

Probability a test statistic is not in the rejecting region, when H_1 is true

Test Power - The probability of correctly rejecting the null hypothesis.

Power = 1 – β = 1 – P(T∉ R|H_1) = P(T ∈ R|H_1)

For a given significance level: α = P(T ∈ R|H_0)

A good test statistic T and rejection region R will have a high power, the highest power test under H_1 is called the most powerful.

M3) Testing for Population Mean

We derive a new distribution in terms of the standard normal which we use to compute our confidence interval:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

Example:

A manufacturer sells packets listed as having weight 454g. From a sample size of 50, we get the mean weight of a bag as 451.22g. Assume the variance of bag weights is 70. Is the observed sample consistent with the claim made by the company at the 5% significance?

H_0 : μ = 454g, H_1: μ =/= 454g

We have this information:

$\underline{x}$ = 451.22g, σ² = 70, n = 50, α = 0.05.

Sample variance = 70/50. So, X ~ N (454, 70/50)

Z = $\underline{X}$ – 454/√35/5 ~ N(0, 1) (standard normal dist.)

Critical value = 0.05 two tails = 1.96

Hence in order to accept H_0, X must be in the interval: 451.6809 < X < 456.3191 (Using the confidence interval formula for known variance).

As x = 451.22 we reject H_0. At the 95% significance there is sufficient evidence to reject the company's claim.

If we have unknown variance, then we have to compute the **bias corrected variance.** We then use the student's t distribution instead for our confidence interval as outlined in case 2.

M4) Sample from Two Populations

If we are given two random samples and have two sample means, we do a Hypothesis test for equality.

Paired Data: A special case when $\underline{X}$ and $\underline{Y}$ are paired – each X_i and Y_i are possibly dependent on each other. We consider a sample of the differences, and test if this has mean 0:

$Z_i = X_i - Y_i$ testing H_0 : μZ = 0 versus H_1: μ_Z =/= 0

Example: Heart Rate before and after exercise.

8.2) Known Variance, X, Y are Independent

Given $\underline{X} = (X_1, \dots, X_{n1})$, $X_i \sim N(\mu_X, \sigma^2 X)$, $\underline{X} \sim N(\mu_X, \sigma^2 X/n_1)$ $\underline{Y} = (Y_1, \dots, Y_{n2})$, $Y_i \sim N(\mu_Y, \sigma^2 Y)$, $\underline{Y} \sim N(\mu_Y, \sigma^2 Y/n_2)$

We get the distribution of difference in sample means:

$X - Y \sim N(\mu_X - \mu_Y, \sigma^2 X/n_1 + \sigma^2 Y/n_2)$

We then put this distribution in the standard normal:

$$\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} \sim N(0, 1)$$

For H_0 we assume $u_x = u_Y$ so we end with z = the above formula with that bracket as 0.

8.3) Unknown Variance, X, Y are independent but with equal variance

We can combine their variance again and get an overall variance.

$$\frac{\overline{X} - \overline{Y}}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim N(0,1)$$

Example:

Compiler1: n_1 = 15, $\underline{x}$ = 114s, s^2_{14} = 310

Compiler2: n_2 = 15, $\underline{y}$ = 94s, s^2_{14} = 290

We assume that the variances of the population variances are the same for both.

... (other hypothesis test work done)

We can get the Bias-Corrected Pooled Sample Variance: S28 =

(14 × 310 + 14 × 290) / (14 + 14) = 300

Hence our test statistic is: $\underline{x}$ – $\underline{y}$ / σ √(1/n₁ + 1/n₂) = 20 /(300√(2/15)) = √ 10 ≈ 3.162

We proceed with Welch's T Test in this example.

8.4) Chi Squared Testing

$$X^2 = \sum_{i=1}^{} \frac{(O_i - E_i)^2}{E_i}$$

M5) Chi Squared Test for Model Checking

1. Determine expected distribution
2. Create a hypotheses based some parameters θ: H0 : θ = θ₀ versus H1 : θ =/= θ₀
3. Construct our E_i table
4. Calculate the Chi-Square Test Statistic X².
5. Calculate the degrees of freedom as: v = (number of possible values X can take) – (number of parameters being estimated) - 1.
6. Calculate the Chi Squared Statistic
7. Calculate the significance a, using a table with v, the degrees of freedom.
8. If X² > $\chi^2_{v,\,1-\alpha}$ (test statistic larger than critical value)

Note that:

All expected values must be larger than 5 for a good test. Hence some bins may have to be merged.

The number of values X can take is typically the number of bins.

M6) Chi Squared Test for Independence

This is the variant done back in A levels. We have a contingency table which has each combination of values of x and y. The only change we do is we count df = (rows-1) x (columns - 1). Questions will be worded like "Determine ... a link between...".

9) Maximum Likelihood Estimate

Given a distribution with unknown parameter θ:

X ~ Distribution(...θ...), and a sample of the distribution X: X = (X₁, X₂, ..., Xₙ), we want to determine **the most probable value for parameter θ, given our data.**

9.1) The Likelihood Function (L(θ))

The likelihood of some observations x₁, x₂, ... , xₙ occurring given some θ is:

$L(\theta) = P(x_1, x_2, \dots, x_n|\theta) = \prod_{i=1}^n f(x_i|\theta)$

Works because f is the probability mass function, and as each observation is independent we can multiply their probabilities.

The Log Likelihood Function (l(θ)) - Used more often than likelihood, much easier to work with l(θ) = ln L(θ)

To get this most probable value for θ, we construct the likelihood function, then get the log likelihood function, and differentiate to determine the value of l(θ) for which we have the maximum. This value is known as the **Maximum Likelihood Estimate** (θ').

9.2) Common Maximum Likelihood Estimates

Given a sample x = (x₁, x₂, ..., xₙ), we can use formulas for the maximum likelihood.

1) Exponential Distribution:

$X \sim \text{Exp}(\theta) \Rightarrow f(x) = \theta e^{-\theta x}$

1) Determine the likelihood in terms of θ:

$$L(\theta) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \theta e^{-\theta x_i} = \theta^n \prod_{i=1}^n e^{-\theta x_i} = \theta^n e^{-\theta\sum_{i=1}^n x_i}$$

2) Obtain log likelihood

$$l(\theta) = \ln L(\theta) = \ln\left(\theta^n e^{-\theta\sum_{i=1}^n x_i}\right) = n\ln\theta - \theta\sum_{i=1}^n x_i$$

3) Differentiate and set to 0:

$$\frac{dl(\theta)}{d\theta} = n\frac{1}{\theta} - \sum_{i=1}^n x_i = 0$$

$$0 = \frac{n}{\theta} - \sum_{i=1}^n x_i$$

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

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

4) Hence, the maximum likelihood estimator is the reciprocal of the mean of the sample.

2) Geometric Distribution:

1) Determine the likelihood in terms of θ:

$$X \sim Geo(\theta) \Rightarrow f(x) = \theta(1-\theta)$$

$$L(\theta) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \theta(1-\theta)^{x_i - 1} = \theta^n\prod_{i=1}^n (1-\theta)^{x_i-1} = \theta^n(1-\theta)^{\sum_{i=1}^n (x_i - 1)} = \theta^n(1-\theta)^{(\sum_{i=1}^n x_i) - n}$$

2) Obtain log likelihood

$$l(\theta) = \ln L(\theta) = \ln\left(\theta^n(1-\theta)^{(\sum_{i=1}^n x_i)-n}\right) = n\ln\theta + \left(\left(\sum_{i=1}^n x_i\right) - n\right)\ln(1-\theta)$$

3) Differentiate and set to 0

$$\frac{dl(\theta)}{d\theta} = \frac{n}{\theta} + \left(\left(\sum_{i=1}^n x_i\right) - n\right)\frac{1}{\theta - 1} = 0$$

$$0 = \frac{n(\theta-1)}{\theta(\theta-1)} + \left(\left(\sum_{i=1}^n x_i\right) - n\right)\frac{\theta}{\theta(\theta-1)}$$

$$0 = n(\theta - 1) + \left(\left(\sum_{i=1}^n x_i\right) - n\right)\theta$$

$$0 = n\theta - n + \left(\left(\sum_{i=1}^n x_i\right) - n\right)\theta$$

$$n = \left(\sum_{i=1}^n x_i\right)\theta$$

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

4) Hence, the maximum likelihood estimator is the reciprocal of the mean of the sample.

3) Binomial Distribution

$$X \sim Binomial(m, \theta) \Rightarrow f(x) = \binom{m}{x}\theta^x(1-\theta)^{m-x}$$

$$L(\theta) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \binom{m}{x_i}\theta^{x_i}(1-\theta)^{m-x_i} = \prod_{i=1}^n \binom{m}{x_i} \times \prod_{i=1}^n \theta^{x_i} \times \prod_{i=1}^n (1-\theta)^{m-x_i}$$

$$= \prod_{i=1}^n \binom{m}{x_i} \times \theta^{\sum_{i=1}^n x_i} \times (1-\theta)^{\sum_{i=1}^n m - x_i} = \prod_{i=1}^n \binom{m}{x_i} \times \theta^{\sum_{i=1}^n x_i} \times (1-\theta)^{mn - \sum_{i=1}^n x_i}$$

$$l(\theta) = \ln L(\theta) = \ln\left(\prod_{i=1}^n \binom{m}{x_i} \times \theta^{\sum_{i=1}^n x_i} \times (1-\theta)^{mn-\sum_{i=1}^n x_i}\right)$$

$$= \ln\prod_{i=1}^n \binom{m}{x_i} + \ln\theta^{\sum_{i=1}^n x_i} + \ln(1-\theta)^{mn - \sum_{i=1}^n x_i}$$

$$= \ln\prod_{i=1}^n \binom{m}{x_i} + \sum_{i=1}^n x_i\ln\theta + \left(mn - \sum_{i=1}^n x_i\right)\ln(1-\theta)$$

$$\frac{dl(\theta)}{d\theta} = 0 + \sum_{i=1}^n x_i\frac{1}{\theta} + \left(mn - \sum_{i=1}^n x_i\right)\frac{1}{\theta - 1} = 0$$

$$0 = \sum_{i=1}^n x_i\frac{\theta - 1}{\theta(\theta-1)} + \left(mn - \sum_{i=1}^n x_i\right)\frac{\theta}{\theta(\theta-1)}$$

$$0 = \sum_{i=1}^n x_i(\theta - 1) + \left(mn - \sum_{i=1}^n x_i\right)\theta$$

$$0 = \theta\sum_{i=1}^n x_i - \sum_{i=1}^n x_i + mn\theta - \theta\sum_{i=1}^n x_i$$

$$0 = -\sum_{i=1}^n x_i + mn\theta$$

$$\frac{\sum_{i=1}^n x_i}{mn} = \theta$$

$$\hat{\theta} = \frac{\overline{x}}{m}$$

10) Posterior

MLE has weaknesses: **Sensitive to Sample Size, Does not use any Prior, Returns a single val** rather than a distribution – so we don't how close other θ are, or how strong our estimate is. **Cannot Assess** – confidence intervals also rely on the sample.

Bayes' Theorem:

P(A|B) = P(B|A) × P(A) / P(B)

We can re-express this as:

P(A|B) = P(B|A) × P(A) / P(B|A)×P(A) + P(B|A)(1-P(A))

Intuition about Posterior, Prior and Likelihood:

Variable θ (e.g disease present?)	Variable X (e.g are symptoms present): x_1	x_2	...	x_n
θ_1	$P(x_1 \mid \theta_1)$	$P(x_2 \mid \theta_1)$	...	$P(x_n \mid \theta_1)$
θ_2	$P(x_1 \mid \theta_2)$	$P(x_2 \mid \theta_2)$	...	$P(x_n \mid \theta_2)$
⋮	⋮	⋮	⋱	⋮
θ_m	$P(x_1 \mid \theta_m)$	$P(x_2 \mid \theta_m)$	...	$P(x_n \mid \theta_m)$

$$\underbrace{P(\theta_j \mid x_i)}_{\text{Posterior}} = \frac{\overbrace{P(x_i \mid \theta_j)}^{\text{Likelihood}}\ \overbrace{P(\theta_j)}^{\text{Prior}}}{\underbrace{P(x_i)}_{\text{Evidence}}}$$

The Posterior is the probability we have the disease given we have symptoms.

The Likelihood is the probability we have symptoms given we have the disease. They are not the same. What we really want to know is the probability we have symptoms (P(x=x)), **the evidence** or the probability we have the disease P(θ) **the prior**).

Posterior ∝ Likelihood x Prior

10.1) Maximum a Posterior

Given some prior information (P(θ)) we can effectively get the MLE, but each probability is weighted by the prior information:

$$\theta'MAP = \arg\max_\theta \left[\prod_{i=1}^n P(X = x_i|\theta) \times P(\theta)\right]$$

This does require us to put prior information P(θ) about θ.

10.2) Conjugate Priors and Bayesian Inference

In Bayesian Inference, we compute the posterior distribution P(θ|x) ∝ P(x|θ) * P(θ). We keep updating our Posterior Distribution as we see new data. We feed our prior and likelihood to produce a posterior. This becomes our new prior, to calculate the next posterior, and so on.

Posterior formula:

$$P(\theta|X) = \frac{P(X|\theta) \times P(\theta)}{P(X)} = \frac{P(X|\theta) \times P(\theta)}{\int_{-\infty}^{\infty} P(X|\theta)P(\theta)\ d\theta}$$

Conjugate Prior: When continually inferring new prior distributions, if the prior distribution is in the same family of distributions (i.e parameters can be different, but same distribution) as the posterior, then it is a conjugate prior.

Likelihood	Conjugate Prior
Bernoulli Binomial Geometric	Beta
Poisson Exponential	Gamma
Normal	Normal

10.3) The Beta Prior Distribution

Where α, β > 0 are hyper-parameters that determine the shape of the distribution, the parameter is θ:

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

Where the normalising value (ensures total integral sums to 1 so it is a valid pdf) is:

$$B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\ d\theta$$

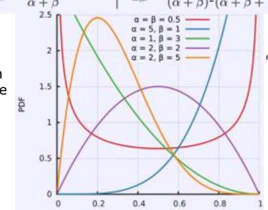
maximal value/θ_{MAP}	mean/bayesian estimate θ_B	variance
$argmax_\theta[Beta(\theta;\alpha,\beta)]$	$E[\theta]$	$E[\theta^2] - (E[\theta])^2$
$m_{\alpha,\beta} = \frac{\alpha - 1}{\alpha+\beta-2}$	$\mu_{\alpha,\beta} = \frac{\alpha}{\alpha+\beta}$	$\sigma^2_{\alpha,\beta} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

1) When α = β it is symmetrical about 0.5

2) Higher values result in steeper/narrower distribution

3) The MAP estimate pulls the estimate towards the prior.

4) As α → 1 and β → 1 Beta(θ; α, β) → U(0, 1) and ^θMAP → ^θMLE.



10.4) Gamma Prior Distribution

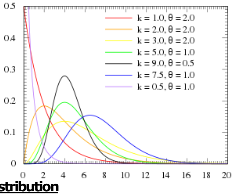
Used for the poisson and exponential.

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}e^{-\beta x}\beta^\alpha}{\Gamma(\alpha)} \quad \text{for } x > 0 \quad \alpha, \beta > 0.$$
$$\Gamma(\alpha) = (\alpha - 1)!.$$

Mean/bayesian estimate = α / β

Variance = α / β²

Maximal value (aka Mode, aka θ_{MAP}) = α -1 / β for α >= 1, 0 for α < 1.



Bayesian Inference Example: Bernoulli Distribution

Let X_i | θ ~ Bernoulli(θ) and θ ~ Beta(θ; a, b) where a > 1 and b > 1.

a) i) Derive the posterior distribution for θ|x₁, x₂, ..., xₙ. What are the formulas for $\hat{\theta}_B$, $\hat{\theta}_{MAP}$, $\hat{\theta}_{MaxPrior}$, $\hat{\theta}_{MeanPrior}$ and $\hat{\theta}_{MLE}$ estimates?

Posterior: Using Bayes Theorem:

$\text{Posterior} = P(\theta|\underline{X}) = \frac{P(\underline{X}|\theta)\,P(\theta)}{P(\underline{X})} = \frac{P(\underline{X}|\theta)\,P(\theta)}{\int_0^1 P(\underline{X}|\theta)\,P(\theta)\,d\theta}$

Denominator:

The denominator, $P(\underline{X}) = \int_0^1 \theta^{n\bar{x}}(1-\theta)^{n(1-\bar{x})} \cdot Beta(\theta; \alpha, \beta)\,d\theta$

Substituting for the numerator:

$P(\underline{X}|\theta)$ is gotten using the Bernoulli PMF:

$P(\underline{X}|\theta) = \prod_{i=1}^n P(x_i|\theta) = \theta^{n\bar{x}}(1-\theta)^{n(1-\bar{x})}$

Thus, $P(\theta|\underline{X}) = \frac{\theta^{n\bar{x}}(1-\theta)^{n(1-\bar{x})} \cdot \theta^{\alpha-1}(1-\theta)}{B(\alpha,\beta)}$ / $\frac{1}{B(\alpha,\beta)}\int_0^1 \theta^{\alpha+n\bar{x}-1}(1-\theta)^{n(1-\bar{x})+\beta-1}\,d\theta$

$= \frac{\theta^{\bar{x}n+\alpha-1}(1-\theta)^{n(1-\bar{x})+\beta-1}}{\int_0^1 \theta^{\bar{x}n+\alpha-1}(1-\theta)^{n(1-\bar{x})+\beta-1}\,d\theta}$

Let's take $A = \bar{x}_n + \alpha$
$B = n(1-\bar{x}) + \beta$

$= \frac{\theta^{A-1}(1-\theta)^{B-1}}{\int_0^1 \theta^{A-1}(1-\theta)^{B-1}\,d\theta} = \frac{\theta^{A-1}(1-\theta)^{B-1}}{B(A,B)}$

$= Beta(\theta; A, B)$

$= Beta(\theta; \bar{x}_n + \alpha, n(1-\bar{x}) + \beta)$ (3) This gives us the posterior distribution for θ|X₁,X₂...Xₙ

ii. If n = 30, ¯x = 0.6, a = 15, b = 10 then compute ^θB, ^θMAP , ^θMaxPrior, ^θMeanPrior and ^θMLE estimates and arrange these in ascending order.

We get the ^θB, ^θMAP, and ^θMLE by simple formula applications. MeanPrior = ^θB, and MaxPrior = MAP for iteration 1. For iteration 2, we take the ^θB and MAP respectively of the previous iteration (of the posterior we fed in).

Normal Distribution Example

1) Single datapoint x sample: Given some x|μ ~ N(μ, σ²) where σ² is known, μ unknown. U We get the likelihood using the normal distribution PDF:

$P(x|u) = f(x|u) = \frac{1}{\sigma\sqrt{2\pi}} \times exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$ where $exp\{n\} = e^n$

We then continue with this process normally, yielding $\sigma_1^2 = \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}$ and $\mu_1 = \sigma_1^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{x}{\sigma^2}\right)$

μ|X ~ N(μ₁, σ₁²)

2) We can extend this for a sample x = x₁, ... , xₙ and distribution xᵢ|μ ~ N(μ, σ²) where σ is known.

We end with:

$$\mu|\underline{x} \sim N(\mu_1, \sigma_1^2)$$

$$\sigma_1^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2} = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} \quad \text{and} \quad \mu_1 = \frac{\mu_0\sigma^2 + \sum_{i=1}^n \sigma_0^2 x_i}{\sigma^2 + n\sigma_0^2} = \sigma_1^2\left(\frac{\mu_0}{\sigma_0^2} + \sum_{i=1}^n \frac{x_i}{\sigma^2}\right)$$

Sufficient Statistic:

A statistic is sufficient for a given model (our chosen distribution) and its associated parameter if no other statistic can be calculated from a sample that provides additional information in computing the value/estimate of the unknown parameter.

For a normal distribution the sufficient statistic is the sample mean: $T(\underline{x}) = \overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$

1. We can calculate the posterior distribution using the likelihood and prior

We now have the Exponential part of the PDF for our normal.

$$P(\mu|\underline{x}) = f(\mu|\underline{x}) = \frac{f(\mu)f(\underline{x}|\mu)}{\int_{-\infty}^{\infty} f(\mu)f(\underline{x}|\mu)\,d\mu} \propto \frac{f(\mu)f(T(\underline{x})|\mu)}{\int_{-\infty}^{\infty} f(\mu)f(\underline{x}|\mu)\,d\mu} \propto f(\mu)f(T(\underline{x})|\mu) = f(\mu)f(\overline{x}|\mu)$$

$$= \frac{1}{\sigma_0\sqrt{2\pi}}exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} \times \frac{1}{\sqrt{2\pi\frac{\sigma^2}{n}}}exp\left\{-\frac{n(\overline{x}-\mu)^2}{2\sigma^2}\right\}$$

$$\propto exp\left\{\frac{-\left(\mu - \frac{\mu_0\sigma^2/n + \overline{x}\sigma_0^2}{\sigma^2/n + \sigma_0^2}\right)^2}{2\frac{\sigma_0^2\sigma^2/n}{\sigma^2/n+\sigma_0^2}}\right\}$$

2. We can now calculate Posterior Distribution μ|x ~ N(μ₁, σ₁²)

$$\sigma_1^2 = \frac{\sigma_0^2\sigma^2/n}{\sigma^2/n + \sigma_0^2} = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}$$

$$\mu_1 = \frac{\mu_0\sigma^2/n + \overline{x}\sigma_0^2}{\sigma^2/n + \sigma_0^2} = \sigma_1^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{\overline{x}n}{\sigma^2}\right)$$