DegreeLive

Bachelor of Commerce (B2001) | Level 2 • Part B Major | Monash University

Statistical thinking

Course Code: ETC2420 • Module 1 Notes

Module 1: Probability Models, Likelihood & Parametric Distribution Fitting

1. Statistical Thinking and the Data-Generating Process

Statistical thinking is the discipline of reasoning rigorously under uncertainty by distinguishing systematic structure from stochastic variation. In commerce, economics, public policy, and science, observed datasets rarely reflect deterministic laws alone. Instead, every observation is treated as a realisation from a **data-generating process (DGP)** that combines stable underlying mechanisms with random fluctuations arising from unmeasured heterogeneity, measurement noise, or genuine indeterminacy. Rather than memorising isolated hypothesis-testing recipes, statistical thinking requires formulating an explicit probabilistic model for how the data arose, quantifying the plausibility of competing parameter values given the observed sample, and checking whether the posited model faithfully represents reality.

Consider a Melbourne retail bank monitoring the time elapsed between consecutive customer service calls at its contact centre, or an Australian general insurer tracking the daily count of motor vehicle claims filed across Victoria. In both cases, individual observations fluctuate unpredictably from minute to minute or day to day, yet aggregate patterns—such as the average arrival rate, tail dispersion, or seasonal shifts—exhibit stable probabilistic regularity. Parametric probability models formalise this regularity using a small vector of interpretable parameters $\theta \in \Theta$. Once a model family is specified, statistical inference reverses the direction of deduction: probability theory asks what data patterns are expected when θ is known, whereas statistical inference uses observed data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ to learn about unknown θ and to forecast future realisations.

1.1 Sample Spaces, Events, and Probability Axioms

A formal probability model rests on a probability triple (Ω, F, P), where Ω denotes the **sample space** of all mutually exclusive elementary outcomes, F is a collection of measurable subsets (events) closed under complements and countable unions, and $P : \mathrm{F} \to [0, 1]$ is a probability measure satisfying Kolmogorov's three axioms:

- **Non-negativity:** For every event $A \in \mathrm{F}$, $P(A) \geq 0$.
- **Normalisation:** The certain event has unit probability, $P(\Omega) = 1$.
- **Countable additivity:** For any sequence of pairwise disjoint events $A_1, A_2, \ldots$, $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

From these three axioms follow the complement rule $P(A^c) = 1 - P(A)$, monotonicity ($A \subseteq B \implies P(A) \leq P(B)$), and the inclusion-exclusion identity $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

1.2 Conditional Probability, Total Probability, and Bayes' Rule

Information updates uncertainty through conditioning. When an event B with $P(B) > 0$ is known to have occurred, the effective sample space contracts to B, and the **conditional probability** of A given B is defined as:

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

Where: $P(A \mid B)$ is the conditional probability of event A given that B has occurred; $P(A \cap B)$ is the joint probability that both A and B occur; and $P(B) > 0$ is the marginal probability of the conditioning event.

Two events A and B are **statistically independent** if and only if $P(A \cap B) = P(A)P(B)$, which is equivalent to $P(A \mid B) = P(A)$ whenever $P(B) > 0$. Independence must never be assumed casually; it is a strong structural claim that knowing B provides zero information about A.

When a sample space is partitioned into mutually exclusive and exhaustive hypotheses $H_1, H_2, \ldots, H_k$ (so that $\bigcup_{j=1}^{k} H_j = \Omega$ and $H_i \cap H_j = \emptyset$ for $i \neq j$), the **Law of Total Probability** expresses the marginal probability of an observed event D as a probability-weighted average across all hypotheses:

DegreeLive

$$P(D) = \sum_{j=1}^{k} P(D \mid H_j)\, P(H_j)$$

Where: H_j represents the j-th partition state with prior probability $P(H_j)$; $P(D \mid H_j)$ is the likelihood of observing evidence D under state H_j; and $P(D)$ is the unconditional marginal probability of D.

Combining the definition of conditional probability with the Law of Total Probability yields **Bayes' Rule**, which inverts the conditioning direction from $P(D \mid H_i)$ to the posterior probability $P(H_i \mid D)$:

$$P(H_i \mid D) = \frac{P(D \mid H_i)\, P(H_i)}{\sum_{j=1}^{k} P(D \mid H_j)\, P(H_j)}$$

Where: $P(H_i \mid D)$ is the posterior probability of hypothesis H_i after observing data D; $P(H_i)$ is the prior probability before observing D; $P(D \mid H_i)$ is the likelihood of D under H_i; and the denominator is the marginal probability $P(D)$.

Worked Example 1.1: Automated Fraud Screening in Australian Card Payments

Suppose an Australian payments processor flags suspicious card-not-present transactions using an automated anomaly detector. Industry audit data indicate that 0.4% of all transactions are genuinely fraudulent (F), so $P(F) = 0.004$ and $P(F^c) = 0.996$. When a transaction is fraudulent, the algorithm triggers an alert (A) with sensitivity $P(A \mid F) = 0.92$. When a transaction is legitimate, the algorithm produces a false alarm with rate $P(A \mid F^c) = 0.03$. What is the probability that a transaction flagged by the system is actually fraudulent?

Step 1: Compute the joint probabilities for each state.

- True positives: $P(A \cap F) = P(A \mid F)P(F) = 0.92 \times 0.004 = 0.00368$.
- False positives: $P(A \cap F^c) = P(A \mid F^c)P(F^c) = 0.03 \times 0.996 = 0.02988$.

Step 2: Apply the Law of Total Probability to obtain the overall alert rate $P(A)$:

$$P(A) = 0.00368 + 0.02988 = 0.03356$$

Where: $P(A)$ is the marginal proportion of all transactions that trigger an automated fraud alert.

Step 3: Apply Bayes' Rule to find the positive predictive value $P(F \mid A)$:

$$P(F \mid A) = \frac{0.00368}{0.03356} \approx 0.1097$$

Where: $P(F \mid A)$ is the posterior probability that a flagged transaction is genuinely fraudulent.

Even with 92% sensitivity and 97% specificity, nearly 89% of alerts are false alarms because genuine fraud has a very low base rate (0.4%). Failing to account for base rates—known as the *base-rate fallacy*—leads decision-makers to overreact to noisy signals.

1.3 Random Variables, Distributions, Expectation, and Variance

A **random variable** $X : \Omega \to \mathrm{R}$ maps elementary outcomes to real numbers so that probabilistic statements can be analysed algebraically. A **discrete random variable** takes values in a countable set X and is characterised by its probability mass function (PMF) $p_X(x) = P(X = x)$, satisfying $p_X(x) \geq 0$ and $\sum_{x \in \mathrm{X}} p_X(x) = 1$. A **continuous random variable** takes values over an interval of real numbers and is characterised by a non-negative probability density function (PDF) $f_X(x)$ satisfying $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$, with interval probabilities given by $P(a \leq X \leq b) = \int_a^b f_X(x)\, dx$. Note that for any continuous random variable, $P(X = x) = 0$ at any single point.

DegreeLive

Both discrete and continuous variables share a unified **cumulative distribution function (CDF)**, defined for all $x \in \mathrm{R}$ as $F_X(x) = P(X \leq x)$. The CDF is non-decreasing, right-continuous, satisfies $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$, and relates to the PDF via $f_X(x) = F_X'(x)$ wherever F_X is differentiable. The inverse CDF, $Q(p) = F_X^{-1}(p) = \inf\{x \in \mathrm{R} : F_X(x) \geq p\}$ for $p \in (0,1)$, defines the **quantile function**, which plays a central role in Q-Q diagnostics and Monte Carlo simulation.

The location and spread of X are summarised by its **expectation** (first moment) and **variance** (second central moment):

$$\mathrm{E}[X] = \begin{cases} \sum_{x\in X} x\, p_X(x) & \text{(discrete)} \\ \int_{-\infty}^{\infty} x\, f_X(x)\, dx & \text{(continuous)} \end{cases}, \qquad \mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$$

Where: $\mathrm{E}[X]$ is the population mean (expected value) of X; $\mathrm{Var}(X)$ is the population variance measuring expected squared deviation from the mean; and $\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$ is the standard deviation expressed in the original units of X. Expectation is a linear operator: $\mathrm{E}[aX + bY + c] = a\mathrm{E}[X] + b\mathrm{E}[Y] + c$ for any constants a, b, c and random variables X, Y, regardless of whether X and Y are independent.

2. Joint, Marginal, and Conditional Distributions

Empirical investigations rarely examine a single variable in isolation. Analysts study how consumer spending varies with disposable income, how ASX equity returns move alongside bond yields, or how hospital length of stay relates to triage severity. Modelling multiple random variables simultaneously requires a joint distribution.

2.1 Joint PMF, Joint PDF, and Marginalisation

For two discrete random variables (X, Y), the **joint PMF** is $p_{X,Y}(x,y) = P(X = x, Y = y) \geq 0$ with $\sum_x \sum_y p_{X,Y}(x,y) = 1$. Summing out one variable recovers the **marginal PMF** of the other:

$$p_X(x) = \sum_y p_{X,Y}(x,y), \qquad p_Y(y) = \sum_x p_{X,Y}(x,y)$$

Where: $p_{X,Y}(x,y)$ is the joint probability mass at (x,y); $p_X(x)$ is the marginal probability mass of X; and $p_Y(y)$ is the marginal probability mass of Y.

Analogously, for jointly continuous random variables (X, Y) with joint PDF $f_{X,Y}(x,y) \geq 0$ and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx\, dy = 1$, probabilities over two-dimensional regions $A \subset \mathrm{R}^2$ are double integrals $\iint_A f_{X,Y}(x,y)\, dx\, dy$. Integrating out the nuisance variable yields the **marginal PDFs**:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx$$

Where: $f_{X,Y}(x,y)$ is the joint density at (x,y); and $f_X(x)$ and $f_Y(y)$ are the marginal probability density functions of X and Y, respectively.

2.2 Conditional Distributions and Conditional Expectation

Whenever $p_X(x) > 0$ (or $f_X(x) > 0$), the **conditional distribution** of Y given $X = x$ normalises the cross-sectional slice of the joint distribution at x so that it integrates or sums to unity:

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x,y)}{p_X(x)}, \qquad f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

DegreeLive

Where: $p_{Y\mid X}(y \mid x)$ and $f_{Y\mid X}(y \mid x)$ denote the conditional PMF and conditional PDF of Y given $X = x$; $p_{X,Y}$ and $f_{X,Y}$ are the joint distributions; and $p_X(x)$ and $f_X(x)$ are the marginal distributions of the conditioning variable.

Random variables X and Y are independent if and only if the joint distribution factors into the product of the marginals everywhere, $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, which implies $f_{Y\mid X}(y \mid x) = f_Y(y)$. When X and Y are dependent, the **conditional expectation** $\mathrm{E}[Y \mid X = x]$ defines the regression function $m(x)$:

$$\mathrm{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y f_{Y\mid X}(y \mid x)\, dy$$

Where: $\mathrm{E}[Y \mid X = x]$ is the expected value of Y within the sub-population where $X = x$. A fundamental optimality property of conditional expectation is that among all functions $g(X)$, $m(X) = \mathrm{E}[Y \mid X]$ uniquely minimises the expected squared prediction error $\mathrm{E}[(Y - g(X))^2]$.

2.3 Covariance, Correlation, and Iterated Moments

Linear co-movement between X and Y is measured by **covariance** and scale-free **Pearson correlation**:

$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y], \qquad \rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \in [-1,1]$$

Where: $\mathrm{Cov}(X,Y)$ measures the direction and magnitude of linear association in product units; and $\rho_{X,Y}$ is the dimensionless correlation coefficient bounded between -1 and $+1$. For any constants a and b, the variance of a linear combination obeys:

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X,Y)$$

Where: a and b are fixed scalar weights; and the cross-term $2ab\,\mathrm{Cov}(X,Y)$ vanishes when X and Y are uncorrelated ($\mathrm{Cov}(X,Y) = 0$). Note carefully that independence implies $\mathrm{Cov}(X,Y) = 0$, but $\mathrm{Cov}(X,Y) = 0$ does *not* imply independence when dependence is non-linear (for example, if $X \sim \mathrm{N}(0,1)$ and $Y = X^2$, then $\mathrm{Cov}(X,Y) = \mathrm{E}[X^3] = 0$ even though Y is a deterministic function of X).

Two powerful identities connect conditional moments to unconditional moments. The **Law of Iterated Expectations** states $\mathrm{E}[Y] = \mathrm{E}[\mathrm{E}[Y \mid X]]$, and the **Law of Total Variance** decomposes total uncertainty into expected conditional variance plus the variance of the conditional mean:

$$\mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}(\mathrm{E}[Y \mid X])$$

Where: $\mathrm{E}[\mathrm{Var}(Y \mid X)]$ is the average within-group variance unexplained by X; and $\mathrm{Var}(\mathrm{E}[Y \mid X])$ is the between-group variance explained by shifts in the conditional mean across values of X.

Worked Example 2.1: Joint Discrete Distribution of Loan Risk Tier and Default

An Australian regional lender classifies small-business loans into Low Risk ($X = 0$) or High Risk ($X = 1$) at origination and observes whether the loan defaults within two years ($Y = 1$) or remains performing ($Y = 0$). Suppose the joint PMF is $p_{X,Y}(0,0) = 0.72, p_{X,Y}(0,1) = 0.03, p_{X,Y}(1,0) = 0.20$, and $p_{X,Y}(1,1) = 0.05$.

Step 1: Compute marginal probabilities for X and Y.

- $p_X(0) = 0.72 + 0.03 = 0.75, p_X(1) = 0.20 + 0.05 = 0.25$.
- $p_Y(0) = 0.72 + 0.20 = 0.92, p_Y(1) = 0.03 + 0.05 = 0.08$. Thus $\mathrm{E}[Y] = 0.08$ and $\mathrm{Var}(Y) = 0.08 \times 0.92 = 0.0736$.

Step 2: Compute conditional default rates $\mathrm{E}[Y \mid X]$ and conditional variances $\mathrm{Var}(Y \mid X)$.

DegreeLive

- Given $X = 0$: $P(Y = 1 \mid X = 0) = 0.03/0.75 = 0.04$. Thus $\mathrm{E}[Y \mid X = 0] = 0.04$ and $\mathrm{Var}(Y \mid X = 0) = 0.04 \times 0.96 = 0.0384$.
- Given $X = 1$: $P(Y = 1 \mid X = 1) = 0.05/0.25 = 0.20$. Thus $\mathrm{E}[Y \mid X = 1] = 0.20$ and $\mathrm{Var}(Y \mid X = 1) = 0.20 \times 0.80 = 0.1600$.

Step 3: Verify the Law of Iterated Expectations and Law of Total Variance.

- $\mathrm{E}[\mathrm{E}[Y \mid X]] = 0.75(0.04) + 0.25(0.20) = 0.03 + 0.05 = 0.08 = \mathrm{E}[Y]$.
- $\mathrm{E}[\mathrm{Var}(Y \mid X)] = 0.75(0.0384) + 0.25(0.1600) = 0.0288 + 0.0400 = 0.0688$.
- $\mathrm{Var}(\mathrm{E}[Y \mid X]) = 0.75(0.04 - 0.08)^2 + 0.25(0.20 - 0.08)^2 = 0.75(0.0016) + 0.25(0.0144) = 0.0012 + 0.0036 = 0.0048$.
- Sum: $0.0688 + 0.0048 = 0.0736 = \mathrm{Var}(Y)$.

3. Parametric Distribution Families for Applied Data

Choosing a parametric family requires aligning the mathematical structure of the distribution—its support, skewness, tail behaviour, and generating mechanism—with the real-world phenomenon under study. Applying a symmetric Normal model to strictly positive, right-skewed insurance claims or inter-arrival times produces nonsensical negative predictions and miscalibrated tail risks.

3.1 Discrete Families: Binomial and Poisson

When an experiment consists of n independent Bernoulli trials, each with constant success probability $p \in (0, 1)$, the total count of successes $Y \sim \mathrm{Binomial}(n, p)$ has PMF:

$$p(y \mid n, p) = \binom{n}{y} p^y (1-p)^{n-y}, \qquad y \in \{0, 1, \dots, n\}$$

Where: n is the fixed number of trials; p is the single-trial success probability; $\binom{n}{y} = \frac{n!}{y!(n-y)!}$ counts the combinations; $\mathrm{E}[Y] = np$; and $\mathrm{Var}(Y) = np(1-p)$.

When events occur randomly and independently at a constant average rate $\lambda > 0$ over a fixed interval of time, space, or exposure —with no upper bound n—the event count follows a **Poisson distribution**, $Y \sim \mathrm{Poisson}(\lambda)$:

$$p(y \mid \lambda) = \frac{e^{-\lambda} \lambda^y}{y!}, \qquad y \in \{0, 1, 2, \dots\}$$

Where: $\lambda > 0$ is the rate parameter representing both the expected count $\mathrm{E}[Y] = \lambda$ and the variance $\mathrm{Var}(Y) = \lambda$. The Poisson distribution arises mathematically as the limit of $\mathrm{Binomial}(n, p)$ as $n \to \infty$ and $p \to 0$ with $np = \lambda$ held fixed. A defining diagnostic feature of the Poisson model is **equidispersion** ($\mathrm{Var}(Y)/\mathrm{E}[Y] = 1$). Real count datasets frequently exhibit **overdispersion** ($\mathrm{Var}(Y) > \mathrm{E}[Y]$) due to unobserved heterogeneity in rates across units; in such cases, a Negative Binomial model (or Gamma-Poisson mixture) is preferred.

3.2 Continuous Families: Exponential, Gamma, and Normal

If arrivals follow a homogeneous Poisson process with rate $\lambda > 0$ per unit time, the waiting time T until the next arrival is continuously distributed on $[0, \infty)$ according to the **Exponential distribution**, $T \sim \mathrm{Exp}(\lambda)$:

$$f(t \mid \lambda) = \lambda e^{-\lambda t}, \qquad F(t \mid \lambda) = 1 - e^{-\lambda t}, \qquad t \geq 0$$

Where: $\lambda > 0$ is the rate parameter (hazard rate); the mean waiting time is $\mathrm{E}[T] = 1/\lambda$; and the variance is $\mathrm{Var}(T) = 1/\lambda^2$. The Exponential distribution is the unique continuous distribution possessing the **memoryless property**: $P(T > s + t \mid T > s) = P(T > t)$. That is, having already waited s minutes without an arrival does not change the remaining waiting-time

DegreeLive

distribution. While appropriate for random arrivals or constant-hazard component failures, the memoryless property is violated when items age or wear out.

Generalising from the first arrival to the α-th arrival in a Poisson process yields the two-parameter **Gamma distribution**, $X \sim \text{Gamma}(\alpha, \beta)$ with shape $\alpha > 0$ and rate $\beta > 0$ (note that $\theta = 1/\beta$ is alternatively called the scale parameter; always check parameterisation):

$$f(x \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x > 0$$

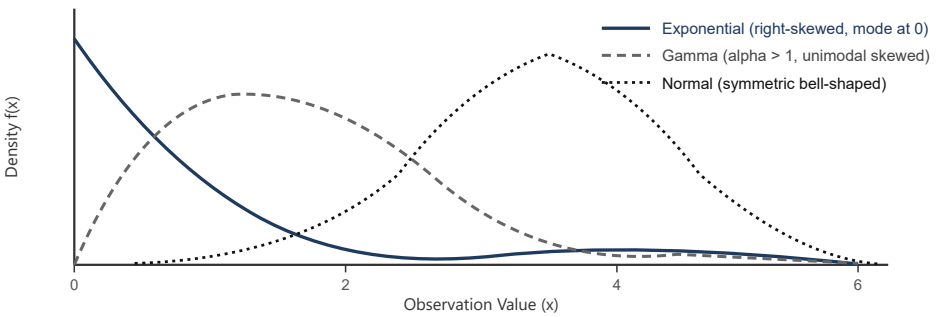
Where: $\Gamma(\alpha) = \int_0^{\infty} u^{\alpha-1} e^{-u}\, du$ is the Gamma function (equal to $(\alpha-1)!$ for positive integers); $\alpha > 0$ controls skewness and shape; $\beta > 0$ is the rate parameter; $\text{E}[X] = \alpha/\beta$; and $\text{Var}(X) = \alpha/\beta^2$. When $\alpha = 1$, the Gamma reduces to $\text{Exp}(\beta)$. When α grows large, skewness diminishes and the Gamma distribution approaches a Normal distribution. Because it accommodates positive support and flexible right skewness, the Gamma distribution is a workhorse for Australian health-care cost claims, rainfall totals, and duration analysis.

The **Normal (Gaussian) distribution**, $X \sim \text{N}(\mu, \sigma^2)$, governs measurements formed by the additive accumulation of many small independent perturbations, as formalised by the Central Limit Theorem:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad x \in \text{R}$$

Where: $\mu \in \text{R}$ is the location (mean, median, and mode); and $\sigma > 0$ is the scale (standard deviation). Standardising via $Z = (X - \mu)/\sigma$ maps any Normal variable to $\text{N}(0, 1)$, whose CDF is denoted $\Phi(z)$.

Distribution	Support	Parameters	Mean $\text{E}[X]$	Variance $\text{Var}(X)$	Applied Context
$\text{Binomial}(n, p)$	$\{0, 1, \ldots, n\}$	$n \in \text{N},\ p \in (0, 1)$	np	$np(1-p)$	Survey respondents supporting a policy; defective units in a batch
$\text{Poisson}(\lambda)$	$\{0, 1, 2, \ldots\}$	$\lambda > 0$	λ	λ	Daily insurance claims; hourly emergency arrivals; web hits
$\text{Exp}(\lambda)$	$[0, \infty)$	$\lambda > 0$ (rate)	$1/\lambda$	$1/\lambda^2$	Inter-arrival times in queues; constant-hazard lifetimes
$\text{Gamma}(\alpha, \beta)$	$(0, \infty)$	$\alpha > 0,\ \beta > 0$ (rate)	α/β	α/β^2	Total claim severities; Australian daily rainfall; service durations
$\text{N}(\mu, \sigma^2)$	$(-\infty, \infty)$	$\mu \in \text{R},\ \sigma^2 > 0$	μ	σ^2	Measurement errors; sample means via CLT; log-asset returns



Degree Live

4. The Likelihood Function and Maximum Likelihood Estimation

Once a parametric family $f(x \mid \theta)$ is chosen, the analyst must estimate the unknown parameter θ from an independent and identically distributed (i.i.d.) sample $x_1, x_2, \ldots, x_n$. The method of **Maximum Likelihood Estimation (MLE)**, formalised by R. A. Fisher, selects the parameter value $\hat{\theta}_{\text{MLE}}$ that renders the observed sample most probable under the assumed model.

4.1 Likelihood, Log-Likelihood, Score, and Fisher Information

For fixed observed data $\mathbf{x} = (x_1, \ldots, x_n)$, the **likelihood function** is the joint density (or joint PMF) viewed explicitly as a function of θ:

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

Where: $L(\theta; \mathbf{x})$ is the likelihood of parameter θ; $f(x_i \mid \theta)$ is the marginal PDF or PMF evaluated at observation x_i; and the product structure relies on the i.i.d. sampling assumption. Note a crucial conceptual distinction: $L(\theta; \mathbf{x})$ is *not* a probability distribution over θ (it need not integrate to one over Θ). Rather, ratios $L(\theta_1; \mathbf{x})/L(\theta_2; \mathbf{x})$ measure the relative evidential support that the data provide for θ_1 over θ_2.

Because products of many small densities easily underflow numerically and are awkward to differentiate, we maximise the monotonic transformation given by the **log-likelihood function** $\ell(\theta; \mathbf{x}) = \ln L(\theta; \mathbf{x})$:

$$\ell(\theta; \mathbf{x}) = \sum_{i=1}^{n} \ln f(x_i \mid \theta)$$

Where: $\ell(\theta; \mathbf{x})$ is the log-likelihood; and $\ln$ denotes the natural logarithm. Assuming $\ell(\theta)$ is twice differentiable and attains its maximum at an interior point of Θ, the MLE $\hat{\theta}$ satisfies the **score equation** and concavity check:

$$S(\theta) = \frac{\partial \ell(\theta; \mathbf{x})}{\partial \theta} = 0, \qquad \left.\frac{\partial^2 \ell(\theta; \mathbf{x})}{\partial \theta^2}\right|_{\theta=\hat{\theta}} < 0$$

Where: $S(\theta)$ is the score function; and the negative second derivative $J_n(\theta) = -\ell''(\theta; \mathbf{x})$ is the **observed Fisher information**, measuring the local curvature (sharpness) of the log-likelihood peak around $\hat{\theta}$. Its expectation across repeated samples, $I_n(\theta) = \mathrm{E}[-\ell''(\theta; \mathbf{X})] = nI_1(\theta)$, is the **expected Fisher information**.

Greater curvature corresponds to a narrower log-likelihood peak and therefore higher precision in estimating θ. Under standard regularity conditions, as $n \to \infty$, the MLE is consistent ($\hat{\theta} \xrightarrow{p} \theta_0$), asymptotically efficient (achieving the Cramér-Rao lower bound), and asymptotically Normal:

$$\hat{\theta}_{\text{MLE}} \overset{\cdot}{\sim} \mathrm{N}\left(\theta_0, \frac{1}{I_n(\theta_0)}\right), \qquad \mathrm{SE}(\hat{\theta}_{\text{MLE}}) \approx \frac{1}{\sqrt{J_n(\hat{\theta}_{\text{MLE}})}}$$

Where: θ_0 is the true data-generating parameter; $I_n(\theta_0)$ is the expected Fisher information; and $\mathrm{SE}(\hat{\theta}_{\text{MLE}})$ is the asymptotic standard error obtained from the observed curvature at the MLE.

4.2 Analytical MLE Derivations and Worked Numerical Examples

Exponential Rate Parameter

Suppose $x_1, \ldots, x_n$ are i.i.d. waiting times from $\mathrm{Exp}(\lambda)$. The log-likelihood is:

DegreeLive

$$\ell(\lambda) = \sum_{i=1}^{n} \ln(\lambda e^{-\lambda x_i}) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

Where: n is the sample size; $\lambda > 0$ is the rate parameter; and $\sum x_i = n\bar{x}$ is the total observed waiting time. Differentiating with respect to λ gives the score $S(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$, yielding $\hat{\lambda}_{\text{MLE}} = 1/\bar{x}$. The second derivative is $\ell''(\lambda) = -n/\lambda^2 < 0$ for all $\lambda > 0$, confirming a unique global maximum, with observed information $J_n(\hat{\lambda}) = n/\hat{\lambda}^2$ and asymptotic standard error $\text{SE}(\hat{\lambda}) = \hat{\lambda}/\sqrt{n}$.

Worked Example 4.1: Fitting an Exponential Model to Melbourne Call-Centre Arrivals

A Melbourne customer-support desk records the following $n = 10$ inter-arrival times (in minutes) between consecutive inbound calls during a steady weekday morning window: $1.8, 0.4, 3.2, 1.1, 0.7, 2.5, 0.9, 4.1, 1.5, 2.8$.

Step 1: Compute the sufficient statistic $\sum_{i=1}^{10} x_i$ and sample mean $\bar{x}$.

$$\sum_{i=1}^{10} x_i = 1.8 + 0.4 + 3.2 + 1.1 + 0.7 + 2.5 + 0.9 + 4.1 + 1.5 + 2.8 = 19.0 \text{ minutes}, \qquad \bar{x} = \frac{19.0}{10} = 1.90 \text{ minutes}$$

Where: $\bar{x}$ is the sample mean inter-arrival duration.

Step 2: Compute the MLE of the arrival rate λ and its asymptotic standard error.

$$\hat{\lambda}_{\text{MLE}} = \frac{1}{1.90} \approx 0.5263 \text{ calls/min}, \qquad \text{SE}(\hat{\lambda}_{\text{MLE}}) = \frac{0.5263}{\sqrt{10}} \approx 0.1664 \text{ calls/min}$$

Where: $\hat{\lambda}_{\text{MLE}}$ estimates the per-minute call rate; and $\text{SE}(\hat{\lambda}_{\text{MLE}})$ approximates its sampling standard error.

Step 3: Evaluate the maximised log-likelihood and predict the probability of a gap exceeding 3 minutes:

$$\ell(\hat{\lambda}) = 10 \ln(0.5263) - 0.5263(19.0) = -6.4185 - 10.000 = -16.4185, \qquad P(X > 3) = e^{-0.5263 \times 3} \approx 0.2062$$

Where: $\ell(\hat{\lambda})$ is the maximised log-likelihood value used later for model comparison; and $P(X > 3)$ is the plug-in survival probability.

Poisson Rate Parameter

For i.i.d. counts $y_1, \ldots, y_n \sim \text{Poisson}(\lambda)$, the log-likelihood is $\ell(\lambda) = \sum_{i=1}^{n} (y_i \ln \lambda - \lambda - \ln(y_i!))$. Setting $\ell'(\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} y_i - n = 0$ gives $\hat{\lambda}_{\text{MLE}} = \bar{y}$, with second derivative $\ell''(\lambda) = -\sum y_i/\lambda^2 < 0$ and standard error $\text{SE}(\hat{\lambda}_{\text{MLE}}) = \sqrt{\hat{\lambda}/n}$.

Normal Mean and Variance

For i.i.d. observations $x_1, \ldots, x_n \sim \text{N}(\mu, \sigma^2)$, the two-parameter log-likelihood is:

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

Where: μ is the population mean; and $\sigma^2 > 0$ is the population variance. Solving the two partial score equations $\partial \ell / \partial \mu = 0$ and $\partial \ell / \partial \sigma^2 = 0$ simultaneously yields:

DegreeLive

$$\hat{\mu}_{\text{MLE}} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Where: $\hat{\mu}_{\text{MLE}}$ is the sample mean; and $\hat{\sigma}^2_{\text{MLE}}$ is the maximum likelihood variance estimator. Notice that $\hat{\sigma}^2_{\text{MLE}}$ divides by n rather than $n-1$. Because $\mathrm{E}[\hat{\sigma}^2_{\text{MLE}}] = \frac{n-1}{n}\sigma^2$, the MLE slightly underestimates σ^2 in finite samples, whereas the classical sample variance $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$ is unbiased. This illustrates a central theme of statistical thinking: maximum likelihood estimators are not guaranteed to be unbiased in finite samples, even though bias vanishes at rate $O(1/n)$ as n increases.

4.3 Likelihood Ratio Tests and Model Selection Criteria

When comparing a restricted null model $H_0 : \theta = \theta_0$ (or a nested sub-model with k fewer free parameters) against an unrestricted model with MLE $\hat{\theta}$, Wilks' theorem states that under H_0 as $n \to \infty$, the **Likelihood Ratio Test (LRT)** statistic follows a chi-squared distribution:

$$\Lambda = 2\left[\ell(\hat{\theta}) - \ell(\theta_0)\right] \overset{\cdot}{\sim} \chi^2_k$$

Where: Λ is twice the log-likelihood gain; $\ell(\hat{\theta})$ is the maximised log-likelihood of the full model; $\ell(\theta_0)$ is the maximised log-likelihood under the null restriction; and k is the number of independent constraints imposed by H_0. Large values of Λ indicate that the data are substantially more plausible under the unrestricted model than under H_0.

For comparing non-nested parametric families (such as Gamma versus Log-Normal versus Weibull on positive expenditure data), the LRT does not apply. Instead, analysts penalise the maximised log-likelihood for model complexity p (number of estimated parameters) using **Akaike's Information Criterion (AIC)** or the **Bayesian Information Criterion (BIC)**:

$$\text{AIC} = -2\ell(\hat{\theta}) + 2p, \qquad \text{BIC} = -2\ell(\hat{\theta}) + p\ln(n)$$

Where: $\ell(\hat{\theta})$ is the maximised log-likelihood; p is the number of freely estimated parameters; and n is the sample size. Lower AIC or BIC indicates a superior trade-off between goodness-of-fit and parsimony. For $n \geq 8$, $\ln(n) > 2$, so BIC imposes a heavier complexity penalty than AIC and favours simpler models in large datasets.

5. Visual Diagnostics, Q-Q Plots, and Goodness-of-Fit Assessment

Fitting a parametric distribution by MLE always produces parameter estimates, even if the chosen family is completely misspecified. Model criticism—checking whether the fitted distribution actually aligns with the data—is therefore mandatory before drawing substantive conclusions or computing tail probabilities.

5.1 Empirical CDF and Quantile-Quantile (Q-Q) Plots

Given ordered observations $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, the **empirical cumulative distribution function (ECDF)** is the step function $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(x_i \leq x)$, which jumps by $1/n$ at each distinct data point. Comparing $\hat{F}_n(x)$ against the fitted parametric CDF $F(x;\hat{\theta})$ visually reveals broad location or scale mismatches, yet subtle discrepancies in the extreme tails are compressed near 0 and 1 on the probability scale.

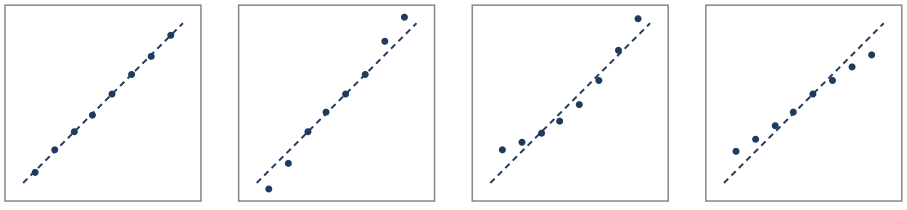
A **Quantile-Quantile (Q-Q) plot** overcomes this compression by plotting the ordered sample quantiles $x_{(i)}$ on the vertical axis directly against the theoretical quantiles $q_i = F^{-1}(p_i;\hat{\theta})$ on the horizontal axis, where standard plotting positions are $p_i = (i - 0.5)/n$ for $i = 1, \ldots, n$. If the data truly originate from $F(\cdot;\theta)$, points should cluster tightly along the 45° reference line $y = x$ (or a straight line $y = \hat{\mu} + \hat{\sigma}z_i$ when plotting against standard $\mathrm{N}(0,1)$ quantiles $z_i = \Phi^{-1}(p_i)$).

Systematic curvature away from the straight line diagnoses specific forms of misspecification:

- **S-shape curving above on the right and below on the left:** The empirical distribution has *heavy tails* (leptokurtic, more extreme values than predicted by the model), common in daily ASX stock returns compared against a Normal fit.

degreelive.org

- **S-shape curving below on the right and above on the left:** The empirical distribution has *light tails* (platykurtic, fewer extremes than the model, such as bounded or uniform-like measurements).
- **Concave upward (U-shaped) bow above the line at both ends:** The empirical distribution is *right-skewed* relative to the reference model (both small and large sample quantiles exceed the symmetric predictions).
- **Concave downward bow below the line at both ends:** The empirical distribution is *left-skewed* relative to the reference model.



5.2 Formal Goodness-of-Fit Hypothesis Tests

Visual inspection can be complemented by formal goodness-of-fit tests. Care is required in interpreting these tests: with small samples, goodness-of-fit tests have low power and may fail to reject a genuinely poor model; with massive samples, tiny practically negligible deviations trigger rejection at $\alpha = 0.05$.

Pearson's Chi-Squared Goodness-of-Fit Test

Designed primarily for discrete counts or binned continuous data partitioned into K mutually exclusive categories, Pearson's statistic compares observed cell counts O_j against expected cell counts $E_j = n\hat{p}_j$:

$$X^2 = \sum_{j=1}^{K} \frac{(O_j - E_j)^2}{E_j} \dot{\sim} \chi^2_{K-1-m}$$

Where: O_j is the observed count in category j; $E_j = n\hat{p}_j$ is the expected count under the fitted model; K is the number of categories after pooling any sparse tail cells so that $E_j \geq 5$; and m is the number of parameters estimated from the data via MLE (for instance, $m = 1$ for Poisson, $m = 2$ for Normal).

Worked Example 5.1: Testing Poisson Fit on Workplace Safety Incidents

An Australian logistics warehouse tracks weekly minor safety incidents over $n = 100$ weeks and records the following frequency table: 0 incidents in 28 weeks; 1 incident in 38 weeks; 2 incidents in 22 weeks; 3 incidents in 8 weeks; and 4 or more incidents in 4 weeks (specifically 3 weeks with 4 incidents and 1 week with 5 incidents). Test at $\alpha = 0.05$ whether a Poisson model fits these counts.

Step 1: Compute the MLE $\hat{\lambda}$ from total incidents divided by 100 weeks.

$$\sum_{i=1}^{100} y_i = 0(28) + 1(38) + 2(22) + 3(8) + 4(3) + 5(1) = 123, \qquad \hat{\lambda} = \frac{123}{100} = 1.23$$

Where: $\hat{\lambda}$ is the estimated mean weekly incident rate.

Step 2: Compute expected Poisson probabilities $\hat{p}_j = e^{-1.23}(1.23)^j / j!$ and expected counts $E_j = 100\hat{p}_j$ (using $e^{-1.23} \approx 0.29229$).

Weekly Incidents (j)	Observed (O_j)	Poisson Prob ($\hat{p}_j$)	Expected (E_j)	$(O_j - E_j)^2/E_j$
0	28	0.2923	29.23	0.0518

Degreelive

Weekly Incidents (j)	Observed (O_j)	Poisson Prob ($\hat{p}_j$)	Expected (E_j)	$(O_j - E_j)^2/E_j$
1	38	0.3595	35.95	0.1169
2	22	0.2211	22.11	0.0005
3	8	0.0907	9.07	0.1262
≥ 4 (pooled)	4	0.0364	3.64	0.0356

Wait: the pooled cell ≥ 4 has $E_5 = 3.64 < 5$. To satisfy the standard Cochran rule ($E_j \geq 5$), we pool $j = 3$ and $j \geq 4$ into a single category ≥ 3, giving $O_{\geq 3} = 8 + 4 = 12$ and $E_{\geq 3} = 9.07 + 3.64 = 12.71$. Its chi-squared contribution is $(12 - 12.71)^2/12.71 = 0.0397$.

Step 3: Sum the contributions across the $K = 4$ pooled categories $(0, 1, 2, \geq 3)$ and determine degrees of freedom.

$$X^2 = 0.0518 + 0.1169 + 0.0005 + 0.0397 = 0.2089, \qquad \text{df} = K - 1 - m = 4 - 1 - 1 = 2$$

Where: X^2 is Pearson's goodness-of-fit statistic; $K = 4$ is the number of post-pooling cells; and $m = 1$ accounts for estimating $\hat{\lambda}$.

Step 4: Compare against $\chi^2_{2,\,0.95} = 5.991$. Because $0.2089 \ll 5.991$ (p-value ≈ 0.901), there is no evidence against the Poisson specification; the warehouse incident counts are consistent with a constant-rate Poisson process.

Kolmogorov-Smirnov and Anderson-Darling Tests

For unbinned continuous data, the **Kolmogorov-Smirnov (KS) statistic** measures the maximum vertical gap between the empirical CDF and the hypothesised continuous CDF $F_0(x)$:

$$D_n = \sup_{x \in \mathrm{R}} \left| \hat{F}_n(x) - F_0(x) \right|$$

Where: D_n is the Kolmogorov-Smirnov supremum distance; $\hat{F}_n(x)$ is the ECDF; and $F_0(x)$ is the reference continuous CDF. Note an important subtlety: when F_0 is completely specified in advance, the standard KS null distribution applies. When parameters $\hat{\theta}$ are estimated from the same sample (such as testing Normality using $\bar{x}$ and s), standard KS critical values are overly conservative, requiring the **Lilliefors correction** or parametric bootstrap simulation. Moreover, because D_n is most sensitive near the median where $F_0(x) \approx 0.5$, the **Anderson-Darling test** applies inverse-variance weighting $[F_0(x)(1 - F_0(x))]^{-1}$ inside a squared discrepancy integral, granting far greater sensitivity to discrepancies in the extreme tails—precisely where financial and risk models most often fail.

DegreeLive