Statistical Inference: Introduction and refresher

Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch

September 4, 2025

- We start with a concrete question, e.g.,
  - Does the Higgs boson exist?
  - Is fraud taking place at this factory?
  - Are these two satellites likely to collide soon?
  - Do lockdowns reduce Covid transmission?
- We aim
  - to use **data**
  - to provide **evidence** bearing on the question,
  - to draw a **conclusion** or reach a **decision** to guide future actions.
- Here we mostly discuss how to express the evidence, but the choice and quality of the data, and how they were obtained, affect the evidence and the clarity of any decision.
- The data typically display both **structure** and haphazard **variation**, so any conclusion reached is uncertain, i.e., is an **inference**.

- Theoretical discussion generally takes observed data as given, but
  - to get the data we may need to **plan an investigation**, perhaps **design an experiment** largely controlled by the investigator — not considered here but often crucial to obtaining strong data and hence secure conclusions; or
  - to use data from an **observational study** (the investigator has little or no control over data collection).
- In both cases the data used may be selected from those available, and especially if we have 'found data' we must ask
  - why am I seeing these data?
  - what exactly was measured, and how?
  - can the observations actually shed light on the problem?
  - will using a function of the available data give more insight?
- For now we suppose these questions have satisfactory answers . . .

- Conventionally divided into
  - **design of investigations** — how do we get reliable data to answer a question efficiently and securely?
  - **descriptive statistics/exploratory data analysis** — how can we get insight into a specific dataset?
  - **inference** — what can we learn about the properties of a 'population' underlying the data?
  - **decision analysis** — what is the optimal decision in a given situation?

  to which we nowadays add

  - **machine learning** — algorithms, generally complex and computationally demanding, often used for prediction/decision-making.

- In principle concerns **only the data available**, mainly involving
  - **graphical summaries** — histograms, boxplots, scatterplots, ...
  - **numerical summaries** — averages, variances, medians, ...
- Some summaries presuppose the existence of 'population' quantities (e.g., a density).
- We use probability models to analyse the properties of these summaries (e.g., formulation of a boxplot, 'is that difference significant?', ...).
- Even when we have 'all the data' (e.g., loyalty card transactions) we may want to ask 'what if?' questions, and these require further assumptions (e.g., temporal stability, future and current customers are similar, ...).

- Use observed data to draw conclusions about a 'population' from which the data are assumed to be drawn, or about future data.
- The 'population' and observed data are linked by concepts of probability.
- Two distinct roles of probability in statistical analysis:
  - as a description of **variation** in data ('chance'), treating the observed data y as an outcome of a random process/probability model, perhaps
    - suggested by the context, or
    - imposed by the investigator (via some sampling procedure);
  - to formulate **uncertainty** ('epistemic probability') about the reality modelled in terms of the random experiment, based on y.
- Most of the course concerns the formulation and expression of uncertainty.
- We first revise some concepts from probability and basic statistics.

- Planning of investigations
- Obtaining reliable data
- Exploratory data analysis/visualisation
- **Model formulation**
- **Point estimation** of a population parameter
- **Interval estimation** for a population parameter
- **Hypothesis testing** to assess whether observed data support a particular model
- **Prediction** of a future or unobserved random variable
- **Decision analysis** to choose an action based on data and the costs of potential actions

This course covers some aspects of those activities listed above.
Many inferential tasks can be formulated in decision-theoretic terms, but we shall mostly avoid this.

- Use observed data to draw conclusions about a 'population', i.e., a model from which the data are assumed to be drawn, or about future data.
- A **statistical model** is a family of probability distributions for data y in a sample space $\mathcal{Y}$.
- A **parametric model (family of models)** $f \equiv f(y;\theta)$ or equivalently $F \equiv f(y;\theta)$ is determined by **parameters** $\theta \in \Theta \subset \mathbb{R}^d$, for fixed finite d.
- If no such θ exists, F is **nonparametric**, and then the parameter is often determined by F through a **statistical functional** $\theta = t(F)$, e.g.,

$$\mu = t_1(F) = \int y \, \mathrm{d}F(y), \quad \sigma^2 = t_2(F) = \int y^2 \, \mathrm{d}F(y) - \left\{ \int y \, \mathrm{d}F(y) \right\}^2.$$

- Parameters have different roles (which can change during an investigation):
  - **interest parameters** represent targets of inference (e.g., the mean of a population, the slope of a line, a baseline blood pressure) with direct substantive interpretations;
  - **nuisance parameters** are needed to complete a model specification, but are not themselves of main concern (e.g., variance of population requires estimating the mean).
- A parametric model should have a 1–1 map from θ to $f(\cdot;\theta)$, so parameters identify models.

- Two broad types of statistical model:
  - **substantive** — based on fundamental subject-matter theory (e.g., quantum theory, Mendelian genetics, Navier–Stokes equations);
  - **empirical** — a convenient, adequately realistic, representation of data variation;
  - and of course a broad spectrum between them.
- We aim that
  - primary questions/issues are encapsulated in interest parameters;
  - secondary aspects can be accounted for, often via nuisance parameters;
  - variation in the data is modelled well enough to give realistic assessments of uncertainty;
  - any special feature of the data or data collection process is represented;
  - different approaches to analysis can if necessary be compared.
- Such models are always provisional and should if possible be checked against data.

Some (important) notation

- Vectors are always column vectors, with row vectors denoted using the transpose $^{\mathrm{T}}$.
- By convention we (try to) use
  - letters like $c, d, \ldots$ for (known) constants,
  - Roman letters for random variables $X, Y, \ldots$ and their realisations $x, y, \ldots$,
  - Greek letters $\mu, \nu, \psi, \lambda, \Omega, \Delta, \ldots$ for unknown parameters, and
  - α is mostly reserved for significance levels.
- We distinguish the data actually observed, y^{o}, from other possible values y, and likewise for estimators $\widehat{\theta}^{\mathrm{o}}$, probabilities $p^{\mathrm{o}} = \Pr(Y \geq y^{\mathrm{o}})$, ..., based on y^{o}.
- We write $\nabla \cdot = \partial \cdot / \partial \varphi$ and $\nabla^2 \cdot = \partial^2 \cdot / \partial \varphi \partial \varphi^{\mathrm{T}}$ for differentiation with respect to a parameter, and ∇_y etc., for other derivatives. Hence if $g(\varphi)$ is a scalar function of a $d \times 1$ parameter φ, then $\nabla g(\varphi)$ is a $d \times 1$ vector and $\nabla^2 g(\varphi)$ is a $d \times d$ matrix, and if $h(\varphi)$ is a $n \times 1$ vector function of φ, then $\nabla h^{\mathrm{T}}(\varphi)$ is a $d \times n$ matrix and $\nabla^{\mathrm{T}} h(\varphi)$ is an $n \times d$ matrix.
- In general discussion we often suppose that data Y come from some unknown 'true' density g, but we fit a candidate density $f(y; \theta)$ that may be different from g.

- Ordered triples $(\Omega, \mathcal{F}, \Pr)$ consisting of
  - a set Ω of **elementary outcomes** ω corresponding to distinct potential outcomes of a random experiment;
  - an **event space** $\mathcal{F}$ of subsets of Ω that satisfy (a) $\Omega \in \mathcal{F}$, (b) if $\mathcal{A} \in \mathcal{F}$, then $\mathcal{A}^c \in \mathcal{F}$, and (c) if $\mathcal{A}_1, \mathcal{A}_2, \ldots \in \mathcal{F}$, then $\bigcup A_j \in \mathcal{F}$;
  - a **probability measure** $\Pr : \mathcal{F} \to [0, 1]$ that satisfies (i) if $\mathcal{A} \in \mathcal{F}$, then $0 \leq \Pr(\mathcal{A}) \leq 1$, (ii) $\Pr(\Omega) = 1$, (iii) if $\mathcal{A}_1, \mathcal{A}_2, \ldots \in \mathcal{F}$ satisfy $\mathcal{A}_j \cap \mathcal{A}_k = \emptyset$ for $j \neq k$, then $\Pr(\bigcup \mathcal{A}_j) = \sum \Pr(\mathcal{A}_j)$.
- We call $(\Omega, \mathcal{F})$ a **measure space** and any $\mathcal{A} \in \mathcal{F}$ an **event (measurable set)**.
- From these we deduce
  - the **inclusion-exclusion formulae**, and
  - computation of probabilities with **combinatorial formulae**.
- If $\Pr(\mathcal{B}) > 0$ we define **conditional probabilities** $\Pr(\mathcal{A} \mid \mathcal{B}) = \Pr(\mathcal{A} \cap \mathcal{B})/\Pr(\mathcal{B})$, and derive
  - a new **conditional probability distribution** $\Pr_{\mathcal{B}}(\mathcal{A}) = \Pr(\mathcal{A} \mid \mathcal{B})$ for $\mathcal{A} \in \mathcal{F}$,
  - the **law of total probability**,
  - **Bayes' theorem**, and
  - the notion of **independent events**, for which $\Pr(\mathcal{A} \cap \mathcal{B}) = \Pr(\mathcal{A})\Pr(\mathcal{B})$.

- Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space and $(\mathcal{X}, \mathcal{G})$ a measurable space. A **random function** X from Ω into $\mathcal{X}$ has the property that $X^{-1}(\mathcal{C}) = \{\omega : X(\omega) \in \mathcal{C}\} \in \mathcal{F}$ for any $\mathcal{C} \in \mathcal{G}$, so $\Pr(X \in \mathcal{C}) = \Pr\{X^{-1}(\mathcal{C})\}$ is well-defined. Such a function is called **measurable**.
- If $\mathcal{X} = \mathbb{R}$ or $\mathbb{R}^n$ we call X a **random variable** and there exists a **cumulative distribution function (CDF)** F such that $\Pr\{X \in (-\infty, x_1] \times \cdots \times (-\infty, x_n]\} = F(x_1, \ldots, x_n)$.
- A CDF increases from 0 when any of its arguments increases from $-\infty$ to $+\infty$.
- F can be written as a sum of (sub-)distributions $F_{\mathrm{ac}} + F_{\mathrm{dis}} + F_{\mathrm{sing}}$, where
  - F_{ac} is absolutely continuous, i.e., there exists a non-negative **probability density function (PDF)** $f_{\mathrm{ac}}(x) = \mathrm{d}F_{\mathrm{ac}}(x)/\mathrm{d}x$,
  - F_{dis} is discrete, i.e., its **probability mass function (PMF)** $f_{\mathrm{dis}}(x)$ is positive only on a finite or countable set $\mathcal{S}$, and
  - F_{sing} is singular, and can be ignored (look up 'Cantor distribution' if interested).
- We call X **continuous** or **discrete** respectively if F_{dis} or F_{ac} is absent.
- If necessary we use **Lebesgue–Stieltjes integration**, whereby

$$\Pr(X \in \mathcal{C}) = \int_{\mathcal{C}} \mathrm{d}F(x) = \int_{\mathcal{C}} f_{\mathrm{ac}}(x)\,\mathrm{d}x + \sum_{x \in \mathcal{C} \cap \mathcal{S}} f_{\mathrm{dis}}(x), \quad \mathcal{C} \subset \mathcal{X};$$

  the notation $\int_a^b$ is unwise because it doesn't distinguish $\mathcal{C} = [a, b]$ from $\mathcal{C} = (a, b)$.

- We define the **conditional distribution** of X given an event $\mathcal{B} \in \mathcal{F}$ by

$$\Pr(X \in \mathcal{A} \mid \mathcal{B}) = \Pr(\{X \in \mathcal{A}\} \cap \mathcal{B})/\Pr(\mathcal{B}).$$

- If $Y = g(X) \in \mathcal{Y}$ and we write $g^{-1}(\mathcal{B}) = \{x : g(x) \in \mathcal{B}\}$ for $\mathcal{B} \subset \mathcal{Y}$, then

$$\Pr(Y \in \mathcal{B}) = \Pr\{g(X) \in \mathcal{B}\} = \Pr\{X \in g^{-1}(\mathcal{B})\}.$$

- If X is continuous and $Y = g(X)$ with g a smooth bijection, then (in obvious notation)

$$f_Y(y) = f_X\{g^{-1}(y)\} \left| \frac{\partial g^{-1}(y)}{\partial y} \right|,$$

  where the last term is the Jacobian of the transformation.

- If $X = (X_1, X_2)$ is continuous, we obtain **marginal** and **conditional** densities

$$f_{X_2}(x_2) = \int f_{X_1,X_2}(x_1, x_2)\, \mathrm{d}x_1, \quad f_{X_1|X_2}(x_1 \mid x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)},$$

  with corresponding formulae in the discrete and mixed cases.

- X_1 and X_2 are **independent ($X_1 \perp\!\!\!\perp X_2$)** iff $f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$, $\forall x_1, x_2$.

Exchangeability is weaker than independence, often used to model variables that are indistinguishable in probabilistic terms, even if not independent.

Definition

Random variables $U_1, \ldots, U_n$ *are* **finitely exchangeable** *if their density satisfies*

$$f(u_1, \ldots, u_n) = f\left(u_{\xi(1)}, \ldots, u_{\xi(n)}\right)$$

for any permutation ξ *of the set* $\{1, \ldots, n\}$. *An infinite sequence* $U_1, U_2, \ldots,$ *is called* **infinitely exchangeable** *if every finite subset of it is finitely exchangeable.*

de Finetti proved that such variables must be constructed as $U_1, \ldots, U_n \mid \theta \overset{\text{iid}}{\sim} F_\theta$, where $\theta \sim G$ for distributions F_θ and G. The simplest theorem to this effect is the one below.

Theorem (de Finetti)

If $U_1, U_2, \ldots$, is an infinitely exchangeable sequence of binary variables taking values in $\{0, 1\}$, then for any n there is a distribution G such that

$$f(u_1, \ldots, u_n) = \int_0^1 \prod_{j=1}^{n} \theta^{u_j} (1-\theta)^{1-u_j} \, G(\mathrm{d}\theta) \tag{1}$$

where

$$G(\theta) = \lim_{m \to \infty} \Pr\left\{ m^{-1}(U_1 + \cdots + U_m) \leq \theta \right\}, \quad \theta = \lim_{m \to \infty} m^{-1}(U_1 + \cdots + U_m).$$

- PDFs and PMFs are not the same but we henceforth use the term **density** for both.
- $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$ means that the X_j are independent and all have density f, and we then call the $\{X_j\}_{j=1}^n$ a **random sample (of size n) from f**.
- $X_1, \ldots, X_n \overset{\text{ind}}{\sim} f_1, \ldots, f_n$ means that the X_j are independent and $X_j \sim f_j$.
- $X_1, \ldots, X_n \overset{\text{ind}}{\sim} (\mu, \sigma^2)$ means that the X_j are independent with mean μ and variance σ^2 (with $0 < \sigma^2 < \infty$). The X_j need not be normal or have the same distribution.
- $X_1, \ldots, X_n \overset{\text{ind}}{\sim} (\mu_1, \ldots, \mu_n, \sigma_1^2, \ldots, \sigma_n^2)$ means that the X_j are independent with means μ_j and variances σ_j^2 (where $0 < \sigma_j^2 < \infty$).
- The p **quantile** of the distribution F of a scalar random variable X is

$$x_p = \inf\{x : F(x) \geq p\}, \quad 0 < p < 1.$$

  Usually $x_p = F^{-1}(p)$ for continuous X, but not for discrete (or mixed) X.
- A **standard normal** variable $Z \sim \mathcal{N}(0, 1)$ has PDF and CDF

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \qquad \Phi(z) = \int_{-\infty}^{z} \phi(u)\,\mathrm{d}u, \quad z \in \mathbb{R}.$$

  and p quantile $z_p = \Phi^{-1}(p)$, so $X = \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$ has p quantile $\mu + \sigma z_p$.

- The **order statistics** of $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$ are the ordered values
$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)}.$$

- In particular, the **minimum** is $X_{(1)}$, the **maximum** is $X_{(n)}$, and the **median** is
$$X_{(m+1)} \quad (n = 2m+1, \text{ odd}), \qquad \tfrac{1}{2}(X_{(m)} + X_{(m+1)}) \quad (n = 2m, \text{ even}).$$
The median is the central value of $X_1, \ldots, X_n$.
- If f is continuous then the X_j must be distinct, and for $r = 1, \ldots, n$ we have
$$\begin{aligned} \Pr(X_{(r)} \leq x) &= \sum_{j=r}^{n} \binom{n}{j} F(x)^j \{1 - F(x)\}^{n-j}, \\ f_{X_{(r)}}(x) &= \frac{n!}{(r-1)!\,1!\,(n-r)!} F(x)^{r-1} f(x) \{1 - F(x)\}^{n-r}. \end{aligned}$$
- Joint densities can be obtained using the argument that gives $f_{X_{(r)}}(x)$, and in particular
$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = n! f(x_1) \cdots f(x_n), \quad x_1 < \cdots < x_n.$$

Example

Find the joint density of $X_{(2)}, \ldots, X_{(n-1)}$ given that $X_{(1)} = x_1$ and $X_{(n)} = x_n$.

- The **expectation** $\mathrm{E}\{g(X)\}$ of $g(X)$ is defined if $\mathrm{E}\{|g(X)|\} < \infty$ as

$$\mathrm{E}\{g(X)\} = \int_{\mathcal{X}} g(x)\,\mathrm{d}F(x).$$

- For scalar X we define **moments** $\mathrm{E}(X^r)$, **mean** $\mu = \mathrm{E}(X)$ and **variance**

$$\mathrm{var}(X) = \mathrm{E}[\{X - \mathrm{E}(X)\}^2] = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \mathrm{E}\{X(X-1)\} + \mathrm{E}(X) - \mathrm{E}(X)^2.$$

- $\mathrm{var}(X) = 0$ iff X is constant with probability one.
- For vector X we define the **mean vector** and **(co)variance matrix**

$$\mu = \mathrm{E}(X), \quad \mathrm{cov}(X_1, X_2) = \mathrm{E}(X_1 X_2^{\mathrm{T}}) - \mathrm{E}(X_1)\mathrm{E}(X_2)^{\mathrm{T}},$$

  and write $\mathrm{var}(X) = \mathrm{cov}(X, X) = E\{(X-\mu)(X-\mu)^{\mathrm{T}}\}$.
- The **correlation**, $\mathrm{corr}(X_1, X_2) = \mathrm{cov}(X_1, X_2)/\{\mathrm{var}(X_1)\mathrm{var}(X_2)\}^{1/2}$, is a measure of dependence between variables that does not depend on their units of measurement.
- Expectation $\mathrm{E}(\cdot)$ is a linear operator, so it is easy to check that

$$\mathrm{E}(a + BX) = a + B\mathrm{E}(X), \quad \mathrm{cov}(a + BX, c + DX) = B\mathrm{var}(X)D^{\mathrm{T}}.$$

- The **conditional expectation** of $g(X, Y)$ given $X = x$ is

$$\mathrm{E}\{g(X, Y) \mid X = x\} = \int_{\mathcal{Y}} g(x, y)\, \mathrm{d}F(y \mid x),$$

  which in the continuous and discrete cases equals

$$\int_{\mathcal{Y}} g(x, y) f_{Y|X}(y \mid x)\, \mathrm{d}y, \quad \sum_{y \in \mathcal{Y}} g(x, y) f_{Y|X}(y \mid x),$$

  and other conditional moments are defined likewise.
- This is a function of X, so it defines a random variable $\tilde{g}(X) = \mathrm{E}\{g(X, Y) \mid X\}$.
- The **law of total expectation (tower property)** gives

$$\begin{aligned} \mathrm{E}\{g(X, Y)\} &= \mathrm{E}_X\left[\mathrm{E}\{g(X, Y) \mid X = x\}\right], \\ \mathrm{var}\{g(X, Y)\} &= \mathrm{E}_X\left[\mathrm{var}\{g(X, Y) \mid X = x\}\right] + \mathrm{var}_X\left[E\{g(X, Y) \mid X = x\}\right], \end{aligned}$$

  where E_X denotes expectation with respect to the marginal distribution of X, etc., with a similar expression for $\mathrm{cov}\{g(X, Y), h(X, Y)\}$(a good ex. to work through).
- We ignore mathematical issues arising from conditioning on events of probability zero — look up 'Borel–Kolmogorov paradox' if interested.

A random variable $X_{n\times 1}$ with real components has the **multivariate normal distribution,** $X \sim \mathcal{N}_n(\mu, \Omega)$**,** if $a^{\mathrm{T}}X \sim \mathcal{N}(a^{\mathrm{T}}\mu, a^{\mathrm{T}}\Omega a)$ for every constant vector $a_{n\times 1}$, and then

- $M_Y(t) = \exp(t^{\mathrm{T}}\mu + \frac{1}{2}t^{\mathrm{T}}\Omega t)$ and the mean vector and covariance matrix of X are

$$\mathrm{E}(X) = \mu_{n\times 1}, \quad \mathrm{var}(X) = \Omega_{n\times n},$$

  where Ω is symmetric semi-positive definite with real components;

- for any constants $a_{m\times 1}$ and $B_{m\times n}$,

$$a + BX \sim \mathcal{N}_m\left(a + B\mu, B\Omega B^{\mathrm{T}}\right);$$

- $a + BX$ and $c + DX$ are independent iff $B\Omega D^{\mathrm{T}} = 0$;
- X has a density on $\mathbb{R}^n$ iff Ω is positive definite (i.e., has rank n), and then

$$f(x; \mu, \Omega) = \frac{1}{(2\pi)^{n/2}|\Omega|^{1/2}} \exp\left\{-\tfrac{1}{2}(x-\mu)^{\mathrm{T}}\Omega^{-1}(x-\mu)\right\}, \quad x \in \mathbb{R}^n; \tag{2}$$

- if $X^{\mathrm{T}} = (X_1^{\mathrm{T}}, X_2^{\mathrm{T}})$, where X_1 is $m \times 1$, and μ and Ω are partitioned correspondingly, then the marginal and conditional distributions of X_1 are also multivariate normal:

$$X_1 \sim \mathcal{N}_m(\mu_1, \Omega_{11}), \quad X_1 \mid X_2 = x_2 \sim \mathcal{N}_m\left\{\mu_1 + \Omega_{12}\Omega_{22}^{-1}(x_2 - \mu_2), \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\right\}.$$

- The **moment-generating function (MGF)** and **cumulant-generating function (KGF)** of a scalar random variable X are

$$M_X(t) = \mathrm{E}\left(e^{tX}\right), \quad K_X(t) = \log M_X(t), \quad t \in \mathcal{N} = \{t : M_X(t) < \infty\}.$$

- $\mathcal{N}$ is non-empty, because $M_X(0) = 1$, but the MGF and KGF are non-trivial only if $\mathcal{N}$ contains an open neighbourhood of the origin, since then

$$M_X(t) = \mathrm{E}\left(\sum_{r=0}^{\infty} \frac{t^r X^r}{r!}\right) = \sum_{r=0}^{\infty} \frac{t^r}{r!}\mathrm{E}(X^r), \quad K_X(t) = \sum_{r=1}^{\infty} \frac{t^r}{r!}\kappa_r,$$

  and one can obtain the **moments** $\mathrm{E}(X^r)$ and **cumulants** κ_r by differentiation.

- In the vector case we define

$$M_X(t) = \mathrm{E}\left(e^{t^{\mathrm{T}} X}\right), \quad K_X(t) = \log M_X(t),$$

  and differentiation with respect to the elements of $t = (t_1, \ldots, t_n)^{\mathrm{T}}$ gives the mean vector and covariance matrix of X.

- There is a 1–1 mapping between distributions and MGFs/KGFs (if the latter are non-trivial).

- KGFs for linear combinations are computed as $K_{a+BX}(t) = a^{\mathrm{T}} t + K_X(B^{\mathrm{T}} t)$.

- A baseline density f_0 with a non-trivial MGF can be used to construct a family of densities by **exponential tilting**, i.e.,

$$f(y;\varphi) = f_0(y)\exp\left\{\varphi^{\mathrm{T}} s(y) - k(\varphi)\right\}, \quad y \in \mathcal{Y}, \varphi \in \mathcal{N},$$

  where

$$\mathcal{N} = \{\varphi : k(\varphi) < \infty\}$$

  and individual members of the family are determined by the value of φ.

- Hölder's inequality gives

$$M\{\alpha\varphi_1 + (1-\alpha)\varphi_2\} \leq M(\varphi_1)^{\alpha} M(\varphi_2)^{1-\alpha} < \infty, \quad 0 \leq \alpha \leq 1,$$

  for any $\varphi_1, \varphi_2 \in \mathcal{N}$, so the set $\mathcal{N}$ and the function k are both convex.

- This implies that $f(y;\varphi)$ is log-concave in φ, which is a very useful property in statistics.

- This construction leads to an elegant general theory putting many well-known distributions (Poisson, binomial, normal, ...) under the same roof.

- If $\theta \in \Theta \subset \mathbb{R}^d$, where $\dim \Theta = d$, and there exists a $d \times 1$ function $s = s(y)$ of data y and a **parametrisation** (i.e., a 1–1 function) $\varphi \equiv \varphi(\theta)$ such that

$$f(y;\theta) = m(y) \exp\left\{s^{\mathrm{T}}\varphi - k(\varphi)\right\} = m(y) \exp\left[s^{\mathrm{T}}\varphi(\theta) - k\{\varphi(\theta)\}\right], \quad \theta \in \Theta, y \in \mathcal{Y},$$

  then this is an (d, d) **exponential family** of distributions, with

  - **canonical statistic** $S = s(Y)$,
  - **canonical parameter** φ,
  - **cumulant generator** k, which is convex on $\mathcal{N} = \{\varphi : k(\varphi) < \infty\}$, and
  - **mean parameter** $\mu \equiv \mu(\varphi) = \mathrm{E}(S;\varphi) = \nabla k(\varphi)$, where $\nabla \cdot = \partial \cdot / \partial \varphi$.

- We suppose that there is no vector $a \neq 0$ such that $a^{\mathrm{T}}S$ is constant, and call the model a **minimal representation** if there is no vector $a \neq 0$ such that $a^{\mathrm{T}}\varphi$ is constant.

- The cumulant-generating function for S is

$$K_S(t) = \log M_S(t) = k(\varphi + t) - k(\varphi), \quad t \in \mathcal{N}' \subset \mathbb{R}^d,$$

  where $0 \in \mathcal{N}'$. On writing $\nabla^2 \cdot = \partial^2 \cdot / \partial\varphi\partial\varphi^{\mathrm{T}}$, one can check that

$$\mathrm{E}(S) = \nabla k(\varphi), \quad \mathrm{var}(S) = \nabla^2 k(\varphi).$$

Example (Poisson sample)

If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Poiss}(\theta)$, find the corresponding exponential family.

Example (Satellite conjunction)

A simple model for the position Y of a satellite in $\mathbb{R}^2$ relative to the origin is

$$Y \sim \mathcal{N}_2 \left\{ \begin{pmatrix} \psi \cos\lambda \\ \psi \sin\lambda \end{pmatrix}, \begin{pmatrix} d_1^{-1} & 0 \\ 0 & d_2^{-1} \end{pmatrix} \right\},$$

where $d_1, d_2 > 0$ are known and $\psi > 0$, $0 < \lambda \leq 2\pi$. Write the corresponding density

$$f(y_1, y_2; \psi, \lambda) = \frac{(d_1 d_2)^{1/2}}{2\pi} \exp\left[-\tfrac{1}{2} \left\{ d_1 (y_1 - \psi\cos\lambda)^2 + d_2 (y_2 - \psi \sin\lambda)^2 \right\} \right], \quad y_1, y_2 \in \mathbb{R},$$

as an exponential family.

- **NB:** avoid confusion — exponential family $\neq$ exponential distribution! The exponential distribution is just one example of an exponential family.

- When $\dim s = d' > \dim \theta = d$ the model is called a (d', d) **curved exponential family**, and the $d' \times 1$ vector $\varphi(\theta)$ gives a d-dimensional sub-manifold of $\mathbb{R}^{d'}$.
- Exponential families are **closed under sampling**: the joint density of independent observations $Y_1, \ldots, Y_n$ from an exponential family with the same $s(Y_j)^{\mathrm{T}}\varphi = S_j^{\mathrm{T}}\varphi$ is

$$\prod_{j=1}^{n} f(y_j;\theta) = \prod_{j=1}^{n} m(y_j) \exp\left\{s_j^{\mathrm{T}}\varphi - k_j(\varphi)\right\} = \prod_{j=1}^{n} m(y_j) \exp\left\{\left(\sum_{j=1}^{n} s_j\right)^{\mathrm{T}} \varphi - \sum_{j=1}^{n} k_j(\varphi)\right\},$$

so with $k_S(\varphi) = \sum_j k_j(\varphi)$, the density of $S = \sum_j S_j = \sum_j s(Y_j)$ is

$$f(s;\theta) = m^*(s) e^{s^{\mathrm{T}}\varphi - k_S(\varphi)}, \quad \text{with} \quad m^*(s) = \int_{\left\{y : \sum_j s(y_j) = s\right\}} \prod_{j=1}^{n} m(y_j)\, \mathrm{d}y.$$

This is an exponential family, with canonical statistic S, canonical parameter φ and cumulant generator $k_S(\varphi)$.

Example (Satellite conjunction)

Show that taking ψ known in Example 4 gives a $(2, 1)$ exponential family.

- A real-valued **convex function** g defined on a vector space $\mathcal{V}$ has the property that for any $x, y \in \mathcal{V}$,
$$g\{tx + (1-t)y\} \leq tg(x) + (1-t)g(y), \quad 0 \leq t \leq 1.$$
Equivalently, for all $y \in \mathcal{V}$, there exists a vector $b(y)$ such that
$$g(x) \geq g(y) + b(y)^{\mathrm{T}}(x-y)$$
for all x. If $g(x)$ is differentiable, then we can take $b(y) = g'(y)$.
- If X is a random variable, $a > 0$ a constant, h a non-negative function and g a convex function, then
$$\begin{aligned} \Pr\{h(X) \geq a\} &\leq \mathrm{E}\{h(X)\}/a, \quad \textbf{(basic inequality)} \\ \Pr(|X| \geq a) &\leq \mathrm{E}(|X|)/a, \quad \textbf{(Markov's inequality)} \\ \Pr(|X| \geq a) &\leq \mathrm{E}(X^2)/a^2, \quad \textbf{(Chebyshev's inequality)} \\ \mathrm{E}\{g(X)\} &\geq g\{\mathrm{E}(X)\}. \quad \textbf{(Jensen's inequality)} \end{aligned}$$
- On replacing X by $X - \mathrm{E}(X)$, Chebyshev's inequality gives
$$\Pr\{|X - \mathrm{E}(X)| \geq a\} \leq \mathrm{var}(X)/a^2.$$

- Let $X, X_1, X_2, \ldots$ have CDFs $F, F_1, F_2, \ldots$ and let $\varepsilon > 0$ be arbitrary. Then
  - X_n converges to X **almost surely**, $X_n \xrightarrow{\text{a.s.}} X$, if $\Pr(\lim_{n\to\infty} X_n = X) = 1$;
  - X_n converges to X **in probability**, $X_n \xrightarrow{P} X$, if $\lim_{n\to\infty} \Pr(|X_n - X| > \varepsilon) = 0$;
  - X_n converges to X **in distribution**, $X_n \xrightarrow{D} X$, if $\lim_{n\to\infty} F_n(x) = F(x)$ at each point x where $F(x)$ is continuous.
  - A sequence $X_1, X_2, \ldots$ of estimators of a parameter θ is **strongly consistent** if $X_n \xrightarrow{\text{a.s.}} \theta$ and **(weakly) consistent** if $X_n \xrightarrow{P} \theta$.
- $\xrightarrow{\text{a.s.}}$ and $\xrightarrow{P}$, but not $\xrightarrow{D}$, require joint distributions of (X_n, X) for every n.
- Let x_0, y_0 be constants, $X, Y, \{X_n\}, \{Y_n\}$ be random variables and $g(\cdot)$ and $h(\cdot,\cdot)$ continuous functions. Then

$$
\begin{aligned}
X_n \xrightarrow{\text{a.s.}} X \quad &\Rightarrow \quad X_n \xrightarrow{P} X \;\Rightarrow\; X_n \xrightarrow{D} X, \\
X_n \xrightarrow{D} x_0 \quad &\Rightarrow \quad X_n \xrightarrow{P} x_0, \\
X_n \xrightarrow{\text{a.s.}} X \quad &\Rightarrow \quad g(X_n) \xrightarrow{\text{a.s.}} g(X), \\
X_n \xrightarrow{D} X \text{ and } Y_n \xrightarrow{D} y_0 \quad &\Rightarrow \quad h(X_n, Y_n) \xrightarrow{D} h(X, y_0).
\end{aligned}
$$

  The last two lines are called the **continuous mapping theorem** (usually used with $\xrightarrow{P}$) and **Slutsky's theorem**.

Theorem (Weak law of large numbers (WLLN))

If $X, X_1, X_2, \ldots \overset{\text{iid}}{\sim} F$ *and* $\mathrm{E}(X)$ *is finite, then* $\overline{X} = n^{-1}(X_1 + \cdots + X_n) \xrightarrow{P} \mathrm{E}(X)$.

Theorem (Strong law of large numbers (SLLN))

If $X, X_1, X_2, \ldots \overset{\text{iid}}{\sim} F$ *and* $\mathrm{E}(X)$ *is finite, then* $\overline{X} = n^{-1}(X_1 + \cdots + X_n) \xrightarrow{\text{a.s.}} \mathrm{E}(X)$.

Theorem (Central limit theorem (CLT, Lindeberg-Levy))

If $X_1, X_2, \ldots \overset{\text{iid}}{\sim} (\mu, \sigma^2)$ *and* $0 < \sigma^2 < \infty$, *then*

$$Z_n = \frac{n^{1/2}(\overline{X} - \mu)}{\sigma} \xrightarrow{D} Z \sim \mathcal{N}(0, 1), \quad n \to \infty.$$

Theorem (Delta method)

If $a_n(X_n - \mu) \xrightarrow{D} Y$, $a_n, \mu \in \mathbb{R}$, $a_n \to \infty$ *as* $n \to \infty$, *and* g *is continuously differentiable at* μ *with* $g'(\mu) \neq 0$, *then* $a_n\{g(X_n) - g(\mu)\} \xrightarrow{D} g'(\mu)Y$.

- An **estimator** of a parameter $\theta \in \Theta$ based on data Y is a random variable $\tilde{\theta} = \tilde{\theta}(Y)$ taking values in Θ. A specific value is an **estimate** $\tilde{\theta}(y)$.
- An **M**(aximisation)-**estimator** is computed using a function $\rho(y; \theta)$ as

$$\tilde{\theta} = \text{argmax}_\theta \frac{1}{n} \sum_{j=1}^{n} \rho(Y_j; \theta).$$

  Often $\tilde{\theta}$ can be identified by solving

$$\frac{1}{n} \sum_{j=1}^{n} \nabla \rho(Y_j; \theta) = 0$$

  and is then called a **Z**(ero)-**estimator**.
- Equivalently we could minimise the **loss function** $-\rho$ with respect to θ.
- If the true underlying model g is known, then $\tilde{\theta}$ is replaced by θ_g, where

$$\theta_g = \text{argmax}_\theta \int \rho(y; \theta) g(y) \, \mathrm{d}y, \quad \int \nabla \rho(y; \theta_g) g(y) \, \mathrm{d}y = 0.$$

  Clearly if $g(y) = f(y; \theta)$, then we want $\theta_g = \theta$, uniquely.

- Some examples (for a d-dimensional parameter θ):
  - **maximum likelihood estimation** (MLE) has $\rho(y;\theta) = \log f(y;\theta)$;
  - **method of moments estimation** (MoM/MM) has $h(y) = (y, y^2, \ldots, y^d)^{\mathrm{T}}$, $\mu(\theta) = \mathrm{E}\{h(Y)\}$, and
$$-\rho(y;\theta) = \{h(y) - \mu(\theta)\}^{\mathrm{T}}\{h(y) - \mu(\theta)\};$$
  - **generalized method of moments estimation** (GMM) (widely used in econometrics) also has a symmetric positive definite $d \times d$ matrix $w(\theta)$ and
$$-\rho(y;\theta) = \{h(y) - \mu(\theta)\}^{\mathrm{T}} w(\theta)\{h(y) - \mu(\theta)\};$$
  - **least squares estimation** (LS) is method of moments estimation with $h(y_j) = y_j$ and $\mu_j(\theta) = \mathrm{E}(Y_j) = x_j^{\mathrm{T}}\theta$;
  - **score-matching estimation** with $Y \sim g$ has
$$-\rho(y;\theta) = \|\nabla_y \log f(y;\theta) - \nabla_y \log g(y)\|_2^2.$$
- There are many (many!) other approaches to estimation.

Example

Discuss maximum likelihood estimation of the parameters of the normal distribution.

Example

Discuss moment estimation of the parameters of the Weibull distribution.

Example

Show that under mild (but not entirely trivial) conditions on the density g, the population version of the score-matching estimator is

$$\operatorname{argmin}_\theta \mathrm{E}\left[\{\nabla_y \log f(Y;\theta)\}^2 + 2\nabla_y^2 \log f(Y;\theta)\right],$$

and give the sample version.

- There are two generic bases for comparing point estimators:
  - **asymptotic** — what happens when $n \to \infty$?
  - **finite-sample** — what happens for sample sizes in practice ($n < \infty$)?
- **Consistency** is a key asymptotic criterion: does $\tilde{\theta}$ approach θ_g when $n \to \infty$?

Definition

An estimator $\tilde{\theta}$ of θ_g is **(weakly) consistent** *if $\tilde{\theta} \xrightarrow{P} \theta_g$ as $n \to \infty$.*

- Consistency is necessary but not sufficient for an estimator to be good, because

$$\tilde{\theta} \xrightarrow{P} \theta_g \quad \Rightarrow \quad \tilde{\theta}^* = \tilde{\theta} + 10^6 / \sqrt{\log \log n} \xrightarrow{P} \theta_g, \quad n \to \infty,$$

  but $\tilde{\theta}^*$ is (probably) useless: consistency can be considered a 'safety net'.
- Obviously we would like $\tilde{\theta}$ to be 'suitably close' to θ_g, by minimising

$$\mathrm{MSE}(\tilde{\theta}; \theta_g) = \mathrm{E}\left\{(\tilde{\theta} - \theta_g)^2\right\}, \quad \mathrm{MAD}(\tilde{\theta}; \theta_g) = \mathrm{E}\left(|\tilde{\theta} - \theta_g|\right),$$

  or other measures of distance (loss functions), asymptotically or in finite samples.

- Using the **bias** $b(\tilde{\theta}; \theta_g) = \mathrm{E}(\tilde{\theta}) - \theta_g$, the **mean square error** can be expressed as

$$\mathrm{MSE}(\tilde{\theta}; \theta_g) = b(\tilde{\theta}; \theta_g)^2 + \mathrm{var}(\tilde{\theta}),$$

  so we must balance ('trade off') the bias and the variance when choosing $\tilde{\theta}$.

- In simple problems we could insist that the estimator is **unbiased**, i.e., $b(\tilde{\theta}; \theta_g) \equiv 0$, but this is usually artificial because
  - many good estimators are biased, and some unbiased estimators are useless;
  - it may be impossible to find an unbiased estimator; and
  - other properties may be more desirable (e.g., robustness).

  An exception is **meta-analysis**, which involves combining different estimators with possibly very varied sample sizes, in which case we want them to estimate the same thing!

Example

The method of moments estimator of a scalar θ based on a random sample $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} (\mu, \sigma^2)$ with sample average $\overline{Y}$ solves the equation $\mu(\theta) = \overline{Y}$. Show that if $\mu(\cdot)$ has two smooth derivatives and is 1–1, then the estimator is consistent and asymptotically normal, with bias and variance both of order n^{-1}.

Definition

If $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are estimators of scalar θ, then the **relative efficiency** *of $\tilde{\theta}_1$ compared to $\tilde{\theta}_2$ can be defined as*

$$\frac{\mathrm{MSE}(\tilde{\theta}_2;\theta)}{\mathrm{MSE}(\tilde{\theta}_1;\theta)}.$$

In large samples the squared bias is often negligible compared to the variance, and we define the **asymptotic relative efficiency** *as $\mathrm{var}(\tilde{\theta}_2)/\mathrm{var}(\tilde{\theta}_1)$. Similar expressions apply if the parameter has dimension d.*

- Under mild conditions on the underlying model, a scalar estimator $\tilde{\theta}$ based on $Y \sim f(y;\theta)$ satisfies the **Cramèr–Rao lower bound**,

$$\mathrm{var}(\tilde{\theta}) \geq \frac{\{1+\nabla b(\tilde{\theta};\theta)\}^2}{\imath(\theta)},$$

  where $\imath(\theta)$ is defined on the next slide. This bound applies for any sample size n. Moreover
  - as $n \to \infty$ the lower bound $\to 1/\imath(\theta)$, the asymptotic variance of the maximum likelihood estimator, which hence is most efficient in large samples; and
  - a similar result applies for vector θ.

- For data $Y \sim f(y;\theta)$ we define the **log likelihood function** $\ell(\theta) = \log f(Y;\theta)$ and $d \times 1$ **score vector** $U(\theta) = \nabla\ell(\theta)$.
- If we can differentiate with respect to θ under the integral sign, we get the **Bartlett identities**:

$$\begin{aligned} 0 &= \int \nabla \log f(y;\theta) \times f(y;\theta)\,\mathrm{d}y, \\ 0 &= \int \nabla^2 \log f(y;\theta) \times f(y;\theta)\,\mathrm{d}y + \int \nabla \log f(y;\theta)\,\nabla^{\mathrm{T}} \log f(y;\theta) \times f(y;\theta)\,\mathrm{d}y, \\ 0 &= \cdots \end{aligned}$$

  giving the moments of $U(\theta)$, viz

$$\mathrm{E}\{U(\theta)\} = 0, \quad \mathrm{var}\{U(\theta)\} = \mathrm{E}\left\{\nabla\ell(\theta)\nabla^{\mathrm{T}}\ell(\theta)\right\} = \mathrm{E}\left\{-\nabla^2\ell(\theta)\right\}, \quad \ldots$$

  where $\mathrm{var}\{U(\theta)\} = \imath(\theta)$ is the $d \times d$ **Fisher (or expected) information** matrix.
- We write $\imath_1(\theta)$ for the Fisher information for a single observation of a random sample $Y_1, \ldots, Y_n$, and then that in the sample is $\imath(\theta) = n\imath_1(\theta)$.
- Later we shall see that in large samples, the maximum likelihood estimator $\widehat{\theta}$ satisfies

$$\widehat{\theta} \sim \mathcal{N}_d\left\{\theta, \imath(\theta)^{-1}\right\}.$$

- Point estimation does not express uncertainty — we need to assess how well the observed data y^{o} support different possible values of a parameter.
- We aim to find subsets of the parameter space that contain the 'true' parameter with a specified probability — when the parameter of interest is scalar, these subsets are usually intervals.
- Pivots are useful in finding such subsets.

Definition

If Y has density $f(y;\theta)$, then a **pivot (or pivotal quantity)** *$Q = q(Y,\theta)$ is a function of Y and θ that has a known distribution (i.e., one that does not depend on θ).*

Example

If $M = \max(Y_1, \ldots, Y_n)$, where $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} U(0,\theta)$, show that $Q_1 = M/\theta$ is a pivot and find a pivot based on $\overline{Y}$.

Definition

Let $Y = (Y_1, \ldots, Y_n)$ be data from a parametric statistical model with scalar parameter θ. A **confidence interval (CI)** *(L, U)* **for** *θ with lower confidence bound L and upper confidence bound U is a random interval that contains θ with a specified probability, called the* **(confidence) level** *of the interval.*

- $L = l(Y)$ and $U = u(Y)$ are computed from the data. They do not depend on θ.
- In a continuous setting (so $<$ gives the same probabilities as $\leq$), and if we write the probabilities that θ lies below and above the interval as

$$\Pr(\theta < L) = \alpha_L, \quad \Pr(U < \theta) = \alpha_U,$$

then (L, U) has confidence level

$$\Pr(L \leq \theta \leq U) = 1 - \Pr(\theta < L) - \Pr(U < \theta) = 1 - \alpha_L - \alpha_U.$$

- Often we seek an interval with equal probabilities of not containing θ at each end, with $\alpha_L = \alpha_U = \alpha/2$, giving an **equi-tailed** $(1 - \alpha) \times 100\%$ **confidence interval**.
- We often take standard values of α, such that $1 - \alpha = 0.9, 0.95, 0.99, \ldots$
- A weaker requirement is $\Pr(L \leq \theta \leq U) \geq 1 - \alpha$, giving confidence level *at least* $1 - \alpha$.

- We use pivots to construct CIs:
  - find a pivot $Q = q(Y, \theta)$;
  - obtain the quantiles q_{α_U}, $q_{1-\alpha_L}$ of Q;
  - then transform the equation

$$\Pr\{q_{\alpha_U} \leq q(Y, \theta) \leq q_{1-\alpha_L}\} = (1 - \alpha_L) - \alpha_U$$

    into the form

$$\Pr(L \leq \theta \leq U) = 1 - \alpha_L - \alpha_U,$$

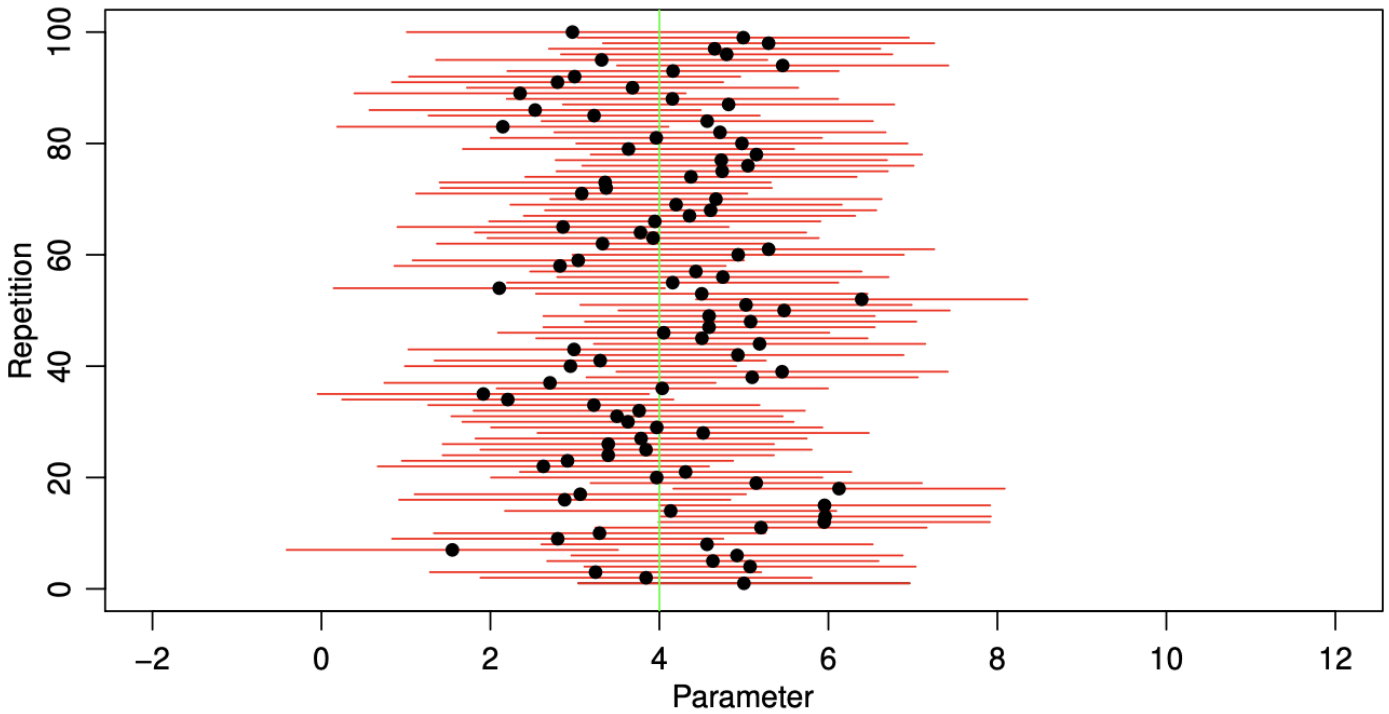
    where the bounds $L = l(Y; \alpha_L, \alpha_U)$, $U = u(Y; \alpha_L, \alpha_U)$ do not depend on θ;
  - then replace Y by its observed value y^{o} to get a realisation of the CI.
- Going from quantiles of Q to L, U is known as **inverting the pivot** — it is convenient if Q is monotone in θ for each Y.
- Often we have an approximate pivot $(\widehat{\theta} - \theta)/V^{1/2} \dot{\sim} \mathcal{N}(0, 1)$, where V estimates $\text{var}(\widehat{\theta})$ and $V^{1/2}$ is called a **standard error**. The resulting (approximate) 95% interval is $\widehat{\theta} \pm 1.96 V^{1/2}$.

Example

In Uniform distribution example on slide 36, find CIs based on Q_1 and on Q_2.

Interpretation of a CI

- (L, U) is a random interval that contains θ with probability $1 - \alpha$.
- We imagine an infinity of possible datasets from the experiment that resulted in (L, U).
- Our CI based on y° is regarded as randomly chosen from the resulting infinity of CIs.
- Although we do not know if $\theta \in (l(y^{\circ}; \alpha_L, \alpha_U), u(y^{\circ}; \alpha_L, \alpha_U))$, the event $\theta \in (L, U)$ has probability $1 - \alpha$ across these datasets.
- In the figure below, the parameter θ (green line) is contained (or not) in realisations of the 95% CI (red). The black points show the corresponding estimates.



- Almost invariably CIs are **two-sided** and **equi-tailed**, i.e., $\alpha_L = \alpha_U = \alpha$, but **one-sided** CIs of form $(-\infty, U)$ or (L, ∞) are sometimes required:
  - compute a two-sided interval with $\alpha_L = \alpha_U = \alpha$, then replace the unwanted limit by $\pm\infty$ (or another value if required in the context).
- For a two-sided CI we define the **lower- and upper-tail errors**

$$\Pr(\theta < L), \quad \Pr(U < \theta)$$

  and if these equal the required value for each possible α_L, α_U, then the **empirical coverage** of the CI exactly equals the desired value:
  - this occurs when the distribution of the corresponding pivot is known, but in practice this distribution is usually approximate, and then we use simulation to assess if and when CIs are adequate;
  - these errors are properties of the CI procedure, not of individual intervals!

- Prediction refers to 'estimation' of unobserved (future, latent, ...) random variables Y_+.
- In parametric cases we often base **prediction (or tolerance) intervals** on existing data Y by finding a pivot that depends on both Y_+ and Y, and predicting Y_+ using this pivot, e.g., using its mean or median.

Example

If $Y_1, \ldots, Y_n, Y_+ \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, give prediction limits and a predictor for Y_+ based on the other variables.

- A **statistical hypothesis** is an assertion about the population underlying some data, or equivalently a restriction on possible models for the data, such as:
  - the population has mean μ_0;
  - the population is $\mathcal{N}(\mu_0, \sigma_0^2)$, with both parameters specified;
  - the population is $\mathcal{N}(\mu, \sigma^2)$, with the parameters unspecified;
  - the data are sampled from the discrete uniform distribution on $\{1, \ldots, 9\}$;
  - the population density is symmetric about some μ;
  - the population mean $\mu(x)$ increases when a covariate x increases.
- These are assertions about populations, not about data, but they have implications for data.
- Sometimes the distribution is fully specified, but not always.
- Some, but not all, hypotheses concern parameters.
- A **hypothesis test** uses a stochastic *argument by contradiction* to make an inference about a statistical hypothesis: we assume that the hypothesis is true, and attempt to use our data to disprove it.

- A **null hypothesis** H_0 to be tested (implicitly defines and **alternative hypothesis** H_a or H_1).
- A **test statistic** T, large values of which suggest that H_0 is false, and with observed value t_{obs}.
- A **P-value**
$$p_{\text{obs}} = \Pr_0(T \geq t_{\text{obs}}),$$
where the **null distribution** $\Pr_0(\cdot)$ denotes a probability computed under H_0.
- The smaller p_{obs} is, the more we doubt that H_0 is true.
- Tests on parameters are often based on pivots: if $\theta = \theta_0$, then $T = |q(Y; \theta_0)|$ has a known distribution G_0, say, and observing a value $t_{\text{obs}} = |q(y^{\text{o}}; \theta_0)|$ that is unusual relative to G_0 'contradicts' H_0.
- In other cases we choose a test statistic that seems plausible, such as Pearson's statistic,
$$T = \sum_{k=1}^{K} (O_k - E_k)^2 / E_k,$$
used to check whether observed counts O_k in K categories agree with their expectations $E_k = \mathrm{E}(O_k)$ computed under H_0.
- In any case we need to know (or be able to approximate) the distribution of T under H_0.

- Essentially three bases for statements of uncertainty:
  - a **frequentist (sampling theory) inference** compares y with a set $\mathcal{S} \subset \mathcal{Y}$ of other data that might have been observed in a hypothetical sampling experiment;
  - a **Bayesian (inverse probability) inference** expresses uncertainty via a prior probability density and uses Bayes' theorem to update this in light of the data;
  - in a designed experiment, clinical trial, sample survey or similar the investigator uses **randomisation** to generate a distribution against which y is compared.
- There are many variants of the first two approaches.

Example (Measuring machines)

A physical quantity θ can be measured with two machines, both giving normal observations $Y \sim \mathcal{N}(\theta, \sigma_m^2)$. A measurement from machine 1 has variance $\sigma_1^2 = 1$, and one from machine 2 has variance $\sigma_2^2 = 100$. A machine is chosen by tossing a fair coin, giving $M = 1, 2$ with equal probabilities. Thus $\mathcal{Y} = \{(y, m) : y \in \mathbb{R}, m \in \{1, 2\}\}$.
If we observe $(y, m) = (0, 1)$, then clearly we can ignore the fact that we might have observed $m = 2$, i.e., we should take $\mathcal{S}_1 = \{(y, 1) : y \in \mathbb{R}\}$ rather than $\mathcal{S}_2 = \{(y, 2) : y \in \mathbb{R}\}$ or $\mathcal{S} = \mathcal{Y}$.

- We assume that y° is just one of many possible datasets $y \in \mathcal{S}$ that might have been generated from $f(y;\theta)$, and the probability calculations are performed with respect to $\mathcal{S}$.
- We choose $\mathcal{S}$ to ensure that the probability calculation is **relevant** to the data actually observed. For example, if y° has n observations, we usually insist that every element of $\mathcal{S}$ also has n observations.
- The repeated sampling principle ensures that (if we use an exact pivot) inferences are **calibrated**, for example, a $(1-\alpha)$ confidence interval (L, U) satisfies

$$\Pr(L < \theta \leq U) = 1 - \alpha,$$

for every $\theta \in \Theta$ and every $\alpha \in (0, 1)$. Hence if such intervals are used infinitely often, then

  - although any particular interval either does or does not contain θ,
  - it was drawn from a population of intervals with error probability exactly α.

Example

What would the confidence intervals look like in the example on slide 44? How would the image on slide 39 change? What hypothetical repetitions form the reference sets?

- Our observed data y^{o} are assumed to be a realisation from a density $f(y \mid \theta)$.
- If we can summarise information about θ, separately from y^{o}, in a **prior density** $f(\theta)$, then we base all our uncertainty statements on the **posterior density** given by Bayes' theorem,

$$f(\theta \mid y^{\mathrm{o}}) = \frac{f(y^{\mathrm{o}} \mid \theta) f(\theta)}{\int f(y^{\mathrm{o}} \mid \theta) f(\theta) \, \mathrm{d}\theta}.$$

- For example, if θ_p satisfies $\Pr(\theta \leq \theta_p \mid y^{\mathrm{o}}) = p$ for any $p \in (0, 1)$, we could give a $(1 - 2\alpha)$ **posterior credible interval** $\mathcal{I}_{1-2\alpha} = (\theta_\alpha, \theta_{1-\alpha})$ such that

$$\Pr(\theta \in \mathcal{I}_{1-\alpha} \mid y^{\mathrm{o}}) = 1 - 2\alpha;$$

  here θ is regarded as random and y^{o} as fixed.
- A point estimate $\tilde{\theta}(y^{\mathrm{o}})$ of θ is obtained by minimising a **posterior expected loss**, i.e.,

$$\tilde{\theta}(y^{\mathrm{o}}) = \operatorname{argmin}_{\tilde{\theta}} \mathrm{E}\left\{ L(\theta, \tilde{\theta}) \mid y^{\mathrm{o}} \right\} = \operatorname{argmin}_{\tilde{\theta}} \int L(\theta, \tilde{\theta}) f(\theta \mid y^{\mathrm{o}}) \, \mathrm{d}\theta,$$

  where the **loss function** $L(\theta, \tilde{\theta}) \geq 0$ measures the loss when θ is estimated by $\tilde{\theta}$.

- Often Bayesian models are formulated using a judgement that some variables/observations are exchangeable, as de Finetti theorems then imply that we can write

$$Y_1, \ldots, Y_n \mid \theta \overset{\text{iid}}{\sim} f(y; \theta), \qquad \theta \sim f(\theta).$$

- In general, Bayesian inference
  - requires the specification of a prior distribution on unknowns, separate from the data;
  - implies that we regard prior information as equivalent to data, putting uncertainty and variation on the same footing;
  - reduces inference to computation of probabilities, so *in principle* is simple and direct.
- Objectively specifying prior 'ignorance' is problematic and can lead to paradoxes, especially in high dimensions.
- (Approximate) Bayesian computation can be performed using
  - conjugate prior distributions (exact computations in simple cases),
  - integral approximations (e.g., Laplace's method),
  - deterministic methods (e.g., variational approximation),
  - simulation, especially Markov chain Monte Carlo.

- To compare how **treatments** affect a **response**, they are **randomised** to experimental **units**:
  - **treatments** are clearly-defined procedures, one of which is applied to each unit;
  - a **unit** is the smallest division of the raw material such that two different units might receive two different treatments;
  - the **response** is a well-defined variable measured for each unit-treatment combination.
- Examples are agricultural trials, industrial experiments, clinical trials, ...
- The experiment is 'under the control' of the investigator, making strong inferences possible.
- Main goals of randomisation:
  - avoidance of systematic error (eliminating bias);
  - estimation of baseline variation (e.g., by use of replication and/or blocking);
  - realistic statement of uncertainty of final conclusions;
  - providing a basis for exact inferences using the randomisation distribution.

- Shoe wear in an paired comparison experiment in which materials A (expensive) and B (cheaper) were randomly assigned to the soles of the left (L) or right (R) shoe of each of $m = 10$ boys.
- A unit is a foot, a treatment is the type of sole, and the response is the amount of wear.
- The $m = 10$ differences $d_1, \ldots, d_m$ have average $\overline{d} = 0.41$.

Boy	Material		Difference
	A	B	d
1	13.2 (L)	14.0 (R)	0.8
2	8.2 (L)	8.8 (R)	0.6
3	10.9 (R)	11.2 (L)	0.3
4	14.3 (L)	14.2 (R)	–0.1
5	10.7 (R)	11.8 (L)	1.1
6	6.6 (L)	6.4 (R)	–0.2
7	9.5 (L)	9.8 (R)	0.3
8	10.8 (L)	11.3 (R)	0.5
9	8.8 (R)	9.3 (L)	0.5
10	13.3 (L)	13.6 (R)	0.3

- This is **paired comparison** experiment, as there are **blocks** of two similar units, each of which is given one treatment at random, according to the scheme

Treatment for boy j	Left foot	Right foot
A	l_j	r_j
B	$\theta + l_j$	$\theta + r_j$

- We observe either $(\theta + l_j, r_j)$ or $(l_j, r_j + \theta)$ so the difference D_j of B and A for boy j is $\theta + l_j - r_j$ or $\theta + r_j - l_j$. These are equally likely, so we can write $D_j = \theta + t_j c_j$, where
  - θ is the unknown (extra wear) effect of B compared to A,
  - $t_j = 1$ if the left shoe of boy j has material B and otherwise equals -1, and
  - $c_j = l_j - r_j$ is the unobserved baseline difference in wear between the left and right feet of boy j.
- If we observe $(\theta + l_j, r_j)$ for boy j, then we cannot observe $(l_j, \theta + r_j)$, which is said to be **counterfactual**.

Example: Shoe data III

- There are 2^m equally-likely treatment allocations, and the observed $\overline{d}$ is a realisation of the random variable

$$\overline{D} = \frac{1}{m}\sum_{j=1}^{m} D_j = \frac{1}{m}\sum_{j=1}^{m} \theta + t_j c_j = \theta + \frac{1}{m}\sum_{j=1}^{m} t_j c_j,$$

  where $t_j = \pm 1$ with equal probabilities, so

$$\mathrm{E}(t_j) = 0, \quad \mathrm{var}(t_j) = 1.$$

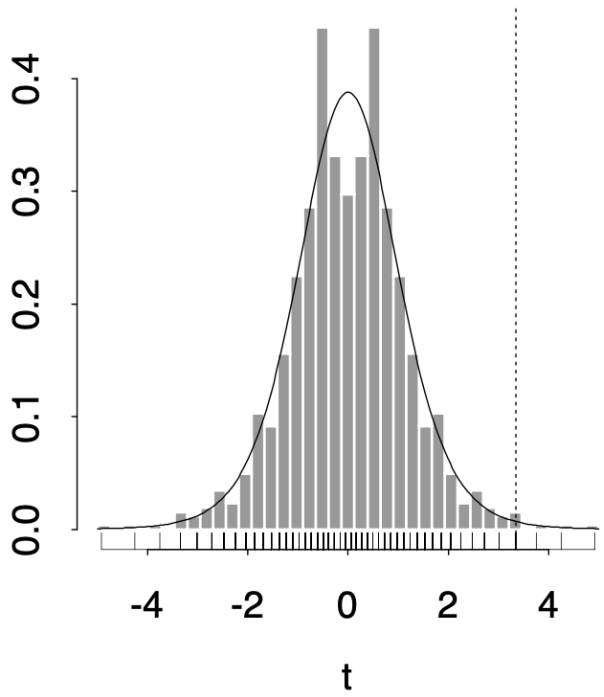
- Hence $\mathrm{E}(\overline{D}) = \theta$ and $\mathrm{var}(\overline{D}) = m^{-2}\sum_{j=1}^{m} c_j^2$, which is unknown because the c_j are unknown, is estimated by (exercise)

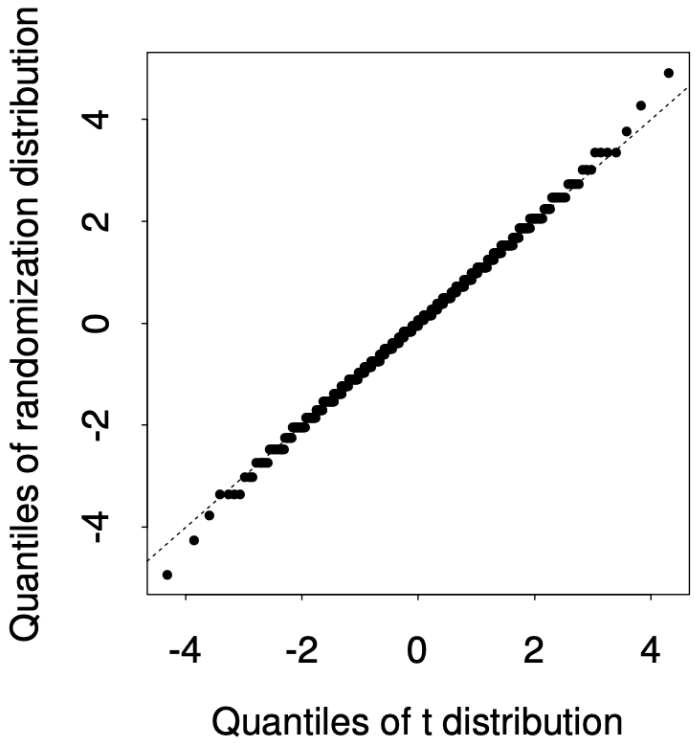
$$S^2 = \frac{1}{m(m-1)}\sum_{j=1}^{m}(D_j - \overline{D})^2.$$

- $\overline{D}$ and S^2 can be computed from the observed data, so the standardized quantity $Z = (\overline{D} - \theta)/S$ is an approximate pivot.
- If there was no difference between B and A (i.e., $\theta = 0$), then $T = \overline{D}/S$ would be symmetrically distributed, as positive and negative values of $\overline{D}$ would be equally likely.

Example: Shoe data IV

Randomization distribution of $T = \overline{D}/S$ for the shoes data, i.e., setting $\theta = 0$, together with a t_9 distribution.
Left: histogram for the values of T, with the t_9 density overlaid; the observed value is given by the vertical dotted line.
Right: probability plot of the randomization distribution against t_9 quantiles.





- If $\theta = 0$, then the observed value of $\overline{D}$ is highly unlikely: just 3 values of $\overline{D}$ exceed $\overline{d} = 0.41$, so if $\theta = 0$ then **exact calculation** gives

$$\Pr(\overline{D} \geq \overline{d}) = 7/2^{10} \doteq 0.007,$$

  which seems unlikely enough to suggest that $\theta > 0$.

- Normal distribution theory suggests that $Z \dot{\sim} t_9$, and the QQ-plot shows that this would work well even here. The symmetry induced by randomisation justifies the widespread use of normal errors in designed experiments.

- **Systematic error** is reduced by randomisation,
  - but if material A had by chance been allocated to all the left feet, then we might have re-randomised;
  - we could have used a design in which A appeared on left feet exactly 5 times.

- **Baseline variation** was reduced by blocking, i.e., using two treatments for each boy, and is estimated by S^2, based only on the observed values $D_1, \ldots, D_m$.

- S^2 also allows a statement of **uncertainty** for $\overline{D}$ and hence for estimates of θ.

Big picture summary

- Statistical inference involves (a family of) **probability models** from which observed data are assumed to be drawn.
- These models express **variation** inherent in the data, but we also wish to express our **uncertainty** about the underlying situation.
- Uncertainty is formulated using
  - a **repeated sampling (frequentist) approach**, which invokes hypothetical repetitions of the data-generating mechanism, or
  - a **Bayesian approach**, which requires that 'prior information' on unknown quantities be expressed as a probability distribution, or
  - a **randomisation approach**, in which the model and hypothetical repetitions are controlled by the investigator.
- Randomisation is the strongest approach, but it is not always applicable.