Probability and Statistics

Contents

The aim of these notes is to summarize a large number of concepts in statistics, both as a short reference and as a quick overview of the subject for people familiar with mathematical notation. It is simply a summary and reorganization from various classical books on these topics.

Chapter 1

Probability

1.1 Basic definitions

Let $\varnothing \neq \Omega$ be a set. A *σ-algebra* (on Ω) is a system $\mathscr{F} \subset \mathscr{P}(\Omega)$ with $\Omega \in \mathscr{F}$, closed under complements and countable unions. $(\Omega, \mathscr{F})$ is an *event space*.

(We can't take the whole $\mathscr{P}(\Omega)$ because it's too big, Vitali's theorem: we can't assign a probability.)

To build our model we start from simple sets whose probabilities we can assign easily, then we generate a σ-algebra: a subset $\mathscr{G} \subset \mathscr{P}(\Omega)$ *generates* a unique σ-algebra $\sigma(\mathscr{G})$ on Ω such that $\mathscr{G} \subset \sigma(\mathscr{G})$.

The choice is often canonical:

1. When Ω is at most countable, we will choose $\mathscr{F} = \mathscr{P}(\Omega)$.
2. When $\Omega = \mathbb{R}^n$ we will use the *Borel σ-algebra*
$$\mathscr{B}_{\mathbb{R}^n} \coloneqq \sigma\left(\left\{\prod_{i=1}^{n}[a_i, b_i] \;\middle|\; a_i < b_i \in \mathbb{Q}\right\}\right).$$
Every $A \in \mathscr{B}_{\mathbb{R}^n}$ is called a *Borel set*. If $\Omega \subset \mathbb{R}^n$, we set $\mathscr{B}_\Omega \coloneqq \{A \cap \Omega \mid A \in \mathscr{B}_{\mathbb{R}^n}\}$.
3. Given $(E_i, \mathscr{E}_i)_{i\in I}$, we define the *product* σ-algebra on $\prod_{i\in I} E_i$ as
$$\bigotimes_{i\in I} \mathscr{E}_i \coloneqq \sigma\left(\left\{\pi_i^{-1}(A_i) \;\middle|\; i \in I,\ A_i \in \mathscr{E}_i\right\}\right).$$

Probability measure (or probability distribution) P on an event space $(\Omega, \mathscr{F})$ is a function $P\colon \mathscr{F} \to [0,1]$ satisfying the normalization property $P(\Omega) = 1$ and σ-additivity for pairwise disjoint events. $(\Omega, \mathscr{F}, P)$ is a *probability space*.

Theorem 1.1.1. P satisfies the following: $P(\varnothing) = 0$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, monotonicity, σ-subadditivity and σ-continuity $P(A_n) \to P(A)$ if $A_n \nearrow A$ or $A_n \searrow A$.

Theorem 1.1.2. Let $(\Omega, \mathscr{F} = \sigma(\mathscr{G}), P)$ where $\mathscr{G} \subset \mathscr{P}(\Omega)$ is an intersection-stable system. Then P is uniquely determined by its restriction $P|_{\mathscr{G}}$ to $\mathscr{G}$.

By Carathéodory's theorem, a function P defined on a suitable system $\mathscr{G}$ may be extended to a probability measure on $\sigma(\mathscr{G})$.

Probability density (function) ρ for P is a function $\rho\colon \Omega \to [0,\infty)$ such that, for all A in $\mathscr{F}$, if Ω is discrete (resp. Borel), one has

$$P(A) = \sum_{\omega\in A} \rho(\omega) \qquad \Big(\text{resp. } P(A) = \int_A \rho(x)dx\Big).$$

It is called respectively a discrete and a Lebesgue density. A discrete density is also called probability mass function (PMF).

In the discrete case there is a 1:1 correspondence between the set of probability measures on $(\Omega, \mathscr{P}(\Omega))$ and the set of $\rho\colon \Omega \to [0,1]$ such that $\sum_{\omega\in\Omega} \rho(\omega) = 1$. In the continuous case $\Omega \subset \mathbb{R}^n$, every measurable $\rho\colon \Omega \to [0,\infty)$ with $\int_\Omega \rho(x)dx = 1$ determines a unique probability measure on $(\Omega, \mathscr{B}_\Omega)$, but not conversely (not every continuous measure has a density; also, two densities differing on a measure-zero set give the same probability).

Distribution of a random variable A *random variable* from $(\Omega, \mathscr{F})$ to $(\Omega', \mathscr{F}')$ is a map $X\colon \Omega \to \Omega'$ which is measurable, i.e. satisfying $X^{-1}(A') \in \mathscr{F}$ for all $A' \in \mathscr{F}'$ (enough to check for a generator of $\mathscr{F}'$). A random variable X from $(\Omega, \mathscr{F}, P)$ to $(\Omega', \mathscr{F}')$ defines a probability measure P_X on $(\Omega', \mathscr{F}')$ by the rule

$$P_X(A') \coloneqq P(X^{-1}(A')) = P(X \in A') \quad \text{ for } A' \in \mathscr{F}',$$

called the *distribution of X under P*. Two random variables are *identically distributed* if they have the same distribution.

A probability measure also called distribution, indeed it's the distribution of the random variable id_Ω.

Random variables can be used to reduce observation depth.

Distribution density A density ρ of P_X is called *distribution density* and links these notions together: $P(X \leqslant c) = F_X(c) = \int_{-\infty}^{c} \rho(x)dx$.

(Cumulative) distribution function For $(\mathbb{R}, \mathscr{B}, P)$,

$$F_P\colon \mathbb{R} \to [0,1], \quad F_P(c) = P((-\infty, c]).$$

For $X\colon \Omega \to \mathbb{R}$ random variable on $(\Omega, \mathscr{F}, P)$,

$$F_X = F_{P\circ X^{-1}}\colon \mathbb{R} \to [0,1], \quad F_X(c) = P(X \leqslant c).$$

Conversely, every increasing right-continuous $F\colon \mathbb{R} \to [0,1]$ with $F(-\infty) = 0$ and $F(+\infty) = 1$ is F_X for $X\colon (0,1) \to \mathbb{R}$, $u \mapsto \inf\{c \in \mathbb{R} \mid F(c) \geqslant u\}$. So, every probability measure P on $\mathbb{R}$ is uniquely determined by F_P.

1.2 Main probability distributions

Bose-Einstein distribution n indistinguishable particles to be distributed in N distinguishable cells. $\Omega = \{(k_1, \dots, k_N) \in \mathbb{Z}_+^N\colon \sum_{j=1}^N k_j = n\}$ has cardinality $\binom{n+N-1}{n}$.

Continuous uniform distribution $\Omega \subset \mathbb{R}^n$ a Borel set with n-dimensional volume $\lambda^n(\Omega) \coloneqq \int 1_\Omega(x)dx =: \int_\Omega dx$. The measure with constant density $\rho(x) = 1/\lambda^n(\Omega)$.

Urn models can be labeled/colored - ordered/unordered - with/without replacement. A useful trick is reducing the observation depth via random variables to compute the probability on the target space as the distribution of the random variable. N total balls, sample of size n.

Discrete uniform Replacement, ordered, labeled.
$\Omega = \{1, \dots, N\}^n \ni (\omega_1, \dots, \omega_n) = \omega$.
Discrete uniform distribution U_Ω has density $\rho(\omega) = 1/N^n$.

n-fold product, Bernoulli Replacement, ordered, colored.
If the set E of color classes $C_i \subset \{1, \dots, N\}$ has discrete density ρ, then the *n-fold product density* of ρ is the density on $\overline{\Omega} = E^n$ given by

$$\rho^{\otimes n}(C_1, \dots, C_n) = \prod_{i=1}^n \rho(C_i) = \prod_{i=1}^n \frac{|C_i|}{N}.$$

If $E = \{0, 1\}$ and $\rho(1) = p \in [0, 1]$ we call it the *Bernoulli distribution* for n trials, when and takes the form $p^k(1-p)^{n-k}$ where k is the number of 1's.

To see this we use a random variable

$$\begin{aligned} X \colon \{1, \dots, N\}^n = \Omega &\to \overline{\Omega} = E^n \\ \omega &\mapsto (C_{\omega_1}, \dots, C_{\omega_n}) \end{aligned}$$

and the probability $\overline{P}$ on $\overline{\Omega}$ which we are interested in is given by the following P probability (the uniform probability on Ω):

$$P(X = (C_{i_1}, \dots, C_{i_n})) = P(C_{i_1} \times \cdots \times C_{i_n}) = \frac{|C_{i_1}| \cdots |C_{i_n}|}{|\Omega|}.$$

Multinomial, binomial... Replacement, unordered, colored.
$\mathcal{C} = \{C_1, \dots, C_r\}$ set of color classes.
$\widehat{\Omega} = \{\mathbf{k} = (k_1, \dots, k_r) \in \mathbb{Z}^{\mathcal{C}} \mid \sum k_i = n\}$, find its probability measure. We start from the n-fold and forget the ordering:

$$\begin{aligned} S \colon \mathcal{C}^n = \overline{\Omega} &\to \widehat{\Omega} \\ \overline{\omega} &\mapsto (S_C(\overline{\omega}))_{C \in \mathcal{C}} = \left(\sum_{i=1}^n 1_{\{C\}}(\overline{\omega}_i) \right)_{C \in \mathcal{C}} \end{aligned}$$

is the histogram of the sample $\overline{\omega}$ and S_C counts the occurrences of color C.

$$\overline{P}(S = \mathbf{k}) = \sum_{\overline{\omega} : S(\overline{\omega}) = \mathbf{k}} \prod_{i=1}^n \frac{|\overline{\omega}_i|}{N} = \frac{n!}{k_1! \cdots k_r!} \prod_{i=1}^r \left(\frac{|C_i|}{N} \right)^{k_i} =$$

$$=: \binom{n}{\mathbf{k}} \prod_{i=1}^r \rho(C_i)^{k_i} =: \mathcal{M}_{n,\rho}(\{\mathbf{k}\})$$

is the *multinomial* distribution (note the multinomial coefficient above).

Special case: the binomial distribution. $\mathcal{C} = \{0,1\}$, $\widehat{\Omega} = \{(k, n-k) \mid k \in \mathbb{Z}\}$ is identified with $\Omega' = \{0, \dots, n\}$. Let $\rho(1) =: p$, then the distribution becomes

$$\mathcal{B}_{n,p}(\{k\}) \coloneqq \binom{n}{k} p^k (1-p)^{n-k}.$$

Replacement, unordered, labeled: it is a special case of the above, when $r = n$ and $k_i = 1$ for all i. We get $n!/N^n$.

No replacement, ordered, labeled. $\Omega_{\neq} = \{\omega \in \{1, \dots, N\}^n \mid \omega_i \neq \omega_j \text{ for } i \neq j\}$ simply has the uniform distribution, $\rho(\omega) = (N-n)!/N!$.

No replacement, unordered, labeled.
$\widetilde{\Omega} = \{\omega \subset \{1, \dots, N\} \mid |\omega| = n\}$
The distribution turns out to be the uniform one:

$$\rho(\omega) = n! \frac{(N-n)!}{N!} = \frac{1}{\binom{N}{n}} = \frac{1}{|\Omega|}$$

(it is obvious, formally can be seen using the random variable $\Omega_{\neq} \to \widetilde{\Omega}$).

Hypergeometric No replacement, unordered, colored.
$\widehat{\Omega} = \{\mathbf{k} = (k_1, \dots, k_r) \in \mathbb{Z}^{\mathcal{C}} \mid \sum k_i = n\}$,
$\mathcal{C} = \{C_1, \dots, C_r\}$ with $|C_i| = N_i$, $\sum N_i = N$.

$$T \colon \{\zeta \subset \{1, \dots, N\} \mid |\omega| = n\} = \widetilde{\Omega} \to \widehat{\Omega}$$
$$\zeta \mapsto (|\zeta \cap C_i|)_{i=1}^r$$

Since $\zeta \mapsto (\zeta \cap C_i)_{i=1}^r$ is a bijection $\{T = \mathbf{k}\} \to \prod_{i=1}^r \{\zeta_i \subset C_i \mid |\zeta_i| = k_i\}$, we get

$$\mathcal{U}_{\widetilde{\Omega}}(T = \mathbf{k}) = \frac{\prod_i \binom{N_i}{k_i}}{\binom{N}{n}} =: \mathcal{H}_{n,\mathbf{N}}(\{\mathbf{k}\}),$$

the *hypergeometric* distribution of parameters n, $\mathbf{N} = (N_i)_{i=1}^r$.

Multinomial approximation of hypergeometric distributions Let E be finite and ρ be a discrete density on E. If $N \to \infty$, $N_a \to \infty$ and $N_a/N \to \rho(a)$ for all $a \in E$, then pointwise $\mathcal{H}_{n,N} \to \mathcal{M}_{n,\rho}$. Intuitively: as the sample space increases it becomes irrelevant whether there is replacement or not; easy proof using asymptotic approximation.

Poisson approximation of binomial distributions Let $p_n \in [0,1]$ be a sequence with $np_n \to \lambda > 0$. Then for every $k \in \mathbb{N}$ we have

$$\mathcal{P}_\lambda(\{k\}) \coloneqq \lim_{n\to\infty} \mathcal{B}_{n,p_n}(\{k\}) = e^{-\lambda}\lambda^k/k!$$

which we call the *Poisson distribution* on $\Omega = \mathbb{Z}_+$ (by writing the exponential series expansion one sees the limit sums to 1 and thus defines a discrete density on $\mathbb{Z}_+$).

Proven similarly as the multinomial approximation, plus using the usual exponential approximation formula.

The Poisson probability measure $\mathcal{P}_{\alpha t}$ measures the probability of k insurance claims in the interval $(0,t]$ ($p_n = \alpha t/n$, so $np_n \to \alpha t =: \lambda$) with α the average number of claims per unit of time.

Theorem 1.2.1. For $n \in \mathbb{N}$ and $p \in (0,1)$ the following error bound holds,

$$\sum_{k \geqslant 0} |\mathcal{B}_{n,p}(\{k\}) - \mathcal{P}_{np}(\{k\})| \leqslant 2np^2$$

Normal distribution Let $v > 0$, $\Omega_N \coloneqq B(0, \sqrt{vN}] \subset \mathbb{R}^N$, $P_N = U_{\Omega_N}$, and the projection $\pi_1 \colon \Omega_N \to \mathbb{R}$ onto the first coordinate. Then

$$\lim_{N\to\infty} P_N(a \leqslant X_1 \leqslant b) = \int_a^b \frac{e^{-x^2/2v}}{\sqrt{2\pi v}} dx$$

and we define the *normal (or Gauss) distribution* $N_{m,v}$ with mean $m \in \mathbb{R}$ and variance $v > 0$ as the probability measure on $(\mathbb{R}, \mathscr{B})$ with density (shown to be a density)

$$\varphi_{m,v}(x) = \frac{e^{-(x-m)^2/2v}}{\sqrt{2\pi v}}$$

The mean (resp. variance) of a sum is the sum of the means (resp. variances): if $X \sim \mathcal{N}(m_X, v_X)$ and $Y \sim \mathcal{N}(m_Y, v_Y)$, then $(X+Y) \sim \mathcal{N}(m_X+m_Y, v_X+v_Y)$.

The following areas will be used in statistics:
$\int_{-\sigma}^{\sigma} \varphi_{0,1} dx \approx 68\%$ (meaning that roughly 68% of the data are within 1 SD from the average)
$\int_{-2\sigma}^{2\sigma} \varphi_{0,1} dx \approx 95\%$
$\int_{-3\sigma}^{3\sigma} \varphi_{0,1} dx \approx 99.7\%$.

Exponential families An exponential family with parameters η is

$$p(\mathbf{x}|\eta) = g(\eta) h(\mathbf{x}) \exp(\eta^T \mathbf{u}(\mathbf{x}))$$

where $\mathbf{u}(\mathbf{x})$ is a function of $\mathbf{x}$ and $g(\eta)$ normalizes the distribution making it sum to 1. Examples include:

- Bernoulli

  $$f(x,p) = p^x (1-p)^{1-x} = (1-p) \exp\left(\ln\left(\frac{p}{1-p}\right) x\right) = \sigma(-\eta) \exp(\eta x)$$

  where $\sigma(\eta) = \frac{1}{1+\exp(-\eta)}$ is the logistic sigmoid function.

- Multinomial...

We have $-\nabla \ln g(\eta) = \mathbb{E}(\mathbf{u}(\mathbf{x}))$ and the maximum likelihood estimator η_{ML} satisfies

$$-\nabla \ln g(\eta_{\mathrm{ML}}) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n) \to \mathbb{E}(\mathbf{u}(\mathbf{x}))$$

so $\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)$ is a sufficient statistic of the distribution.

Mixture distributions If p_i are distribution densities (or cumulative distribution functions), then the sum $\sum_i w_i p_i$ with weights summing to 1 also represents a distribution. The uncountable case is written as $f(x) = \int_\Theta p(x|\theta)q(\theta)$ where $p(x|\theta)$ are a family of distributions parametrized by θ, and θ is itself a random variable with distribution $q(\theta)$.

1.3 Waiting time distributions

Negative binomial (Pascal), geometric The waiting time until the nth success, i.e. number of failures before the nth success is

$$T_n \colon \{0,1\}^{\mathbb{N}} \to \Omega = \mathbb{N}, \quad T_n(\omega) = \min\{k \in \mathbb{N} \mid \textstyle\sum_{i=1}^{k} \omega_i = n\} - n$$

We want the probability of the nth success on the $(n+k)$th draw; we can compute it as the distribution of T_n or directly as

$$\binom{n+k-1}{k} p^n (1-p)^k =: \binom{-n}{k} p^n (p-1)^k =: \overline{B}_{n,p}(\{k\})$$

where p is the probability of success and the binomial coefficient is the number of choices for the k failure times (the -1 comes from the last draw having to be a success). This is called the *negative binomial distribution* or the *Pascal distribution*. When $n = 1$ we call it the *geometric distribution*.

Gamma, exponential For a fixed α, the probability of at least k claims in the interval $(0, t]$ as a function of t is given by

$$\Gamma_{\alpha,k}((0,t]) = 1 - P_{\alpha t}(\{0,\dots,k-1\}) = \int_0^t \frac{\alpha^k}{(k-1)!} s^{k-1} e^{-\alpha s} ds =: \int_0^t \gamma_{\alpha,k}(s) ds$$

hence we get a probability measure on $(0,\infty)$ with Lebesgue density $\gamma_{\alpha,k}$ (use Euler's Γ function

$$\Gamma(r) = \int_0^\infty s^{r-1} e^{-s} ds, \qquad r > 0$$

to check that it is a density). We call this the *Gamma distribution* $\Gamma_{\alpha,k}$. It can be generalized to $\Gamma_{\alpha,r}$ for every real $r > 0$, replacing $(k-1)!$ with $\Gamma(r)$. If $r = 1$ we call it the *exponential* distribution (the first point i.e. "claim" is exponentially distributed).

Beta β_{n-r+1} is the distribution of the rth order statistic of an n-sample from a continuous uniform distribution. If we are to be given n fixed points in the unit interval $(0,1)$, i.e. some $\omega \coloneqq (\omega_1,\dots,\omega_n) \in \Omega \coloneqq (0,1)^n$ with the uniform distribution, where is the rth point in order? (E.g. if we are expecting n deliveries, how long does it take to receive the rth?) The projection $\pi_i(\omega)$ is the instant of the i-th point. We see π_i as random variables on Ω. Define the random variables $\pi_{r:n}$ so that

$$\pi_{1:n} < \pi_{2:n} < \cdots < \pi_{n:n} \quad \{\pi_{1:n},\dots,\pi_{n:n}\} = \{\pi_1,\dots,\pi_n\}.$$

The *Beta distribution* is defined as

$$P(\pi_{r:n} \leqslant c) = \int_{-\infty}^{c} \beta_{r,n-r+1}(s) ds$$

where

$$\beta_{a,b}(s) = \frac{s^{a-1}(1-s)^{b-1}}{B(a,b)} \quad 0 < s < 1$$

where $B(a,b)$ is Euler's Beta function. In particular the distribution density of the first point is $\beta_{1,n}(s) = n(1-s)^{n-1}$ and the probability measure given by the density $\beta_{1,1}$ is the uniform probability on $[0,1]$.

1.4 Conditional probability

$P(B|A) \coloneqq \frac{P(B\cap A)}{P(A)}$; $P(-|A)$ is the unique probability measure such that A is certain and it's proportional to P on subevents of A.

It is not related to causality, it can be interpreted as our evaluation of the situation after we have been *informed* that A has happened (but not after it *has* happened, in fact B could happen before A! For example: draw a ball without looking, then draw another one and see its color, then we can re-evaluate the probability of the color of the first one.

Let $\Omega = \bigcup_{i\in I} B_i$ be a countable partition.

- Case-distinction formula: $P(A) = \sum_{i\in I} P(B_i)P(A|B_i)$.
- Bayes' formula: $P(B|A) = \frac{P(B)P(A|B)}{P(A)}$ or, more in general,

$$P(B_k|A) = \frac{P(B_k)P(A|B_k)}{\sum_{i\in I} P(B_i)P(A|B_i)}.$$

- Multiplication rule: given events $A_1, \ldots, A_n$, we have

$$P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)\cdots P(A_n|A_1 \cap \cdots \cap A_{n-1}).$$

  This helps us study the case when the distribution of a first random variable X_1 is known and X_k is known when $X_1, \ldots, X_{k-1}$ are known. Along with the next results it is useful in the construction of multi-stage models.

Evaluation of medical tests Ω = population, $P = U_\Omega$, $D \subset \Omega$ people with the disease, $P \subset \Omega$ people testing positive.

$P(D)$ = prevalence of the disease, $P(P|D)$ = test sensitivity, $P(P|D^c)$ = (1 − specificity).

Positive correctness: $P(D|P) = \frac{P(D)P(P|D)}{P(D)P(P|D)+P(D^c)P(P|D^c)}$ can be low if the prevalence is low, even though the sensitivity is high.

Negative correctness: $P(D^c|P^c) = \ldots$.

Three doors problem Probability is 2/3 if we assume the host will not open the chosen door, or 1/2 otherwise.

Theorem 1.4.1 (Construction of probability measures via conditional probabilities)**.** Let $\Omega_1, \ldots, \Omega_n$ be countable, ρ_1 a density on Ω_1, $\rho_{k|\omega_1,\ldots,\omega_{k-1}}$ a density on Ω_k for $k \geqslant 2$ and $\omega_i \in \Omega_i$ for $i = 1, \ldots, k-1$. Let $\Omega = \prod_{i=1}^n \Omega_i$, π_i be the i-th projection.

Then, there is a unique probability measure P on $(\Omega, \mathscr{P}(\Omega))$ with the properties:

1. $P(\pi_1 = \omega_1) = \rho_1(\omega_1)$ for all $\omega_1 \in \Omega_1$,
2. $P(\pi_k = \omega_k \mid \pi_1 = \omega_1, \ldots, \pi_{k-1} = \omega_{k-1}) = \rho_{k|\omega_1,\ldots,\omega_{k-1}}(\omega_k)$.

This P is given by

$$P(\{(\omega_1, \ldots, \omega_n)\}) = \rho_1(\omega_1)\rho_{2|\omega_1}(\omega_2) \cdots \rho_{n|\omega_1,\ldots,\omega_{n-1}}(\omega_n).$$

In the case of an infinite product space $\Omega = \prod_{i \geqslant 1} \Omega_i$, replace $\mathscr{P}(\Omega)$ by $\bigotimes_{i \geqslant 1} \mathscr{P}(\Omega_i)$. Then the theorem becomes: there exists a unique probability measure P such that, for all $k \geqslant 1$ and $\omega_i \in \Omega_i$,

$$P(\pi_1 = \omega_1, \ldots, \pi_k = \omega_k) = \rho_1(\omega_1)\rho_{2|\omega_1}(\omega_2) \cdots \rho_{n|\omega_1,\ldots,\omega_{n-1}}(\omega_n).$$

The finite case: the stated expression is identical to the multiplication rule above, under the conditions of the theorem, proving uniqueness. Then we check that it satisfies the two properties.

Independence

- A family $(A_i)_{i \in I}$ of events is called independent if, for every finite subset $\varnothing \neq J \subset I$, we have
$$P(\bigcap_{i \in J} A_i) = \prod_{i \in J} P(A_i).$$
- Since we are also interested in independence of subexperiments in addition to events, we define a family $(X_i)_{i \in I}$ of random variables $\Omega \to \Omega_i$ independent if the family $(\{X_i \in B_i\})_{i \in I}$ is independent for every choice $B_i \in \mathscr{F}_i$.

Note this is stronger than pairwise independence. Toss two coins and consider events: first toss is heads, second toss is heads, both tosses give same result. They are pairwise independent but not independent as a family.

Independence despite causality: e.g., rolling two dice, A = sum of points is 7, B = first dice shows 6 are independent although the sum depends causally on the value of the first dice.

As a consequence of the uniqueness theorem, to check independence of random variables it is enough to choose $B_i \in \mathcal{G}_i$ where $\mathcal{G}_i$ is an intersection-stable generator of $\mathscr{F}_i$.

A family of events is independent iff the corresponding family of indicator functions is independent.

Criterion for independence of finitely many random variables. Consider a finite sequence of random variables $Y_i \colon \Omega \to \Omega_i$ $(i = 1, \ldots, n)$ on a probability space $(\Omega, \mathscr{F}, P)$. In the discrete case of countable Ω_i's, $(Y_i)_i$ is independent iff $P(Y_1 = \omega_1, \ldots, Y_n = \omega_n) = \prod_{i=1}^n P(Y_i = \omega_i)$ for all $\omega_i \in \Omega_i$. In the real case $\Omega_i = \mathbb{R}$, the family is independent iff $P(Y_1 \leqslant c_1, \ldots, Y_n \leqslant c_n) = \prod_{i=1}^n P(Y_i \leqslant c_i)$ for all $c_i \in \mathbb{R}$. (The "only if" direction is trivial in both cases).

For example, the n-fold product measure $\rho^{\otimes n}$ on E^n makes the projection random variables independent.

1.5 Product measures

We want to construct a model for infinite coin tosses, so we need to prove existence of infinitely many independent random variables and work with infinite product σ-algebras.

Theorem 1.5.1 (Construction of independent random variables with prescribed distributions)**.** Let $(\Omega_i, \mathscr{F}_i, P_i)_{i\in I}$ be probability spaces, I countable. Then there exists $(\Omega, \mathscr{F}, P)$ and independent $Y_i \colon \Omega \to \Omega_i$ such that $P \circ Y_i^{-1} = P_i$ for all $i \in I$.

Corollary 1.5.2 (Existence of infinite product measures)**.** Let $(\Omega_i, \mathscr{F}_i, P_i)_{i\in I}$ be probability spaces, I countable. Then there exists a unique probability measure P on $(\Omega \coloneqq \prod_i \Omega_i, \mathscr{F} \coloneqq \bigotimes_i \mathscr{F}_i)$ such that for all finite $J \subset I$ and all $A_i \in \mathscr{F}_i$ one has

$$P(\pi_i \in A_i \text{ for all } i \in J) = \prod_{i\in J} P_i(A_i)$$

i.e. the projections $\pi_i \colon \Omega \to \Omega_i$ are independent with distribution P_i. We denote this P by $\bigotimes_i P_i$ and call it the *product measure* of the P_i's.

The proof of the above corollary also shows that independence and product measures are closely related: a countable family Y_i of random variables is independent iff $P \circ (Y_i)_{i\in I}^{-1} = \bigotimes_{i\in I} P \circ Y_i^{-1}$ (the joint distribution equals the product of marginal distributions).

The density of a finite product measure is the product of the single densities, if they exist. An infinite product measure never admits a density, since there is no Lebesgue measure on $\mathbb{R}^{\mathbb{N}}$.

Convolution Measures Q_1, Q_2 on $\mathbb{R}$, then the distribution of the sum Q_1+Q_2 is the convolution $Q_1 * Q_2 \coloneqq (Q_1 \otimes Q_2) \circ A^{-1}$ where $A \colon \mathbb{R}^2 \to \mathbb{R}$ sends $(x_1+x_2) \mapsto x_1 + x_2$.

1.6 Moments

Expectation of a real random variable is defined as the first moment (when it exists)

$$\mu \coloneqq \mathbb{E}(X) \coloneqq \int_\Omega X dP = \int_{\mathbb{R}} x dP_X$$

where the equality follows from the change of variable theorem

$$\int_\Omega (g \circ X) dP = \int_{\mathbb{R}} g dP_X$$

(this is the law of the unconscious statistician, LOTUS, by which we can write the expectation of $\mathbb{E}(gX)$ without the distribution of gX). In the Lebesgue case we can further write $dP_X = dF_X = f_X dx$.

If $E|X| < \infty$ we say X has an expectation and write $X \in \mathscr{L}^1$.

If P on $(\mathbb{R}, \mathscr{B})$ then we define $\mathbb{E}(P) = \mathbb{E}_P(\mathrm{Id}_{\mathbb{R}})$ and $\mathbb{E}(P) = \mathbb{V}_P(\mathrm{Id}_{\mathbb{R}})$ if they exist.

$\mathbb{E}$ is linear and commutes with products for independent variables.

Variance If $X \in \mathscr{L}^2$ we define its variance as the second central moment

$$\mathbb{V}(X) \coloneqq \sigma^2 \coloneqq \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2,$$

and σ is called standard deviation. We define the covariance of $X, Y \in \mathscr{L}^2$ as

$$\mathrm{Cov}(X, Y) \coloneqq \mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y)) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

We have $\mathrm{Cov}(aX+b, cY+d) = ac\mathrm{Cov}(X, Y)$, in particular $\mathbb{V}(aX+b) = a^2\mathbb{V}(X)$, and

$$\mathbb{V}\left(\sum X_i\right) = \sum \mathbb{V}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)$$

(for X_i pairwise uncorrelated, i.e. when there are no covariances, this is called Bienaymé's identity).

The correlation is defined as $\rho(X, Y) = \mathrm{Cov}(X, Y)/\sigma_X\sigma_Y \in [-1, 1]$; it is a measure of association (it is the slope of linear regression line; cfr. SD line, axis of symmetry of the scatter plot). Independent implies uncorrelated.

Law of total expectation and variance $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$ is the law of total expectation. This can be used to prove the law of total variance $\mathbb{V}(X) = \mathbb{E}(\mathbb{V}(X|Y)) + \mathbb{V}(\mathbb{E}(Y|X))$.

General moments The n-th moment of X is $\mathbb{E}X^n$; the n-th central moment is $\mu_n \coloneqq \mathbb{E}(X - \mu)^n$; the n-th standardized (or normalized) moment is $\frac{\mathbb{E}(X-\mu)^n}{\sigma^n}$. The moment generating function (mgf) is $M_X(t) = \mathbb{E}\exp(tX)$ and then the nth moment is its nth derivative evaluated at 0. If X and Y are independent, then $M_{X+Y}(t) = M_X(t)M_Y(t)$ whence we get, for example, that the mean and variance of a sum of normal distributions are the sum of the means and sum of the variances. The sequence of moments does not determine a distribution uniquely in general, unless the random variables have bounded support.

The following are examples of higher-order statistics (using the third or higher power of a sample).

Skewness is the third standardized moment $\frac{\mathbb{E}(X-\mu)^3}{\sigma^3}$, which measures the symmetry of the distribution: negative (or left) skew means the left tail is longer and the mass of the distribution is concentrated to the right, and symmetrically for positive (or right) skew.

Skewness is not directly related to the relationship between mean and median (though often the mean is to the left of the median in negative skew), unlike the old definition of non-parametric skew: $(\mu - \nu)/\sigma$ where ν is the median. Another possible definition is Pearson's mode skewness $(\mu - \mathrm{mode})/\sigma$.

Kurtosis is the fourth standardized moment $K(X) \coloneqq \frac{\mathbb{E}(X-\mu)^4}{\sigma^4}$; excess kurtosis is $K - 3$ since $K = 3$ for a normal distribution. A distribution with positive excess kurtosis is called leptokurtic (fat-tailed, has heavy tails, more mass on the tails of its support than a normal distribution, more extreme outliers, otherwise platykurtic (flatter, e.g. uniform).

Mode A mode of a distribution is any local maximum of the density function (or probability mass function for discrete distributions). The distribution is unimodal if the maximum is unique, multimodal otherwise. For symmetric unimodal distributions, the mean (if defined), median and mode coincide.

1.7 Stochastic processes

A stochastic process on a probability space $(\Omega, \mathcal{F}, P)$ is a collection

$$X = \{X_t \colon (\Omega, \mathcal{F}) \to (S, \Sigma) \mid t \in T\}.$$

A filtration of $\mathcal{F}$ is a collection $(\mathcal{F}_t)_t$ of sub-σ-algebras such that $\mathcal{F}_s \subset \mathcal{F}_t$ if $s \leqslant t$ (the index set T is totally ordered). Standard example:

$$\mathcal{F}_t = \sigma(X_s \mid s \leqslant t) = \sigma(X_s^{-1}(A) \mid A \in \Sigma, s \leqslant t)$$

called the natural filtration of $\mathcal{F}$ wrt X: it records past behavior of X.

X is adapted (or non-anticipating) to a filtration $(\mathcal{F}_t)_t$ if it cannot see into the future: $X_t \colon \Omega \to S$ is $(\mathcal{F}_t, \Sigma)$-measurable for each t.

A stopping time wrt $(\mathcal{F}_t)_t$ is a random variable $\tau \colon \Omega \to T \cup \{+\infty\}$ such that $\{\tau \leqslant t\} \in \mathcal{F}_t$ for all $\tau \leqslant t$.

A process X adapted to $(\mathcal{F}_t)_t$ has the Markov property if, for all $A \in \Sigma$ and $s < t$, one has

$$P(X_t \in A \mid \mathcal{F}_s) = P(X_t \in A \mid X_s),$$
$$\text{alternatively } E(f(X_t) \mid \mathcal{F}_s) = E(f(X_t) \mid \sigma(X_s))$$

Intuitively, the future only depends on the present state and not the past.

Memorylessness for a random variable is the property $P(X > s + t | X > s) = P(X > t)$ for all s, t integer or real numbers (depending on whether X is discrete or continuous). The only such distributions are the geometric (discrete) and exponential (continuous): defining the survival function $S(t) = P(X > t)$, the property can be written as $S(t + s) = S(t)S(s)$ whence $S(q) = S(1)^q$ for rational q, and the only continuous S satisfying this is $S(x) = \exp(-\ln S(1)x)$. (See survival analysis). It's weaker than the Markov property.

Martingale A stochastic process Y_i is a martingale wrt another one X_i if Y_i has finite expectation and the conditional expected value of the next observation given the previous observations, is equal to the latest observation:

$$E(Y_{n+1} \mid X_1, \ldots, X_n) = Y_n \text{ (discrete)}$$
$$E(Y_t \mid \{X_{s'}, s' \leqslant s\}) = Y_s \text{ for all } s \leqslant t \text{ (continuous)}$$
$$E(Y_t \mid \mathcal{F}_s) = Y_s \text{ for all } s \leqslant t \text{ (continuous)}$$

In the case $X = Y$, we just say X is a martingale. Examples: unbiased random walk, gambler's capital in fair games.

1.8 Convergence

- Convergence in distribution: $F_n(x) \to F(x)$ for all x where $F(x)$ is continuous (the cdf's of X_n, X resp.), denoted by $X_n \rightsquigarrow X$. If F is continuous this is equivalent to uniform convergence $\|F_n - F\|_\infty \to 0$.
- Convergence in probability: for every ε, $P(|X_n - X| > \varepsilon) \to 0$, denoted by $X_n \xrightarrow{P} X$.
- Convergence in $\mathscr{L}^p$: $\mathbb{E}|X_n - X|^p \to 0$.
- Almost sure convergence: $P(\{\omega \in \Omega \mid X_n(\omega) \to X(\omega)\}) = 1$

A random variable is asymptotically normal if it converges in distribution to a normally distributed random variable.

Weak law of large numbers (WLLL) $\mathscr{L}^2$ version: if $X_i \in \mathscr{L}^2$ are pairwise uncorrelated with variance bounded by v, then $P(|\frac{1}{n}\sum_1^n (X_i - E(X_i))| \geqslant \varepsilon) \leqslant \frac{v}{n\varepsilon^2}$ for all $\varepsilon > 0$. If they are iid, we may use the following version:

$\mathscr{L}^1$ version: if $X_i \in \mathscr{L}^1$ pairwise iid, write $E(X_i) =: m$, then $\frac{1}{n}\sum_1^n X_i \xrightarrow{P} m$ (sample means tends to mean in probability).

$\mathscr{L}^2$ proof: the argument inside the absolute value is $\mathscr{L}^2$ with mean 0 and variance $\frac{1}{n^2}\sum \mathbb{V}(X_i) \leqslant v/n$, so the theorem follows immediately from Chebyshev's inequality: $P(|Y - \mathbb{E}Y| \geqslant \varepsilon) \leqslant \mathbb{V}(Y)/\varepsilon^2$ for all $Y \in \mathscr{L}^2$ and $\varepsilon > 0$.

Note: Chebyshev's inequality implies that the proportion of samples within k standard deviations is $\geqslant 1 - 1/k^2$.

Strong law of large numbers (SLLL) If $X_i \in \mathscr{L}^2$ are pairwise uncorrelated with variance bounded by v, then $\frac{1}{n}\sum_1^n (X_i - E(X_i)) \to 0$ almost surely (i.e. pointwise almost everywhere; this implies convergence in probability too).

Central limit theorem $X_i \in \mathscr{L}^2$ iid with $\mathbb{E}(X_i) = \mu$ and $\mathbb{V}(X_i) = \sigma^2 > 0$, then

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}_{0,1}$$

that is, the sample mean of iid's tends to a normal distribution with mean and variance of the sample mean, regardless of the particular distribution.

Applying this to X_i Bernoulli we get as a special case the normal approximation of $\mathcal{B}$ by De Moivre-Laplace, also known as local limit theorem. In this case we have convergence not only in distribution but even uniformly.

Normal approximation of $\mathcal{B}$ Let $p \in (0,1)$, $q = 1-p$, $c > 0$, $x_n(k) = \frac{k-np}{\sqrt{npq}}$.

$$\max_{k:|x_n(k)|\leqslant c} \left| \frac{\mathcal{B}_{n,p}(\{k\})}{\varphi(x_n(k))/\sqrt{npq}} - 1 \right| \xrightarrow{n\to\infty} 0$$

This says that the histogram tends to a Gaussian curve; the integral form of this theorem states the same in terms of the area and is useful to avoid binomial computations: for $0 \leqslant k \leqslant l \leqslant n$ one has

$$\mathcal{B}_{n,p}(\{k, \cdots, l\}) \xrightarrow{n\to\infty} \Phi(x_n(l + 1/2)) - \Phi(x_n(k - 1/2))$$

Chapter 2

Statistics

Statistical model A collection of probability spaces $(X, \mathscr{F}, P_\theta : \theta \in \Theta)$, where the P_θ are probability measures of the same class. It is *parametric* if $\Theta \subset \mathbb{R}^d$, *discrete* if X is discrete, *continuous* if $X \in \mathscr{B}_{\mathbb{R}^n}, \mathscr{F} = \mathscr{B}_X$. A *standard* model is a model that is either discrete or continuous.

Statistics studies the following problem: by looking at some observations (represented by elements of X), how can we determine the right θ, i.e. the probability measure P_θ that controls our phenomenon?

Estimators A *statistic* is a random variable on $(X, \mathscr{F})$.

A *(point) estimator* of θ is a statistic $T\colon X \to \Theta$ (Θ must have a σ-algebra); its values are the *estimates*. Clearly, for this to be useful, we must choose a T that "resembles" θ in some way.

Maximum Likelihood Estimation Given an outcome x, we can estimate θ as the $T(x)$ for which x has the largest $P_{T(x)}$-probability: we say that T is a *maximum likelihood estimator (MLE)* for θ if

$$\rho_{T(x)}(x) = \max_{\theta \in \Theta} \rho_\theta(x)$$

i.e. $T(x)$ maximizes the likelihood function $\mathcal{L}_x\colon \theta \mapsto \rho_\theta(x)$ for the outcome x.

More in general, if $\tau(\theta)$ is a characteristic of θ given by $\tau\colon \Theta \to \Sigma$, then we define an estimator of τ as a statistic $T\colon (X, \mathscr{F}) \to (\Sigma, \mathscr{S})$; a MLE of $\tau(\theta)$ is $\tau \circ T$ with T a MLE of θ.

Sample mean and variance are the Gaussian MLEs The MLE for the n-fold Gaussian model $(\mathbb{R}^n, \mathscr{B}^n, {\mathcal{N}_{m,v}}^{\otimes n} : m \in \mathbb{R}, v > 0)$ is $T = (\overline{X}, V)\colon \mathbb{R}^n \to \mathbb{R} \times (0, \infty)$ where

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

are called the *sample mean* and *sample variance* respectively.

We should maximize $\rho(m, v, x) = (2\pi v)^{-n/2} \exp[-\sum_{i=1}^n \frac{(x_i - m)^2}{2v}]$ wrt (m, v): we find $\overline{X}$ by minimizing the mean squared error inside exp, which is obtained

for the sample mean by Pythagoras' theorem

$$\frac{1}{n}\sum(x_i - m)^2 = \frac{1}{n}\sum(x_i - \overline{X})^2 + (\overline{X} - m)^2;$$

then we find V by differentiating $\log \rho_x(M, v)$ wrt v...

Bias What are some desirable requirements for an estimator? An estimator T (or a sequence of estimators T_n depending on observations $X_1, \ldots, X_n$) is *unbiased* if

$$\text{Bias} \coloneqq E_\theta(T) - \theta = 0 \quad \text{for all } \theta \in \Theta.$$

$\{T_n\}$ is *asymptotically unbiased* if $E_\theta(T_n) \to \theta$ as $n \to \infty$, for all $\theta \in \Theta$. For example, if T_n is such that $\mathbb{E}_\theta(T_n) = \frac{n}{n+1}\theta$ then clearly it's biased but asymptotically unbiased.

In a series of measurements, the bias is an error affecting all measurements in the same way, unlike chance errors. Unbiasedness assures that the values are typically centered around θ, but for this to be useful we should also know that they don't fluctuate too much, that is, that T has small variance. Note that a small variance is not useful in the case of a very biased estimator (because then the values would be centered around a wrong value).

A MLE is not necessarily unbiased: we saw that the sample variance is a Gaussian MLE, but we'll now check that it's biased.

Unbiased estimators for mean and variance In a general n-fold product model $(\mathbb{R}^n, \mathscr{B}^n, P_\theta{}^{\otimes n} : \theta \in \Theta)$ with expectation and variance defined for all θ, the sample mean $\overline{X}$ is an unbiased estimator of $m(\theta) \coloneqq \mathbb{E}(P_\theta)$ (immediate by linearity). Reordering Pythagoras' theorem above and applying $\mathbb{E}$, we compute the expectation of the sample variance

$$\begin{aligned}
\mathbb{E}(V) &= \frac{1}{n}\sum \mathbb{E}\left((X_i - m)^2\right) - \mathbb{E}\left((\overline{X} - m)^2\right) = \\
&= \frac{1}{n}\sum \left(\mathbb{V}(X_i - m) - \mathbb{E}(X_i - m)^2\right) - \mathbb{V}(\overline{X} - m) = \\
&= \frac{1}{n}\sum \mathbb{V}(X_i) - \frac{\mathbb{V}(X_i)}{n} = \frac{n-1}{n}\sigma^2 :
\end{aligned}$$

therefore, to get an unbiased estimator of $v(\theta) \coloneqq \mathbb{V}(P_\theta)$, we must take the corrected sample variance

$$V^* \coloneqq \frac{n}{n-1}V = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Properties of M, V Distributions... (thm 5.3.1 Casella Berger page 248); sum of normal distributions is normal with... $\overline{X}$ and V^* are independent random variables. $\overline{X} \sim \mathcal{N}(m, v/n)$ (using distribution of sum of normal variables), $(n-1)V^*/v \sim \chi^2_{n-1}$.

Decision theory How to choose between different estimators, or measure the quality of an estimator? A loss function (or error function) $L\colon \Theta \times \Theta \to \mathbb{R}$ measures the distance $L(\theta, \widehat{\theta}(x))$ e.g. squared error, absolute error, etc... The risk function is defined as

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta(L(\theta, \widehat{\theta})) = \int L(\theta, \widehat{\theta}(x)) f(x, \theta) dx$$

($\mathbb{E}_\theta$ is the expectation under the probability P_θ). This is the mean squared error (MSE) when the loss function is the square; in this case we can derive the expression for the bias-variance tradeoff:

$$\begin{aligned}\mathrm{MSE}(T) \coloneqq \mathbb{E}_\theta((T-\theta)^2) = \mathbb{V}_\theta(T) + (\mathbb{E}_\theta(T) - \theta)^2 = \\ = \mathbb{V}_\theta(T) + \mathrm{Bias}_T(\theta)^2\end{aligned}$$

and this is the quantity that we aim to minimize, even at the cost of admitting some bias. For example, if we know that θ belongs to some range where the error of T' is smaller than the error of T, then we use T' even if its bias is bigger than that of T. (We will always suppose that our estimators belong to $\bigcap_{\theta\in\Theta} \mathscr{L}^2(P_\theta)$ so that the variances are defined.)

An unbiased estimator T is called a *best estimator* if $\mathbb{V}_\theta(T) \leqslant \mathbb{V}_\theta(S)$ for all other unbiased estimators S and all $\theta \in \Theta$.

We may choose $\widehat{\theta}$ so as to minimize either the maximum risk $\sup_\theta R(\theta, \widehat{\theta})$ (a minimax rule) or the Bayes risk $\int R(\theta, \widehat{\theta}) f(\theta) d\theta$ (a Bayes rule, $f(\theta)$ is a prior distribution).

In the regression setting we further average over "θ", we write:

$$\mathbb{E}(L) = \iint L(y, \widehat{y}(\mathbf{x})) f(\mathbf{x}, y) d\mathbf{x} dy;$$

when L is the square function, this is minimized by the conditional mean $\widehat{y}(\mathbf{x}) \coloneqq \int y f(y|\mathbf{x}) dy = \mathbb{E}_y[y|\mathbf{x}]$, the conditional expectation (chapter 3).

In the classification setting, we should minimize

$$\mathbb{E}(L) = \int \sum_k L(C_k, \widehat{C}_k(\mathbf{x})) f(\mathbf{x}, C_k) d\mathbf{x}.$$

If L is the 0-1 loss ("accuracy"), the minimum is given pointwise by the *Bayes classifier*

$$\widehat{C}(\mathbf{x}) = \mathrm{argmax}_{i=1}^k P(C_i \mid X = \mathbf{x}),$$

which we can estimate e.g. by K-nearest neighbors, Naive Bayes Classifier...

Method of moments Another way to estimate θ is to define $\widehat{\theta}_n$ such that each moment $\int x^j dF_{\widehat{\theta}_n}$ equals the sample moment $\frac{1}{n}\sum_{i=1}^n X_i^j$. These estimators are often biased.

Standard error The distribution of $\widehat{\theta}_n$ is called the sampling distribution; its standard deviation is called the standard error $\mathrm{se}(\widehat{\theta}_n) = \sqrt{\mathbb{V}(\widehat{\theta}_n)}$. If we estimate it, we use notation $\widehat{\mathrm{se}}(\widehat{\theta}_n)$. We estimate $\mathbb{V}_F(\widehat{\theta}_n)$ (depends on the unknown distribution F) by $\mathbb{V}_{\widehat{F}_n}(\widehat{\theta}_n)$; if we still don't have a formula to compute this,

we must do a further bootstrap estimate: we create B data sets $x_{1i}, \dots, x_{ni}$ ($i = 1, \dots, B$) by drawing observations with replacement from the data, which simulates random variables $\sim \widehat{F}_n$; we get estimates $\widehat{\theta}_n(x_{1i}, \dots, x_{ni})$ for $i = 1, \dots, B$ of which we compute the sample variance.

Regular models, Fisher information Define the score function $U_\theta(x) \coloneqq \frac{\partial}{\partial\theta} \log \rho(x, \theta) = \rho'_x(\theta)/\rho_x(\theta)$ (suppose likelihood function positive, differentiable and some other nice properties, so-called *regular models*). The Fisher Information of a model is the function $I\colon \theta \mapsto \mathbb{V}_\theta(U_\theta)$. This is used to approximate standard error of a MLE.

I vanishes on an interval iff ρ_x is constant on Θ_0 for almost all x, so no observation can distinguish between the parameters in Θ_0. Additivity: an n-fold product model has Fisher information $I^{\otimes n} = nI$.

Asymptotic normality of MLE Under appropriate regularity conditions, $\sqrt{\mathbb{V}(\widehat{\theta}_n)} \approx \sqrt{1/I_n(\theta)} =: \widehat{\mathrm{se}}$ and the statistic $(\widehat{\theta}_n - \theta)/\widehat{\mathrm{se}}$ (or $(\widehat{\theta}_n - \theta)/\mathrm{se}$) converges to $\mathcal{N}(0, 1)$ in distribution. Consequently we have the approximate $1 - \alpha$ confidence interval

$$C_n = \widehat{\theta}_n \mp z_{\alpha/2}\widehat{\mathrm{se}}, \quad P_\theta(C_n \ni \theta) \to 1 - \alpha.$$

Delta method For $\tau = g(\theta)$ with g differentiable and $g'(\theta) \neq 0$, the MLE is $\widehat{\tau} = g(\widehat{\theta})$ (equivariance) and we have asymptotic normality as above, with $\mathrm{se}(\widehat{\tau}_n) = |g'(\widehat{\theta})|\widehat{\mathrm{se}}(\widehat{\theta}_n)$.

Parametric bootstrap Alternatively, one can estimate standard errors using the bootstrap method sampling from $f(x, \widehat{\theta}_n)$ (instead of $\widehat{F}_n$ as in the non-parametric bootstrap).

Cramér-Rao inequality (or information inequality). Repeating a regular experiment n times, the variance of an unbiased estimator of τ has order at least $1/n$, but there is a condition to get a best estimator, an estimator that is *Cramér-Rao efficient*:

Take a regular model, $\tau\colon \Theta \to \mathbb{R}$ a $\mathscr{C}^1$ function, $\tau' \neq 0$, T regular estimator of τ, i.e. satisfying

$$\frac{d}{d\theta}\mathbb{E}_\theta(T) = \int T(x)\frac{d}{d\theta}\rho(x, \theta)dx.$$

Then

$$\mathbb{V}_\theta(T) \geqslant \tau'(\theta)^2/I(\theta)$$

with equality iff $T - \tau(\theta) = \tau'(\theta)U_\theta/I(\theta)$ for all $\theta \in \Theta$, i.e.

$$\rho(x, \theta) = \exp[a(\theta)T(x) - b(\theta)]h(x)$$

for a primitive of I/τ', $h\colon X \to (0, \infty)$ measurable function,

$$b(\theta) = \log \int_X \exp[a(\theta)T(x)]h(x)dx.$$

Exponential models Models with this likelihood function are called *exponential models* wrt the statistic T. Exponential models of one parameter have the increasing likelihood property which will be useful in tests. Examples: a family of

binomial distributions $\{B_{n,\theta} : \theta \in (0,1)\}$ for fixed n, $T(x) = x/n$;

Poisson distributions $\{P_\theta : \theta > 0\}$, $T(x) = x$;

normal distributions $\{N_{\theta,v} : \theta \in \mathbb{R}\}$ for fixed variance v, $T(x) = x$;

the n-fold product of exponential models for $T(x)$, by setting $T_n(x) = \frac{1}{n}\sum_{i=1}^n T \circ X_i$.

Consistency A sequence of estimators $T_n(X_1, \ldots, X_n)$ is *consistent* if $T_n \xrightarrow{P_\theta} \theta$ as $n \to \infty$ (this is analogous to the WLLL).

We restrict to the case of independent observations, so to the infinite product model $(X = E^{\mathbb{N}}, \mathscr{F} = \mathscr{E}^{\otimes\mathbb{N}}, P_\theta = {Q_\theta}^{\otimes\mathbb{N}} : \theta \in \Theta)$.

Consistency of $\mathbb{E}$ and v in real product models The sequences $(M_n)_{n\geqslant 1}$, $(V_n^*)_{n\geqslant 2}$ of sample mean and corrected variance are consistent.

Consistency of MLE's Consider a one-parameter standard model with Θ open interval, $Q_\theta \neq Q_{\theta'}$ for $\theta \neq \theta'$ (identifiable), and n-fold likelihood function $\rho^{\otimes n}(x,\theta) = \prod_{i=1}^n \rho(x_i,\theta)$ *unimodal* in θ for all x and $n \geqslant 1$, i.e. there is a MLE $T_n \colon E^{\mathbb{N}} \to \mathbb{R}$ such that the function is increasing for $\theta < T_n(x)$ and decreasing for $>$. Then the sequence T_n is consistent.

Long proof, uses relative entropy.

Entropy The relative entropy (or Kullback-Leibler divergence) is

$$H(P,Q) \coloneqq KL(\sigma||\rho) = \mathbb{E}_P\left(\log\frac{\rho}{\sigma}\right) = -\int p(x)\ln\left(\frac{q(x)}{p(x)}\right)$$

measures how distinguishable the two probability measures are (it's $\geqslant 0$ with $=$ if $P = Q$). The entropy

$$H(X) = -\mathbb{E}(\log p(X)) = -\int p(x)\ln p(x)dx$$

measures how "scattered" the distribution is (is it always well defined?), or the average amount of information given by a random variable X with density p. Given a joint density $p(x,y)$, the conditional entropy

$$H(y|x) = -\iint p(y,x)\ln p(y|x)dydx$$

is the average additional information needed to specify y once x is known (because the additional information is $-\ln p(y|x)$ and satisfies $H(x,y) = H(y|x) + H(x)$. The mutual information $I(X,Y) = KL(p(X,Y)||p(X)p(Y))$ measures how far X and Y are from being independent, and satisfies $I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$.

Poisson, Exponential, Uniform...

2.1 Data reduction

We let $\mathbf{X} = (X_1, \ldots, X_n)$ denote the random vector and $\mathbf{x} = (x_1, \ldots, x_n)$ a particular sample.

A statistic $T(\mathbf{X})$ defines a form of data reduction or data summary of the variables $X_1, \ldots, X_n$. The image of T defines a partition of the sample space according to the value of the statistic.

Sufficiency $T(\mathbf{X})$ is a sufficient statistic for θ if $P_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t)$ does not depend on θ. (The data processing inequality is an equality in this case). Note: we assume $T(\mathbf{X})$ has a discrete distribution otherwise the condition has probability 0.

The sufficiency principle states that in this case, any inference about θ should depend on $T(\mathbf{X})$ and not on the particular sample: the statistic contains all the useful information about θ.

Indeed, the distribution on $T^{-1}(t)$ can be computed from the model without knowing θ, and we can use it to generate observations $\mathbf{Y}$ with the same conditional distribution as $\mathbf{X}$. We want to show that they have the same unconditional distribution:

$$\begin{aligned} P_\theta(\mathbf{X} = \mathbf{x}) &= P_\theta(\mathbf{X} = \mathbf{x} \wedge T(\mathbf{X}) = T(\mathbf{x})) = \\ &= P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))P_\theta(T(\mathbf{X}) = T(\mathbf{x})) = \\ &= P(\mathbf{Y} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))P_\theta(T(\mathbf{X}) = T(\mathbf{x})) = \\ &= P_\theta(\mathbf{Y} = \mathbf{x} \wedge T(\mathbf{X}) = T(\mathbf{x})) = P_\theta(\mathbf{Y} = \mathbf{x}) \end{aligned}$$

for all $\mathbf{x}, \theta$.

To verify that a statistic is sufficient, we must only verify that $P_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$ does not depend on θ. By the equalities written above, this equals $p(\mathbf{x}|\theta)/q(T(\mathbf{x})|\theta)$, the ratio of the pmf's, so the condition is that this ratio is constant in θ.

2.2 Bayesian statistics

Instead of minimizing the mean squared error of an estimator uniformly for θ, we minimize an average over θ. Consider a standard model $(X, \mathscr{F}, P_\theta \colon \theta \in \Theta)$; we choose a prior distribution density $f(\theta)$ on $(\Theta, \mathscr{B}_\Theta^d)$ expressing our belief for the parameter before any data; we choose a statistical model $f(x \mid \theta)$ (instead of $f(x, \theta)$!) expressing our belief for x given θ; observe data and calculate the posterior distribution $f(\theta \mid X_1, \ldots, X_n)$:

$$\begin{aligned} f(\theta \mid x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n \mid \theta) f(\theta)}{\int f(x_1, \ldots, x_n \mid \theta) f(\theta) d\theta} &= \frac{\prod f(x_i \mid \theta) f(\theta)}{\int \prod f(x_i \mid \theta) f(\theta) d\theta} \\ &=: \frac{L_n(\theta) f(\theta)}{c_n} \propto L_n(\theta) f(\theta) \end{aligned}$$

(first equality is a continuous version of Bayes' theorem). Now we can compute the posterior mean

$$\overline{\theta}_n = \int \theta f(\theta \mid x_1, \ldots, x_n) d\theta$$

and a $1-\alpha$ posterior confidence interval (a,b) by finding a and b such that the tails up to a and from b both have area $\alpha/2$, indeed

$$P(\theta \in (a,b) \mid x_1,\dots,x_n) = \int_a^b f(\theta \mid x_1,\dots,x_n)d\theta = 1-\alpha.$$

Bayesian approach to curve fitting We introduce a prior distribution over the coefficients $\mathbf{w}$; then we could find $\mathbf{w}$ so as to maximize the posterior distribution $f(\mathbf{w}|\mathbf{x},\mathbf{y},...) \propto f(\mathbf{y}|\mathbf{x},\mathbf{w},...)f(\mathbf{w}|...)$, but this is equivalent to the old approach. Instead, in the Bayesian approach the predictive distribution is

$$f(y|x,\mathbf{x},\mathbf{y}) = \int f(y|x,\mathbf{w})f(\mathbf{w}|\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{w}$$

2.3 Confidence regions

A *confidence region* with error level α (or confidence level $1-\alpha$) for a characteristic $\tau\colon \Theta \to \Sigma$ to be estimated is a map $C\colon X \to \mathscr{P}(\Sigma)$ such that for all $\theta \in \Theta$

$$P_\theta\Big(x \in X \mid \tau(\theta) \in C(x)\Big) \geqslant 1-\alpha.$$

(we assume $\{s \in C(-)\} \in \mathscr{F}$ for all $s \in \Sigma$). C is also called an *interval estimator* when $\Sigma = \mathbb{R}$ and every $C(x)$ is an interval.

Intuition ($\Sigma = \Theta$, $\tau = \mathrm{id}$): given x, what θ's are "most likely" to be the true one? Answer: those for which x is most likely to be observed (with probability at least $1-\alpha$). These are contained in $C(x)$. In other words, the probability of observing an x (i.e. of generating an interval) such that $\theta \notin C(x)$ is less than α; resampling in the long run, a percentage tending to $100(1-\alpha)\%$ of the confidence intervals will contain the true value.

Constructing a confidence region for θ: for each $\theta \in \Theta$, define a set of most likely outcomes as $C_\theta \coloneqq \{x \in X \mid \rho_\theta(x) \geqslant c_\theta\}$ for the smallest $c_\theta > 0$ needed to exceed the confidence level $1-\alpha$: this determines how many x we should include in order to exceed the confidence level. Then set $C(x) \coloneqq \{\theta \in \Theta \mid x \in C_\theta\}$.

In the prototypical case, the density ρ_θ is *unimodal*, i.e. it has a single peak: then C_θ is an interval around the peak, large enough to make sure the tails together have probability at most α.

Quantiles and fractiles Let $(\mathbb{R},\mathscr{B},P)$ be a probability space, $\alpha \in (0,1)$. A number $q \in \mathbb{R}$ is called an *α-quantile of P* if $P((-\infty,q]) \geqslant \alpha$ and $P([q,\infty)) \geqslant 1-\alpha$. The quantile of a real-valued random variable is defined as the quantile of its distribution.

When the data is divided in 4 groups we call them quartiles: a *lower quartile*, *median*, *upper quartile* is respectively a 1/4-, 1/2-, 3/4- quantile. When divided in 100 groups they are called percentiles: 25th percentile = 0.25-quantile = first quartile. An $(1-\alpha)$-quantile is also called an *α-fractile*.

Percentiles are used to summarize histograms that don't follow a normal distribution, and hence can't be satisfactorily summarized by giving average and SD. Conversely, if we are given a score (an observation), its percentile rank is the percentage of data below that score.

Look at the graphical representation of a quantile in the cumulative distribution function and the density function, respectively. One has:

$$\alpha = F_P(q) = P((-\infty, q]) = \int_{\infty}^{q} \rho(x)dx$$

so q is a point where the value of the distribution crosses the level α, and the area subtended by the density in $(-\infty, q]$ is α. This q is unique if F_Q is strictly increasing in a neighborhood of q.

2.4 Statistical decision

We decompose $\Theta = \Theta_0 \sqcup \Theta_1$ where $\theta \in \Theta_0$ is the *null hypothesis* (the normal case) and $\theta \in \Theta_1$ is the *alternative* (the problematic case).

Tests A *test* of Θ_0 against Θ_1 is a statistic $\varphi\colon X \to [0,1]$ used to set a decision rule so that, on the basis of the observed x, we decide as follows: if $\varphi(x) = 0$ we accept the null hypothesis; if $\varphi(x) = 1$ we reject it; if $\varphi(x) \in (0,1)$ we decide at random, with probability $\varphi(x)$ for rejection of the null hypothesis.

It is called *non-randomized* if $\varphi(x) \in \{0,1\}$ for all x: then $R \coloneqq \varphi^{-1}(1)$ is the *rejection region* or *critical region*, $\varphi^{-1}(0)$ is the *acceptance region*.

One-sided (left-sided) test: $H_0\colon \theta \leqslant \theta_0$. Two-sided test $H_0\colon \theta = \theta_0$ (simple hypothesis) i.e. a rejection region of the form $R = \{x\colon T(x) > c\}$ where T is an appropriate test statistic (an estimator of θ?) and a critical value c which is chosen so as to get the desired size α (below).

Significance and power of a test The *power function* of φ is

$$\beta_\varphi\colon \Theta \to [0,1],\ \theta \mapsto E_\theta(\varphi) = P_\theta(X \in R)$$

For $\theta \in \Theta_0$, $\beta_\varphi(\theta)$ is the probability of false rejection of H_0 (type 1 error).

For $\theta \in \Theta_1$, $\beta_\varphi(\theta)$ is called the *power* of φ at θ and is $1-$ the probability of type 2 error (false acceptance of H_0).

The *size* of φ is $\sup_{\theta\in\Theta_0} \beta_\varphi(\theta)$, the worst case probability of a type 1 error. If a test has size $\leqslant \alpha$ we say it is a test of significance level α. So, a test φ should have small level and big power, i.e. $\beta_\varphi(\theta)$ as large as possible for all $\theta \in \Theta_1$.

A test φ is a *uniformly most powerful (UMP) test* of level α if its level is α and its power exceeds that of any other test ψ of level α. The problem is finding the right balance between significance level and power of a test.

A test φ is *unbiased* of level α if, for all $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$, $\beta_\varphi(\theta_0) \leqslant \alpha \leqslant \beta_\varphi(\theta_1)$ (one decides for the alternative with a higher probability when it is right than when it is false).

Relation between confidence regions and tests For $\theta_0 \in \Theta$, an α-confidence region $\{\theta_0 \notin C(-)\}$ is the rejection region of a test for $H_0\colon \theta = \theta_0$ against $H_1\colon \theta \neq \theta_0$, and conversely, given a such test for every $\theta_0 \in \Theta$, they can be combined to contruct an α-confidence region.

Terminology in binary classification

$$P + N = (TP + FN) + (TN + FP) = \\ = (TP + FP) + (TN + FN) = \text{class}(P) + \text{class}(N)$$

- type 1 error = false positive;

  type 2 error = false negative

- false positive rate = FP/N (corresponds to the significance level, which is however an a-priori property of the test)

  the complement of

  specificity = true negative rate TN/N.

- sensitivity = recall = true positive rate = TP/P (corresponds to the statistical power of the test),

  the complement of false negative rate FN/P.

- positive predictive value, or precision: $TP/\text{class}(P)$;

  false discovery rate: its complement $FP/\text{class}(P)$

- negative predictive value: $TN/\text{class}(N)$

  false omission rate: its complement $FN/\text{class}(N)$

- prevalence: $P/\text{total} = P/(P + N)$;

  accuracy: $(TP + TN)/(P + N)$

If prevalence is low we may have high specificity but low positive predictive value.

Experiment procedure First we choose a level α and the type of test φ; φ may depend on a parameter $c \in X$ (e.g. the number of times something should be guessed correctly for us to conclude extrasensory power), which is chosen so that $\beta_\varphi(\theta) \leqslant \alpha$ for all $\theta \in \Theta_0$.

The threshold c (hence, the size) has to be chosen before the experiment. If we modify it after the experiment we are essentially using a test $\psi = 1$ identically (unless we repeat the experiment...).

Note that depending on the probability P_θ involved, certain thresholds may give a better power function even if they require a lower percentage of successes.

To avoid fixing a threshold, we use p-values:

p-value

- Abstract definition: a statistic with $p(X) \in [0, 1]$ and $\sup_{\theta\in\Theta_0} P_\theta(p(X) \leqslant \alpha) \leqslant \alpha$. Given this, we can construct a level α test as a test that rejects H_0 iff $p(X) \leqslant \alpha$.

- If we are given tests T_α of level $\alpha = \sup_{\theta\in\Theta_0} P_\theta(X \in R_\alpha)$ for every $\alpha \in (0, 1)$ with rejection regions R_α, then $p(x) = \inf\{\alpha \mid T_\alpha(x) \in R_\alpha\}$ is the smallest level at which H_0 is rejected.

- Most common case: $\{T > c\}$, $p(x) = \sup_{\theta\in\Theta_0} P_\theta(T \geqslant T(x))$. Probability of observing $T \geqslant T(x)$ under the null hypothesis: the smaller it is, the stronger the evidence for H_1.

Don't confuse it with the test size: p-value is a property of the result, whereas size is a property of the test. A result is statistically significant when the p-value is $\leqslant \alpha$.

p-hacking or data snooping is the practice of testing multiple hypotheses on a single data set in order to find a statistically significant result which is in fact a spurious correlation.

Neyman-Pearson test We can find the most powerful test of level α for a standard model of the form $(X, \mathscr{F}; P_0, P_1)$. A Neyman-Pearson test has the form $\varphi(x) = 1$ if $R(x) \coloneqq \rho_1(x)/\rho_0(x) > c$ (*likelihood ratio*) for some threshold c, and 0 if $R(x) < c$. Choose c an α-fractile of $P_0 \circ R^{-1}$. What value if $R(x) = c$? We want α to equal $\mathbb{E}_0(\varphi) = P_0(R > c) + \gamma P_0(R = c)$ for some γ.

$$\begin{aligned} P_0(R = c) &= -P_0(R > c) + P_0(R \geqslant c) \\ &\geqslant -P_0(R > c) + \alpha \end{aligned}$$

If $P_0(R = c) = 0$ then $P_0(R > c) = \alpha$ and we can take $\varphi(x) = 0$, otherwise take $\gamma \coloneqq (\alpha - P_0(R > c))/P_0(R = c)$. Every such Neyman-Pearson test is a most powerful test of level α, and every most powerful test of this level is indistinguishable from a Neyman-Pearson test.

How quickly does the power increase for independently repeated observations? Consider the infinite product model $(X = E^{\mathbb{N}}, \mathscr{F} = \mathscr{E}^{\otimes\mathbb{N}}, P_\theta = Q_\theta^{\otimes\mathbb{N}} : \theta \in \{0, 1\})$ associated to a simple model as above, and φ_n be size-α Neyman-Pearson tests using only observations $X_1, \ldots, X_n$, where X_i are the projections. Then $\mathbb{E}_1(\varphi_n) \approx 1 - e^{-n}H(Q_0, Q_1)$, more precisely

$$\lim_n \frac{1}{n} \log(1 - \mathbb{E}_1(\varphi_n)) = -H(Q_0, Q_1)$$

.

One-sided tests Neyman-Pearson can be generalized for models with increasing likelihood ratios wrt a statistic $T\colon X \to \mathbb{R}$, i.e. $T(x) < T(y)$ implies $R_{\theta':\theta}(x) < R_{\theta':\theta}(y)$. We can test the one-sided (left-sided) test problem $H_0\colon \theta \leqslant \theta_0$ with a UMPT φ of level α of the form

$$\varphi(x) = \begin{cases} 1 & T(x) > c \\ \gamma & T(x) = c \\ 0 & T(x) < c \end{cases}$$

with c and γ uniquely determined by the condition $G_\varphi(\theta_0) = \alpha$. Moreover, G_φ is increasing.

Construct φ as in the simple case. To show it's a UMPT we use the increasing property just mentioned: take some $\theta < \theta'$, then $R_{\theta':\theta}(x) < R_{\theta':\theta}(c)$ implies $T(x) < T(c)$ hence $\varphi(x) = 0$ and $>$ implies $\varphi(x) = 1$, so φ is a N-P test of the simple null hypothesis $\{\theta\}$ against $\{\theta'\}$; for $\theta = \theta_0 < \theta' \in \Theta_1$ this shows it's a UMPT. To show the level is α, we should look at all other $\theta \in \Theta_0$ too;

it's enough to show G_φ is increasing. Since the simple test of θ against θ' is more powerful than the constant test ψ, then $G_\varphi(\theta') > G_\psi(\theta') = G_\varphi(\theta)$ and we conclude.

The right-sided hypothesis can be tested multiplying both T and θ by -1.

2.5 Common tests

First some useful distributions that will show up.

χ, F, t-distributions Let $X_i \sim N_{0,1} \sim Y_j$ iid. Then:

The chi-square χ^2_n is $\Gamma_{1/2,n/2}$, which is shown to be the distribution of $\sum_{i=1}^n X_i^2$.

Fisher's $F_{m,n}$-distribution is the distribution of $(\frac{1}{m}\sum_1^m X_i^2)/(\frac{1}{n}\sum_1^n Y_i^2)$.

Student's t_n-distribution is the distribution t_n of $T = X_1/\sqrt{\frac{1}{n}\sum_{j=1}^n Y_j^2}$. Note that $T^2 \sim F_{1,n}$.

n-fold Gaussian product model $(\mathbb{R}^n, \mathscr{B}^n, N_{m,v}^{\otimes n} \colon m \in \mathbb{R}, v > 0)$. Student's theorem:

M and V^* are independent; $M \sim N_{m,v/n}$; $\frac{n-1}{v}V^* \sim \chi^2_{n-1}$;

$T_m \coloneqq (M-m)\sqrt{n/V^*} \sim t_{n-1}$.

We compute a $(1-\alpha)$-confidence interval for the expectation m as $(M \mp t\sqrt{V^*/n})$ where t is the $\alpha/2$-fractile of t_{n-1}. Indeed, $P_\theta \circ T_{m(\theta)}^{-1}$ is shown to equal $N_{0,1}^{\otimes n} \circ T_0^{-1} = t_{n-1}$ so it does not depend on θ and it is symmetric; moreover it has a Lebesgue density which is decreasing on $[0,\infty)$, hence the shortest interval with probability $\geqslant 1-\alpha$ is centered around the origin and it's $(-t,t)$, so we set $C_m = T_m^{-1}(-t,t)$.

The 95% CI for a Gauss distribution is given by $t\sqrt{V^*} = t\sigma^*$ where t is the $1-\alpha/2 = 0.975$-quantile of $N_{0,1}$ i.e. $t = \Phi^{-1}(0.975) = 1.96$.

Tests for the n-fold Gaussian model The following are UMP tests of level α.

General procedure: we measure by how many SD's the observed value of the statistic (score) differs from a benchmark (e.g. μ_0, chosen depending on the null hypothesis) and calculate the p-value; if it's low enough, we can reject the null hypothesis.

- One-sided χ^2-test of $H_0 \colon v \leqslant v_0$ with rejection region

$$\sum_{i=1}^n (X_i - \overline{X})^2 > v_0 \chi^2_{n-1,1-\alpha}$$

  is a UMP test of level α. If μ is known replace $\overline{X}$ by μ; then the statistic is $\sim \chi^2_n$ (so we should use $\chi^2_{n,1-\alpha}$) and the product model is exponential wrt it, so by the above theory we get a UMPT.

  If μ is unknown we replace it with $\overline{X}$ and by the above theory the $nV/v \sim \chi^2_{n-1}$ so the one in the statement is a test of level α. To check that it's a most powerful test is more complicated.

The symmetric statement does not hold: a UMPT does not exist in this case; the test of $H_0 \colon v \geqslant v_0$ with rejection region

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 < v_0 \chi^2_{n-1,\alpha}$$

is a UMP test only within the class of *unbiased* level-α tests.

- One-sided Student's t-test of $H_0 \colon \mu \leqslant \mu_0$ with rejection region

$$T(X) \coloneqq \frac{\overline{X} - \mu_0}{S^*/\sqrt{n}} > t_{n-1,1-\alpha}$$

(the α-fractile of the t_{n-1} distribution) is a UMP test within the class of unbiased level-α tests (a UMPT does not exist). Invert inequalities for the opposite test.

One-sided Gauss test: if σ is known, we can replace S^* by σ and the statistic has a $\mathcal{N}(0,1)$ distribution so we replace $t_{n-1,1-\alpha}$ by $\Phi^{-1}(1-\alpha)$. This is a UMP test.

- Two-sided Student's t-test of $H_0 \colon \mu = \mu_0$ with rejection region

$$\frac{|\overline{X} - \mu_0|}{S^*/\sqrt{n}} > t_{n-1,1-\alpha/2}$$

is a UMP test within the class of unbiased level-α tests. If σ is known this is just the Wald test below.

Z-test The Z-score or standard score of an observation (raw score) x is $(x - \mathbb{E}(X))/\sigma(X)$, the distance measured in standard deviations of x from the population mean (this means converting the value to standard units). The p-values are $\Phi(Z)$, $\Phi(-Z)$, $2\Phi(-|Z|)$ for left-, right-, two-sided tests respectively.

When μ, σ are not known we use $(x - \overline{X})/S$ as above. If we know the actual σ we replace S^* in the t-tests above and we use normal quantiles instead of t-quantiles and we call them Z-tests.

To compare whether two means are the same, 2-sample Z-test:

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 + \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Also 2-sample t-test when variances unknown.

Wald test For asymptotically normal $\widehat{\theta}$ the Wald test of size α (asymptotically) of $H_0 \colon \theta = \theta_0$ rejects when $|(\widehat{\theta} - \theta_0)/\widehat{\mathrm{se}}| > z_{\alpha/2}$. Equivalently we are testing whether θ_0 is in the asymptotic $1-\alpha$ confidence interval $\widehat{\theta} \mp \widehat{\mathrm{se}} z_{\alpha/2}$. If the true value is $\theta_* \neq \theta_0$ the power $\beta(\theta_*)$ is approximately... We approximate the p-value using Φ as above.

Permutation test $X_1, \ldots, X_m \sim F_X$ and $Y_1, \ldots, Y_n \sim F_Y$, we want to test $H_0 \colon F_X = F_Y$. Consider $T(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = |\overline{X}_m - \overline{Y}_n|$. Compute $T_1, \ldots, T_B$ by applying $B \leqslant N!$ different permutations to the arguments of T: under the null hypothesis they are uniformly distributed, we compute the approximate p-value $P_0(T > t) = \frac{1}{B}\sum_{i=1}^{B} I(T_j > t)$.

Likelihood ratio test for $H_0\colon \theta \in \Theta_0$; the likelihood ratio statistic is $\lambda = 2\log\left(\frac{\sup_\Theta \mathcal{L}(\theta)}{\sup_{\Theta_0}\mathcal{L}(\theta)}\right) = 2\log\left(\frac{\mathcal{L}(\widehat{\theta})}{\mathcal{L}(\widehat{\theta}_0)}\right)$ (hats denote MLE's).

If $\theta = (\theta_1, \dots, \theta_r)$ and Θ_0 is a subset of vectors having the last coordinates fixed, $\Theta_0 = \{\theta\colon (\theta_{q+1}, \dots, \theta_r) = (\theta_{0,q+1}, \dots, \theta_{0,r})\}$, then $\lambda \rightsquigarrow \chi^2_{r-q,\alpha}$ and the p-value is $P(\lambda < \chi^2_{r-q})$.

Multiple testing We have m tests of hypotheses H_{0i} with p-values P_i, $i = 1, \dots, m$. The chance of at least one false rejection is higher than that of each single test. The Bonferroni method is to reject H_{0i} if $P_i < \alpha/m$; then the probability of falsely rejecting any null hypothesis is at most α (by subadditivity of probability).

2.6 Goodness of fit and independence tests

Empirical distribution The absolute frequency $h_n(i) = \#\{k \in \{1, \dots, n\} \mid X_k = i\}$; the vector of relative frequencies $L_n = (\frac{h_n(1)}{n}, \dots, \frac{h_n(s)}{n})$ is the *histogram*. The *empirical distribution* is $L = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and it's the unique MLE of θ; empirical distribution function $\widehat{F}_n\colon x \mapsto \frac{1}{n}\sum_{i=1}^n I(X_i \leqslant x)$. The quantiles of L are called *sample quantiles* (order statistics below are a special case). The Glivenko-Cantelli theorem states that $\sup_x|\widehat{F}_n(x) - F(x)|$ converges to 0 almost surely. (note: histograms as a representation can also have bars with variable width where vertical axis is frequency density and the bars have area proportional to frequency).

A statistical functional is any statistic $T(F)$. Its plug-in estimator is $T(\widehat{F}_n)$. For a linear functional $\int r(x)dF(x)$ the plug-in estimator becomes $\frac{1}{n}\sum r(X_i)$.

Likelihood-ratio χ^2-goodness-of-fit test Test of the null hypothesis $H_0\colon \theta = \rho$ to test, after n trials, if the observed distribution is the one we expect. The test should reject for big enough values of the log-likelihood ratio

$$\log\frac{\sup_{\theta\neq\rho}\prod_{i=1}^s \theta(i)^{h_n(i)}}{\prod_{i=1}^s \rho(i)^{h_n(i)}} = \log\frac{\sup_{\theta\in\Theta}\cdots}{\cdots} = n\sum_{i=1}^s L_n(i)\log\frac{L_n(i)}{\rho(i)} = nH(L_n,\rho)$$

which has the quadratic approximation

$$2nH(L_n,\rho) \xrightarrow{P_\rho} D_{n,\rho} \coloneqq n\sum_{i=1}^s \frac{L_n(i)^2}{\rho(i)} - n$$

and under H_0 this $\rightsquigarrow \chi^2_{s-1}$ i.e. $\lim_n P_\rho(D_{n,\rho} \leqslant c) = \chi^2_{s-1}([0,c])$. So the test with rejection region

$$D_{n,\rho} > \chi^2_{s-1,1-\alpha}$$

has approximately size α for n large enough (for small n one should use the exact distribution of $D_{n,\rho}$ which can be derived from the multinomial distribution). (write it in terms of p-value?)

Kolmogorov-Smirnov test One-sample K-S test for H_0: (the sample comes from F); statistic $D_n = \sup_x |\widehat{F}_n(x) - F(x)|$; under H_0 then $D_n \to 0$ almost surely (Glivenko-Cantelli theorem) and $\sqrt{n}D_n \rightsquigarrow \sup_t |B(F(t))|$; the Kolmogorov distribution is that of

$$K = \sup_{t\in[0,1]} |B(t)| \qquad \text{with cdf} \qquad F(x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}\exp(-2k^2x^2)$$

so the test rejects at level α if $\sqrt{n}D_n > K_{1-\alpha}$ where $P(K \leqslant K_{1-\alpha}) = 1-\alpha$.

Two-sample K-S test for H_0: $F_1 = F_2$. $D_{n,m} = \sup_x |\widehat{F}_{1,n}(x) - \widehat{F}_{2,m}(x)|$, we reject when $\sqrt{\frac{nm}{n+m}}D_{n,m} > K_{1-\alpha}$.

Pearson's χ^2 multinomial test $X = (X_1, \ldots, X_k) \sim$ multinomial $(n, \mathbf{p})$, the MLE of $\mathbf{p}$ is $\widehat{\mathbf{p}} = (X_1/n, \ldots, X_k/n)$. We want to test H_0: $\mathbf{p} = \mathbf{p}_0$.

$$T(X) = \sum_{j=1}^{k} \frac{(X_j - \mathbb{E}(X_j))^2}{\mathbb{E}(X_j)} = \sum_{j=1}^{k} \frac{(X_j - np_{0j})^2}{np_{0j}}$$

and if H_0 is true then $T \rightsquigarrow \chi^2_{k-1}$ so the test rejecting if $T > \chi^2_{k-1,\alpha}$ has asymptotic level α.

Pearson's χ^2 independence test Under the hypothesis H_0 of independence,

$$\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(X_{ij} - E_{ij})^2}{E_{ij}} \rightsquigarrow \chi^2_{(I-1)(J-1)}$$

where $E_{ij} = n\widehat{p}_{ij} = n\widehat{p}_{i\bullet}\widehat{p}_{\bullet j} = \frac{X_{i\bullet}X_{\bullet j}}{n}$ (we used H_0) is the expected number of observations in cell ij.

Likelihood-ratio χ^2-test of independence A contingency table (or cross tabulation) records $h_n(ij)$ for $ij \coloneqq (i,j) \in E = A \times B = \{1,\ldots,a\} \times \{1,\ldots,b\}$. Consider all strictly positive densities on E

$$\Theta = \left\{ \theta = \theta(ij)_{ij\in E} \in (0,1)^E \colon \sum_{ij\in E} \theta(ij) = 1 \right\},$$

null hypothesis H_0: $\theta = \theta^A \otimes \theta^B$ product of the marginal distributions, $\theta^A(i) = \sum_{j\in B}\theta(ij)$ (denote Θ_A the set of marginal distributions on A...), so

$$\Theta_0 = \left\{ \alpha \otimes \beta = (\alpha(i)\beta(j))_{ij\in E} : \alpha \in \Theta_A, \beta \in \Theta_B \right\}$$

The empirical joint distribution is $L_n = (L_n(ij))_{ij\in E} = (h_n(ij)/n)_{ij\in E}$. Accept H_0 if L_n is sufficiently close to $L_n^A \otimes L_n^B$, with the distance measured similarly as before computing the log-likelihood ratio

$$\sum_{ij\in E} h_n(ij) \log\left(\frac{L(ij)}{L^A(i)L^B(j)}\right) = nH(L, L^A \otimes L^B)$$

and finding an analogous statistic $\widetilde{D}_n$ (as for the goodness-of-fit test) which $\rightsquigarrow \chi^2_{(a-1)(b-1)}$ under P_ρ, so a test with approximate size α is the one with rejection region

$$\widetilde{D}_n > \chi^2_{(a-1)(b-1);1-\alpha}$$

.

Fisher's exact test Above situation with $A = B = \{1, 2\}$ (e.g. positive or negative outcome of medical test). For $\theta \in \Theta_0$ we have

$$P_\theta(h_n 11 = h_{11} \mid h_n^A 1 = a_1, h_n^B 1 = b_1) =$$
$$= \mathcal{H}_{b_1;a_1,a_2}(\{h_{11}\}) = \frac{\binom{a_1}{h_{11}}\binom{a_2}{h_{12}}}{\binom{n}{b_1}} = \frac{\binom{b_1}{h_{11}}\binom{b_2}{h_{21}}}{\binom{n}{a_1}} = \mathcal{H}_{a_1;b_1,b_2}(\{h_{11}\})$$

(the probability that h_{11} elements are positive in a random sample, without replacement, of b_1 elements chosen from a set of n of which a_1 are positive: it's a hypergeometric distribution).

	$A1$	$A2$	
$B1$	h_{11}	h_{12}	b_1
$B2$	h_{21}	h_{22}	b_2
	a_1	a_2	n

2.7 Epidemiology

Odds ratio The odds of an event A are defined as $P(A)/(1 - P(A))$. (Then $P(A) = \text{odds}/(1 + \text{odds})$. Trick: if odds are m/n, probability is $m/(m + n)$; if probability is m/n, odds are $m/(n - m)$).

The odds of $A|B$ are defined in the same way, the odds ratio is then defined as $\psi = \text{odds}(D|E)/\text{odds}(D|E^c)$, the log-odds ratio is the log of this quantity. ψ measures the association between the events, e.g. odds of a disease conditioned to the event of being exposed to something are the same (independent events) iff $\psi = 1$. With notation of a contingency table, $\psi = p_{11}p_{00}/p_{01}p_{10}$. Note it's the same as the odds ratio of $B|A$.

How do we estimate it? 3 methods yielding the same result:

- Multinomial sampling: we draw a sample from the population, then $X = (X_{ij})_{ij} \sim \mathcal{M}(n, p)$, use $\widehat{p}_{ij} = X_{ij}/n$ hence

$$\widehat{\psi} = X_{11}X_{00}/X_{01}X_{10}.$$

- Cohort sampling: we are given some exposed and non-exposed people and we count the number of people with the disease in both groups:

  $X_{01} \sim \text{binom}(x_{0\bullet}, P(D|E^c))$,

  $X_{11} \sim \text{binom}(x_{1\bullet}, P(D|E))$

  so we can estimate $\widehat{P}(D|E) = X_{11}/x_{1\bullet}$, $\widehat{P}(D|E^c) = X_{01}/x_{0\bullet}$ which is enough to compute $\widehat{\psi}$ which is the same as above.

- Case-control sampling: we are given some people with and without disease and we count how many were exposed.

  $X_{10} \sim \text{binom}(x_{\bullet 0}, P(E|D^c))$,

  $X_{11} \sim \text{binom}(x_{\bullet 1}, P(E|D))$

  and, analogously as above, we find estimates of $P(E|D^c)$, $P(E|D)$; we compute the odds ratio of $E|D$ which is the same as for $D|E$ and yields the same $\widehat{\psi}$ as before.

  Note: $P(D|E) - P(D|E^c)$ can't be estimated in this setting because, using Bayes' theorem, we would need an estimate of $P(D)$ which we don't have. However we can estimate the relative risk if the disease is rare (rare disease assumption):

$$\text{RR} = \frac{P(D|E)}{P(D|E^c)} \to \psi \text{ as } P(D) \to 0$$

  .

Note: both cohort and case-control studies can be retrospective or prospective, depending on whether we look back at existing data (post hoc) or we enroll new participants during the study.

Epidemiology terminology In an experiment with an experimental group of E subjects and control group with C subjects, we have:

- the number of events in each group: EE and CE; the event rates for each group: EER = EE/E and CER = CE/C (this, which estimates $P(D|E^c)$, is the baseline risk); the term incidence (in a given group) may refer to either the number or the rate of events in that group; do not confuse it with prevalence, which is $P(D)$, the rate of the disease in the population, whereas incidence has to do with the occurrence of new cases.
- the absolute risk reduction ARR = CER − EER (resp. increase changing the sign); the general term risk difference can mean either ARR or ARI.
- the number needed to treat NNT = ARR^{-1} (resp. needed to harm);
- relative risk (or risk ratio) RR = EER/CER
- relative risk reduction RRR = (CER−EER)/CER = 1−RR (resp. relative risk increase changing the sign);
- odds ratio OR = $\frac{\text{EE}/(1-\text{EE})}{\text{CE}/(1-\text{CE})}$.

Efficacy?? $V + N = P$. Efficacy $E = \frac{M_N/N - M_V/V}{M_N/N} = 1 - \frac{M_V}{M_N}\frac{N}{V}$ so $M_V = (1-E)\frac{V}{N}M_N$.

$$\frac{M_V}{M_P} = \frac{M_V}{M_V + M_N} = \frac{(1-E)\frac{V}{N}M_N}{M_N + (1-E)\frac{V}{N}M_N} = \frac{(1-E)\frac{V}{N}}{1 + (1-E)\frac{V}{N}}$$

(if $V = N$ this is EER/(CER+EER)).

Conversely,

$$E = 1 - \frac{\frac{M_V}{M_P}\frac{N}{V}}{1 - \frac{M_V}{M_P}}.$$

Causal inference If Y measures the health status of a patient, let Y^1, Y^0 equal Y respectively when the patient is treated or not (counterfactual model). Of course in real life we can only observe one of the two values for each patient. The average causal effect or average treatment effect

$$\theta = \text{ATE} = \mathbb{E}(Y^1) - \mathbb{E}(Y^0) = \frac{1}{n}\sum_{i=1}^{n}(Y_i^1 - Y_i^0)$$

where n is the total number of patients, can only be estimated.

If Y is binary we may alternatively use causal odds ratio $\frac{P(Y^1=1)/P(Y^1=0)}{P(Y^0=1)/P(Y^0=0)}$, causal relative risk $P(Y^1 = 1)/P(Y^0 = 1)$.

Define association as $\alpha = \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0)$ (where $X = 1, 0$ means receiving treatment or not, respectively). In general $\alpha \neq \theta$ (association is not causation); we may have $\alpha > 0$ with $\theta = 0$ or even $\theta < 0$. An association between X and Y can be created by an association between (Y^0, Y^1) and X, e.g. healthy user bias. However, if the subjects are assigned to treatment randomly, then $\alpha = \theta$ because X and (Y^0, Y^1) are independent; by the law of large numbers we have a consistent estimator

$$\begin{aligned}\widehat{\theta} = \widehat{\alpha} &= \widehat{\mathbb{E}}(Y|X = 1) - \widehat{\mathbb{E}}(Y|X = 0) \\ &= \frac{1}{n_T}\sum_{i=1}^{n} Y_i X_i - \frac{1}{n - n_T}\sum_{i=1}^{n} Y_i(1 - X_i)\end{aligned}$$

where $n_T = \sum_1^n X_i$ is the number of treated patients.

Suppose the treatment is not binary but $X \in \mathbb{R}$ represents the dose of a drug. We replace (Y^0, Y^1) with a function Y^x, the causal regression function is $\theta(x) = \mathbb{E}(Y^x)$, the regression function is $r(x) = \mathbb{E}(Y|X = x)$. Random assignment implies $\theta(x) = r(x)$; otherwise we may have constant θ (no causation) with non-constant r (association).

Adjusting for confounding If treatment is not assigned randomly, the study is called observational: the patients select their own value X of treatment. In this case we should control for confounding variables, i.e. find groupings of the subjects such that, within each group, subjects have similar characteristics and hence we may assume that X is essentially random, that is, X and Y^x are independent. Denote these groups with a random variable Z: then

$$\theta(x) = \int \mathbb{E}(Y|X = x, Z = z)dF_Z(z)$$

(compare with $r(x) = \mathbb{E}(Y|X = x) = \int \mathbb{E}(Y|X = x, Z = z)dF_{Z|X}(z|x)$).

(A confounder is a variable causing the first and second variables. On the other hand, a mediator is caused by the first variable and causes the second.)

Simpson's paradox A treatment that is beneficial for men and beneficial for women seems to be harmful overall if we have $P(Y = 1|X = 1) < P(Y = 1|X = 0)$; the explanation is that this formula does not mean that the treatment is harmful, which should be written as $P(Y^1 = 1) < P(Y^0 = 1)$.

If in group i ($i = 1, 2$) there are m_i admitted males over M_i male applicants, and analogously for females, then we may have $m_i/M_i < f_i/F_i$ in both groups

and still $\frac{m_1+m_2}{M_1+M_2} > \frac{f_1+f_2}{F_1+F_2}$ if more men apply to the group with highest admission rate. Analogously: let 1 and 2 be the groups of respectively young and old people, T_i, N_i denote respectively treated and non-treated status against a given disease, and t_i, n_i be the numbers of subjects in the two categories who have died from the disease: if treated people tend to be old ($N_1 \ll T_2, n_1 \ll t_2$), we may have $\frac{t_1+t_2}{T_1+T_2} > \frac{n_1+n_2}{N_1+N_2}$ even if the treatment works, and controlling for group will indeed reveal that it works.

2.8 Non-parametric methods

We don't assume a particular class of distributions.

Order and rank tests We only make use of the natural ordering of $\mathbb{R}$. Take iid $X_1, \dots, X_n \sim Q$ unknown but assumed continuous, i.e. $Q(\{x\}) = 0$ for all $x \in \mathbb{R}$, equiv. F_Q continuous (true if Q has a density). This ensures that there are no ties, i.e. that $P(X_i = X_j) = 0$ for $X_i \neq X_j$.

The *order statistics* $X_{1:n}, \dots, X_{n:n}$ are the X_i put in increasing order. If there are ties they can be defined more generally as $X_{j:n} = \min\{c \in \mathbb{R} \mid \sum_{i=1}^n 1_{X_i \leqslant c} \geqslant j\}$ (the smallest number exceeding at least j of the other outcomes, the smallest j/n quantile of the empirical distribution L; the collection of order statistics is in 1-1 correspondence with L).

The median only depends on the ordering: if μ a median of Q, $T\colon \mathbb{R} \to \mathbb{R}$ order-preserving, then $T(\mu)$ is a median of $Q \circ T^{-1}$.

CI for the median Let $b_n(\alpha)$ be the largest α-quantile of the binomial distribution $B_{n,1/2}$; then

$$[X_{k:n}, X_{n-k+1:n}] \quad \text{with } k = b_n(\alpha/2)$$

is a confidence interval for $\mu(Q)$ of level α. The *sample median* is

$$\mu(L) = \begin{cases} X_{k+1:n} & n = 2k+1 \\ \frac{X_{k:n}+X_{k+1:n}}{2} & n = 2k \end{cases}$$

Sign test for the median Iid $X_1, \dots, X_n \sim Q$ continuous, $\mu_0 \in \mathbb{R}$. A test of level α for the two-sided problem $H_0\colon \mu(Q) = \mu_0$, the acceptance region $\{X_{k:n} \leqslant \mu_0 \leqslant X_{n-k+1:n}\}$ with $k \coloneqq b_n(\alpha/2)$;

for the one-sided problem $H_0\colon \mu(Q) \leqslant \mu_0$, the acceptance region $\{X_{k:n} \leqslant \mu_0\}$ with $k \coloneqq b_n(\alpha/2)$ whose power function is a strictly increasing function of $p(Q) \coloneqq Q(\mu_0, \infty)$.

One-parameter exponential family with strictly increasing statistic and coefficient function: then $P_\theta \prec P'_\theta$ for $\theta < \theta'$.

Rank statistics of a sample sequence $X_1, \dots, X_n$ are $R_1, \dots, R_n$ given by $R_i = \#\{j \mid X_j \leqslant X_i\}$ is the position of X_i in the ordered statistics, hence $X_i = X_{R_i:n}$.

Wilkoxon signed-rank test ...

Mann-Whitney U test (or Wilcoxon rank-sum test) Suppose we have $n = k+l$ observations with $X_1, \ldots, X_k \sim P$ and $X_{k+1}, \ldots, X_n \sim Q$. We want to test $H_0 \colon P = Q$ against $H_1 \colon P \prec Q$ where $P \preceq Q$ means stochastic dominance, i.e. $P(c, \infty) \leqslant Q(c, \infty)$ for all $c \in \mathbb{R}$. This can be used as a test for a difference in medians, i.e. to check if one group has a significantly higher median than the other.

Set $S_P \coloneqq R_1 + \ldots R_k$ and $S_Q \coloneqq R_{k+1} + \cdots + R_{k+l}$. If $S_P < S_Q$ the majority of outcomes of the P-sample is further to the left of the real axis. How much smaller should it be to conclude $P \prec Q$? Sufficient to consider $S \coloneqq S_P$ since $S_P + S_Q = n(n+1)/2$. We're interested in the U statistic:

$$S = U + \frac{k(k+1)}{2}, \quad U = U_{k,l} \coloneqq \sum_{i=1}^{k} \sum_{j=k+1}^{k+l} 1_{X_i > X_j}$$

which can also be defined as

$$U = \min_{i=1,2} \left(n_1 n_2 + \frac{n_i(n_i+1)}{2} - S_i \right).$$

Then the test has rejection region $\{U < c\} = \{W < c + k(k+1)/2\}$, where U, under the hypotheses $P = Q$ continuous, has the distribution:

$$P^{\otimes n}(U = m) = N(m; k, l) \Big/ \binom{n}{k}$$

where N is the number of partitions of $\sum_{i=1}^{k} m_i = m$ into $m_1 \leqslant m_2 \cdots \leqslant m_k$ from the set $\{0, \ldots, l\}$. In particular $P^{\otimes n}(U = m) = P^{\otimes n}(U = n - m)$.

$N(m; k, l)$ can be computed recursively as $\sum_{j=0}^{k} N(m - j; j, l - 1)$.

For large values of k, l, a normal approximation exists.

The corresponding two-sided problem has rejection region $\{U_P < c\} \cup \{U_Q < c\}$.

Density estimation methods

Histogram method Out of N observations, n_i of them fall into bin i of width Δ_i; then we estimate $p_i = n_i / N\Delta_i$. The width determines the smoothness of the model (from biased to overfitting). Computational advantage: the data set can be discarded once the histogram has been computed. Issues: the estimated density is discontinuous; curse of dimensionality: for M bins and D variables we have M^D total bins.

General idea: an idea of local neighborhood based on some distance (here Euclidean) and a smoothing parameter. We start from an approximation $p(\mathbf{x}) = K/NV$ of the probability density of $\mathbf{x}$ falling into a region of volume V containing K points. Fixing V will give kernel methods, fixing K will give nearest-neighbor methods.

Note: in the following methods the training set must be stored in order to be also used during the prediction phase (memory-based methods).

Kernel methods we estimate the density at $\mathbf{x}$ with

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} K_\lambda (\mathbf{x} - \mathbf{x}_n)$$

where K_λ is a kernel function with bandwidth λ; for example, the basic idea is to use $K_\lambda = \mathbb{1}_{[-\lambda/2,\lambda/2]}/\lambda^D$ so that we get the number of data points inside a cube of side λ centered at $\mathbf{x}$, divided by the region volume. However, this is still discontinuous, so we can choose smoother kernels, as long as they satisfy $K(\mathbf{u}) \geqslant 0$ and $\int K(\mathbf{u}) d\mathbf{u} = 1$. The Gaussian kernel is

$$K_\sigma(\mathbf{u}) = \varphi_{\mathbf{u},\sigma^2}$$

Issue: we want to be able to change λ depending on the data density of the region.

Nearest-neighbors Now we fix the parameter K and let V be the volume of the sphere containing exactly K points.

It can be extended to classification: the estimates $p(\mathbf{x}|C_k) = K_k/N_k V$, $p(\mathbf{x}) = K/NV$, $p(C_k) = N_k/N$, and using Bayes' theorem the posterior probability is

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{p(\mathbf{x})} = \frac{K_k}{K}$$

so the classifier assigns to the class with highest probability.

Chapter 3

Regression analysis

To study the relationship between X and Y we want to estimate the regression function, or conditional expectation

$$r(x) \coloneqq \mathbb{E}(Y|X=x) = \int yf(y|x)dy$$

using data $(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{X,Y}$. Note that r is found by minimizing $\mathbb{E}(Y - f(X))^2 = \int (y - f(x))^2 dF_{X,Y}(x, y)$ pointwise for each x.

The name *regression* was used by Galton and refers to the tendency of the response variable to "regress" towards the mean as the covariate approaches extreme values on either side.

3.1 Linear models

A linear model for n observations is, written in vector form,

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

(we may use a more general linear basis function model replacing $\mathbf{X}$ with $\mathbf{\Phi}$). $\mathbf{X} \in \mathbb{R}^{n \times s}$ is the design matrix with full rank $s < n$ (this is required to invert $\mathbf{X}^t\mathbf{X}$, it means that there is no perfect multicollinearity i.e. dependence between the predictors; the rank is the number of degrees of freedom; if $\mathbf{X}$ close to singular use singular value decomposition?), $\beta = (\beta_1, \ldots, \beta_s)^t$ parameter vector to be estimated, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^t$ error vector with $\varepsilon_i \in \mathscr{L}^2$ with known distribution and mean zero, so

$$\mathbb{E}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\beta.$$

Assumptions The following assumptions are possible:

- Normal (or Gaussian) linear models: ε_i iid $\sim \mathcal{N}(0, \sigma^2)$.
- Gauss-Markov assumptions are more general:
  - zero-mean $\mathbb{E}(\varepsilon_i) = 0$
  - homoscedastic, i.e. same finite variance $\mathbb{V}(\varepsilon_i) = v$

  - non-autocorrelated $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ if $i \neq j$, or equivalently $\mathbb{E}(\varepsilon_i \varepsilon_j) = 0$ (though not necessarily independent).

- Generalized Gauss-Markov: $\mathbb{E}(\varepsilon_i) = 0$, general covariance matrix (possibly correlated or heteroscedastic). Used in generalized least squares.

In a non-experimental science like econometrics we use instead the conditional means in our assumptions: $\mathbb{E}(\varepsilon_i | \mathbf{x}_1, \ldots, \mathbf{x}_n) = 0$ for $i = 1, \ldots, n$ (exogeneity), which implies: $\mathbb{E}(\varepsilon_i) = 0$ (by the law of total expectation), orthogonality $\mathbb{E}(\mathbf{x}_j \cdot \varepsilon_i) = \mathbf{0}$, and $\mathrm{Cov}(\varepsilon_i, x_{jk}) = 0$.

Note on multicollinearity Perfect multicollinearity is defined as a relationship $\lambda_0 + \sum_{i=1}^{p} \lambda_i X_i = 0$ where X_i is the vector of observations for the i-th parameter; general multicollinearity is defined adding an error term to the equation so that the relation can only hold approximately, which gives rise to invertible but possibly ill-defined matrices in the OLS model.

Signs of multicollinearity are big changes when modifying the predictors, insignificant coefficients for the affected variables despite positive joint F-tests or significant coefficients in the corresponding simple regression.

The statistical model is

$$(\mathbb{R}^n, \mathscr{B}^n, P_{\beta,v} \mid \beta \in \mathbb{R}^s, v > 0)$$

where $P_{\beta,v}$ is the distribution of the random vector $A\beta + \varepsilon$.

Different design matrices $\mathbf{X}$ and error vectors give different models: a trivial example is the Gaussian product model $\mathcal{N}_{m,v}^{\otimes n}$ obtained when $\mathbf{X} = (1, \ldots, 1)^t$, $\beta = m \in \mathbb{R}$, $\varepsilon \sim \mathcal{N}_n(0, \mathbb{1}v)$. Simple linear regression is given by $s = 2$,

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

and analogously for multiple linear regression, polynomial regression. A Gaussian linear model is any linear model with $\varepsilon \sim \mathcal{N}_n(0, \mathbb{1}v)$.

(For multiple regression with k outputs to predict we replace $\mathbf{y}$ with an $n \times k$ matrix $\mathbf{Y}$ (each row being a transposed observation) and the vector β with a $s \times k$ matrix $\mathbf{B}$).

Ordinary least squares (OLS) An estimate $\widehat{\beta}$ defines the fitted line $\widehat{r}(X) = (1, X_1, \ldots, X_p)\widehat{\beta}$, the fitted values $\widehat{y}_i = \widehat{r}(\mathbf{x}_i)$, the residuals $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$ and the residual sum of squares (also called sum of squared errors)

$$\mathrm{RSS} = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \|\mathbf{y} - \widehat{\mathbf{y}}\|^2 = \|\mathbf{y} - \mathbf{X}\beta\|^2$$

(or equivalently the MSE $= \mathrm{RSS}/n$, or the RMSE $= \sqrt{\mathrm{RSS}/n}$). The least squares estimate is the vector $\widehat{\beta}$ minimizing RSS, i.e. the projection of $\mathbf{y}$ onto the s-dimensional subspace $\mathrm{Im}\mathbf{X} \subset \mathbb{R}^n$. This projection is $\Pi = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ and the solution is

$$\widehat{\beta} \coloneqq (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

found by solving: $\frac{\partial \mathrm{RSS}}{\partial \beta_j} = 0$ for all j iff $-2^t(\mathbf{y} - \mathbf{X}\beta) = 0$ iff $\mathbf{X}^t\mathbf{X}\beta = \mathbf{X}^t\mathbf{y}$. Without a full statistical model, we only have a mathematical solution but not a statistical solution: we can't derive statistical properties (e.g. confidence intervals) of the estimator. So, we need to start adding assumptions.

Gauss-Markov assumptions (same finite unknown variance σ and uncorrelatedness of the ε_i). Then: $\widehat{\beta}_i$ are consistent and the least-squares estimator is a BLUE (best linear unbiased estimator) of β, easily seen from the definition.

$$\mathbb{E}(\widehat{\beta}) = \beta \qquad \mathbb{V}(\widehat{\beta} \mid \mathbf{x}_1, \dots, \mathbf{x}_n) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$$

Gaussian linear model If ε_i are normal, the least squares estimator is the MLE which can be found by maximizing the (log-)likelihood function.

$$f(\mathbf{y}|\beta, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}_{\beta\mathbf{x}_i, \sigma^2}(y_i)$$

(note the use of independence).

If we estimate σ^2, we can get the $\widehat{\mathrm{se}}(\widehat{\beta}_i)$ from the covariance matrix. The sample variance RSS/n is a MLE for σ^2, but in order to get an unbiased one we must correct it to

$$\widehat{\sigma}^2 \coloneqq \frac{\mathrm{RSS}}{n-s} = \frac{\|\mathbf{y} - \mathbf{X}\widehat{\beta}\|^2}{n-s} \qquad (\text{for } \mathbf{X} \text{ as above } n - s = n - p - 1)$$

indeed, we have $(n-s)V^* = |\mathbf{y} - \mathbf{X}\widehat{\beta}| = v|\xi - \Pi\xi|^2$. Using an orthonormal basis change of the argument vector, $\eta = O^t\xi$, this can be written $v\sum_{s+1}^{n} \eta_k^2$. Since the ξ_j are uncorrelated and standardized, we have $\mathbb{E}(\eta_k^2) = \mathbb{E}(\sum_{i,j} O_{ik}O_{jk}\xi_i\xi_j) = \sum_{i=1}^{n} O_{ik}^2 = 1$.

Tests for Gaussian linear models For the Gaussian model $\xi \sim N_n(0, \mathbb{1})$ so that $P_\theta = N_n(A\gamma, v\mathbb{1})$, so we can generalize Student's theorem 2.5:

$$\widehat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}) \qquad \frac{n-s}{v}V^* \sim \chi^2_{n-s}$$

so we have the usual approximate $1-\alpha$ confidence intervals for β_i using $\widehat{\mathrm{se}}(\widehat{\beta}_i)$, and we have t-tests for $H_0\colon \beta_j = 0$ using the t-statistic

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}\sqrt{v_j}} \sim t_{n-p-1} \qquad (v_j = [(\mathbf{X}^t\mathbf{X})^{-1}]_{jj})$$

- a confidence region (ellipsoid) for β with error level α, and an approximate one, are respectively

$$C(-) = \{\beta \in \mathbb{R}^s : |A(\beta - \widehat{\beta})|^2 < sV^* f_{s,n-s;1-\alpha}\} \tag{3.1}$$

$$C(-) = \{\beta \in \mathbb{R}^s : |A(\beta - \widehat{\beta})|^2 < V^* \chi^2_{s;1-\alpha}\} \tag{3.2}$$

Indeed, $|\mathbf{X}(\widehat{\beta} - \beta)|^2/v \sim \chi^2_s$ is independent of V^*, and therefore $|\mathbf{X}(\widehat{\beta} - \beta)|^2/sV^* \sim F_{s,n-s}$.

- χ^2-tests: an α-confidence interval for the variance is
$$C(-) = \left(\frac{n-s}{\chi^2_{n-s;1-\alpha/2}} V^*, \frac{n-s}{\chi^2_{n-s;\alpha/2}} V^* \right)$$
(follows immediately from the distribution of the rescaled variance above) and a test for $H_0 \colon v \leqslant v_0$ (resp. $\geqslant$) has rejection region $(n-s)V^* > v_0 \chi^2_{n-s;1-\alpha}$ (resp. ?)

- t-tests: an α-confidence interval for a linear characteristic $\tau(\beta) = c \cdot \beta$ is
$$C(-) = (c \cdot \widehat{\beta} \mp \delta \sqrt{V^*})$$
and a level-α test of $H_0 \colon c \cdot \beta = m_0$ has rejection region $|c \cdot \widehat{\beta} - m_0| > \delta \sqrt{V^*}$, where $\delta = t_{n-s;1-\alpha/2} \sqrt{c^t (A^t A)^{-1} c}$.

- F-test for $H_0 \colon A\gamma \in H$, with $H \subset \operatorname{Im} A =: L$ subspace of dimension $r < s$. Then $|\Pi_L X - \Pi_H X|^2 / v \sim \chi^2_{s-r}$ is independent of V^*, so the Fisher statistic
$$F_{H,L} \coloneqq \frac{n-s}{s-r} \frac{|\Pi_L Y - \Pi_H Y|^2}{|Y - \Pi_L Y|^2} = \frac{|A\widehat{\gamma} - \Pi_H Y|^2}{(s-r)V^*} \sim F_{s-r,n-s}$$
and a level-α test has rejection region $F_{H,L} > f_{s-r,n-s;1-\alpha}$.

Generalized least squares (GLS) We are in the generalized Gauss-Markov assumptions, covariance matrix $\sigma^2 \Sigma$ where Σ is a known positive defininte matrix and σ^2 is an unknown scale parameter; transform the data to $Y^* = \Sigma^{-1/2} Y$, $A^* = \Sigma^{-1/2} A$, then the new model falls into the ordinary case and we find its $\widehat{\beta}$.

Model fit Too few covariates = high bias = underfitting. Too many covariates = high variance = overfitting.

R^2 coefficient of determination The most general definition is
$$R^2 = 1 - \frac{RSS}{TSS}.$$
For a linear regression model including a constant (i.e. design matrix has a column of 1's), we have
$$\mathrm{TSS} = \mathrm{RSS} + \mathrm{ESS}$$
$$\sum (y_i - \overline{y})^2 = \sum (y_i - \widehat{y}_i)^2 + \sum (\widehat{y}_i - \overline{y})^2$$
Compare with ANOVA partition (the cross-term is zero by direct computation). $R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}}$ equals the squared sample correlation between the y_i and $\widehat{y}_i$.

For simple linear regression we also have $\mathrm{ESS} = S_{xy}^2 / S_{xx}$ where we defined $S_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y})$, which provides the link to the t-test; $R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$ equals the squared sample correlation between the y_i and x_i.

Adjusted R^2 to select between models containing different numbers of variables, since the usual R^2 tends to decrease as we add more variables, we define the correction $1 - \frac{\mathrm{RSS}/(n-d-1)}{\mathrm{TSS}/(n-1)}$ (essentially we have replaced biased estimates of the variance, i.e. numerator and denominator both divided by n, with their unbiased corrections).

Assumptions We can check whether a linear model is appropriate by plotting the residuals and checking if they look like iid random variables. We should not discern patterns (e.g. a U-shape), heteroscedasticity (seen as an increase in magnitude of residuals; can be fixed by fitting first-order differences, log-transformed data or Box-Cox transformation) or autocorrelation (e.g. daily stock prices are autocorrelated, stock returns aren't).

Complements

- Regularization (weight decay/parameter shrinkage): one can control overfitting by limiting the size of the coefficients, if we replace the error function with $E_D(w) + \lambda E_W(w)$ i.e. we add a penalty term (governed by a to-be-chosen coefficient), usually of the form $\lambda \sum |w_j|^q$ for some $q > 0$.
- MLE might by difficult with big amounts of data, so we may want to use sequential learning (or on-line learning) techniques like stochastic descent to find the correct weights.
- Functional regression
- Error-in-variables (EIV) models: assume there are errors in the independent variables too, so X becomes a random variable like Y.
- Orthogonal least squares: minimizes orthogonal distances rather than vertical distances. This is not a MLE in general.

3.2 Analysis of variance (ANOVA)

Denotes methods to analyze the difference between means from different groups of observations, generalizing the t-test to more than two means. The most common form is based on linear models.

Observations, X_{ik} partitioned into groups from $G = \{1, \dots, s\}$ (one single family of groups = one-factor ANOVA); the index set is therefore $B = \{ik := (i,k) \colon i \in G, 1 \leqslant k \leqslant n_i\}$ with cardinality $n = \sum_{i \in G} n_i$.

We set up a model with an unknown mean vector $\beta = (m_i)_{i \in G}$ and an unknown parameter v such that $X_{ik} = m_i + \sqrt{v}\xi_{ik}$ which we write in matrix form as $X = A\beta + \sqrt{v}\xi$ where

$$A\beta = \begin{pmatrix} \mathbb{1}_{n_1} & & \\ & \ddots & \\ & & \mathbb{1}_{n_s} \end{pmatrix} \beta = (m_1, \dots, m_1, m_2, \dots, m_2, \dots, m_s, \dots, m_s)^T$$

The unbiased estimator $\widehat{\beta}$ of $\beta = (m_i)_{i \in G}$ is just the vector of sample means:

$$\widehat{\beta} = (A^T A)^{-1}(A^T X) = \begin{pmatrix} n_1 & & \\ & \ddots & \\ & & n_s \end{pmatrix}^{-1} \begin{pmatrix} n_1 M_1 & & \\ & \ddots & \\ & & n_s M_s \end{pmatrix} = \begin{pmatrix} M_1 \\ \vdots \\ M_s \end{pmatrix}$$

so that $A\widehat{\beta} = (M_1, \dots, M_1, M_2, \dots, M_2, \dots, M_s, \dots, M_s)^T$ and

- $\widehat{\sigma}^2 = \frac{\|X - A\widehat{\beta}\|^2}{n-s} = \frac{1}{n-s}\sum_{ik\in B}(X_{ik} - M_i)^2 = \frac{1}{n-s}\sum_{i\in G}(n_i - 1)\widehat{\sigma}_i^2 =: V_{\mathrm{wg}}$, the average sample variance within the groups;
- the total sample variance $V_{\mathrm{tot}} = \frac{\|X - M\mathbb{1}_n\|^2}{n-1} = \frac{1}{n-1}\sum_{ik\in B}(X_{ik} - M)^2$;
- the variance between the groups $V_{\mathrm{bg}} = \frac{\|A\widehat{\beta} - M\mathbb{1}_n\|^2}{s-1} = \frac{1}{s-1}\sum_{i\in G} n_i(M_i - M)^2$ is the empirical variance of the group means (with a weighting by the number of observations).

By Pythagoras' theorem we have the following (equivalent) relations

$$\|X - M\mathbb{1}_n\|^2 = \|X - A\widehat{\beta}\|^2 + \|A\widehat{\beta} - M\mathbb{1}\|^2$$
$$(n-1)V_{\mathrm{tot}} = (n-s)V_{\mathrm{wg}} + (s-1)V_{\mathrm{bg}}$$

We also note that, unless the means of each group are all the same, V_{tot} is not an unbiased estimator of v:

$$\mathbb{E}_{\beta,v}(V_{\mathrm{tot}}) = v + \frac{1}{n-1}\sum_{i\in G} n_i(m_i - \overline{m})$$

We assume ξ_{ik} are iid standard normal so we have the results obtained for Gaussian linear models: the confidence ellipsoid eq. (3.1) for $\beta = (m_i)_{i\in G}$, a t-test for the two-sample problem $m_1 = m_2$ is obtained for $s = 2$, $H_0\colon c\cdot\beta = 0$, $c = (1, -1)^T$, an F-test for a multi-sample problem $H_0\colon m_1 = \cdots = m_s$ (we don't do $\binom{s}{2}$ tests because if α were chosen small enough to compensate, the power would be too small) can be written as $H_0\colon A\beta \in H \coloneqq \langle(1, \ldots, 1)^T\rangle \subset \mathbb{R}^n$, $F_{H,L} = V_{\mathrm{bg}}/V_{\mathrm{wg}}$ so the rejection region is $V_{\mathrm{bg}} > f_{s-1,n-s;1-\alpha}V_{\mathrm{wg}}$.

Chapter 4

Survival analysis

From a *failure density* $f(\tau)$ we define the failure distribution (its CDF, probability of failure before t) and the survival function (its complementary, the probability of survival up to t):

$$F(t) = P(T \leqslant t) = \int_0^t f(\tau)\mathrm{d}\tau = 1 - S(t)$$

The instantaneous hazard rate (or failure rate), (also called force of mortality, instantaneous rate of mortality...), is

$$h(t) \coloneqq \lim_{\Delta t \to 0} \frac{P(t \leqslant T \leqslant t + \Delta t \mid T \geqslant t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{\Delta t \cdot S(t)} = \frac{f(t)}{S(t)}.$$

Examples of common models

- Exponential density: $f(\tau) = \lambda e^{-\lambda\tau}$ has CDF $F(t) = 1 - e^{-\lambda t}$ and constant hazard rate $h(t) = \lambda$, showing the memorylessness of the exponential distribution.
- Gompertz' law of mortality: $h(t) = Bc^t$ with $B \in (0, 1)$ and $c > 1$.

Assumption If T_x denotes the future lifetime of a subject of age x, define F_x and S_x as above but with T_x instead of T. We use the assumption $P(T_x \leqslant t) = P(T_0 \leqslant x + t | T_0 > x)$, whence $S_0(x + t) = S_0(x)S_x(t)$. We have $h(x + t) = f_x(t)/S_x(t)$.

Actuarial notation

- ${}_tp_x \coloneqq S_x(t) = P(T_x > t)$;
- ${}_tq_x \coloneqq F_x(t) = P(T_x \leqslant t) = \int_0^t {}_sp_x\, \mu_{x+s}\mathrm{d}s$, $\mu_x \coloneqq h(x)$;
  $q_x \coloneqq {}_1q_x$ is called the mortality rate at age x.
- ${}_{u|t}q_x \coloneqq P(u < T_x \leqslant u + t)$ (deferred mortality probability)
- The curtate future lifetime for a life aged x is the random variable $K_x \coloneqq \lfloor T_x \rfloor$. By definition $P(K_x = k) = P(k \leqslant T_x < k + 1)$.

- $\mathring{e}_x \coloneqq \mathbb{E}(T_x)$ (complete expectation of life);
  $e_x \coloneqq \mathbb{E}(K_x) = \sum_{k=0}^{\infty} kP(K_x = k) = \sum_{k=1}^{\infty} S_x(k)$.

Chapter 5

Misc

Zipf-Mandelbrot law states that the relative frequency f of an item of rank k (the kth most frequent) in a corpus of N elements is often distributed according to the power-law distribution

$$f(k, N, q, s) = \frac{1/(k+q)^s}{H_{N,q,s}}$$

where $H_{N,q,s} = \sum_{i=1}^{N} (q+i)^s$ and $s > 0$ and q are parameters. For $q = 0$ and finite N this is Zipf's law. For $q = 0$ and $N = \infty$ this is a zeta distribution (if $s > 1$).

Benford's law is satisfied when, in a set of numbers, the leading digit (first digit) d is distributed according to

$$P(\{d\}) = \log_{10}(d+1) - \log_{10}(d).$$

This happens when the numbers are log-normally distributed, for example when they arise from multiplicative fluctuations, e.g. stock prices (whereas additive fluctuations give rise to normal distributions).