I - Probability

Probability spaces

A probability space is an ordered triple $(\Omega, \mathcal{F}, P)$ in which

- ⋆ the universe Ω is a non-empty set containing elementary outcomes ω
- ⋆ $\mathcal{F}$ is a set of subsets of Ω that is (i) non-empty, (i) contains complements of its members, and (iii) contains countable unions of its members
- ⋆ Probabilities $P(A) \in [0,1]$ are defined for all $A \in \mathcal{F}$, (i) $P(\Omega) = 1$ and (ii) the probability of a countable union of pairwise disjoint elements of $\mathcal{F}$ equals the sum of their individual probabilities.
- ⋆ If $A \in \mathcal{F}$ then $P(A^c) = 1 - P(A)$, so $P(\emptyset) = 0$ because $P(\Omega) = 1$
- ⋆ inclusion-exclusion: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, more generally
$$P(\cup_{i=1}^n A_i) = \sum_{r=1}^n (-1)^{r+1} \sum_{1 \le i_1 < \cdots < i_r \le n} P(A_{i_1} \cap \cdots \cap A_{i_r})$$
- ⋆ Boole's inequality: $P(\cup_{i\in\mathbb{N}} A_i) \le \sum_{i\in\mathbb{N}} P(A_i)$
- ⋆ Continuity of P as a set function: $\lim_{n\to\infty} P\left(\cup_{i=1}^n A_i\right) = P\left(\lim_{n\to\infty} \cup_{i=1}^n A_i\right)$ and $\lim_{n\to\infty} P\left(\cap_{i=1}^n A_i\right) = P\left(\lim_{n\to\infty} \cap_{i=1}^n A_i\right)$

Conditional probabilities and partitions of Ω

- ⋆ Conditional probability of A given B: $P(A \mid B) = \dfrac{P(A\cap B)}{P(B)}$
- ⋆ $P(A\cap B) = P(A)P(B) \iff$ A and B are independent
- ⋆ If $\{B_i\}_{i\in\mathbb{N}}$ partition Ω (i.e., $B_i \cap B_j = \emptyset$ when $i \ne j$ and $\Omega = \cup_i B_i$), then (law of total probability) $P(A) = \sum_i P(A \mid B_i)P(B_i)$, giving (Bayes' theorem)
$$P(B_j \mid A) = \frac{P(A\mid B_j)P(B_j)}{\sum_i P(A\mid B_i)P(B_i)}.$$

Random variables

A random variable is a function $X : \Omega \to \mathbb{R}$. The support of X is $S_X := \{x \in \mathbb{R} : \exists\, \omega \in \Omega \text{ s.t. } X(\omega) = x\}$. Then

- ⋆ $F_X(x) = P(X \le x)$
- ⋆ $F_X \nearrow$ on $\mathbb{R}$
- ⋆ $F_X(b) - F_X(a) = P(a < X \le b)$
- ⋆ $1 - F_X(x) = P(X > x)$
- ⋆ The p quantile is $x_p = \inf\{x \in \mathbb{R} : F_X(x) \ge p\}$, for $p \in (0,1)$.

Discrete random variables: S_X is finite or countable

- ⋆ Probability mass function (PMF): $f_X(x) = P(X = x)$
- ⋆ $P(X \in B) = \sum_{x\in B} f_X(x)$
- ⋆ $f_{X\mid X\in B}(x \mid X \in B) = P(X = x \mid X \in B) = \dfrac{P(X = x \cap X \in B)}{P(X\in B)}$
- ⋆ $\mathrm{E}\{g(X)\} = \sum_{x\in S_X} g(x) f_X(x)$, $\mathrm{Var}(X) = \mathrm{E}[\{X - \mathrm{E}(X)\}^2] = \mathrm{E}(X^2) - \mathrm{E}(X)^2$

Continuous random variables: S_X is uncountable

- ⋆ Probability density function (PDF): $f_X(x) = \dfrac{\mathrm{d}F_X}{\mathrm{d}x}(x) \ne P(X = x)$
- ⋆ $P(X\in B) = \int_{x\in\mathbb{R}} f_X(x) I_B(x)\,\mathrm{d}x = \int_{x\in B} f_X(x)\,\mathrm{d}x$
- ⋆ $f_{X\mid X\in B}(x) = \dfrac{f_X(x) I_B(x)}{P(X\in B)}$
- ⋆ $\mathrm{E}\{g(X)\} = \int_{x\in\mathbb{R}} g(x) f_X(x)\,\mathrm{d}x$, $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$

Law of total expectation

$\mathrm{E}\{g(X)\} = \sum_{i\in\mathbb{N}} \mathrm{E}\{g(X) \mid B_i\} P(B_i)$ where $\{B_i\}_{i\in\mathbb{N}}$ partitions Ω

Generating functions

- ⋆ The moment generating function of X is $M_X(t) = \mathrm{E}\left(e^{tX}\right)$
- ⋆ The cumulant generating function of X is $K_X(t) = \ln M_X(t)$

Notable variables and properties

- ⋆ **Bernoulli variable** For $A \in \Omega$ we have $I_A(y) = 1$ if $y \in A$ and $I_A(y) = 0$ if $y \notin A$. We have $f_X(x) = P(A)^x(1 - P(A))^{1-x}$, $\mathrm{E}(I_A) = P(A)$, $\mathrm{Var}(I_A) = P(A)\,(1 - P(A))$ and $S_X = \{0, 1\}$
Can also be defined with a probability p, I_p, with $f_X(x) = p^x(1-p)^{1-x}$, $\mathrm{E}(I_p) = p$ and $\mathrm{Var}(I_p) = p\,(1-p)$.
- ⋆ **Binomial** For $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} I_p$ we define $X = Z_1 + \cdots + Z_n$ and write $X \sim B(n,p)$, $n \in \mathbb{N}, p \in (0,1)$. We have $f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$, $\mathrm{E}(X) = np$, $\mathrm{Var}(X) = np(1-p)$ and $S_X = \{0, 1, \ldots, n\}$
- ⋆ **Negative binomial** If $X \sim \mathrm{NegBin}(n,p)$, $n \in \mathbb{N}, p \in (0,1)$. We have $f_X(x) = \binom{x-1}{n-1} p^n (1-p)^{x-n}$, $\mathrm{E}(X) = \dfrac{n}{p}$, $\mathrm{Var}(X) = \dfrac{n(1-p)}{p^2}$ and $S_X = \{n, n+1, \ldots, \infty\}$. If $n = 1$ we write $X \sim \mathrm{Geom}(p)$.
- ⋆ **Hypergeometric** If $X \sim \mathrm{HypGeom}(k,b,n)$, $k,b,n \in \mathbb{N}$. We have $f_X(x) = \dfrac{\binom{b}{x}\binom{k}{n-x}}{\binom{k+b}{n}}$, $\mathrm{E}(X) = n\dfrac{b}{b+k}$, $\mathrm{Var}(X) = n\dfrac{bk(b+k-n)}{(b+k)^2(b+k-1)}$ and $S_X = \{\max(0, b+k-n), \ldots, \min(b,n)\}$
- ⋆ **Poisson** If $X \sim \mathrm{Poiss}(\lambda)$, $\lambda \in \mathbb{R}^{+*}$. We have $f_X(x) = e^{-\lambda}\dfrac{\lambda^x}{x!}$, $\mathrm{E}(X) = \lambda$, $\mathrm{Var}(X) = \lambda$ and $S_X = \{0, 1, \ldots, \infty\}$.
Note that for $X_n \sim B(n, p_n)$ with $\lim_{n\to\infty} np_n = \lambda$, then $X_n \overset{n\to\infty}{\to} X \sim \mathrm{Poiss}(\lambda)$
- ⋆ **Discrete uniform** If $X \sim \mathrm{Unif}\{1, .., n\}$, $n \in \mathbb{N}$. We have $f_X(x) = \dfrac{1}{n}$, $\mathrm{E}(X) = \dfrac{n+1}{2}$, $\mathrm{Var}(X) = \dfrac{n^2-1}{12}$ and $S_X = \{1, \ldots, n\}$
- ⋆ **Continous Uniform** If $X \sim U(a,b)$, $a < b \in \mathbb{R}$. We have $f_X(x) = \dfrac{1}{b-a}$, $F_X(x) = \dfrac{x-a}{b-a}$, $\mathrm{E}(X) = \dfrac{b+a}{2}$, $\mathrm{Var}(X) = \dfrac{(b-a)^2}{12}$ and $S_X = (a,b)$
- ⋆ **Exponential** If $X \sim \exp(\lambda)$, $\lambda \in \mathbb{R}^{+*}$. We have $f_X(x) = \lambda e^{-\lambda x}$, $F_X(x) = 1 - e^{-\lambda x}$, $\mathrm{E}(X) = \dfrac{1}{\lambda}$, $\mathrm{Var}(X) = \dfrac{1}{\lambda^2}$ and $S_X = (0, +\infty)$
- ⋆ **Pareto** If $X \sim \mathrm{Pareto}(\alpha,\beta)$, $\alpha,\beta \in \mathbb{R}^+$. We have $f_X(x) = \dfrac{\alpha\beta^\alpha}{x^{\alpha+1}}$, $F_X(x) = 1 - \left(\dfrac{\beta}{x}\right)^\alpha$, $\mathrm{E}(X) = \dfrac{\alpha\beta}{\alpha-1}$ defined only for $\alpha > 1$, $\mathrm{Var}(X) = \dfrac{\beta^2\alpha}{(\alpha-1)^2(\alpha-2)}$ defined only for $\alpha > 2$ and $S_X = (\beta, +\infty)$
- ⋆ **Laplace** If $X \sim \mathrm{Laplace}(\lambda,\eta)$, $\lambda \in \mathbb{R}^+, \eta \in \mathbb{R}$. We have $f_X(x) = \dfrac{\lambda}{2} e^{-\lambda|x-\eta|}$, $F_X(x) = \dfrac{1}{2} e^{\lambda(x-\eta)}$ for $x \le \eta$ and $F_X(x) = 1 - \dfrac{1}{2} e^{-\lambda(x-\eta)}$ for $x > \eta$, $\mathrm{E}(X) = \eta$, $\mathrm{Var}(X) = \dfrac{2}{\lambda^2}$ and $S_X = \mathbb{R}$
- ⋆ **Gamma** If $X \sim \mathrm{Gamma}(\alpha,\beta)$, $\alpha,\beta \in \mathbb{R}^{+*}$, then $f_X(x) = \dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$, $F_X(x) = \dfrac{\int_0^{\beta x} t^{\alpha-1} e^{-t}\,\mathrm{d}t}{\Gamma(\alpha)}$, $\mathrm{E}(X) = \dfrac{\alpha}{\beta}$, $\mathrm{Var}(X) = \dfrac{\alpha}{\beta^2}$ and $S_X = (0, +\infty)$
- ⋆ **Gaussian** If $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}, \sigma \in \mathbb{R}^{+*}$. We have $f_X(x) = \dfrac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sqrt{2\pi\sigma^2}}$, $F_X(x) = \Phi\left(\dfrac{x-\mu}{\sigma}\right)$, $\mathrm{E}(X) = \mu$, $\mathrm{Var}(X) = \sigma^2$ and $S_X = \mathbb{R}$
- ⋆ χ^2 For $Z_1, \ldots, Z_\nu \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ we define $X = Z_1^2 + \cdots + Z_\nu^2$ and write $X \sim \chi^2_\nu$ for $\nu \in \mathbb{N}$. We have $f_X(x) = \dfrac{x^{\nu/2-1} e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}$, $F_X(x) = \dfrac{\int_0^{x/2} t^{\nu/2-1} e^{-t}\,\mathrm{d}t}{\Gamma(\nu/2)}$ for $x > 0$, and $\mathrm{E}(X) = \nu$, $\mathrm{Var}(X) = 2\nu$ and $S_X = \mathbb{R}^{+*}$.

Random vectors

- ⋆ $\mathbb{R}^n \ni X = (X_1, \ldots, X_n)$ is a random vector.
- ⋆ $f_X(x) = f_{X_1,\ldots,X_n}(x) : \mathbb{R}^n \to \mathbb{R}$ is its joint density function.
- ⋆ If $A \subset \{1, \ldots, n\}$ satisfies $|A| = p$, B is its complement, and we write $x = (x_A, x_B)$, then $X_A = X_{i:i\in A}$ has **marginal density function** $f_{X_A}(x_A) = \int_{x_B\in\mathbb{R}^{n-p}} f_X(x)\,\mathrm{d}x_B = \int_{x_B\in\mathbb{R}^{n-p}} f_X(x_A, x_B)\,\mathrm{d}x_B$ and the **conditional density function** of X_B given that $X_A = x_A$ is
$$f_{X_B\mid X_A}(x_B \mid x_A) = \frac{f_X(x_A, x_B)}{f_{X_A}(x_A)}, \quad x_A \in \mathbb{R}^p, \quad x_B \in \mathbb{R}^{n-p}.$$
- ⋆ If $f_{X_B\mid X_A}(x_B \mid x_A) = f_{X_B}(x_B)$ for all possible values of x_A and x_B, then X_A and X_B are independent and we can write $f_X(x) = f_{X_A}(x_A) f_{X_B}(x_B)$.
- ⋆ We define $\mathrm{E}\{g(X)\} = \int_{x\in\mathbb{R}^n} g(x) f_X(x)\,\mathrm{d}x$, where $g(x) : \mathbb{R}^n \to \mathbb{R}$.
- ⋆ We define the **mean vector** as $\mathrm{E}(X) = \mu = (\mathrm{E}(X_1), \ldots, \mathrm{E}(X_n))^T \in \mathbb{R}^n$.
- ⋆ For $1 \le k, l < n$, the covariance is $\mathrm{Cov}(X_k, X_l) = \mathrm{E}(X_k X_l) - \mathrm{E}(X_k)\mathrm{E}(X_l)$. Note that $X_k \perp\!\!\!\perp X_l \implies \mathrm{Cov}(X_k, X_l) = 0$.
- ⋆ The **correlation** between X_k and X_l is $\mathrm{Corr}(X_k, X_l) := \dfrac{\mathrm{Cov}(X_k, X_l)}{\sqrt{\mathrm{Var}(X_k)\mathrm{Var}(X_l)}}$. Note that $\mathrm{Corr}(X_k, X_l) \in [-1, 1]$
- ⋆ The **covariance matrix** of X is $\mathrm{Cov}(X) = \Omega \in \mathbf{R}^{n\times n}$, where $\Omega_{k,l} = \mathrm{Cov}(X_k, X_l)$. As $\mathrm{Cov}(X_k, X_l) = \mathrm{Cov}(X_l, X_k)$, $\Omega = \Omega^T$, and Ω is symmetric positive semi-definite.

Multivariate Gaussian distribution

- ⋆ For $X = (X_1, \ldots, X_n)^T \in \mathbb{R}^{n\times 1}$ with $\mathrm{E}(X) = \mu$ and $\mathrm{Cov}(X) = \Omega$, if $\forall u \in \mathbb{R}^n$, $u^T X \sim \mathcal{N}\left(u^T\mu, u^T\Omega u\right)$, we write $X \sim \mathcal{N}_n(\mu, \Omega)$.
- ⋆ If $\mathrm{Rank}(\Omega) = n$ then $f_X(x) = \{(2\pi)^n|\Omega|\}^{-1/2} e^{-(x-\mu)^T\Omega^{-1}(x-\mu)/2}$, $x \in \mathbb{R}^n$.
- ⋆ With $X \sim \mathcal{N}_n(\mu, \Omega)$ and A, B, $X_A \in \mathbb{R}^p$ and $X_B \in \mathbb{R}^{n-p}$ defined as above, we define $\mu_A = \mathrm{E}(X_A)$, $\mu_B = \mathrm{E}(X_B)$, $\Omega_{AA} = \mathrm{Cov}(X_A) \in \mathbb{R}^{p\times p}$, $\Omega_{BB} = \mathrm{Cov}(X_B) \in \mathbb{R}^{(n-p)\times(n-p)}$, $\mathrm{Cov}(X_A, X_B) = \Omega_{AB} \in \mathbb{R}^{p\times(n-p)}$ and $\Omega_{BA} = \Omega_{AB}^T$.
Then $X_A \sim \mathcal{N}_p(\mu_A, \Omega_{AA})$ and
$$X_\mathcal{A} \mid X_\mathcal{B} = x_\mathcal{B} \sim \mathcal{N}_p\left(\mu_\mathcal{A} + \Omega_{\mathcal{AB}}\Omega_{\mathcal{BB}}^{-1}(x_\mathcal{B} - \mu_\mathcal{B}), \Omega_{\mathcal{AA}} - \Omega_{\mathcal{AB}}\Omega_{\mathcal{BB}}^{-1}\Omega_{\mathcal{BA}}\right).$$

Transformation of variables

For $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ and a function $g(x) : \mathbb{R}^n \to \mathbb{R}$, the distribution of $Y := g(X)$ is $P(Y \le y) = F_Y(y) = \int_{x:g(x)\le y} f_X(x)\,\mathrm{d}x$.

The case $n = 1$ with $g \nearrow$ and $g^{-1}(y) = x$ bijective on S_X gives $F_X(x) = P(X \le x) = P(g(X) \le g(x)) = F_Y(g(x))$
$\implies F_Y(y) = F_X(g^{-1}(y))$.
A similar calculation applies for $g \searrow$ and $g^{-1}(y) = x$ bijective on S_X.

II - Data fitting and statistics

Approximations and convergence

- ⋆ Below $a > 0$, $h(x) > 0$ for all $x \in \mathbb{R}$ and $g(x)$ is a convex function on $\mathbb{R}$.
- ⋆ Basic inequality: $P\{h(X) \geq a\} \leq \mathrm{E}\{h(X)\}/a$.
- ⋆ Markov's inequality $P(|X| \geq a) \leq \mathrm{E}(|X|)/a$.
- ⋆ Chebyshev's inequality: $P(|X| \geq a) \leq \mathrm{E}(X^2)/a^2$ or $P\{|X - \mathrm{E}(X)| \geq a\} \leq \dfrac{\mathrm{Var}(X)}{a^2}$.
- ⋆ Jensen's inequality: $g\{\mathrm{E}(X)\} \leq \mathrm{E}\{g(X)\}$.
- ⋆ Quadratic mean convergence $X_n \overset{2}{\to} X \iff \lim_{n\to\infty} \mathrm{E}\left((X_n - X)^2\right) = 0$ with $E(X_n^2), E(X^2) < \infty$
- ⋆ Convergence in probability: $X_n \overset{P}{\to} X \iff \forall \varepsilon > 0, P\left(\lim_{n\to\infty} |X_n - X| > \varepsilon\right) = 0$
- ⋆ Convergence in distribution/law: $X_n \overset{D}{\to} X \iff \lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ at all x where $F(x)$ is continuous.
- ⋆ $X_n \overset{2}{\to} X \implies X_n \overset{P}{\to} X \implies X_n \overset{D}{\to} X$
- ⋆ $X_n \overset{D}{\to} x_0 \implies X_n \overset{P}{\to} x_0$ for constant x_0
- ⋆ $X_n \overset{P}{\to} x_0 \implies g(X_n) \overset{P}{\to} g(x_0)$ with $g(x)$ continuous at x_0.
- ⋆ **Slutsky's lemma**: $X_n \overset{D}{\to} X, Y_n \overset{P}{\to} y_0 \implies X_n + Y_n \overset{D}{\to} X + y_0$ and $X_n Y_n \overset{D}{\to} X y_0$
- ⋆ **Law of small numbers**: If $X_n \sim B(n, p_n)$ with $\lim_{n\to\infty} np_n = \lambda$, then $X_n \overset{D}{\to} X$ where $X \sim \mathrm{Poiss}(\lambda)$.
- ⋆ **Weak law of large numbers**: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ where $E(X_i) = \mu < \infty$ $\implies \overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i \overset{P}{\to} \mu \left(\iff \forall \varepsilon > 0, P\left(\lim_{n\to\infty} |X_n - \mu| > \varepsilon\right) = 0\right)$
- ⋆ **Strong law of large numbers**: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ where $E(X_i) = \mu < \infty$ $\implies P\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1$

Central limit theorem

- ⋆ **CLT**: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ such that $\mathrm{E}(X_i) = \mu$ and $0 < \mathrm{Var}(X_i) = \sigma^2 < \infty$ $\implies Z_n = \dfrac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \overset{D}{\to} Z \sim \mathcal{N}(0, 1)$
- ⋆ **Delta method**: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ such that $\mathrm{E}(X_i) = \mu$ and $0 < \mathrm{Var}(X_i) = \sigma^2 < \infty$ with $g(x)$ such as $g'(\mu) \neq 0 \implies Z_n = \dfrac{g\left(\overline{X}_n\right) - g(\mu)}{\sqrt{g'(\mu)^2\sigma^2/n}} \overset{D}{\to} Z \sim \mathcal{N}(0, 1)$
- ⋆ **Quantiles**: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, $p \in (0, 1)$ where $x_p = F^{-1}(p)$ and $f(x_p) > 0$ $\implies Z_n = \dfrac{X_{(\lceil np \rceil)} - x_p}{\sqrt{p(1-p)/nf(x_p)^2}} \overset{D}{\to} Z \sim \mathcal{N}(0, 1)$

Statistics

- ⋆ A statistic G is a function that depends only on the data $y = (y_1, \ldots, y_n)$: $G = g(y)$
- ⋆ A random sample is a set of independent identically distributed data $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F$
- ⋆ From the order statstics $y_{(i) \leq y_{(j)}}$, $1 \leq i \leq j \leq n$, the empirical quantiles/quantiles of the sample $y_{(\lceil np \rceil)}$ for $p \in (0, 1)$ can be defined
- ⋆ The breakdown point of a statistic is $p \times 100\%$, where $p \in (0, 1)$ is the (asymptotically as $n \to \infty$) smallest value such that sending $x_1, \ldots, x_{\lceil np \rceil} \to \pm\infty$ sends the statistic to $\pm\infty$.

Notable statistics

- ⋆ Summaries of location: the average (arithmetic mean) $\overline{y} := n^{-1}\sum_{i=1}^{n} y_i$ and the sample median $y_{(\lceil n/2 \rceil)}$.
- ⋆ Summaries of scale / dispersion: the inter-quartile range $\mathrm{IQR} := y_{(\lceil 3n/4 \rceil)} - y_{(\lceil n/4 \rceil)}$, the range $y_{(n)} - y_{(1)}$, and the sample standard deviation

$$s(y) := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}({y_i}^2 - n\overline{y}^2)}$$

- ⋆ Summary of dependence for $(x_1, y_1), \ldots, (x_n, y_n)$: the sample correlation coefficient

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\left\{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2\right\}^{1/2}}$$

Useful plots

- ⋆ **Boxplot**: defined by five numbers, the central line $y_{(\lceil n/2 \rceil)}$, the limits of the box $y_{(\lceil n/4 \rceil)}$ and $y_{(\lceil 3n/4 \rceil)}$, the limits of the "whiskers", at the y_i that are most extreme but inside $y_{(\lceil n/4 \rceil)} - 1.5\,\mathrm{IQR}$ and $y_{(\lceil 3n/4 \rceil)} + 1.5\,\mathrm{IQR}$. Points outside the whiskers are shown individually.
- ⋆ **Q-Q plot**: Assuming that $y_1, \ldots, y_n \overset{\text{iid}}{\sim} F$, the order statistics are plotted against the corresponding quantiles of F, i.e., we plot $\left(F^{-1}\left(i/(n+1)\right), y_{(i)}\right)$. A line close to $x = y$ suggests that the data come from F. If F is parametric, a modified plot (depending on F) can be used to estimate some parameters.

Hypothesis testing

- ⋆ Hypothesis testing is 'proof by stochastic contradiction': we suppose that a null hypothesis H_0 about reality is true and attempt to disprove H_0 using data. Distributions of data Y under H_0 are denoted by subscript 0; these are 'null distributions'.
- ⋆ A test requires a test statistic $T = t(Y)$, large values of which suggest that H_0 is false.
- ⋆ The observed value of T, t_{obs}, is used to compute a p-value $p_{\text{obs}} = P_0\left(T \geq t_{\text{obs}}\right)$, small values of which cast doubt on H_0.
- ⋆ If a clear decision is required, a 'significance level' $\alpha \in (0, 1)$ is chosen (e.g., $\alpha = 0.05, 0.01, 0.001$), and H_0 is rejected iff $p_{\text{obs}} < \alpha$, or equivalently if $t_{\text{obs}} > t_{1-\alpha}$, where $t_{1-\alpha}$ is the $1-\alpha$ quantile of the null distribution of T.
  In a decision setting the possible outcomes of a statistical test are:

	State of nature	
Decision on H_0	H_0 is true	H_0 is false
Not rejected (negative)	True negative	False negative (type II error)
Rejected (positive)	False positive (type I error)	True positive

- ⋆ We may have a clearly-specified alternative/counter hypothesis H_1 that is true when H_0 is false.
- ⋆ With H_1 and α specified, the probability of a true positive is $P(\text{rejecting } H_0 \text{ at significance level } \alpha \text{ when } H_1 \text{ is true}) = P_1(T \geq t_{1-\alpha})$, where $\alpha = P_0(T \geq t_{1-\alpha})$ is the probability of a false positive.
- ⋆ α is called the size and $P_1(T \geq t_{1-\alpha}) =: \beta(\alpha)$ the power of the test.
- ⋆ A test is said to be optimal if it maximizes $\beta(\alpha)$ for all $\alpha \in (0, 1)$.
- ⋆ **Pearson's statistic** When data follow a multinomial distribution (under a certain hypothesis) with denominator n and k categories (it models an experiment with k possible outcomes repeated independently n times, generalising the binomial law), then Pearson's statistic $T = \sum_{i=1}^{k}\dfrac{(O_i - E_i)^2}{E_i}$ follows a χ^2_{k-1} if $\sum E_i/k \geq 5$. This is widely used for tests of fit.

Point estimation

- ⋆ An estimator of the parameter θ of a parametric model is a function of the data $T = t(Y)$ that estimates θ. An estimate is a specific value $t = t(y)$ of $T = t(Y)$.
- ⋆ The bias of an estimator $\tilde{\theta}$ is $b_{\tilde{\theta}(\theta)} = \mathrm{E}(\tilde{\theta}) - \theta$.
- ⋆ The mean square error of an estimator $\tilde{\theta}$ is $\mathrm{MSE}_{\tilde{\theta}(\theta)} = \mathrm{E}\left(\tilde{\theta} - \theta\right)^2 = b^2_{\tilde{\theta}(\theta)} + \mathrm{Var}(\tilde{\theta})$
- ⋆ For two unbiased estimators of θ, $\tilde{\theta}_1$ and $\tilde{\theta}_2$, we say that $\tilde{\theta}_1$ is more efficient than $\tilde{\theta}_2$ if $\mathrm{Var}(\tilde{\theta}_1) \leq \mathrm{Var}(\tilde{\theta}_2)$

Types of estimators

- ⋆ **Moment estimator**: For $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f_{Y_1}(\theta)$ where $\theta \in \mathbb{R}^p$, and moments $\mathrm{E}(Y_j^r) = \mu_r(\theta)$ for $r \leq p$: $\dfrac{1}{n}\sum_{i=1}^{n} Y_i^r \overset{n\to\infty}{\to} \mu_r(\theta)$. This gives a set of p equations in θ whose solution gives an estimator for θ.
- ⋆ **Maximum likelihood**: For $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f_{Y_1}(\theta)$ where $\theta \in \mathbb{R}^p$, and $Y = (Y_1, \ldots, Y_n) \sim f_Y(\theta)$ the likelihood function $L(\theta) = f_Y(y, \theta) = \prod_{i=1}^{n} f(y_i, \theta)$ and the log-likelihood $l(\theta) = \log L(\theta)$ functions can be defined.
  The value $\hat{\theta}$ such that $L(\hat{\theta}) \geq L(\theta)$ (or $l(\hat{\theta}) \geq l(\theta)$) $\forall\theta$ is the maximum likelihood estimator. From this, the observed information $J(\theta) = -\dfrac{\mathrm{d}^2 l(\theta)}{\mathrm{d}\theta^2}$ and expected / Fisher information $I(\theta) = \mathrm{E}(J(\theta))$ can be defined for later use.

Interval estimation

- ⋆ If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f_{Y_1}(\theta)$ where $\theta \in \mathbb{R}^p$, a confidence interval for θ is a statistic in the form of an interval that contains θ with a given probability (called the confidence level of the interval).
- ⋆ An interval of the form (L, U) is called bilateral and an interval of the form $(-\infty, U)$ or $(L, +\infty)$ is called unilateral.
- ⋆ If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f_{Y_1}(\theta)$ where $\theta \in \mathbb{R}^p$ and $\tilde{\theta}$ is an estimator of θ with V an estimator of $\mathrm{Var}(\tilde{\theta})$. Then $V^{1/2}$ is called a standard deviation of $\tilde{\theta}$

Construction of an interval

- ⋆ **Using the CLT**: If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f_{Y_1}(\theta)$ where $\theta \in \mathbb{R}^p$ and if $\tilde{\theta}$ is an estimator of θ with a standard deviation $V^{1/2}$ with $\tilde{\theta} \;\dot{\sim}\; \mathcal{N}(\theta, V)$, then $(L, U) = \left(\tilde{\theta} - V^{1/2} z_{1-\alpha_L}, \tilde{\theta} - V^{1/2} z_{\alpha_U}\right)$ is a confidence interval with approximate confidence level $1 - \alpha_L - \alpha_U$.
  For a bilateral interval we usually chose $\alpha_L = \alpha_U = \alpha/2$ to have a symmetrical interval.
  For a unilateral interval we chose $\alpha_L = \alpha_U = \alpha$ and replace the unwanted limit by $\pm\infty$.
- ⋆ **Limit law of the MLE**: If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f_{Y_1}(\theta)$ where $\theta \in \mathbb{R}^p$ and if $\hat{\theta}$ is the maximum likelihood estimator of θ, then under mild regularity conditions, $J(\hat{\theta})^{1/2}(\hat{\theta} - \theta) \overset{D}{\to} \mathcal{N}_p(0, I_p)$. We then use the method above.

By Adam Avedissian
Content based on the lecture slides of Professor Anthony Davison
Format based on the template added by Fingal Mychkine Nagel Persoud