$N(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$	μ, σ^2	$\overline{X}, S_n^2$	
$\chi_n^2 = \sum Z_i^2$	$\frac{1}{\Gamma(\frac{n}{2})2^{n/2}} x^{n/2-1} e^{-x/2}, x > 0$	$n, 2n$	N/A	
$\text{Exp}(\beta)$	$\frac{1}{\beta} e^{-x/\beta}, x > 0$	β, β^2	$\overline{X}$	$1/\beta^2$
$\Gamma(\alpha, \beta)$	$\frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, x > 0$	$\alpha\beta, \alpha\beta^2$	N/A	
$\beta(\alpha, \beta)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, x \in (0,1)$	$\frac{\alpha}{\alpha+\beta}, \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	N/A	
$U(a, b)$	$\frac{1}{b-a} I(x \in (a, b))$	$\frac{a+b}{2}, \frac{1}{12}(b-a)^2$	$X_{(1)}, X_{(n)}$	
$\text{Bernoulli}(p)$	$p^k(1-p)^k, k = 0, 1$	$p, p(1-p)$	$\overline{X}$	$\frac{1}{p(1-p)}$
$\text{Bin}(n, p)$	$\binom{n}{k} p^k (1-p)^{(n-k)}, k \in [0, n]$	$np, np(1-p)$	$\hat{p} = X/n$	$\frac{1}{np(1-p)}$
$\text{Geo}(p)$	$p(1-p)^{k-1}, k \geq 1$	$\frac{1}{p}, \frac{1-p}{p^2}$	$\frac{n}{\sum X_i}$	$\frac{1}{p^2} + \frac{1}{p(1-p)}$
$\text{Poisson}(\lambda)$	$e^{-\lambda} \frac{\lambda^k}{k!}, k \geq 0$	λ, λ	$\overline{X}$	$1/\lambda$

$z_\alpha = \Phi^{-1}(1-\alpha)$

$\mathbb{V} X = \mathbb{E} X^2 - \mu^2$, $\mathbb{E}_X g(X) = \int f(x) g(x) dx$

Law of Total Exp. $\mathbb{E} Y = \mathbb{E}_X \mathbb{E}(Y|X)$

$M(t) = \mathbb{E} e^{t^T X}$ so $\frac{\partial M(t)}{\partial t}|_{t=0} = \mathbb{E} \frac{\partial}{\partial t} e^{t^T X}|_{t=0} = \mathbb{E} X e^{t^T X}|_{t=0} = X$

Stein's identity for moments. If $X \sim N(\mu, \sigma^2)$, $\mathbb{E}[g(X)(X-\mu)] = \sigma^2 \mathbb{E} g'(X)$

Concentration inequalities

Chebyshev. $P(|X - \mathbb{E} X| \geq k\sigma) \leq \frac{1}{k^2}$

Chernoff. $P(e^{t(x-\mu)} \geq e^{tu}) \leq \frac{\mathbb{E} e^{t(x-\mu)}}{e^{tu}}$

Gaussian mgf. $M_X(t) = \exp(t\mu + t^2\sigma^2/2)$

Gaussian tail bound. $P(X - \mu \geq t) \leq \exp\left(\frac{-t^2}{2\sigma^2}\right)$

Hoeffding's. $P(|\overline{X} - \mu| \geq t) \leq 2\exp\left(\frac{-2nt^2}{(b-a)^2}\right)$

Bernstein's. $P(|\overline{X} - \mu| \geq t) \leq 2\exp\left(\frac{-nt^2}{2(\sigma^2 + (b-a)t)}\right)$

McDiarmid's. Lipschitz functions of bdd RVs
$|f(X^n) - f(Y^n)| \leq L\|X - Y\|_2$
$P(|f(X^n) - \mathbb{E} f| \geq t) \leq 2\exp(\frac{-t^2}{2L^2})$

Levy's. If $X^n \sim Z$ and $|f(X) - f(Y)| \leq L\|X - Y\|_2$,
$P(|f(X^n) - \mathbb{E} f(X^n)| \geq t) \leq 2\exp(-t^2/(2L^2))$

$Y^n \sim \chi^2$ tail bound:
$P(|\overline{Y} - 1)| \geq t) \leq \exp(-nt^2/8)$ only for $t \in (0, 1)$
JL. $\|X_i - X_j\|_2^2$ preserved within $1 \pm \varepsilon$ $\forall i, j$
by $ZX_i/\sqrt{m}$ for $m \geq 16\log(n/\delta)/\varepsilon^2$

Convergence

Almost sure $P(\lim_{n\to\infty} X_n = X) = 1$. Implies $\xrightarrow{p}$.

Conv in qm. $\mathbb{E}(X_n - X)^2 \to 0$. Implies $\xrightarrow{p}$.

Convergence in probability, **consistent** est.
$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$. Implies $\xrightarrow{d}$.

Convergence in dist, $\lim_{n\to\infty} F_n(t) = F(t)$
for t s.t. F is continuous. Implies $\xrightarrow{p}$ for constant X.

WLLN. $\overline{X}_n \xrightarrow{p} \mu$ by Chebyshev

CMT. $h(X_n) \xrightarrow{d} h(X)$ if $X_n \xrightarrow{d} X$

Also $X_n \xrightarrow{d} X, Y_n \xrightarrow{d} Y \Rightarrow X_n + Y_n, X_n Y_n$

Slutsky's. $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} c \Rightarrow X_n + Y_n, X_n Y_n$

CLT. $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z$ for independent X^n with μ, σ^2

Can use $\hat{\sigma_n}$ by Slutsky's, CMT, WLLN

Lyapunov CLT. $s_n^2 = \sum \sigma_i^2$. $\frac{1}{s_n} \sum X_i - \mu \xrightarrow{d} Z$ if $\lim_{n\to\infty} \frac{1}{s^3} \sum |X_i - \mu|^3 = 0$

Berry-Esseen. avg $\to Z$ quickly if μ_3 is small
$\sup_t |P(Z_n \leq t) - P(Z \leq t)| \leq \frac{9\mu_3}{\sigma_3\sqrt{n}}$, $\mu_3 = |X_i - \mu|^3$

Delta method. $g(X_n) \approx g(\mu) + g'(\mu)(X_n - \mu)$
so $\frac{X_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z \Rightarrow \frac{g(X_n) - g(\mu)}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, g'(\mu)^2)$
$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} N(0, \nabla_x g(\mu)^T \Sigma \nabla_x g(\mu))$

Uniform laws

Glivenko-Cantelli. $\Delta = \sup_t |\hat{F}(t) - F(t)| \xrightarrow{p} 0$
Uniform LLN, any distribution, test F_0:
Kolmogorov-Smirnov test $\sup_x |\hat{F}_n(x) - F_0(x)|$ small

Shattering $N_{\mathcal{A}}(z^n) = |z^n \cap A : A \in \mathcal{A}| \leq 2^n$

Shattering coeff $s(\mathcal{A}, n) = \max_{z^n} N_{\mathcal{A}}(z^n)$

Sauer's lemma. $s(\mathcal{A}, n) \leq (n+1)^{VC(\mathcal{A})}$

$\Delta(\mathcal{A}) = \sup_{A \in \mathcal{A}} |P_n(A) - P(A)|$

VC thm. $P(\Delta(A) \geq t) \leq 8s(\mathcal{A}, n)\exp(\frac{-nt^2}{32}) = \delta$

Functions. $\Delta(\mathcal{F}) = \sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum f(X_i) - \mathbb{E} f\right|$

Rademacher complexity.
$\mathcal{R}(\mathcal{F}) = \mathbb{E}_{\varepsilon, X} \sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum \varepsilon_i f(X_i)\right|$

Rademacher thm. $\mathbb{E}\,\Delta(\mathcal{F}) \leq 2\mathcal{R}(\mathcal{F})$

Decision theory, Bayes

Minimax bound $\mathop{\mathbb{E}}_{\theta\sim\pi} R(\theta,\hat\theta_\pi) \le \mathop{\mathbb{E}}_{\theta\sim\pi} R(\theta,\hat\theta_m) \le \sup_\theta R(\theta,\hat\theta_m) \le \sup_\theta R(\theta,\hat\theta_{\text{up}})$

If $\sup_\theta R(\theta,\hat\theta_\pi) \le \mathop{\mathbb{E}}_{\theta\sim\pi} R(\theta,\hat\theta_\pi)$ e.g. $R(\theta,\hat\theta_\pi)=c$, then π is a *least favorable prior* and $\hat\theta_\pi$ is minimax

Posterior risk $R(\hat\theta|X^n) = \mathbb{E}_{\theta\sim\pi} L(\theta,\hat\theta(X^n))$

Posterior dist $\pi(\theta|X^n) = \frac{p(X^n|\theta)\pi(\theta)}{m(X^n)}$

Bayes risk $\mathbb{E}_{\theta\sim\pi} R(\hat\theta) = \int R(\hat\theta|X^n)m(X^n)dX^n$

MSS

MSS. $p(X^n;\theta)/p(Y^n;\theta) \perp\!\!\!\perp \theta$ iff $T(X^n)=T(Y^n)$

Factorization theorem. T sufficient iff $p(X^n;\theta)=h(X^n)g(T;\theta)$

Rao-Blackwell. $R(E[\hat\theta|T]) \le R(\hat\theta)$ for sufficient T

MLE

Asymp normal $\sqrt{n}(\hat\theta_{mle}-\theta) \xrightarrow{d} N(0, I_1(\theta)^{-1})$

Equivariant: if $\eta=g(\theta)$ then $\hat\eta=g(\hat\theta)$

Score $s(\theta) = \sum \nabla_\theta \log p(X_i;\theta)$

Fisher info
$I_n(\theta) = -\mathbb{E}\sum \nabla^2 \ell\ell(\theta) = \mathbb{E}\, s(\theta)s(\theta)^T = \text{Cov}(s)$

Asymp efficient: Cramér-Rao $\mathbb{V}\hat\theta \ge \frac{1}{nI_1(\theta)}$

KL dist

$KL(p||q) = \mathbb{E}_p \log\frac{p(x)}{q(x)} \ge 0$ w/eq if $p=q$

$KL(p||q) = \mathbb{E}_p \log p - \mathbb{E}_p \log q = -H(p) + H(p,q)$

$\hat{\mathbb{E}}_p \log q = \frac{1}{n}\sum \log q(X_i) = \frac{1}{n}\ell_q(X_i) \Rightarrow$ MLE

MLE consistent: $KL = (KL - KL_n) + KL_n \le 0$ by a uniform law $\sup_{\hat\theta}|KL_n - KL| \xrightarrow{p} 0$, so by strong identifiability $|\hat\theta-\theta| \le 0 \Rightarrow \hat\theta \xrightarrow{p} \theta$

Testing and confidence

Valid if $P_0(X^n \in R) \le \alpha$

Power $P_1(X^n \in R) = 1 - \text{Type II}$

p-value $p = \min\{\alpha | T(X^n) \in R_\alpha\}$

p-values test over α, CI test over Θ

CI-test duality. $C(X^n) = \{\theta | X^n \in \text{accept}(\theta)\}$

Pivot. dist of $Q(\theta, X^n) \perp\!\!\!\perp \theta$, e.g. $\overline{X}-\mu$ or $X_{(n)}/b$
$\Rightarrow$ CI: $C(X^n) = \{\theta | Q \in S\}$ s.t. $P(Q\in S) = 1-\alpha$

Tests

NP test simple/simple: $P_0\left(\frac{f_1(X^n)}{f_0(X^n)} \ge c\right) \le \alpha$
uniformly most powerful

Wald test simple/composite. Given $\hat\theta \xrightarrow{d} N(\theta_0, \sigma_0^2)$ under H_0, reject if $|T_n| = |\frac{\hat\theta-\theta_0}{\sigma_0}| \ge z_{\frac{\alpha}{2}}$. $\hat\sigma_0$ ok too.

For MLE, $T_n = \sqrt{I_n(\hat\theta)}(\hat\theta-\theta_0)$
Power $1-\Phi(\Delta+z_{\alpha/2}) + \Phi(\Delta - z_{\alpha/2})$
$\Delta = \sqrt{nI_1(\theta_1)}(\theta_0-\theta_1)$

GLRT composite/composite. $\lambda = \frac{L(\hat\theta_0)}{L(\hat\theta)}$

Wilks: asymp dist of GLRT
$T = -2\log\lambda(X_i) \xrightarrow{d} \chi^2_\nu$, reject if $T > \chi^2_{\nu,\alpha}$
$\ell(\theta) \approx \ell(\hat\theta) + \ell''(\hat\theta)\frac{(\theta-\hat\theta)^2}{2}$
$T = 2\ell(\hat\theta) - 2\ell(\theta_0) \approx -\ell''(\hat\theta)(\theta_0-\hat\theta)^2$
$T = \frac{-\frac{1}{n}\ell''(\hat\theta)}{I_1(\theta_0)}(\sqrt{I_n(\theta_0)}(\theta_0-\hat\theta))^2$. WLLN, Slutsky's

Pearson χ^2 test for k-cat multinomial
$T = \sum^k \frac{(Y_i - np_{0,i})^2}{np_{0,i}} \xrightarrow{d} \chi^2_{k-1}$ under H_0, counts

Permutation test. Under $H_0: P=Q$, $T(X^n,Y^n) = |\overline{X} - \overline{Y}| \sim T(\text{random permutations})$, so $P_0(T \ge T_{obs}) = 1 - \hat\Phi(T_{obs})$ i.e. reject if $T_{obs} \ge \hat\Phi^{-1}(1-\alpha)$

Bootstrap

Bootstrap: compute $\hat\theta$ on B samples $X_i^* \sim P_n$
$\Rightarrow$ dist $\Rightarrow$ mean, sample var, CIs, etc
MC est draws new samples

Exp family

$p(x;\theta) = h(x)\exp\left(\sum \eta_i(\theta)T_i(x) - A(\theta)\right)$
$p(X^n) = \prod h(X_i)\exp\left(\eta(\theta)\sum T(X_i) - nA(\theta)\right)$
or $\sum_i \theta_i \sum_j T_i(x_j)$

$N(x,\theta) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{\mu}{\sigma^2}x + \frac{-1}{2\sigma^2}x^2 + \frac{-\mu^2}{2\sigma^2}\right)$
Poisson $\frac{1}{x!}\exp(x\log\theta - \theta)$
Binom $\binom{n}{x}\exp\left(x\log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right)$

$\eta(\theta)$ are the natural parameters
Sufficient statistics T or $(\sum T_1(X_i), \ldots, \sum T_n(X_i))$
minimal if T is full rank

$A(\theta)$ log normalizer ensures area 1
domain $\{A | A(\theta) \text{ finite}\}$

$\partial A(\theta)/\partial\theta_i = \mathbb{E}\, T_i(x)$
$\partial^2 A(\theta)/\partial\theta_i\partial\theta_j = \text{Cov}(T_i(X), T_j(X) \succeq 0$
A convex so $\ell\ell$ concave so easy to compute MLE

$I_1(\theta) = \nabla^2 A(\theta) = n\,\mathbb{V}\,T(X)$

For exp, Bregman divergence = KL:
$\rho(\theta_1,\theta_2) = A(\theta_2) - A(\theta_1) - \nabla A(\theta)^T(\theta_2-\theta_1)$

Multiple testing

FWER $\le P_0(\cup \text{reject } H_i)$. FDR $= \mathbb{E}\,\text{FDP} = \mathbb{E}\,V/R$. FWER $\ge$ FDR w/eq under global null

Sidak correction. For indep. p_i, FWER $\le 1-(1-\alpha_i)^n = a \Rightarrow \alpha_i = 1-(1-\alpha)^{1/n}$

Bonferroni correction. UB: FWER $\le n\alpha_i = \alpha \Rightarrow \alpha_i = \alpha/n$, $P(\text{reject } H_i) = \alpha_i$

Holm's procedure. Reject H_i while $p_{(i)} \le \frac{\alpha}{n-i+1}$. $V = 0$ if $\min_{i\in I_0} p_i > \frac{\alpha}{n_0}$. Else FWER $\le n_0 \cdot \frac{\alpha}{n_0}$

BH procedure. Reject up to $\max p_{(i)} < \frac{i}{d}\alpha$
For indep p_i, $FDR \le \frac{d_0 i_{\max}}{d i_{\max}} = \frac{d_0\alpha}{d}$

Causal inference

Estimate **ATE** $\tau = \mathbb{E}\,Y(1) - \mathbb{E}\,Y(0)$

Randomized trial $W \perp\!\!\!\perp Y(1), Y(0)$
$\tau = \mathbb{E}[Y(1)|W=1] - \mathbb{E}[Y(0)|W=0] \Rightarrow \hat{\mathbb{E}}$ on Y^{obs}

Fisher's exact p-value test for sharp null $Y_i(0) = Y_i(1) \forall i$, randomize W, empirical $\hat\tau$ dist, compare τ_{obs}

No unmeasured confounding for obs. study
$W \perp\!\!\!\perp Y(1), Y(0)|X$
$\tau = \mathbb{E}_X\,\mathbb{E}[Y(1) - Y(0)|X]$
$= \mathbb{E}_X\,\mathbb{E}[Y(1)|X, W=1] - \mathbb{E}_X\,\mathbb{E}[Y(0)|X, W=0]$
$\hat\tau = \hat{\mathbb{E}}_X[\hat\mu_1(X) - \hat\mu_2(X)]$

Inverse **propensity score** or Horvitz-Thompson
$\pi(x) = P(W=1|x) = \mathbb{E}(W|x)$
$\tau = \mathbb{E}_X\,\mathbb{E}_W\,\mathbb{E}[\frac{Y(1)W}{\pi(X)}|X] - \mathbb{E}_X\,\mathbb{E}_W\,\mathbb{E}[\frac{Y(0)(1-W)}{1-\pi(X)}|X]$
$\hat\tau = \frac{1}{n}\sum \frac{Y_i^{\text{obs}} W_i}{\pi(X_i)} - \frac{Y_i^{\text{obs}}(1-W_i)}{1-\pi(X_i)}$

Gaussian seq

$y = \theta + \varepsilon$, $\varepsilon_i \sim N(0, \sigma^2/n)$
MSE/lsq $\hat\theta = y$ with $R = \mathbb{E}\|\varepsilon\|^2 = \sigma^2 d/n$ or $\sigma^2 s/n$

Hard thresh. $\hat\theta_i = y_i I(|y_i| \ge t)$ minimizes $\frac{1}{2}\|y-\theta\|_2^2 + \frac{t^2}{2}\|\theta\|_0$.
W.p. $1-\delta$, $\max \varepsilon_i \le \sigma\sqrt{2\log(2d/\delta)/n} = \frac{t}{2}$
$\|\hat\theta - \theta\|_2^2 \le 9\sum \min(\theta_i^2, t^2/4)$
$R \lesssim \sum \min(\theta_i^2, \sigma^2(\log d)/n)$, sum over s entries if sparse
$\|\theta\|_1 \le R \Rightarrow R \lesssim R\sigma\sqrt{(\log d)/n}$
Split on $|\theta_i| \ge \frac{t}{2}$ and bound k

Soft thresh. $\hat\theta_i = \text{sgn}(y_i)\max(|y_i| - t, 0)$ minimizes $\frac{1}{2}\|y-\theta\|_2^2 + t\|\theta\|_1$

Regression

$L(\hat r) = \int (\hat r(x) - r(x))^2 dx$
$R = \mathbb{E}\,L = \mathbb{E}(Y - \hat r(X))^2 = \int b(x)^2 + v(x)\,dx$
$v(x) = \mathbb{E}(\hat r(x) - r(x))^2$
$r(x) = \mathbb{E}[Y|X=x]$ minimizes R
$y_i = \beta^{*T} X_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$
$\hat\beta = (X^TX)^{-1}X^Ty = \frac{1}{n}\hat\Sigma^{-1}X^Ty$ mins $\frac{1}{2n}\|y - X\beta\|_2^2$
Mean in-sample error $\sigma^2 d/n$ looks like mean est.
$\mathbb{E}\|\hat\beta - \beta^*\|_2^2 \le \frac{\sigma^2 d}{n\lambda_{(1)}(X^TX)^{-1}}$
$\frac{1}{n}\|X\hat\beta - X\beta^*\|_2^2 \le 4\sigma\|\beta^*\|_1\sqrt{\frac{2\log(2d/\delta)}{n}}$
$\cdot \le \langle \frac{1}{n}X^T\varepsilon, \hat\beta - \beta^*\rangle \le \|\hat\beta - \beta^*\|_1 \|\frac{1}{n}X^T\varepsilon\|_\infty$

Kernel regression

$\hat r(x) = \sum w_i(x) Y_i$, $w_i(x) \propto K(\frac{x - X_i}{h}) \le \frac{1}{nh}$
assume $K = I(X \in [-1, 1])$
$b(x) \le \sum w_i(x)|r(X_i) - r(x)| \le Lh$
$v(x) \le \mathbb{E}(\sum \varepsilon_i w_i(x))^2 \le \frac{\sigma^2}{Lh}$
$h = n^{-1/3}$, $R = n^{-2/3}$
In dim d, $R = (Lh)^2 + \frac{\sigma^2}{nh^d} \approx n^{\frac{-2}{2+d}}$

Bayesian

Credible sets $\int \pi(\theta|X^n)d\theta = 1 - a$

Frequentist guarantees: π shrinking around θ^*, rate

Consistent $\pi(\{\theta : |\theta - \theta^*| \ge \varepsilon_n\}|X^n) \to 0$

Bernstein-von Mises: valid asymp CI if $\pi(\theta) > 0$ near θ^* (low dim): $\|\pi(\theta|X^n), N(\hat\theta, I_n^{-1}(\hat\theta))\|_{\text{TV}} \to 0$

Model Selection

AIC: $2\ell_j(\hat\theta_j) - 2d_j$ or $\frac{\ell_j(\hat\theta_j)}{n} - \frac{d_j}{n}$

BIC: $2\ell_j(\hat\theta_j) - d_j \log n$

Or hypothesis testing: $H_0 : \theta = 0$, $P(X^n \in R) \le \alpha$ like AIC but slightly higher penalty to control error