Hypothesis Testing

Null hypothesis $H_0 : \theta \in \Theta_0$, **alternative** $H_1 : \theta \in \Theta_1$.
Type I error: If H_0 true but we reject H_0.
To construct a test:

$$\begin{aligned} &1.\ \text{Choose a test statistic } W = W(X_1,\dots,X_n)\\ &2.\ \text{Choose a rejection region } R\\ &3.\ \text{If } W \in R, \text{ reject } H_0 \text{ otherwise retain } H_0 \end{aligned} \tag{1}$$

The **power function** $\beta(\theta) = \mathbb{P}_\theta(W \in R)$ for a rejection region R.
Want **level-**α test ($\sup_{\theta\in\Theta_0}\beta(\theta) \le \alpha$) that maximizes $\beta(\theta \in \Theta_1)$.
A level-α test with power fn β is **uniformly most powerful** if: $\beta(\theta) \ge \beta'(\theta)\ \forall\theta \in \Theta_1\ \forall\beta' \ne \beta$.

Neyman-Pearson Test

For simple $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, reject H_0 if $\frac{L(\theta_1)}{L(\theta_0)} > k$.
where k chosen s.t. $\mathbb{P}(\frac{L(\theta_1)}{L(\theta_0)} > k) = \alpha$.

Wald Test

For $H_0 : \theta = \theta_0$ and $H_1 : \theta \ne \theta_0$, reject H_0 if $\left|\frac{\hat\theta_n - \theta_0}{se(\hat\theta_n)}\right| > z_{\alpha/2}$.
where $z_{\alpha/2}$ is the inverse standard-normal CDF of $1 - \frac{\alpha}{2}$.
and $\hat\theta_n$ an estimator s.t. $(\hat\theta - \theta)/\text{se} \rightsquigarrow \mathcal{N}(0,1)$ eg: $\theta = \hat\theta_{\text{mle}}$
and $se = \sqrt{\text{Var}(\hat\theta_n)}$. Can also use (for eg.) $\hat{se} = \sqrt{S_n^2/n}$.
and if $\hat\theta_n$ efficient, can approx: $se \approx \sqrt{\frac{1}{I_n(\theta)}}$ or $\hat{se} \approx \sqrt{\frac{1}{I_n(\hat\theta_n)}}$.

Likelihood Ratio Test

For $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \notin \Theta_0$, reject H_0 if $\lambda(x^n) = \frac{L(\hat\theta_0)}{L(\hat\theta)} \le c$.
where $L(\hat\theta_0) = \sup_{\theta\in\Theta_0} L(\theta)$ and $L(\hat\theta) = \sup_{\theta\in\Theta} L(\theta)$.
and c chosen s.t. $\mathbb{P}(\lambda(x^n) \le c) = \alpha$.
Thm: under $H_0 : \theta = \theta_0 \implies W_n = -2\log\lambda(X^n) \rightsquigarrow \chi^2_1$
$\implies$ reject H_0 if $W_n > \chi^2_{1,\alpha}$.
Also: for $\theta = (\theta_1,\dots,\theta_k)$, if H_0 fixes some of the parameters
$\implies -2\log\lambda(X^n) \rightsquigarrow \chi^2_\nu$, where $\nu = \dim(\Theta) - \dim(\Theta_0)$.

P-Values

The **p-value** $p(x^n)$ is the smallest α-level s.t. we reject H_0.
Thm: For a test of the form: reject H_0 when $W(x^n) > c$,
$\implies p(x^n) = \sup_{\theta\in\Theta_0}\mathbb{P}_\theta(W(X^n) \ge W(x^n)) = \sup_{\theta\in\Theta_0}[1 - F(W(x^n)|\theta)]$.
Thm: Under $H_0 : \theta = \theta_0$, $p(x^n) \sim \text{Unif}(0,1)$.

Permutation Test

$X^n \sim F$, $Y^m \sim G$, $H_0 : F = G$, $H_1 : F \ne G$
Let $Z = (X^n, Y^m)$ and $L = (1,\dots,1,2,\dots,2)$.
Let $W = g(L,Z) = |(\text{ave of 1 labeled pts}) - (\text{ave of 2 labeled pts})|$.
Let $p = \frac{1}{N!}\sum_\pi \mathbb{I}(g(L_\pi,Z) > g(L,Z)) \implies$ reject H_0 when $p < \alpha$.

Confidence Intervals

We want a $1-\alpha$ **confidence interval** $C_n = [L(X^n), U(X^n)]$ s.t. $\mathbb{P}_\theta(L(X^n) \le \theta \le U(X^n)) \ge 1-\alpha, \quad \forall\theta \in \Theta$.
Generally, a $1-\alpha$ **confidence set** C_n is a random set $C_n \subset \Theta$ s.t. $\inf_{\theta\in\Theta}\mathbb{P}_\theta(\theta \in C_n(X^n)) \ge 1-\alpha$.

Using Probability Inequalities

Prob inequalities give (for eg.) $\mathbb{P}(|\hat\theta_n - \theta| > \epsilon) \le g(\exp^{-f(\epsilon)}) \underset{\text{set to}}{=} \alpha$.
solving for ϵ gives $\epsilon = \tilde f(\alpha) \implies \mathbb{P}\left(|\hat\theta_n - \theta| > \tilde f(\alpha)\right) \le \alpha$
$\implies C_n = \left(\hat\theta - \tilde f(\alpha), \hat\theta + \tilde f(\alpha)\right)$.

Inverting a Test

In level-α tests $\mathbb{P}_{\theta_0}(T(x^n) \in R) \le \alpha \Rightarrow$ let $C_n = \{\theta : T(x^n) \in A(\theta_0)\}$.
where $A(\theta_0) = \{T(x^n) \notin R \mid \theta = \theta_0\}$ (ie the accept region if $\theta = \theta_0$).
For Wald: $C_n = \hat\theta_n \pm (z_{\alpha/2} \times se) = \hat\theta_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt n}$.
For LRT: $C_n = \{\theta : \frac{L(\theta)}{L(\hat\theta)} > c\}$ (for test where reject H_0 if $\frac{L(\hat\theta_0)}{L(\hat\theta)} \le c$).

Pivots

$Q(X^n,\theta)$ a **pivot** if the distribution of Q does not depend on θ.
Find a, b s.t. $\mathbb{P}_\theta(a \le Q(X^n,\theta) \le b) \ge 1-\alpha$, $\forall\theta$.
$\implies C_n = \{\theta : a \le Q(X^n,\theta) \le b) \ge 1-\alpha\}$.

Large Sample Confidence Intervals

For mle $\hat\theta_n$ with se $\approx 1/\sqrt{I_n(\hat\theta_n)}$, approx $1-\alpha$ confidence sets are:
For Wald: $C_n = \hat\theta_n \pm (z_{\alpha/2} \times \text{se})$
For Wald with delta method: $C_n = \tau(\hat\theta_n) \pm (z_{\alpha/2} \times \text{se}(\hat\theta) \times |\tau'(\hat\theta_n)|)$
For LRT: $C_n = \left\{\theta : -2\log\left(\frac{L(\theta)}{L(\hat\theta)}\right) \le \chi^2_{k,\alpha}\right\}$

Nonparametric Inference

The **empirical CDF** is: $\hat F_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \le x)$
Thm (DKW): $\forall\epsilon > 0, \mathbb{P}(\sup_x|\hat F(x) - F(x)| > \epsilon) \le 2e^{-2n\epsilon^2}$
The **kernel density estimator** is: $\hat p_n(x) = \frac{1}{n}\sum_{i=1}^n \frac{1}{h}K\left(\frac{x - X_i}{h}\right)$
where K a symmetric zero-mean density, and bandwidth $h > 0$.
Thm: The risk $R = \mathbb{E}(\mathcal{L}(p,\hat p)) = \int(\text{b}^2(x) + \text{Var}(x))dx = \frac{a}{n^{4/5}}$
for some a, where $\mathcal{L}(p,\hat p) = \int(p(x) - \hat p(x))^2dx$, and $\text{b}^2(x) = \mathbb{E}(\hat p(x)) - p(x)$. And this is minimax.
A **statistical functional** $T(F)$ is any function of the CDF.
A **plug-in estimator** of $\theta = T(F)$ is: $\hat\theta_n = T(\hat F_n)$.
Often, $\hat\theta_n \approx \mathcal{N}(T(F), \hat{\text{se}}^2)$, where $\hat{\text{se}}$ is estimate of $\sqrt{\text{Var}\left(T(\hat F_n)\right)}$.

Bootstrap

The **bootstrap** is a nonparametric way to find standard errors and confidence intervals of estimators of statistical functionals:

$$\begin{aligned} &1.\ \text{Draw } X_1^*,\dots,X_n^* \sim \hat F_n \text{ (via } X_i^* \sim \{X_1,\dots,X_n\} \text{ unif).}\\ &2.\ \text{Compute } T_n^* = g(X_1^*,\dots,X_n^*)\\ &3.\ \text{Do 1. and 2. } B \text{ times to get } T_{n,1}^*,\dots,T_{n,B}^*\\ &4.\ \text{Let } v_{\text{boot}} = \frac{1}{B}\sum_{b=1}^B\left(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^B T_{n,r}^*\right)^2 \end{aligned} \tag{2}$$

Then: $v_{\text{boot}} \xrightarrow{a.s.} \text{Var}_{\hat F_n}(T_n)$ as $B \to \infty$ and $\hat{\text{se}}(T_N) = \sqrt{v_{\text{boot}}}$

Bayesian Inference

Frequentists: probability is long-run frequencies. Procedures are random but parameters are fixed, unknown quantities.
Bayesians: probability is a measure of subjective degree of belief. Everything is random, including parameters.
Using Bayes Thm $\nRightarrow$ Bayesian inference.
For $X_1,\dots,X_n \sim p(x|\theta)$, and prior $\pi(\theta)$, **Bayes Thm** gives:

$$\pi(\theta|X^n) = \frac{p(X^n|\theta)\pi(\theta)}{m(X^n)} = \frac{p(X^n|\theta)\pi(\theta)}{\int p(X^n|\theta)\pi(\theta)d\theta} \tag{3}$$

Prediction

For train-data $(X_i, Y_i)|_{i=1,\dots,n}$, want to predict Y given a new X, where $Y \in \{0,1\}$ (**classification**) or $Y \in \mathbb{R}$ (**regression**).
For prediction rule $h(X)$,
classification risk: $R(h) = \mathbb{P}(Y \ne h(X)) = \mathbb{E}(I(Y \ne h(X)))$
regression risk: $R(h) = \mathbb{E}\left((Y - h(X))^2\right)$
Thm 1: $R(h)$ minimized by $m(x) = \mathbb{E}(Y|X = x)$.
The **Bayes classifier** $h_B(x) = I(m(x) \ge 1/2)$

Model Selection

Consider models $\mathcal{M}_{1,\dots,k}$, $\mathcal{M}_j = \{p(y;\theta_j) : \theta_j \in \Theta_j\}$, $\hat\theta_j = \text{mle}(\mathcal{M}_j)$
AIC: choose $j^* = \arg\max|_j\ \text{AIC}(j) = 2\log L_j(\hat\theta_j) - 2\dim(\Theta_j)$
BIC: choose $j^* = \arg\max|_j\ \text{BIC}(j) = \log L_j(\hat\theta_j) - \left(\frac{\dim(\Theta_j)}{2}\log n\right)$
Cross-validation: For train-data $= Y_{1,\dots,n}$ and test-data $= Y^*_{1,\dots,n}$
choose $j^* = \arg\max|_j \hat K_j = \frac{1}{n}\sum_{i=1}^n \log p(Y_i^*;\hat\theta_j)$