## 1. Key Distributions

| | | $\mu_x$ | $\sigma^2_x$ |
|---|---|---|---|
| A) - Binomial | $f(x) = \binom{n}{p} p^x (1-p)^{n-x}$ | $n \cdot p$ | $n(p(1-p)$ |
| - Poisson | $f(x) = \frac{\lambda^x}{x!} \exp(-\lambda)$, $\lambda > 0$ | $\lambda$ | $\lambda$ |
| - Exponential | $f(x) = \lambda \cdot \exp(-\lambda x)$ | $1/\lambda$ | $1/\lambda^2$ |

B) Exponential family

$$f(y;x) = h(y)\exp(\eta(\theta)\,T(y) - A(\eta(\theta)))$$

canonical form: $\eta(\theta) = c_1\theta + c_2$

## 2. Method of moments

- Set theoretical moment equal to sample moment

$E[X^k]$ set equal to $\frac{1}{n}\sum X_i^k$

→ can also at moment around center

- Expected value

cont. $E[x^k] = \int_{-\infty}^{\infty} x^k \cdot f(x)\,dx$

disc. $E[x^k] = \sum_{\delta} x_i^k\, p(x)$

- $\mathrm{Var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$
- Corr: $\rho = \frac{\mathrm{Cov}(X,Y)^2}{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$ where $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$

## 3. a) Likelihood + log-likelihood

$L(\theta) = f(x,\theta) = \prod_{i=1}^{n} f(x_i;\theta)$, $\theta \in \Theta$ (assumes $x_{1:n}$ i.i.d)

$\ell(\theta) = \log(f(x,\theta)) = \sum_{i=1}^{n} \log(f(x_i;\theta))$, $\theta \in \Theta$

→ (log-) Likelihood without constant is known as kernel

b) MLE (used to find MLE)

$\hat\theta_{ML} = \underset{\theta}{\mathrm{argmax}}\, L(\theta) = \underset{\theta}{\mathrm{argmax}}\, \ell(\theta) = \underset{\theta}{\mathrm{argmax}}\, \tilde L(\theta) = \underset{\theta}{\mathrm{argmax}}\, \tilde\ell(\theta)$

c) Relative Likelihood $\tilde L(\theta) = \frac{L(\theta)}{L(\hat\theta_{ML})}$

$\tilde\ell(\theta) = \ell(\theta) - \ell(\hat\theta_{ML})$

d) One-to-one Transformations (If $\varphi = h(\theta)$, $\theta = h^{-1}(\varphi)$)

- $\hat\varphi_{ML} = h(\hat\theta_{ML})$
- $S_\varphi(\varphi) = S_\theta(\theta)\frac{d\,h^{-1}(\varphi)}{d\varphi}$
- $I_\varphi(\varphi) = I_\theta(\theta)\left[\frac{d\,h^{-1}(\varphi)}{d\varphi}\right]^2 - S_\theta(\theta)\frac{d^2 h^{-1}(\varphi)}{d\varphi^2}$
- $I_\varphi(\hat\varphi_{ML}) = I_\theta(\hat\theta_{ML})\left[\frac{d\,h(\theta_0)}{d\theta}\right]^{-2}$ diff first! then plug in
- $J_\theta(\theta) = J_\theta(\theta)\cdot\left[\frac{d\,h(\theta)}{d\theta}\right]^{-2}$

e) Quadratic approx. of log-Likelihood

$\ell(\theta) \approx \ell(\hat\theta_{ML}) + \underbrace{S_\theta(\hat\theta_{ML})}_{=0}(\theta - \hat\theta_{ML}) - \frac{1}{2} I_\theta(\hat\theta_{ML})(\theta - \hat\theta_{ML})^2$

→ simplification using rel. log-likelihood

$\tilde\ell(\theta) \approx -\frac{1}{2} I_\theta(\hat\theta_{ML})(\theta - \hat\theta_{ML})^2$

f) Likelihood Ratio

$$\Lambda_{x_{1:n}}(\theta_1,\theta_2) = \frac{L(\theta_1; x_{1:n})}{L(\theta_2; x_{1:n})} = \frac{\tilde L(\theta_1, x_{1:n})}{\tilde L(\theta_2, x_{1:n})}$$

↳ if compared to $\tilde L(\hat\theta_{ML}; x_{1:n})$ → can recover $\tilde L(\theta)$

## 4. Sufficient Statistics

$T$ with realization $t = h(x_{1:n})$ is sufficient for $\theta$ iff:

$$f(x_{1:n}|t) = \frac{f(t(x_{1:n}))}{f(t)} \cdot f(x_{1:n}) \quad \perp \theta$$

($f(t(x_{1:n}))$: $0,1$)

b) Factorization Theorem

$$f(x_{1:n};\theta) = g_1(t,\theta)\cdot g_2(x_{1:n})$$

($t(\theta)$; $g_2$ can also $=1$ (e.g. sample))

c) Minimal Sufficiency

- $T$ is min. sufficient if it can be written as a function of any other sufficient statistic
  → $T_1$ & $T_2$ min. sufficient for $\theta$
  → $T_1 = g(T_2)$, $T_2 = g^{-1}(T_1)$
- Necc. and suff. criterion for $T$ to be min. suff. is that $h(x_{1:n}) \overset{t}{=} h(\tilde x_{1:n})$ iff:

$\Lambda_{x_{1:n}}(\theta_1,\theta_2) = \Lambda_{\tilde x_{1:n}}(\theta_1,\theta_2)$ → $L(\theta)$ is min. suff.

## 5. Unbiasedness and Consistency

- Estimator ($T_n = h(X_{1:n})$) vs. Estimate ($t_n = h(x_{1:n})$)
  ↑ suitable statistic whose realization is a point estimate; ↑ sensible value for $\theta$ acc. to data and model
- Unbiasedness: $E(T_n(X)) = \theta$, $\forall\,\theta\in\Theta, n\in\mathbb{N}$
  (asymptotic) $\lim_{n\to\infty}$ " " " "
- Consistency: $T_n(X) \longrightarrow \theta$ for $n\to\infty$
- MS Consistency: $E[(T_n(X) - \theta)^2] \xrightarrow{n\to\infty} 0$ (show: (i) unbiased (ii) consistent)
- Jensen's Inequality: $E[g(X)] = g(E[X])$ linear $g$; $\leq$ concave $g$; $\geq$ convex $g$
- CLT: To derive asympt. distributions for mean ($\bar x$) or sum ($\sum x_i$) of R.Vs

$$\frac{\sqrt{n}(\bar x - E[X])}{\sqrt{\mathrm{Var}[x]}} \xrightarrow{D} N(0,1)$$

## 6. Further Likelihood properties / FI

a) E and Var of Score (under FRC)

$E[S(\theta;X)] = 0$

$\mathrm{Var}[S(\theta;X)] = J(\theta)$

b) Exp. FI from a random sample:

$J_{1:n}(\theta) = n\cdot J(\theta)$

## 7. Test and Confidence Interval

a) Confidence Interval:

$P(T_u \leq \theta \leq T_o) = \gamma \quad \forall\,\theta\in\Theta$

b) Pivot: Statistic depending on data ($x_{1:n}$) and $\theta$ whose distribution is independent of $\theta$ → pivotal distr.
↳ approx Pivot: first show asymptotic MS Consistency!

c) Tests: quantify evidence against $H_0$ (using p-value)

- P-value: $P($ obtaining equal or more extreme result ($T$) than observed $|\ H_0$ is true$)$
- Type I: $P(T(x_{1:n}) \geq c_{crit} \mid H_0) = \alpha$ (sign. level)
- Type II: $P(T(x_{1:n}) \leq c_{crit} \mid H_1) = \beta \to 1-\beta$ (Power)

→ Power curve: shows $P_{1-\beta}$ (power) for diff. levels of $\theta_0$

d) Statistics (Z, Wald, Score, LR)

- Wald: $\sqrt{I(\hat\theta_{ML})}(\hat\theta_{ML} - \theta_0) \overset{a}{\sim} N(0,1)$
  (a) $J(\hat\theta_{ML})$ or (b) $I/J(\theta)$
  (1) CI: $[\hat\theta_{ML} \pm z_{\frac{1+\gamma}{2}}\cdot se(\hat\theta_{ML})]$
  (2) CI: $\left\{\theta : \frac{(\hat\theta_{ML} - \theta)^2}{I(\theta)} \leq z^2_{\frac{1+\gamma}{2}}\right\}$ → solve $\theta$
  ↳ solve quad. eq: $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$
- Score: $\frac{S(\theta;x_{1:n})}{\sqrt{J(\theta;x_{1:n})}} \overset{a}{\sim} N(0,1)$ (asymptotic) (FRC hold)
  ↑ or $I(\theta)$ or $I/J(\hat\theta_{ML})$
  → same CI as (2)
- Likelihood Ratio

$$W = 2\cdot\log\left[\frac{L(\hat\theta_{ML})}{L(\theta)}\right] = -2\,\tilde\ell(\theta) \overset{a}{\sim} \chi^2_{\gamma,p}$$

$$\to CI: \left\{\theta;\ \tilde L(\theta) \geq \exp\left[-\frac{1}{2}\chi^2_{\gamma,p}\right]\right\}$$

$$= \tilde\ell(\theta) \geq -\frac{1}{2}\chi^2_{\gamma,p}$$

↳ use LR to compute CI / Regions for test when we have rel. log-likelihood

e) Common critical values with Z-score:
10%: 1.64 ; 5%: 1.96 ; 1%: 2.58

## 8. Distribution of the MLE → Need FRC

a) $\hat{\theta}_{ML} \overset{a}{\sim} N(\theta_0, J_{1:n}(\theta_0)^{-1})$ ← as $n \uparrow$, Var ↓
or $n \cdot J(\theta_0)^{-1}$, $I(\hat{\theta}_{ML})^{-1}$

→ MLE is asympt. unbiased + consistent

b) Cramér Rao Lower bound → to show efficiency

For an arbitrary consistent estimator, under FRC

$$\mathrm{Var}(T(X_{1:n})) \geq \frac{g'(\theta)^2}{J(\theta)} \quad \left(= \frac{1}{J(\theta)} \text{ if } g(\theta)=\theta\right)$$

↳ also CRLB for $\theta_{ML}$

→ an asympt. unbiased estimator that attains the CRLB is efficient

b) Efficiency: $\mathrm{eff}(T) = \frac{\frac{\partial}{\partial\theta} E(T|X_{1:n})^2}{n\,J(\theta_0)\cdot \mathrm{Var}(T(X_{1:n}))} \leq 1$

($n\,J(\theta_0) = \mathrm{Var}(S(\theta))$)

c) SE of the MLE

$\mathrm{se}(\hat{\theta}_{ML}) = 1/\sqrt{I(\hat{\theta}_{ML})}$ or $1/\sqrt{J(\hat{\theta}_{ML})}$

d) Delta Method (for SE in general)

$$\mathrm{se}(h(\hat{\theta})) = \mathrm{se}(\hat{\theta}) \cdot \left|\frac{d\,h(\hat{\theta})}{d\,\hat{\theta}}\right|$$

↳ can use to calc Wald CI for transformation

e) Variance Stabilizing Transformations

$$\varphi = h(\theta) \propto \int^{\theta} J_\theta(u)^{\frac{1}{2}}\,du$$

is a var. stabilizing trans of $\theta$

## 9. Multiparameter Model

a) $L(\vec{\theta}) \to \ell(\vec{\theta}) \to S(\vec{\theta}) = \frac{\partial \ell(\vec{\theta})}{\partial \vec{\theta}}$ … $\hat{\vec{\theta}}_{ML} = (\hat{\theta}_1 \dots \hat{\theta}_p)^T$

b) Fisher Info Matrix:

$$I(\vec{\theta}) = -\left(\frac{\delta^2 \ell(\vec{\theta})}{\delta\theta_i\,\delta\theta_j}\right), \quad 1 \leq i,j \leq p$$

sym. $p \times p$ matrix + pos. definite

↳ Hess. of the Hessian

c) SE: $\sqrt{[I(\hat{\theta}_{ML})^{-1}]_{ii}}$ (ith diagonal component)

d) SE of a difference: split data into two independent parts (e.g. obs. 0 and 1 with two LCEs)

→ SE of $\hat{\delta}_{ML} = \beta_{ML} - \lambda_{ML}$ has form:

$$\mathrm{se}(\hat{\delta}_{ML}) = \sqrt{I_0(\lambda_{ML})^{-1} + I_1(\beta_{ML})^{-1}}$$

e) Score Statistic: $J_{1:n}(\vec{\theta})^{-\frac{1}{2}} S(\vec{\theta}; x_{1:n})^T \overset{a}{\sim} N_p(0, I_p)$

or $I(\vec{\theta}; x_{1:n})$ or $\hat{\theta}_{ML}$

f) Multivariate Delta Method: Transformation $g(\vec{\theta}_{ML})$, $\mathbb{R}^p \to \mathbb{R}^q$

for $\vec{\theta}_{ML} \overset{a}{\sim} N_p(\vec{\theta}, I(\vec{\theta}_{ML})^{-1})$

→ $g(\vec{\theta}_{ML}) \overset{a}{\sim} N_q(g(\vec{\theta}), D(\vec{\theta}_{ML})\, I(\vec{\theta}_{ML})^{-1} D(\vec{\theta}_{ML})^T)$

where $D(\vec{\theta})$ denotes $q \times p$ Jacobian (1st deriv) of $g(\vec{\theta})$

→ sqrt of Var term are se of components of $g(\hat{\vec{\theta}}_{ML})$

## 10) Profile Likelihood

∫ max the nuisance out

a) $L_p(\vec{\theta}) = \max_{\vec{\eta}} L(\vec{\theta}; \vec{\eta}) = L(\vec{\theta}, \hat{\vec{\eta}}_{ML}(\vec{\theta}))$ of $\theta$

b) Estimated Likelihood of $\theta$: $L_e(\vec{\theta}) = L(\vec{\theta}; \hat{\vec{\eta}}_{ML})$

→ ignores uncertainty in $\vec{\eta}$! also generally not correctly quantifies the uncertainty w.r.t. $\vec{\theta}$

c) + Profile log-Likelihood & relative profile (log-) Likelihood

## 11) Generalized Likelihood Ratio Statistic

- LR stat analogous to above $\omega \sim \chi^2_p$
- For $\vec{\theta}$ ($q$-dim) parameter of interest vector
  $\vec{\eta}$ ($r$-dim) nuisance parameter vector

1. get joint Likelihood $L(\vec{\theta}; \vec{\eta})$
2. Profile Likelihood for parameter of interest $L_p(\vec{\theta}) = \max_{\vec{\eta}} L(\vec{\theta}; \vec{\eta})$
3. LR: $\omega_p := -2\,\tilde{\ell}_p(\vec{\theta})$ (via profile log-L)

$$= 2\log\left[\frac{L_p(\hat{\vec{\theta}}_{ML})}{L_p(\vec{\theta})}\right] \overset{a}{\sim} \chi^2_q$$

↳ can use to test $H_0: \vec{\theta} = \vec{\theta}_0$ → reject if $\omega$ is larger ($> q_{1-\alpha, n-p}$)

## 12) Regression: $\vec{Y} = X\vec{\beta} + \vec{\varepsilon}$ (with $X$ full rank)

- $\beta_{LS} = \mathrm{argmin}_{\vec{\beta}} (\vec{y} - X\vec{\beta})(\vec{y} - X\vec{\beta})^T = (X^TX)^{-1}X^Ty$
  - $\vec{y} \sim N_n(X\vec{\beta}, \sigma^2 I_n)$
  - $\hat{\vec{\beta}} \sim N_{p+1}(\vec{\beta}, \sigma^2 (X^TX)^{-1})$
  - $\hat{\vec{y}} = X\hat{\vec{\beta}} \sim N_n(X\beta, \sigma^2 X(X^TX)^{-1}X^T)$

## 12) Model Selection

a) AIC $= -2\,\ell(\hat{\vec{\theta}}_{ML}) + 2p$ ← total # parameters estimated

b) BIC $= -2\,\ell(\hat{\vec{\theta}}_{ML}) + p\log(n)$

## 13) Prediction

a) Plug-in Prediction (Distribution)

$f(y) = f(y; \hat{\theta}_{ML})$ ← just plug in + ignore uncertainty $\theta$

b) Extended Likelihood

$L(\theta, y) = f(x_{1:n}, y, \theta) = f(y;\theta) \cdot \prod_{i=1}^{n} f(x_i;\theta)$

→ focus on the kernel

c) Predictive Likelihood

$L_{pred}(y) = \max_{\theta} L(\theta, y) = L(\hat{\theta}(y), y)$

↳ corr. distribution $f_p(y) = L_{pred}(y) / \int L_{pred}(a)\,da$

d) Assessment of Predictions

- $PIT(y_0) = \hat{F}(y_0) \sim U(0,1)$
- $CRPS = \int \{\hat{F}(t) - I_{[y_0, \infty]}(t)\}^2\,dt$

lower is better! ↑ indicator with = 1 for $y_0 \leq t$

## 14) Bayesian Inference (parameter is R.V.)

a) $f(\theta|x) = \frac{f(x|\theta)\cdot f(\theta)}{f(x)}$

$= L(\theta)$ $\propto f(x|\theta)\cdot f(\theta)$

b) Point estimates: $E[\theta|X] = \int \theta\, f(\theta|x)\,d\theta$

- Post. mode: $\mathrm{Mod}(\theta) = \mathrm{argmax}_{\theta} f(\theta|x)$
- Post. median: any number that satisfies $\int_{-\infty}^{a} f(\theta|x)\,d\theta = 0.5$ & $\int_{a}^{+\infty} f(\theta|x)\,d\theta = 0.5$

c) Improper Prior: $\int_{\Theta} f(\theta)\,d\theta = \infty$ or $\Sigma$ for discrete

d) Jeffrey's Prior: $f(\theta) \propto \sqrt{J(\theta)}$ → take smallest proportion

e) Credible Interval: $\int_{t_\ell}^{t_u} f(\theta|y_{1:n})\,d\theta = \gamma$

↳ R.V. $\theta|x$ contained with prob. $\gamma$

↳ just take $\frac{1\pm\gamma}{2}$ quantiles of posterior

↳ HPD interval for equi-tailed = Wald Type Interval

f) Bayesian Testing: $\frac{P(H_0|y_{1:n})}{P(H_1|y_{1:n})} = \underbrace{\frac{P(y_{1:n}|H_0)}{P(y_{1:n}|H_1)}}_{\text{Bayes Factor}} \cdot \frac{P(H_0)}{P(H_1)}$

g) Conjugates:

| Likelihood | | Conjugate prior/post |
|---|---|---|
| Bin, Geom | → | $\pi \sim Be(\kappa, \beta)$ |
| Poi & Exp | → | $\lambda \sim G(\kappa, \beta)$ |
| $N$ with $\sigma^2$ known | → | $\mu \sim W(\mu, \tau^2)$ |
| $N$ with $\mu$ known | → | $\mu \sim IG(\alpha, \beta)$ |

h) Normal, Normal model ($y|\mu \sim N$; $\mu \sim N$)

$$f(\mu|y) \propto \exp\left(-\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)\left(\mu - \frac{\left(\frac{n\bar{y}}{\sigma^2} + \frac{\eta}{\tau^2}\right)}{\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)}\right)^2\right)$$

→ precision of data + precision of prior

$(\cdot)^{-1}$ post. var

posterior mean