small aside:

Definition (Absolute Continuity) On $(\mathbb{X}, \mathcal{M})$, a finitely additive set function ϕ is *absolutely continuous* with respect to a measure μ if $\phi(A) = 0$ for each $A \in \mathcal{M}$ with $\mu(A) = 0$. We also say ϕ is *dominated* by μ and write $\phi \ll \mu$. If ν and μ are measures such that $\nu \ll \mu$ and $\mu \ll nu$ then μ and ν are *equivalent*.

ν

Ex: a discrete distribution is dominated by the counting measure
a continuous distribution is dominated by the Lebesgue measure.

Theorem (Lebesgue-Randon-Nikodym) Assume that ϕ is a σ-finite countably additive set function and μ is a σ-finite measure. There exist unique σ-finite countably additive set functions ϕ_s and ϕ_{ac} such that $\phi = \phi_{ac} + \phi_s$, $\phi_{ac} \ll \mu$, ϕ_s and μ are mutually singular and there exists a measurable extended real valued function f such that

$\exists$ a set B s.t. $\phi_s(B) = 0$ and $\mu(B^c) = 0$.

$$\phi_{ac}(A) = \int_A f d\mu, \qquad \text{for all } A \in \mathcal{M}.$$

If g is another such function, then $f = g$ a.e. wrt μ. If $\phi \ll \mu$ then $\phi(A) = \int_A f d\mu$ for all $A \in \mathcal{M}$.

Think about a measure ϕ w/ positive value at 0 and also continuous ≥ 0.

let μ = Lebesgue
ν = counting measure over $\{0\}$.

$\Rightarrow \mu(\{0\}) = 0$ but $\nu(\{0\}^c) = 0$ $\Rightarrow \phi_s = \nu$ and ϕ_{ac} is the rest.

Definition (Radon-Nikodym Derivative) $\phi = \phi_{ac} + \phi_s$ is called the *Lebesgue decomposition*. If $\phi \ll \mu$, then the density function f is called the *Radon-Nikodym derivative* of ϕ wrt μ.

"density"

So what?

let μ = Lebesgue measure
ν = counting measure over $\{0\}$.

$$\Rightarrow P(Y \in A \mid \theta) = \int_A f(y;\theta)\, d\mu(y) + \overbrace{\int_A p(y;\theta)\, d\nu(y)}^{= \text{ sum over elements in } A \cap \{0\}.}$$

let $\lambda = \mu + \nu$ and $f_*(y;\underline{\theta}) = \mathbb{I}(y \notin \{0\}) f(y;\theta) + \mathbb{I}(y \in \{0\}) p(y;\theta)$.

$$\Rightarrow P(Y \in A \mid \theta) = \int_A f_*(y;\theta)\, d\lambda(y).$$

R-N derivative of prob measure on $\mathcal{X}$ $\Rightarrow$ valid density

Let $L_*(\underline{\theta} \mid \underline{Y}) \propto f_*(\underline{Y} \mid \underline{\theta})$

holding Y fixed, function of $\underline{\theta}$

Scaling: We can scale the continuous & discrete parts of the likelihood and still have a valid likelihood.

Let's use dominating measure $\lambda_{**} = \alpha\mu + \beta\nu$ for $\alpha, \beta > 0$.

Then the corresponding RN derivative $f_{**}(y;\theta) = \frac{I(y \notin \{0\})}{\alpha} f(y|\theta) + \frac{I(y \in \{0\})}{\beta} p(y|\theta)$

$$\Rightarrow P(Y \in A|\theta) = \int_A f_{**}(y;\theta)\, d\lambda_{**}(y).$$

and a valid likelihood would be $L_{**}(\underline{\theta}|\underline{y}) \propto f_{**}(\underline{y}|\underline{\theta})$.

$\Rightarrow$ we can scale the continuous and discrete parts of the likelihood however we like and its still valid.

Implications: we can scale discrete & cts parts however we'd like
What to do? Doesn't matter (mostly):

Let's say have a sample w/ n_0 $Y_i = 0$
$m = n - n_0$ $Y_i > 0$
iid.

$$\begin{aligned}
L_{**}(\underline{\theta}|\underline{y}) &= \prod_{i=1}^n f_{**}(Y_i;\underline{\theta}) \\
&= \prod_{Y_i=0} \frac{1}{\beta} p(Y_i;\underline{\theta}) \prod_{Y_i>0} \frac{1}{\alpha} f(y_i|\theta). \\
&= \frac{1}{\beta^{n_0}\alpha^m} \underbrace{\prod_{y_i=0} p(Y_i;\underline{\theta}) \prod_{y_i>0} f(y_i|\underline{\theta})} . \\
&= \frac{1}{\beta^{n_0}\alpha^m} \prod_{i=1}^n f_*(Y_i;\underline{\theta}). \\
&\propto \prod_{i=1}^n f_*(Y_i;\underline{\theta}) = L_*(\underline{\theta}|\underline{y}).
\end{aligned}$$

$\Rightarrow$ scaling can be ignored in MLE applications (more next).

1.2.5 Proportional Likelihoods

Likelihoods are equivalent for point estimation as long as they are proportional and the constant of proportionality does not depend on unknown parameters.

Why?

transformed

Consider if $Y_i, i = 1, \ldots, n$ are iid continuous with density $f_Y(y; \boldsymbol{\theta})$ and $X_i = g(Y_i)$ where g is increasing and continuously differentiable. Because g is one-to-one, we can construct Y_i from X_i and vice versa.

known

$Y_i = g^{-1}(X_i)$

$\Rightarrow \{Y_1, \ldots, Y_n\}$ and $\{X_1, \ldots, X_n\}$ are "equivalent" because they contain the exact same information

(intuition) $\Rightarrow$ likelihood-based inference based on $\{Y_1, \ldots, Y_n\}$ should be identical to inference based on $\{X_1, \ldots, X_n\}$.

More formally, the density of X_i is $f_X(x; \boldsymbol{\theta}) = f_Y(h(x); \boldsymbol{\theta})h'(x)$, where $h = g^{-1}$, and

$$
\begin{aligned}
L(\boldsymbol{\theta}|\boldsymbol{X}) &= \prod_{i=1}^{n} f_Y(h(X_i); \theta)\, h'(X_i) \\
&= \prod_{i=1}^{n} f_Y(Y_i; \theta)\, h'(g(Y_i)) \\
&= \prod_{i=1}^{n} f_Y(Y_i; \theta) \cdot \frac{1}{g'(Y_i)} \\
&= L(\underline{\theta}|\underline{Y}) \underbrace{\left\{ \prod_{i=1}^{n} \frac{1}{g'(Y_i)} \right\}}_{\text{doesn't depend on } \underline{\theta}}
\end{aligned}
$$

aside:

$$h'(x) = \frac{dh(x)}{dx} = \frac{dg^{-1}(x)}{dx} = \frac{1}{g'(g^{-1}(x))}$$

because:

$x = f(f^{-1}(x))$ deriv of both sides, solve.

$\Rightarrow$ MLE are identical whether derived from $L(\underline{\theta}|\underline{Y})$ or $L(\underline{\theta}|\underline{X})$.

Example (Likelihood Principle): Consider data from two different sampling plans:

1. A binomial experiment with $n = 12$. Let $Y_i = 1$ if i^{th} trial is a success and 0 otherwise.

$$L_1(p|\boldsymbol{Y}) = \binom{12}{S} p^S (1-p)^{12-S}, \text{ where } S = \sum_{i=1}^{n} Y_i$$

2. A negative binomial experiment, i.e. run the experiment until three zeroes are obtained.

$$L_2(p|\boldsymbol{Y}) = \binom{S+2}{S} p^S (1-p)^3.$$

The ratio of these likelihoods is

$$\frac{L_1(p|\boldsymbol{Y})}{L_2(p|\boldsymbol{Y})} = \frac{\binom{12}{S} p^S (1-p)^{12-S}}{\binom{S+2}{S} p^S (1-p)^3} = \frac{\binom{12}{S}}{\binom{S+2}{S}} (1-p)^{9-S}$$

Suppose $S = 9$. Is all inference equivalent for these likelihoods? Debatable.

Then $\frac{L_1(p|\underline{y})}{L_2(p|\underline{y})} = \frac{\binom{12}{S}}{\binom{S+2}{S}}$ doesn't depend on p!

Argue

YES both cases $\hat{p}_{MLE} = \frac{9}{12}$

NO Consider the hypothesis test $H_0: p = \frac{1}{2}$ vs. $H_a: p > \frac{1}{2}$

If in experiment ①, sum(dbinom(c(9,10,11,12), size=12, p=$\frac{1}{2}$)) = .0730

② 1 − sum(dnbinom(seq(0,8), size=3, prob=$\frac{1}{2}$)) = .0327 } p-values.

Bayesians: Discrepancy in p-values implies frequentist methods are not logical because inference is not based solely on likelihoods. Maintain proportional likelihoods contain information.

formalized in Berger and Wolpert (1984).

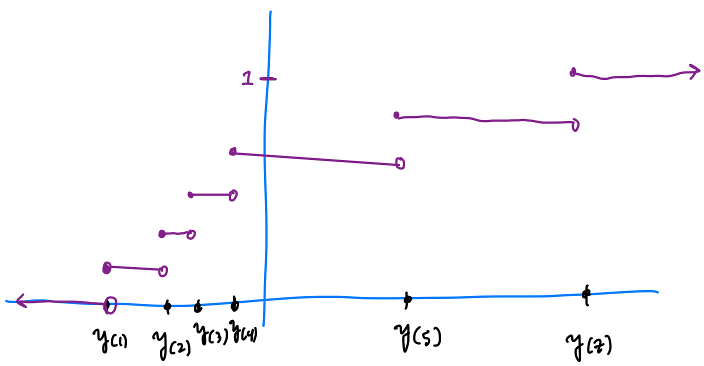
The likelihood principle states all the information about $\boldsymbol{\theta}$ from an experiment is contained in the actual observation $\boldsymbol{y}$. Two likelihood functions for $\boldsymbol{\theta}$ (from the same or different experiments) contain the same information about $\boldsymbol{\theta}$ if they are proportional.

1.2.6 Empirical Distribution Function as MLE

Recall the empirical cdf:

Suppose $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$ are the order statistics of an iid sample from an unknown distribution function F_Y. Our goal is to estimate F_Y.

$$\hat{F}_Y(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(y \geq y_{(i)})$$



Is this a “good” estimator of F_Y?

Maybe not... If you belief F_Y has support on $\mathbb{R}$, having $\hat{F}_Y(y) = 0$ for $y < y_{(1)}$ & $\hat{F}_Y(y) = 1$ for $y \geq y_{(n)}$ could be a problem.

OTOH, $F_Y(y)$ is likely to be pretty close to 0 for $y < y_{(1)}$ and close to 1 for $y \geq y_{(n)}$

hopefully.

otherwise talk to Dan or Ben about "extremes"

Another answer:

Yes, because it's MLE.

Suppose $Y_1, \ldots, Y_n$ are iid with distribution function $F(y)$. Here $F(y)$ is the unknown parameter.

1. $F(y)$ is nonnegative nondecreasing
2. $F(y)$ is right continuous
3. $\lim_{y \to -\infty} F(y) = 0$ and $\lim_{y \to \infty} F(y) = 1$.

$\Rightarrow$ parameter space is set of all distribution functions.

An approximate likelihood for F is

(ignoring $(2h)^{-n}$ factor)

small pos. constant.

$$L_h(F|\mathbf{Y}) = \prod_{i=1}^{n} \{F(Y_i + h) - F(Y_i - h)\}$$

- Assume no ties so that h small enough s.t. $[Y_i - h, Y_i + h]$ does not contain Y_j for $j \neq i$.

can prove this when we have ties in general, similar argument w/ slight change.

Let $\overset{\text{size of jump}}{p_{i,h}} = F(Y_i + h) - F(Y_i - h) \Rightarrow L_h(F|Y) = \prod_{i=1}^{n} p_{i,h}$

Note: this is maximized only if $p_{i,h} > 0$ $i = 1, \ldots, n$.

Since increasing $p_{i,h}$ increases $L_h(F|Y)$ want $p_{i,h} > 0$ to be as large as possible w/ $\overbrace{\sum_{i=1}^{n} p_{i,h} \overset{?}{\leq} 1}^{\text{satisfy 1,3 above.}}$

happens when $= 1$!

$\Rightarrow$ goal: maximize $\prod_{i=1}^{n} p_{i,h}$ subject to $p_{i,h} > 0$ and $\sum_{i=1}^{n} p_{i,h} = 1$.

optimization problem to be solved by Lagrange multipliers, find stationary points of:

$$g(p_{1,h}, \ldots, p_{n,h}, \lambda) = \underbrace{\sum_{i=1}^{n} \log(p_{i,h})}_{\text{likelihood}} + \lambda \underbrace{\left(\sum_{i=1}^{n} p_{i,h} - 1\right)}_{\text{constraint}}$$

get by solving

$$\left.\begin{aligned} \frac{\partial g}{\partial p_{i,h}} &= \frac{1}{p_{i,h}} + \lambda = 0 \quad i = 1, \ldots, n \\ \frac{\partial g}{\partial \lambda} &= \sum_{i=1}^{n} p_{i,h} - 1 = 0 \end{aligned}\right\} \text{ implies } p_{i,h} = -\frac{1}{\lambda} = \frac{1}{n}$$

plug in the second eq.

$\Rightarrow \lambda = -n$

$\Rightarrow$ any function $\hat{F}_h(y)$ which satisfies $\hat{F}_h(Y_i + h) - \hat{F}_h(Y_i - h) = \frac{1}{n}$ $i = 1, \ldots, n$ maximize $L_h(F|Y)$.