Likelihoods

History of the course:

Taught since 2017 or 2018 by

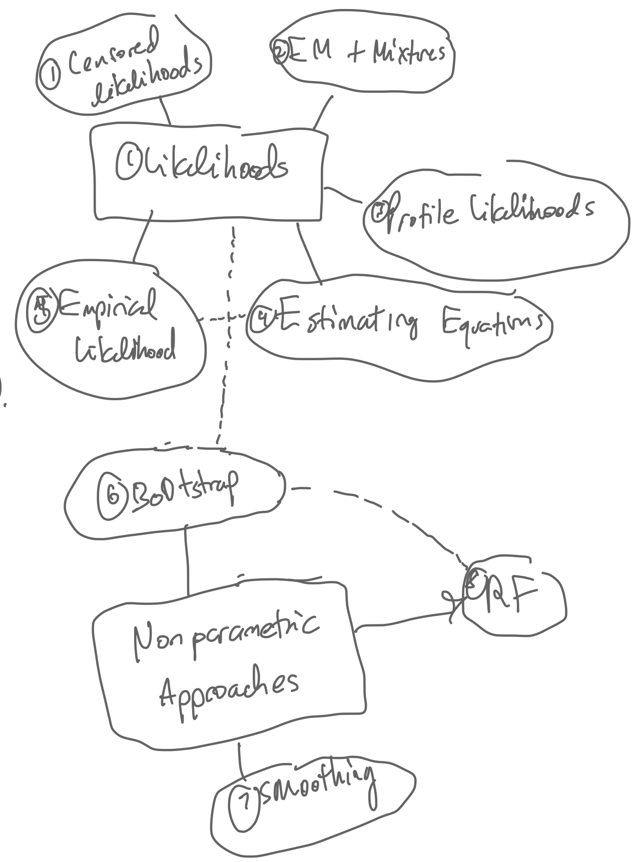
Haonan } Thanks for notes/help preparing class.
Dan }

Now: Me

Idea: Many important topics not taught in other courses but should be "core" to a PhD in statistics.

Not linear!

0.1 Outline



1. Likelihoods
2. EM, k-means
3. Profile likelihood
4. Estimating Equations (M estimation).
5. Empirical likelihoods.

midterm →

6. Bootstrap
7. Smoothing methods
8. Random Forests (?).

1 Likelihood Construction and Estimation

Likelihood-based Methods

MLE
LRT
Likelihood-based uncertainty (CI's).

Why do Statisticians love likelihood-based estimation?

1. Invariance property of MLE: If a distribution is parametrized by θ but interest is in a function of θ, $\tau(\theta)$. If $\hat{\theta}$ is MLE of θ, $\tau(\hat{\theta})$ is MLE of $\tau(\theta)$.

2. Asymptotically unbiased; constistent

   $$\lim_{n \to \infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0$$

3. Asymptotically efficient; variance achieve Cramer-Rao Lower Bound.

   estimator has <u>all</u> the information.

4. Relationship w/ Fisher Information matrix allows for construction of CI's.

   (based on asymptotic properties).

Downsides?

1. Very model based! You are assuming know entire distribution

2. It often require numerical optimization.

Still... we ♡ it!

1.1 Introduction

→ doesn't say iid.

Definition: Suppose random variables $\boldsymbol{Y} = (Y_1, \dots Y_n)^\top$ has joint density or probability mass function $f_{\boldsymbol{Y}}(\boldsymbol{y}, \boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_b)$. Then the *likelihood function* is

$$L(\boldsymbol{\theta}|\boldsymbol{Y}) = f_{\boldsymbol{Y}}(\boldsymbol{Y}, \boldsymbol{\theta}).$$

← in general, likelihood = joint

likelihood is random.

(because it depends on the data!).

Know MLE is random & we quantify its uncertainty.

Given a vector of observations y, the likelihood is a function of θ:
For any (valid) value of θ, it returns a number (the likelihood).

$\hat{\theta}_{MLE}$ obtained by finding value of θ which yields max likelihood value.

Key concept: In all situations, the likelihood is the joint density of the observed data to be analyzed.

Comments:

1. "density" can mean continuous density or pmf.

2. "observed data" will be generalized. E.g, censored data.

1.1.1 Notation

Given $\boldsymbol{y}$, note that $L(\boldsymbol{\theta}|\boldsymbol{y}) : \mathbb{R}^b \to \mathbb{R}$.

$$\underline{\theta} = (\theta_1, \ldots, \theta_b)^T$$

Likelihood $L(\underline{\theta}|\underline{y})$ is scalar valued!

Generally, we optimize $\ell(\boldsymbol{\theta}) = \underbrace{\log}_{\text{monotone increasing}} L(\boldsymbol{\theta}|\boldsymbol{y})$.

$$\arg\max_{\underline{\theta}} L(\underline{\theta}|\underline{y}) = \arg\max_{\underline{\theta}} \ell(\underline{\theta}).$$

How? Take derivatives, set to zero, solve.

Generally convention is the derivative of a function (i.e. $\ell(\underline{\theta})$) wrt a vector ($\underline{\theta}$), is a row vector $\ell'(\underline{\theta}) = \frac{\partial \ell(\underline{\theta})}{\partial \underline{\theta}} = \left(\frac{\partial \ell(\underline{\theta})}{\partial \theta_1}, \ldots, \frac{\partial \ell(\underline{\theta})}{\partial \theta_b} \right)$.

Define <u>score function</u>

$$S(\underline{\theta}) = \ell'(\underline{\theta})^T = \begin{pmatrix} \frac{\partial \ell(\underline{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell(\underline{\theta})}{\partial \theta_b} \end{pmatrix}$$

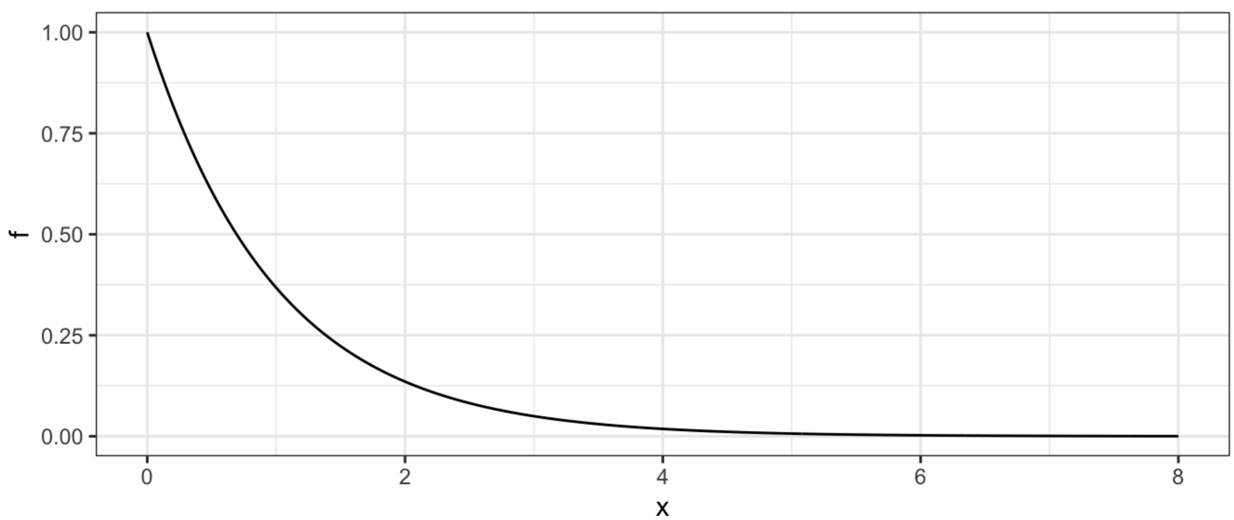
$b \times 1$
column vector.

Example: Suppose we have $Y_1, \ldots Y_n \overset{iid}{\sim} \text{Exp}(\lambda)$. The likelihood function is defined as

$$L(\lambda|\underline{y}) = f_{\underline{y}}(\underline{y}, \lambda).$$
$$= \prod_{i=1}^{n} f_y(y_i;\lambda)$$
$$= \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} y_i} \Rightarrow \ell(\lambda) = n\log\lambda - \lambda\sum_{i=1}^{n} y_i$$

```
# likelihood simulation
n <- 10
lambda <- 1

# plot of exponential(lambda) density
data.frame(x = seq(0, 8, .01)) |>
  mutate(f = dexp(x, rate = lambda)) |>
  ggplot() +
  geom_line(aes(x, f))
```



```
# define likelihood
loglik <- function(lambda, data)
{
  lik <- prod(dexp(data, rate = lambda))
  loglik <- sum(dexp(data, rate = lambda, log = T))

    out <- data.frame(lik = lik, loglik = loglik)
    return(out)
}

# simulate data
data <- rexp(n = n, rate = lambda)

# plot likelihood and loglikelihood
data.frame(lambda = seq(0, 3, by = .01)) |>
  rowwise() |>
  mutate(loglik = loglik(lambda, data)) |>
  unnest(cols = c(loglik)) |>
  pivot_longer(-lambda, names_to = "func", values_to = "vals") |>
  ggplot() +
  geom_vline(aes(xintercept = 1 / mean(data)), lty = 2) + # max
        likelihood estimate is 1/mean
  geom_line(aes(lambda, vals)) +
  facet_wrap(~func, scales = "free")
```

L

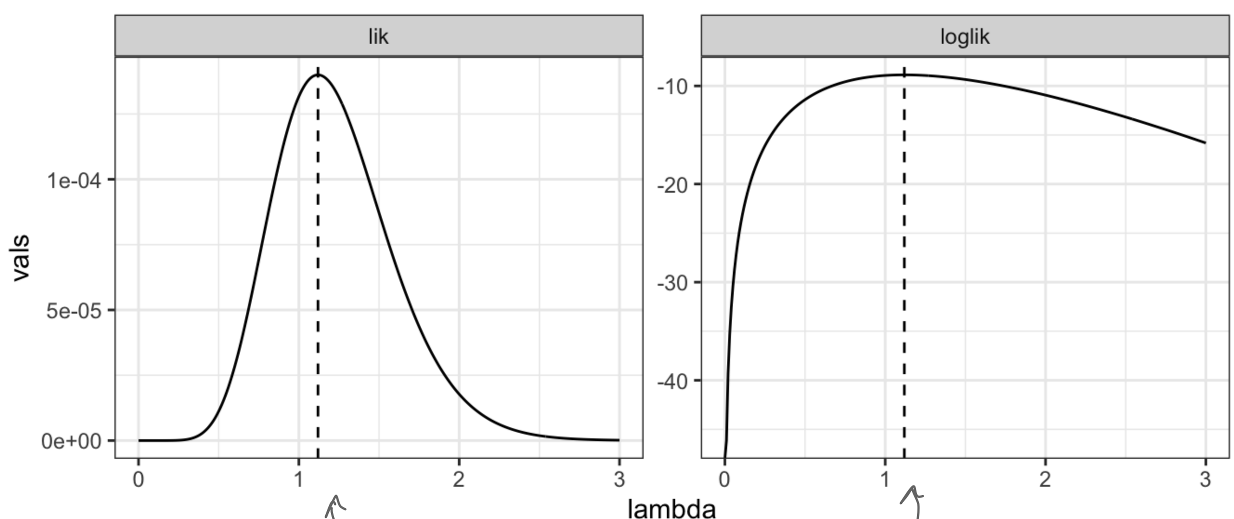
ℓ

$\log(f_Y(y;\lambda))$.

now we have realized data!

MLE

MLE



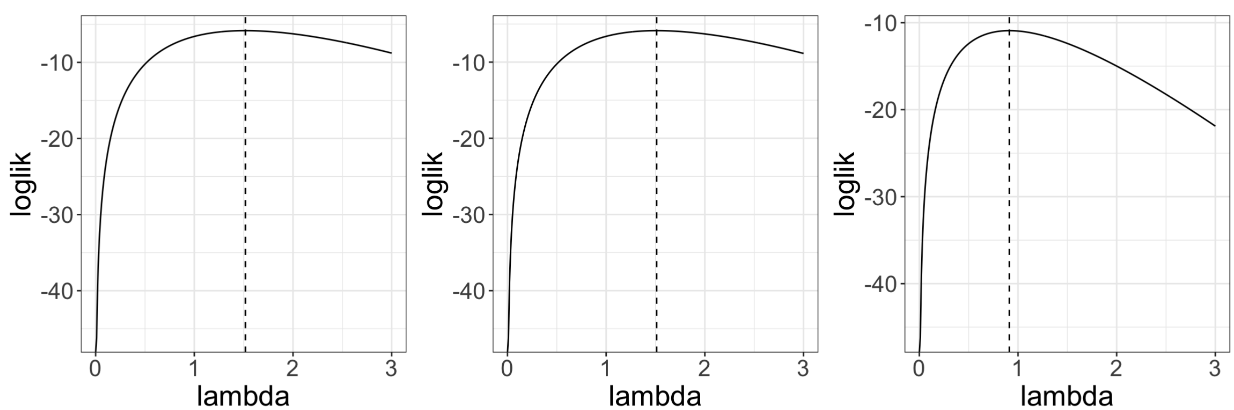
max happens at same place!

The likelihood function is random!

```
for(i in seq_len(3)) {
  # simulate data
  data <- rexp(n = n, rate = lambda)

  # plot likelihood and loglikelihood
  data.frame(lambda = seq(0, 3, by = .01)) |>
    rowwise() |>
    mutate(loglik = loglik(lambda, data)) |>
    unnest(cols = c(loglik)) |>
    ggplot() +
    geom_vline(aes(xintercept = 1 / mean(data)), lty = 2) + # max
        likelihood estimate is 1/mean
    geom_line(aes(lambda, loglik)) +
    theme(text = element_text(size = 20)) -> p ## make legible in
        notes

    print(p)
}
```



Question: What does likelihood integrate to?

Answer: not 1! function of θ, not Y.

Your Turn: What is the effect of sample size on the log-likelihood function? Make a plot showing the log-likelihood function that results from $n = 10$ vs. $n = 100$ with corresponding MLE.

1.2 Construction

The use of the likelihood function in parameter estimation is easiest to understand in the case of discrete iid random variables.

1.2.1 Discrete IID Random Variables

Suppose each of the n random variables in the random sample $Y_1, \ldots, Y_n$ have probability mass function $f(y; \boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(Y_1 = y), y = y_1, y_2, \ldots$. The likelihood is then defined as:

$$L(\boldsymbol{\theta}|\boldsymbol{Y}) = \text{ joint density of observed random variables}$$

$$\overset{iid}{=} \text{product of univariate "densities"}$$

$$= \prod_{i=1}^{n} \underbrace{f(Y_i ; \underline{\theta})}_{pmf}$$

$$= \prod_{i=1}^{n} P_{\theta}(Y_i^* = Y_i | Y_i)$$

where $Y_1^*, \ldots, Y_n^*$ are iid RV's w/ same distribution as $Y_1, \ldots, Y_n$ (but mutually independent of $Y_1, \ldots, Y_n$)

In other words,

the likelihood is the probability of getting sample actually obtained for a given value of $\underline{\theta}$.

(1) In discrete case, can be thought of as a probability. (over what domain?)

(2) Will the likelihood sum to 1 over the parameter space? No.

(3) Probability of finding a particular realization for a given $\underline{\theta}$.

Example (Fetal Lamb Movements): Data on counts of movements in five-second intervals of one fetal lamb ($n = 240$ intervals:)

No. of Movements	0	1	2	3	4	5	6	7
Count	182	41	12	2	2	0	0	1

$= 240.$

Assume a iid Poisson model: $P(Y = y) = f_Y(y; \lambda) = \frac{\exp(-\lambda)\lambda^y}{y!}$. Then the likelihood is

$$L(\lambda \mid \underline{y}) = \prod_{i=1}^{n} f_y(y_i; \lambda) = \prod_{i=1}^{n} \lambda^{y_i} \frac{e^{-\lambda}}{y_i!} = \lambda^{\sum_{i=1}^n y_i} e^{-n\lambda} \left(\prod_{i=1}^{n} y_i! \right)^{-1}$$

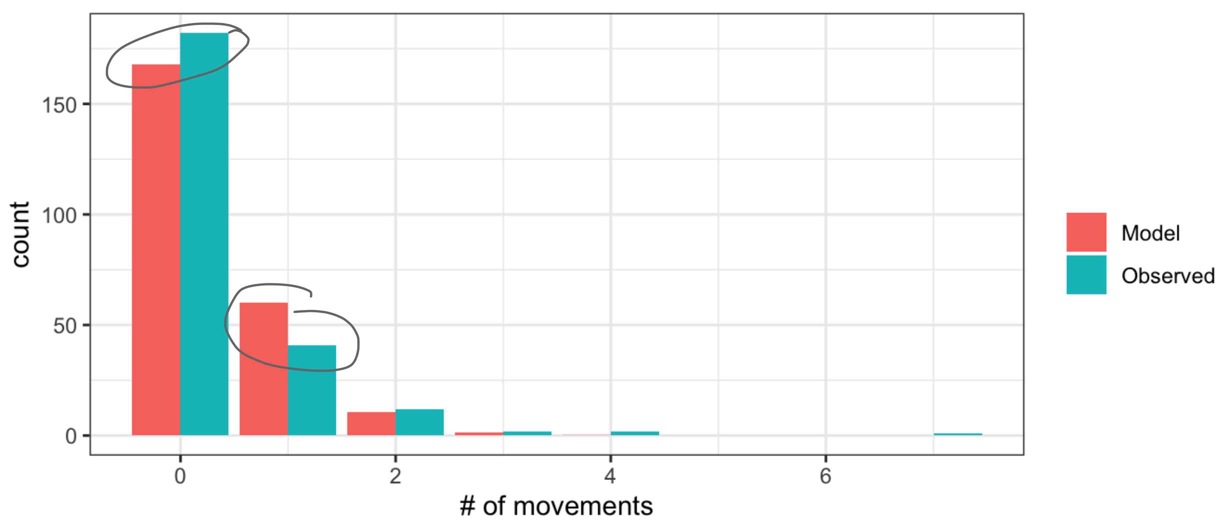
$$\ell(\lambda) = \sum_{i=1}^{n} y_i \log \lambda - n\lambda - \log\left(\prod y_i!\right).$$

$$\ell'(\lambda) = \frac{\Sigma y_i}{\lambda} - n \overset{\text{set}}{=} 0.$$

Equating the derivative of the loglikelihood with respect to λ to zero and solving results in the MLE

$$\hat{\lambda}_{\text{MLE}} = \frac{\Sigma y_i}{n} = \frac{86}{240} = 0.358.$$

This is the best we can do with this model. But is it good?



χ^2 GOF test returns p-value of 0.00025. NOT good.

Illustration of a disadvantage of likelihood-based methods: very model based!

pg 31-32 of ESI extends to Zero-Inflated Poisson.

1.2.2 $\underbrace{\text{Multinomial Likelihoods}}_{\text{more interesting discrete likelihoods.}}$

The multinomial distribution is a generalization of the binomial distribution where instead of 2 outcomes (success or failure), there are now $k \geq 2$ outcomes.

Consider independently tossing n balls into k urns, where p_i is prob of the ball landing in i^{th} urn on each toss $i = 1, \ldots, k$.

$\Rightarrow N_i$ balls in i^{th} urn and $\sum_{i=1}^{k} N_i = n$.

The probability mass function is (n total trials, k categories).

$$P(N_1 = n_1, \ldots, N_k = n_k) = p(n_1, \ldots, n_k; p_1, \ldots, p_k)$$

$$= \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} \quad \text{where} \quad 0 \le p_i \le 1 \text{ and } \underbrace{\sum_{i=1}^{n} p_i = 1}_{\text{note } p_k = 1 - \sum_{i=1}^{k-1} p_i}$$

For $\underbrace{N_1, \ldots, N_k}_{\text{(to be) observed data}}$, $N_i =$ the number of balls in i^{th} urn, $\underbrace{\sum_{i=1}^{k} N_i = n}$ (total balls thrown).

$\Rightarrow N_i$'s not independent!

$$L(p \mid N_1, \ldots, N_k) = \frac{n!}{N_1! \cdots N_k!} p_1^{N_1} \cdots p_k^{N_k}$$

$$= \frac{n!}{N_1! \cdots N_k!} p_1^{N_1} \cdots p_{k-1}^{N_{k-1}} \left(1 - \sum_{i=1}^{k-1} p_i\right)^{N_k}.$$

only have $k-1$ params to estimate.

$$\Rightarrow \ell(p) = \text{const} + N_1 \log p_1 + N_2 \log p_2 + \cdots + N_k \log\left(1 - \sum_{i=1}^{k-1} p_i\right)$$

The maximum likelihood estimator of p_i:

$$\frac{\partial \ell(p)}{\partial p_j} = \frac{N_j}{p_j} - \frac{N_k}{\underbrace{1 - \sum_{i=1}^{k-1} p_i}_{p_k}} \overset{\text{set}}{=} 0$$

$$\Rightarrow \begin{pmatrix} N_k & 0 & \cdots & 0 & -N_1 \\ 0 & N_k & 0 \cdots & 0 & -N_2 \\ 0 & 0 & N_k \cdots & 0 & -N_3 \\ \vdots & & & & \vdots \\ 0 & 0 & \cdots & N_k & N_{k-1} \\ 1 & 1 & \cdots & 1 & 1 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ \vdots \\ p_k \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

$$\Rightarrow N_k p_j - N_j p_k = 0$$

$$\Rightarrow \hat{p}_{j\,MLE} = \frac{N_j}{n} \quad \text{(what you would think)}$$

More interesting multinomial likelihoods arise when the p_i are modeled as a function of a lesser number of parameters $\theta_1, \ldots, \theta_m$, $m < k - 1$.

Example (Capture-Recapture): To estimate fish survival during a specific length of time (e.g., one month), a common approach is to use a removal design.

S = prob of fish surviving one month.

Time 0: catch and tag n fish. fixed.

Time 1: catch and remove some # of tagged fish.

p = prob tagged fish is caught and removed.

N_1 = # tagged fish removed at time 1.

prob a tagged fish is caught at time 1 = $\underset{\text{prob. survival}}{S} \cdot \underset{\text{prob caught}}{p} = p_1$

Time 2: repeat

N_2 = # tagged fish removed at time 2

prob a tagged fish is caught at time 2 = $s^2(1-p)p = p_2$

⋮

Time $k-1$: repeat

N_{k-1} = # tagged fish removed at time $k-1$

prob a tagged fish is caught at time $k-1$ $s^{k-1}(1-p)^{k-2}p = p_{k-1}$

k^{th} : tagged fish is not removed

$$N_k = n - \sum_{i=1}^{k-1} N_i$$

$$p_k = 1 - sp - s^2(1-p)p - s^3(1-p)^2p - \dots - s^{k-1}(1-p)^{k-2}p = 1 - \sum_{i=1}^{k-1} p_i$$

Goal: estimate p and s

Say you catch and remove $N_1, \dots, N_k$ fish @ k times, $\sum_{i=1}^{k} N_i = n$

The likelihood is the probability of catching $N_1, \dots, N_k$ w/ n total tagged. known.

$\Rightarrow$ multinomial w/ $p_i = s^i(1-p)^{i-1}p$ $i = 1, \dots, k-1$ and $p_k = 1 - \sum_{i=1}^{k-1} p_i$.

Recall: $L(p \mid N_1, \dots, N_k) = \dfrac{n!}{N_1! \cdots N_k!} p_1^{N_1} \cdots p_k^{N_k}$

(substitute p, s) $= \dfrac{n!}{N_1! \cdots N_k!} (sp)^{N_1} (s^2(1-p)p)^{N_2} \cdots (s^{k-1}(1-p)^{k-2}p)^{N_{k-1}} p_k^{N_k}$

where $p_k = 1 - sp - s^2(1-p)p - \dots - s^{k-1}(1-p)^{k-1}p$.

↑ complicated, not log friendly.

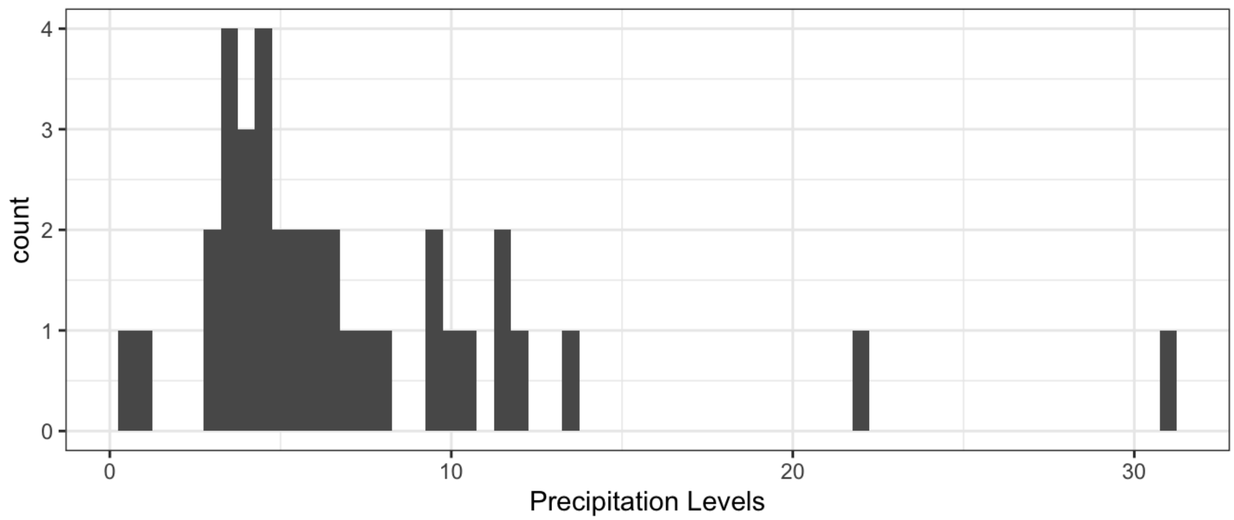
What now? take log + partial derivatives wrt s & p, solve?

(use a computer)

1.2.3 Continuous IID Random Variables

Recall: the likelihood is the joint density of data to be analyzed.

Example (Hurricane Data): For 36 hurricanes that had moved far inland on the East Coast of the US in 1900-1969, maximum 24-hour precipitation levels during the time they were over mountains.



We model the precipitation levels with a gamma distribution, which has density

$$f(y; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} \exp(-y/\beta), \quad y > 0, \alpha, \beta > 0.$$

This leads to the likelihood

$$L(\underline{\theta} \mid \underline{y}) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y_i^{\alpha-1} e^{-y_i/\beta} = \{\Gamma(\alpha)\}^{-n} \beta^{-n\alpha} \{\prod_{i=1}^{n} y_i\}^{\alpha-1} e^{-\sum_{i=1}^{n} y_i/\beta}$$

Of course, this cannot be interpreted as a probability because

$$P(Y = y) = 0 \quad \forall y!$$

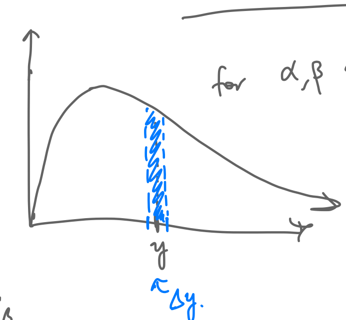
To get a probability, need to go from a density to a measure. i.e. integrate!

But likelihood not necessarily integrate to 1 (won't return value in $[0,1]$ necessarily).

But it may be useful to think of the value of the likelihood as being proportional to a probability.

Given data $y_1, \dots, y_n$

$$L(\alpha, \beta \mid y) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y_i^{\alpha-1} e^{-y_i/\beta} \longrightarrow$$



for α, β given.

prob is approximately $\prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y_i^{\alpha-1} e^{-y_i/\beta} \cdot \Delta y_i$

More formally, begin with the definition of a derivative

$$g'(x) = \lim_{h \to 0^+} \frac{g(x+h) - g(x-h)}{2h}.$$

Let F be the cumulative distribution function of a continuous random variable Y, then (if the derivative exists)

$$\underline{f(y)} = \lim_{h \to 0^+} \frac{F(y+h) - F(y-h)}{2h} = \lim_{h \to 0^+} \frac{P(Y \in (y-h, y+h])}{2h}.$$

If we substitute this definition of a density into the definition of the likelihood

$$\begin{aligned}
L(\underline{\theta} \mid Y) &= \prod_{i=1}^{n} f(Y_i; \underline{\theta}) \\
&= \prod_{i=1}^{n} \lim_{h \to 0^+} \frac{F_\theta(Y_i + h) - F_\theta(Y_i - h)}{2h} \\
&= \lim_{h \to 0^+} \prod_{i=1}^{n} \frac{1}{2h} \left(F_\theta(Y_i + h) - F_\theta(Y_i - h) \right) \\
&= \lim_{h \to 0^+} \underbrace{\left(\frac{1}{2h} \right)^n}_{\text{proportionality "constant" because it doesn't depend on } \theta} \underbrace{\prod_{i=1}^{n} P_\theta(Y_i^* \in (Y_i - h, Y_i + h])}_{Y_i^* \text{ indep of } Y_i \text{ w/ same dsn as } Y_i}.
\end{aligned}$$

For small h, likelihood is proportional to a probability!

as $h \to 0^+$ prob $\downarrow 0$ and $\left(\frac{1}{2h}\right)^n \uparrow$ to balance.

$\Rightarrow$ likelihood is proportional to the probability of obtaining a new sample that is close to the sample we obtained.

Compare this to the iid discrete case:

$$L(\underline{\theta} \mid \underline{y}) = \prod_{i=1}^{n} f(y_i; \theta)$$

$$= \lim_{h \to 0^+} \prod_{i=1}^{n} \{ F(y_i + h; \theta) - F(y_i - h; \theta) \}$$

↖ now we don't need the proportionality constant.

So likelihoods for discrete RVs get weighted differently than likelihoods for continuous RVs.

This has to do w/ underlying dominating measure of these RVs

(discrete: counting measure
continuous: Lebesgue measure)

Thought for later: what about mixtures?

Example (Hurricane Data, Cont'd): Recall with a gamma model, the likelihood for this example is

$$L(\boldsymbol{\theta}|\boldsymbol{Y}) = \{\Gamma(\alpha)\}^{-n}\beta^{-n\alpha}\left\{\prod Y_i\right\}^{\alpha-1}\exp\left(-\sum y_i/\beta\right),$$

take derivs, set = 0
solve?

and log-likelihood

$$\ell(\boldsymbol{\theta}) = -n\log\Gamma(\alpha) - n\alpha\log\beta + (\alpha-1)\sum\log Y_i - \frac{\sum Y_i}{\beta}$$

$\frac{\partial \log\Gamma(\alpha)}{\partial\alpha} = -\gamma + \sum_{k=1}^{\infty}\left(\frac{1}{k} - \frac{1}{k+\alpha-1}\right)$

Euler-Mascheroni constant.

```
## loglikelihood function
neg_gamma_loglik <- function(theta, data) {
  -sum(log(dgamma(data, theta[1], scale = theta[2])))
}
```

so we can use nlm (minimizes).

partial derivs of Gamma likelihood do not result in linear system of eq'ns
⇒ use numerical optimization (common!).

optim

```
## maximize
mle <- nlm(neg_gamma_loglik, c(1.59, 4.458), data = hurr_rain)
mle$estimate
```

$\hat\alpha_{MLE}$ $\hat\beta_{MLE}$

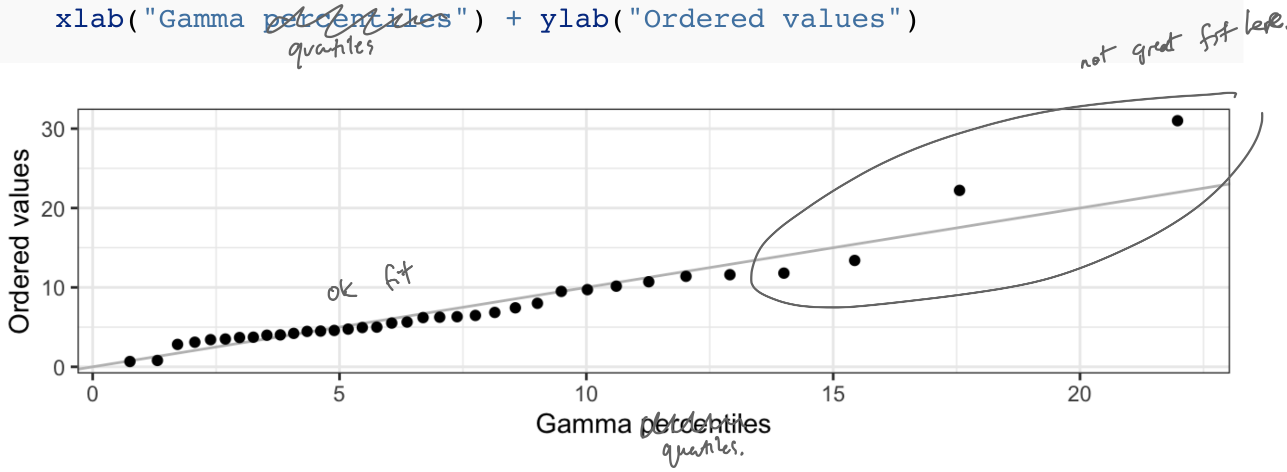
```
## [1] 2.187214 3.331862
```

quantiles based on Gamma fitted model.

```
## Gamma QQ plot
data.frame(theoretical = qgamma(ppoints(hurr_rain), mle$estimate[1],
                 scale = mle$estimate[2]),
           actual = sort(hurr_rain)) |>
  ggplot() +
  geom_abline(aes(intercept = 0, slope = 1), colour = "grey") +
  geom_point(aes(theoretical, actual)) +
  xlab("Gamma percentiles") + ylab("Ordered values")
```

actual data (ordered)

quantiles



1.2.4 Mixtures of Discrete and Continuous RVs

Some data Y often have a number of zeros and the amounts greater than zero are best modeled by a continuous distribution.

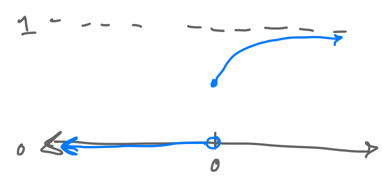
Ex: Rainfall, snowfall, Amt of time I climb in a day.

In other words, they have positive probability of taking a value of exactly zero, but continuous distribution otherwise.

- different from ZIP, which is discrete.
- can be generalized to other cases, point mass doesn't have to be at zero (e.g. record linkage)

A sensible model would assume Y_i are iid with cdf

$$F_Y(y; p, \boldsymbol{\theta}) = \begin{cases} 0 & y < 0 \\ p & y = 0 \\ p + (1-p)F_T(y; \boldsymbol{\theta}) & y > 0 \end{cases}$$



where $0 < p \leq 1$ is $P(Y = 0)$ and $F_T(y; \boldsymbol{\theta})$ is a distribution function for a continuous positive random variable.

Another way to write this:

$$F_y(y; p, \underline{\theta}) = P(Y \le y) = p\,\mathbb{I}(0 \le y) + (1-p)F_T(y; \underline{\theta}).$$

No problems w/ cdf.

How to go from here to get a likelihood?

What is the "density" here?

practical.

One approach: let n_0 be the number of zeroes in the data and $m = n - n_0$ be the number of non-zero Y_i. This leads to an intuitive way to contruct the likelihood for iid $Y_1, \ldots, Y_n$ distributed according to the above distribution: recalling the construction of likelihoods for cts & discrete iid rvs.

$$L(\overset{p,}{\boldsymbol{\theta}}|\boldsymbol{Y}) = \lim_{h \to 0^+} \Big(\frac{1}{2h}\Big)^{\overset{\text{weights for cts part.}}{m}} \prod_{i=1}^{n} \{F_Y(Y_i + h; p, \boldsymbol{\theta}) - F_Y(Y_i - h; p, \boldsymbol{\theta})\}$$

$$= \overbrace{\lim_{h\to 0^+} \{F_Y(h; p, \underline{\theta}) - F_Y(-h; p, \underline{\theta})\}^{n_0}}^{Y_i = 0} \times \lim_{h \to 0^+} \prod_{Y_i > 0} \left\{ \frac{F_Y(Y_i + h; p, \underline{\theta}) - F_Y(Y_i - h; p, \underline{\theta})}{2h} \right\}$$

m Y_i's >0.

$$= \lim_{h\to 0^+} \{p + (1-p) F_T(h; \underline{\theta})\}^{n_0} \times \lim_{h \to 0^+} \prod_{Y_i > 0} \left\{ \frac{p + (1-p) F_T(Y_i + h; \underline{\theta}) - p - (1-p) F_T(Y_i - h; \underline{\theta})}{2h} \right\}$$

$$= \underbrace{p^{n_0 = n-m} \times (1-p)^m}_{\text{Bernoulli component for } n_0 = n-m \text{ zeroes.}} \underbrace{\prod_{Y_i > 0} f_T(Y_i; \underline{\theta})}_{\text{continuous component for } m \text{ nonzero values.}}.$$

$$\ell(p, \underline{\theta}) = (n-m)\log p + m \log(1-p) + \sum_{Y_i > 0} \log f_T(Y_i; \underline{\theta}).$$

Notice: $\hat{p}_{MLE} = \frac{n-m}{n}$ & $\underbrace{\hat{\theta}_{MLE}}_{\text{will only use } Y_i > 0}$ obtained in usual way from m obs. w/ density $f_T(y; \underline{\theta})$

Kind of similar to law of total probability:

$$L(p, \underline{\theta} | \underline{y}) \overset{\text{"proportional"}}{\propto} \prod_{i=1}^{n} P(Y_i = y)$$

$$= \prod_{i=1}^{n} \{ P(Y_i = y_i | Y_i = 0) P(Y_i = 0) + P(Y_i = y_i | Y_i \neq 0) P(Y_i \neq 0) \}.$$

$$= \prod_{i=1}^{n} \{ \delta_{Y_i = 0}(y_i) \cdot p + \delta_{Y_i \neq 0}(y) f_T(y_i | \underline{\theta}) \cdot (1-p) \} = p^{n-m} (1-p)^m \prod_{Y_i > 0} f_T(y_i | \underline{\theta})$$

Feels a little arbitrary in how we are defining different weights on our likelihood for discrete and continuous parts.

Turns out, it doesn't matter! (mostly) (Need some $\underbrace{\text{STAT 630/720}}_{\text{not much!}}$ to see why.)

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~~ ν then μ and ν are *equivalent*.

Ex: continuous dsn is dominated by the Lebesgue measure.

Ex: a discrete dsn is dominated by the counting measure.

Theorem (Lebesgue-Radon-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 $\phi = \phi_{ac} + \phi_s \ll \mu$, 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 $\mu(B^c) = 0$

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

R-N derivative "density"

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 ϕ w/ pos value at 0 and continuous > 0. Let μ = Lebesgue, ν = counting measure on $\{0\}$.

$\mu(\{0\}) = 0$, 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.

So what?

Let μ = Lebesgue measure over $\mathbb{R}^+$

ν = counting measure over $\{0\}$.

sum over elements in $A \cap \{0\}$.

$$\Rightarrow P(Y \in A \mid \underline{\theta}) = \int_A f(y; \theta)\, d\mu(y) + \int_A p(y; \underline{\theta})\, d\nu(y).$$

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

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

R-N derivative of prob measure of Y $\Rightarrow$ valid density.

Let $L_*(\underline{\theta} \mid \underline{Y})$ = joint density of observed data!

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

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

Then corresponding R-N derivative $f_{**}(y;\underline{\theta}) = \frac{\mathbb{I}(y \notin \{0\})}{\alpha} f(y|\underline{\theta}) + \frac{\mathbb{I}(y \in \{0\})}{\beta} p(y|\underline{\theta})$

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

A valid likelihood would be $L_{**}(\underline{\theta}|\underline{y}) = f_{**}(y|\underline{\theta})$.

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

Implications: We can scale the discrete and continuous components however we like. What to do? Mostly doesn't matter.

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

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

$\Rightarrow$ Scaling can be ignored in MLE applications.

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?

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 exactly the same information

(intuition) $\Rightarrow$ likelihood 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

from 530(?).

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

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

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

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

$\Rightarrow$ MLE are identical where 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 w/ $n=12$. Let $Y_i = 1$ if i^{th} trial is successful and 0 otherwise.

```
$$
L_1(p | \boldsymbol Y) = {12 \choose S} p^S (1 - p)^{12 - S}, \text{
where } S = \sum\limits_{i = 1}^n Y_i
$$
```

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

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

Arguing:

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

$$\ell_1(p) = \log\binom{12}{9} + 9\log p + 3\log(1-p)$$

$$\frac{d\ell_1(p)}{d\ell(p)} = \frac{9}{p} - \frac{3}{1-p} \overset{set}{=} 0 \Rightarrow 9 - 9p = 3p \Rightarrow \hat{p}_{MLE} = \frac{9}{12}$$

$$\ell_2(p) = \underbrace{\log\binom{11}{9}}_{\text{same excep constant}} + 9\log p + 3\log(1-p) \Rightarrow \hat{p}_{MLE} = \frac{9}{12}$$

NO: Consider 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 $\frac{1}{2}$)) = 0.0327

$\Big\}$ p-values.

Bayesians: Discrepancy in p-values implies frequentist methods not logical because inferences not based solely on likelihood.

Maintain proportional likelihoods contain same information (formalized by 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.

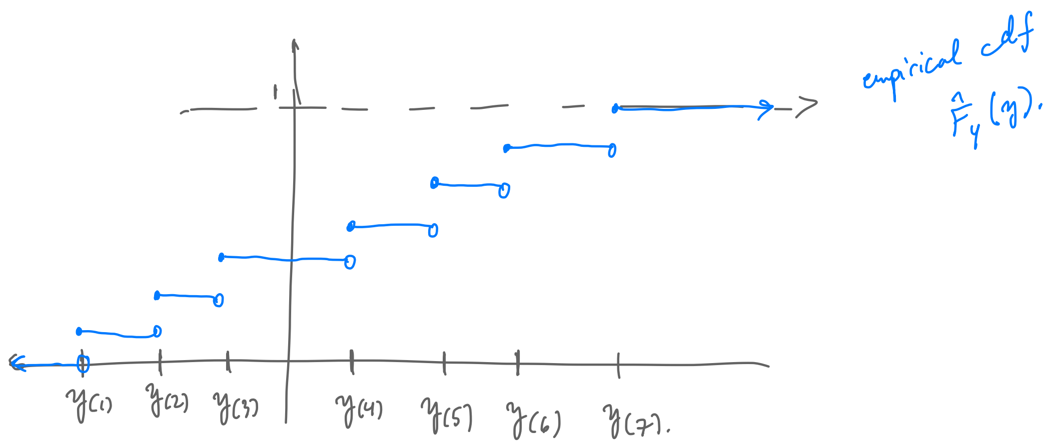
1.2.6 Empirical Distribution Function as MLE

$y_1, \dots, y_n$

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 believe F_y has support on $\mathbb{R}$,

having $\hat{F}_y(y) = 0$ for $y < y_{(1)}$ + $\hat{F}_y(y) = 1$ for $y \geq y_{(n)}$.

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 somebody who does "extremes"

Another answer:
Yes, because it's MLE.

require $F(y)$ to have properties of dsn function

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

$\Longrightarrow$ "parameter space" is set of all distribution functions.

An approximate likelihood for F is

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

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

Assume no ties so that h is small enough such that $[Y_i - h, Y_i + h]$ doesn't contain Y_j for any $j \neq i$.

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

Note: ① increasing $p_{i,h}$ increases $L(F, Y)$.
② $L_h(F|Y)$ is maximized only if $p_{i,h} > 0$ $i=1,\ldots,n$.

$\Rightarrow$ Want $p_{i,h}$ to be as large as possible w/ $p_{i,h} > 0$ & $\sum_{i=1}^{n} p_{i,h} \leq 1$. (3. above). max happens when $=1$.

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

Optimization problem to be solved using Lagrange multipliers (find stationary points of:)

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

get by solving:

$$\frac{\partial g(\ldots)}{\partial p_{i,h}} = \frac{1}{p_{i,h}} + \lambda \overset{\text{set}}{=} 0 \quad i=1,\ldots,n.$$

$$\frac{\partial g(\ldots)}{\partial \lambda} = \sum_{i=1}^{n} p_{i,h} - 1 \overset{\text{set}}{=} 0$$

implies $p_{i,h} = -\frac{1}{\lambda}$

$\Rightarrow \lambda = -n$.

$\Rightarrow$ 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 $L_h(F, Y)$.

As $h \to 0$, $\hat{F}_h(y) \to \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y)$ $\longleftarrow$ this is the empirical dsn f'n $\Rightarrow$ MLE.

1.2.7 Censored Data

Censored data occur when the value is only partially known. This is different from *truncation*, in which the data does not include any values below (or above) a certain limit.

truncation:

For example, we might sample only households that have an income above a limit, L_0. If all incomes have distribution $F(x; \boldsymbol{\theta})$, then for $y > L_0$,

$$P(Y_1 \leq y | Y_1 > L_0) = \frac{P(Y_1 \leq y, Y_1 > L_0)}{P(Y_1 > L_0)} = \frac{F(y; \underline{\theta}) - F(L_0; \theta)}{1 - F(L_0; \underline{\theta})}$$

The likelihood is then

$$L(\underline{\theta} | \underline{y}) = \prod_{i=1}^{n} \left\{ \frac{f(Y_i; \theta)}{1 - F(L_0; \underline{\theta})} \right\}.$$

this is just an iid likelihood w/ densities adjusted to take into account that $Y_i > L_0$.

1.2.7.1 Type I Censoring

censoring:

Suppose a random variable X is normally distributed with mean μ and variance σ^2, but whenever $X \leq 0$, all we observe is that it is less than or equal to 0. If the sample is set to 0 in the censored cases, then define

$$Y = \begin{cases} 0 & \text{if } X \leq 0 \\ X & \text{if } X > 0. \end{cases}$$

Y : observed,

X : latent / hidden.

The distribution function of Y is

$X \sim N(\mu, \sigma^2)$.

at $y = 0$: $F_Y(0) = P(Y = 0) = P(X \leq 0)$.

$= \Phi\left(-\frac{\mu}{\sigma}\right)$, where Φ is standard Normal cdf.

for $y > 0$: $F_Y(y) = P(Y \leq y) = P(X \leq y) = \Phi\left(\frac{y - \mu}{\sigma}\right)$

and $F_Y(y) = 0$ for $y < 0$.

Suppose we have a random sample $Y_1, \ldots, Y_n$ and let n_0 be the number of sample values that are 0. Then $m = n - n_0$ and

$$L_h(\underline{\theta}|\underline{y}) = \left(\frac{1}{2h}\right)^m \prod_{i=1}^{n} \{F_y(y_i + h; \underline{\theta}) - F_y(y_i - h; \underline{\theta})\}$$

$$= \left\{\Phi\left(\frac{h-\mu}{\sigma}\right) - 0\right\}^{n_0} \times \prod_{y_i > 0} \left\{\frac{\Phi((y_i+h-\mu)/\sigma) - \Phi((y_i-h-\mu)/\sigma)}{2h}\right\}$$

$$L(\underline{\theta}|\underline{y}) = \lim_{h \to 0^+} L_h(\underline{\theta}|\underline{y})$$

$$= \underbrace{\left\{\Phi\left(-\frac{\mu}{\sigma}\right)\right\}^{n_0}}_{P_\theta(Y \text{ censored})^{n_0}} \underbrace{\prod_{y_i > 0} \left\{\frac{1}{\sigma}\phi\left(\frac{y_i-\mu}{\sigma}\right)\right\}}_{\prod_{y_i \text{ not censored}} f_y(y_i|\underline{\theta})}$$

same as mixture!

More generally for left censoring @ L_0:

$$L(\underline{\theta}|\underline{y}) = \{F(L_0)\}^{n_0} \prod_{y_i > L_0} f_y(y_i|\underline{\theta}).$$

L_0 fixed = "type I censoring"

Type II censoring = "let's collect until first r censored values"

We might have censoring on the left at L_0 and censoring on the right at R_0, but observe all values of X between L_0 and R_0. Suppose X has density $f(x; \boldsymbol{\theta})$ and distribution function $F(x; \boldsymbol{\theta})$ and

$$Y_i = \begin{cases} L_0 & \text{if } X_i \le L_0 \\ X_i & \text{if } L_0 < X_i < R_0 \\ R_0 & \text{if } X_i \ge R_0 \end{cases}$$

If we let n_L and n_R be the number of X_i values $\le L_0$ and $\ge R_0$ then the likelihood of the observed data $Y_1, \ldots, Y_n$ is

$$L(\underline{\theta}|\underline{y}) = \underbrace{\{F(L_0; \underline{\theta})\}^{n_L}}_{\text{left censored part}} \underbrace{\left\{\prod_{L_0 < y_i < R_0} f_y(y_i; \underline{\theta})\right\}}_{\text{observed}} \underbrace{\{1 - F(R_0; \underline{\theta})\}^{n_R}}_{\text{right censored.}}$$

We could also let each X_i be subject to its own censoring values L_i and R_i. For the special case of right censoring, define $Y_i = \min(X_i, R_i)$. In addition, define $\delta_i = \mathbb{I}(X_i \le R_i)$. Then the likelihood can be written as

$$= \begin{cases} 1 & \text{if observe } X_i \\ 0 & \text{if } X_i \text{ censored.} \end{cases}$$

$$L(\theta|y) = \prod_{i=1}^{n} f(y_i;\theta)^{\delta_i} \underbrace{\left[1 - F(R_i;\theta)\right]^{1-\delta_i}}_{\text{right censored part.}}$$

Example (Equipment failure times): Pieces of equipment are regularly checked for failure (but started at different times). By a fixed date (when the study ended), three of the items had not failed and therefore were censored.

$R_2 = 72 = Y_2$

y	2	72	51	50	33	27	14	24	4	21
delta	1	0	1	0	1	1	1	1	1	0

censored (not failure times).

Suppose failure times follow an exponential distribution $F(x;\sigma) = 1 - \exp(-x/\sigma), x \ge 0$. Then

$$L(\sigma|\boldsymbol{Y}) = \prod_{i=1}^{n} \left[\frac{1}{\sigma} \exp\left(-\frac{y_i}{\sigma}\right)\right]^{\delta_i} \left[\exp(-y_i/\sigma)\right]^{1-\delta_i}$$

$$= \left(\frac{1}{\sigma}\right)^{n-n_R} \exp\left(-n\bar{Y}/\sigma\right)$$

$$\ell(\sigma) = -(n-n_R)\log\sigma - \frac{n\bar{Y}}{\sigma}$$

$$\frac{d\ell(\sigma)}{\sigma} = \frac{-(n-n_R)}{\sigma} + \frac{n\bar{Y}}{\sigma^2} \overset{\text{SET}}{=} 0$$

$$n\bar{Y} = \sigma(n-n_R) \Rightarrow \hat{\sigma}_{MLE} = \frac{n\bar{Y}}{n-n_R}$$

$$= \frac{10 \cdot 30.8}{7} = 44.0$$

1.2.7.2 Random Censoring

So far we have considered censoring times to be fixed. This is not required.

e.g. in medical studies patients to enter at different times (w/ fixed end date), modeled as a random variable.

How is this different than previous example?

This leads to random censoring times, e.g. R_i, where we assume that the censoring times are independent of $X_1, \dots, X_n$ and iid with distribution function $G(t)$ and density $g(t)$.

Again, $Y_i = \min(X_i, R_i)$ and $\delta_i = I(X_i \le R_i)$.

Let's consider the contributions to the likelihood:

due to $(\underbrace{Y_i, \delta_i = 1}_{Y_i \text{ observed.}})$:

$$\frac{P(Y_i \in (y-h, y+h], \delta_i = 1)}{2h} = \frac{P(X_i \in (y-h, y+h], X_i \le R_i)}{2h}$$

$$= \frac{1}{2h} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{I}(y-h < t \le y+h, t \le r) f(t;\underline{\theta}) g(r)\, dt\, dr$$

$$= \frac{1}{2h} \int_{y-h}^{y+h} \left[\int_{-\infty}^{\infty} \mathbb{I}(t \le r) g(r)\, dr \right] f(t;\theta)\, dt$$

(FTC). $$= \frac{1}{2h} \int_{y-h}^{y+h} [1 - G(t)] f(t;\theta)\, dt$$

$$\xrightarrow{h \to 0^+} [1 - G(y)] f(y;\theta).$$

due to $(\underbrace{Y_i, \delta_i = 0}_{Y_i \text{ censored}})$:

$$\frac{P(Y_i \in (y-h, y+h], \delta_i = 0)}{2h} = \frac{1}{2h} P(R_i \in (y-h, y+h], X_i > R_i)$$

$$= \frac{1}{2h} \int_{y-h}^{y+h} \left[\int_{-\infty}^{\infty} \mathbb{I}(t > r) f(t;\theta)\, dt \right] g(r)\, dr.$$

$$= \frac{1}{2h} \int_{y-h}^{y+h} [1 - F(r;\theta)] g(r)\, dr$$

$$\xrightarrow{h \to 0^+} [1 - F(y;\theta)] g(y)$$

which results in

$$L(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{\delta}) = \prod_{i=1}^n f(Y_i;\theta)^{\delta_i} [1 - G(Y_i)]^{\delta_i} \times \prod_{i=1}^n g(Y_i)^{1-\delta_i} [1 - F(Y_i;\theta)]^{1-\delta_i}$$

$$= \prod_{i=1}^n \underbrace{f(Y_i;\theta)^{\delta_i} [1 - F(Y_i;\theta)]^{1-\delta_i}}_{\text{same as for fixed end point}} \underbrace{g(Y_i)^{1-\delta_i} [1 - G(Y_i)]^{\delta_i}}_{\text{no } \theta \Rightarrow \text{not needed for MLE.}}$$