

Likelihoods

History of the course:

Taught since 2017 or 2018 by

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

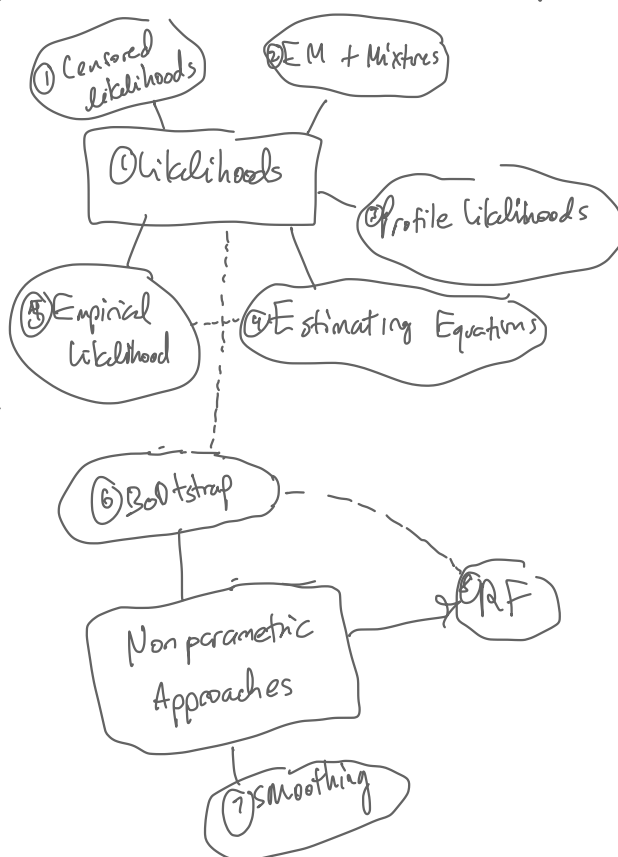
Now: Me

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

midterm



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 , $\psi(\theta)$, if $\hat{\theta}$ is MLE of θ , $\psi(\hat{\theta})$ is MLE of $\psi(\theta)$.
2. Asymptotically unbiased; consistent
$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| > \epsilon) = 0$$
3. Asymptotically efficient; variance achieve Cramer-Rao Lower Bound.
estimator has all 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

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

$$L(\boldsymbol{\theta}|\mathbf{Y}) = f_{\mathbf{Y}}(\mathbf{Y}, \boldsymbol{\theta}). \leftarrow \text{in general, likelihood} = \text{joint}$$

\uparrow likelihood is random.

(because it depends on the data!).

Know MLE is random + we quantify its uncertainty.

Given a vector of observations \mathbf{y} , the likelihood is a function of $\boldsymbol{\theta}$:
For any (valid) value of $\boldsymbol{\theta}$, it returns a number (the likelihood).

$\hat{\boldsymbol{\theta}}_{MLE}$ obtained by finding value of $\boldsymbol{\theta}$ 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 \mathbf{y} , note that $L(\boldsymbol{\theta}|\mathbf{y}) : \mathbb{R}^b \rightarrow \mathbb{R}$.

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_b)^\top$$

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

Generally, we optimize $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}|\mathbf{y})$.

$\underbrace{\hspace{10em}}$
monotone increasing

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

How? Take derivatives, set to zero, solve.

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

Define score function

$$S(\boldsymbol{\theta}) = \ell'(\boldsymbol{\theta})^\top = \begin{pmatrix} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_b} \end{pmatrix}$$

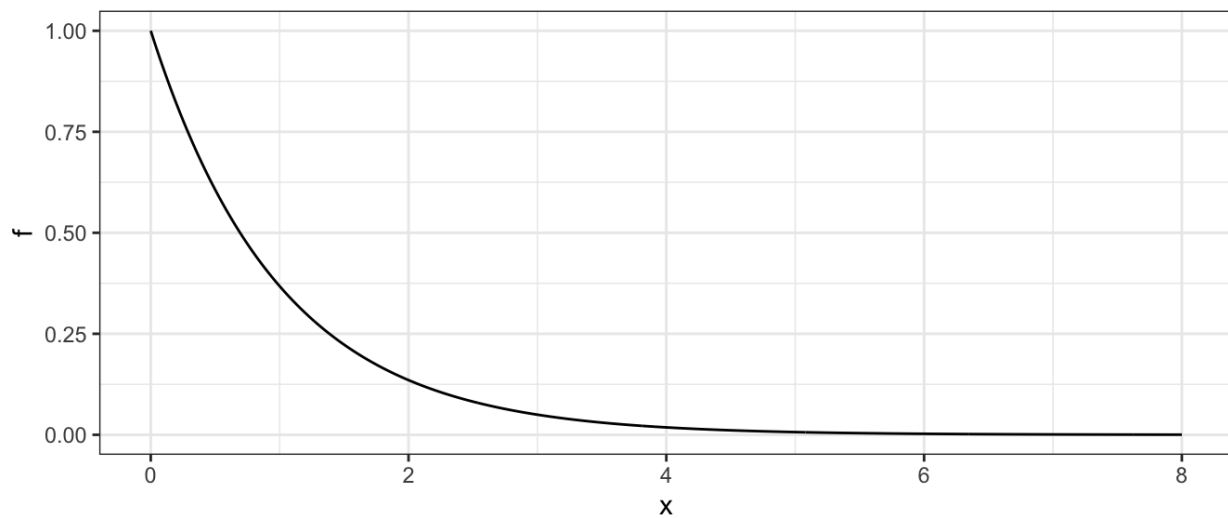
$b \times 1$
column vector.

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

$$\begin{aligned} L(\lambda | Y) &= f_Y(Y, \lambda) \\ &= \prod_{i=1}^n f_Y(y_i; \lambda) \\ &= \prod_{i=1}^n \lambda e^{-\lambda y_i} = \lambda^n e^{-\lambda \sum_{i=1}^n y_i} \Rightarrow \ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^n y_i \end{aligned}$$

```
# 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))
```



```

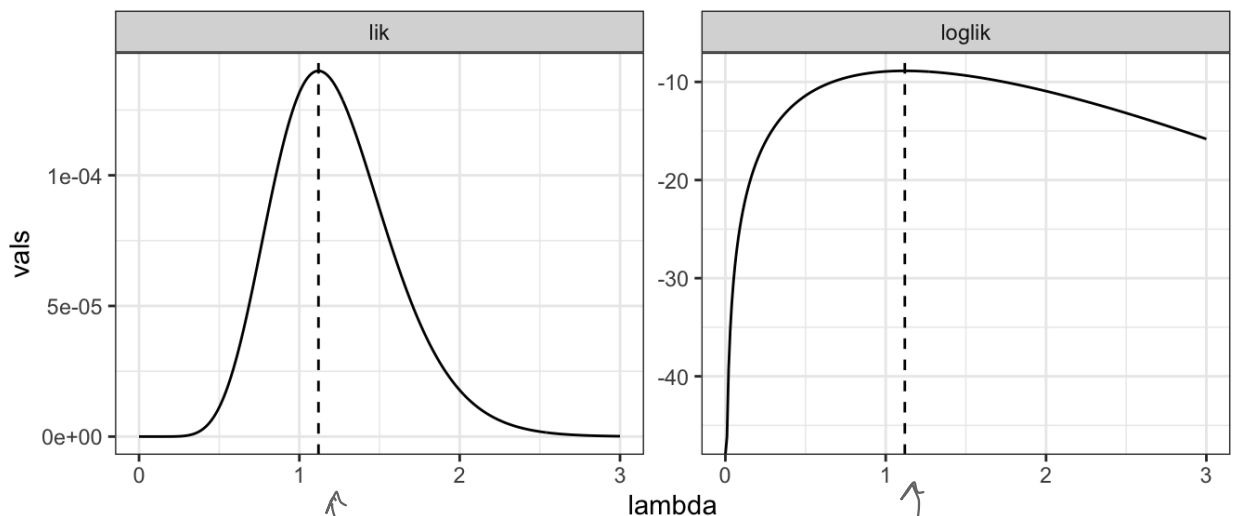
# define likelihood
loglik <- function(lambda, data)
{
  lik <- prod(dexp(data, rate = lambda))
  loglik <- sum(dexp(data, rate = lambda, log = T))
  out <- data.frame(lik = lik, loglik = loglik)
  return(out)
}

# 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 ← MLE
  # likelihood estimate is 1/mean
  geom_line(aes(lambda, vals)) +
  facet_wrap(~func, scales = "free")

```

now we have realized data!



max happens at same place!

The likelihood function is random!

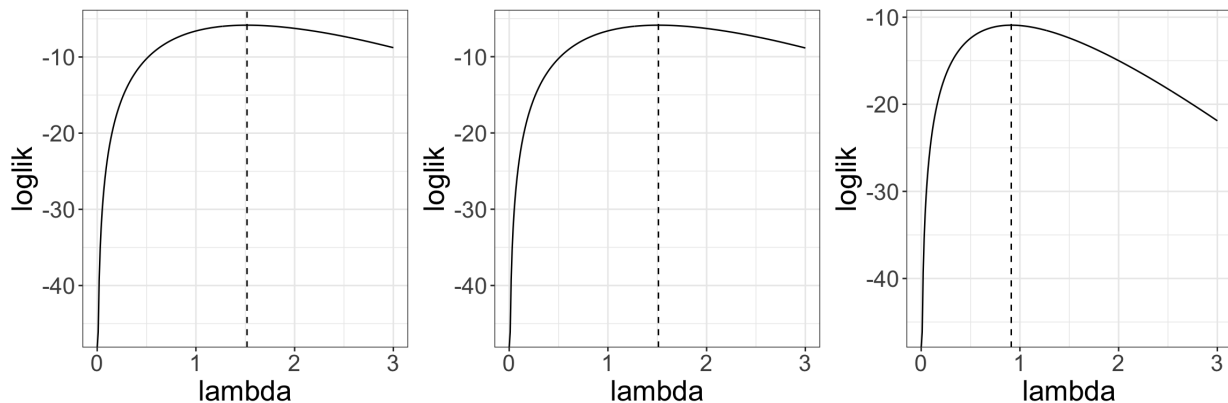
```

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! \rightarrow function of θ , not ψ .

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 sample Y_1, \dots, Y_n have probability mass function $f(y; \theta) = P_\theta(Y_1 = y)$, $y = y_1, y_2, \dots$. The likelihood is then defined as:

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

In other words,

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

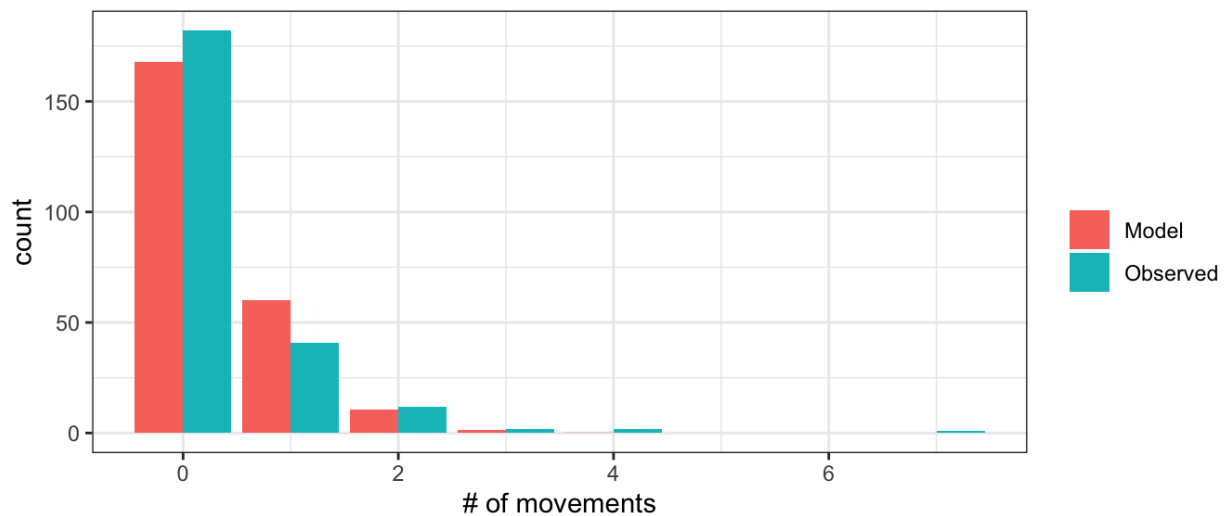
No. of Movements	0	1	2	3	4	5	6	7
Count	182	41	12	2	2	0	0	1

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

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

$$\hat{\lambda}_{\text{MLE}} =$$

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



1.2.2 Multinomial 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.

The probability mass function is

For N_1, \dots, N_k , $N_i =$ the number of balls in i^{th} urn,

The maximum likelihood estimator of p_i :

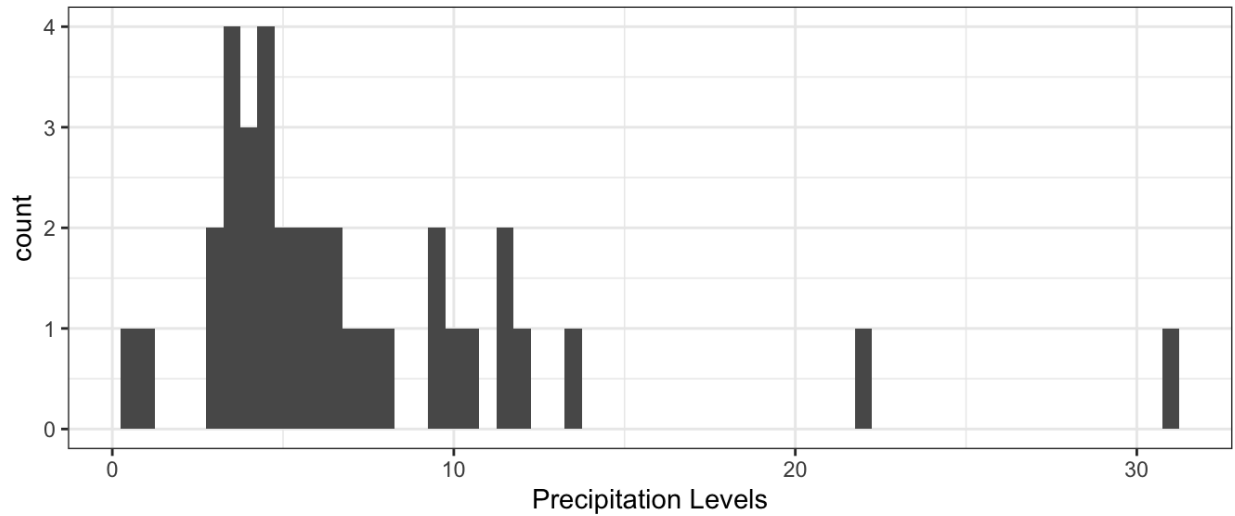
More interesting multinomial likelihoods arise when the p_i are modeled as a function of a lesser number of parameters $\theta_1, \dots, \theta_m$, $m < k - 1$.

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.

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

Of course, this cannot be interpreted as a probability because

To get a probability, need to go from a density to a measure.

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

More formally, begin with the definition of a derivative

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

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

$$f(y) = \lim_{h \rightarrow 0^+} \frac{F(x+h) - F(x-h)}{2h} =$$

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

Compare this to the iid discrete case:

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

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

and log-likelihood

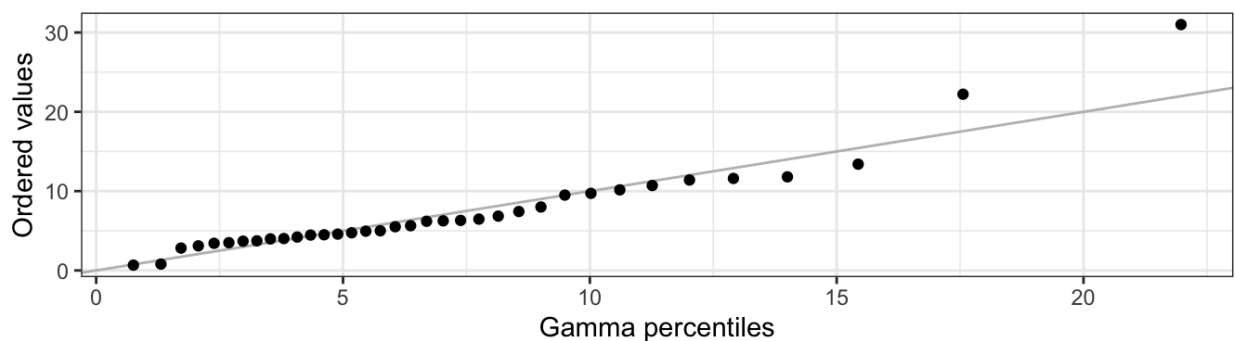
$$\ell(\boldsymbol{\theta}) =$$

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

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

## [1] 2.187214 3.331862

## 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")
```



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:

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

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:

How to go from here to get a likelihood?

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 construct the likelihood for iid Y_1, \dots, Y_n distributed according to the above distribution:

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \lim_{h \rightarrow 0^+} \left(\frac{1}{2h} \right)^m \prod_{i=1}^m \{F_Y(Y_i + h; p, \boldsymbol{\theta}) - F_Y(Y_i - h; p, \boldsymbol{\theta})\}$$

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! (Need some STAT 630/720 to see why.)

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*.

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 $\phi = \phi_{ac} + \phi_s \ll \mu$, ϕ_s and μ are mutually singular and there exists a measurable extended real valued function f such that

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

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 $A \in \mathcal{M}$.

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?

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 $Y_i, i = 1, \dots, n$ 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.

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

$$L(\boldsymbol{\theta}|\mathbf{X}) =$$

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

\$\$

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

\$\$

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

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

The ratio of these likelihoods is

$$\frac{L_1(p|\mathbf{Y})}{L_2(p|\mathbf{Y})} =$$

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

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

1.2.6 Empirical Distribution Function as MLE

Recall the empirical cdf:

Suppose $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$ 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 ?

Yes, because it's MLE.

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

An approximate likelihood for F is

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

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.

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 $y > L_0$,

$$P(Y_1 \leq y | Y_1 > L_0) =$$

The likelihood is then

1.2.7.1 Type I 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

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

The distribution function of Y is

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

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

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

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

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 $\delta_i = \mathbb{I}(X_i \leq R_i)$. Then the likelihood can be written as

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.

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

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

$$L(\sigma|\mathbf{Y}) =$$

1.2.7.2 Random Censoring

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

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 $g(t)$.

Let's consider the contributions to the likelihood:

which results in

$$L(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\delta}) =$$