Chapter 2: Essential Concepts and IID Examples

S352: Data Modeling and Inference

Department of Statistics, Indiana University

Overview

1. Chapter 1: Discrete models with one parameter (one or few data)
   - Ber(p) and Bin(n, p)
   - Geom(p)
   - Poi(λ)
2. Chapter 2: More models and lots of data!
   - 2.1 Discrete models with one or more parameters
     - Poi(λ) vs Negtive Binomial(r, p)
   - 2.2 Continuous models with one or more parameters
     - Unif($0, b$) and Unif(a, b)
     - N(μ, σ_0^2) and N(μ, σ^2)
     - Exp(λ)
     - Gamma(k, λ)
   - 2.3 Zero-inflated mixture models

Bigger data

So far we've dealt with cases where the data sets have been relatively small.

With bigger data sets, unless we have a neat theoretical solution, we need to deal with the data as *vectors* in R.

This is quite manageable (R is good with vectors), but it may require rethinking how we write code...

See file in Canvas: `s352-R-week3-vectors`

1. Operations with vectors
2. Example: Lots of Binomial data.

Example: Lots of binomials

The file `binomdata.txt` contains 1000 independent observations from a Binomial distribution with parameters $m = 100$ (number of Bernoulli trials) and p (unknown probability).

Size of the sample: $n = 1000$!

Can we use `optimize()` or `optim()` to find the MLE for p?

What's the log-likelihood?

Previously we noted that log-likelihood of IID data was

$$l(\theta) = \sum_{i=1}^{n} \log f(y_i|\theta)$$

For IID observations of Binomial$(100, p)$ data $y_1, \ldots, y_n$, this is

$$l(p) = \sum_{i=1}^{n} \left[\log \binom{100}{y_i} + y_i \log p + (100 - y_i) \log(1 - p) \right]$$

Now we want to get R to calculate this if the y's are a vector.

Writing a negative log-likelihood function

Fortunately many R functions are able to take vectors as arguments.

```
binom.nll <- function(p){
        lik <- dbinom(binomdata, size = 100, prob = p) # a vector
        loglik <- log(lik) # a vector
        nloglik <- -sum(loglik) # a number
        return(nloglik)
   }
binom100.nll <- function(p, data = binomdata){
    l1 <- log(choose(100, data))
    l2 <- data * log(p)
    l3 <- (100 - data) * log(1 - p)
    loglik <- l1 + l2 + l3
    nloglik <- -sum(loglik)
    return(nloglik)
}
```

If you want to, you can generalize to any n...

Likelihood with more data

discrete models for count data

Data: School days

The file `schooldays.txt` contains a variable `absent`, which gives the number of days 154 Australian school children were absent from school.

This is count data, so two possible distributions we could use are:

- the Poisson;
- the Negative Binomial.

Is there a reason to prefer one model to the other? Take a look at the summary and plot first ...

Checking a model: Simulation

Let's try simulating 154 observations from a Poisson distribution:

```
sim <- rpois(154, mean(schooldays$absent))
```

Does this look like the observed data? Draw graphs...

The negative binomial distribution

A **negative binomial distribution** is a discrete distribution on the whole numbers. The classical negative binomial distribution has PMF:

$$f(x) = \binom{x+r-1}{x} p^r (1-p)^x, \quad x = 0, 1, \ldots$$

where the parameters are

- r or "size": a positive integer
- p: a real number between 0 and 1

The definition can be extended so that r can be any positive real number. You can get R to calculate the PMF for you:

```
dnbinom(x, size, prob)
```

Using optim() with multiple parameters

Let's say we want to fit a negative binomial. This has two parameters.

To get optim() to work, we need to write a negative log-likelihood function where the argument is a *vector* containing the two parameters:

```
nb.nll <- function(params){
    size <- params[1]
    prob <- params[2]
    lik <- dnbinom(absent, size, prob)
    loglik <- sum(log(lik))
    nll <- -loglik
    return(nll)
}
```

where params is a vector of length 2 contains the size and p parameters.

Using optim() with multiple parameters

By default, optim() uses the "Nelder-Mead" method.

Code: `optim(par = c(1, 0.5), fn = nb.nll, hessian = TRUE)`

Output:

```
$par
[1] 1.10989583 0.06434699

$value
[1] 586.5711

...
```

The MLE is 1.1 for the size parameter and 0.064 for p. (Try simulation.)

Continuous models and likelihood

definitions and notations

Continuous likelihood: The idea

Recall: How did we write a likelihood for a discrete model?

Now, what do we do when the distribution we'd like to use to model the data is continuous?

- Use the probability density function (PDF) instead of the PMF.
- We use the notation $f(x)$ to denote both PMFs and PDFs.
- Is it OK to do this?

Parametric statistical models

A **parametric statistical model** is a collection of joint density functions, $f(\boldsymbol{y}; \boldsymbol{\theta})$, indexed by parameter(s) $\boldsymbol{\theta}$.

The components of $\boldsymbol{\theta}$ are real numbers.

There are also **nonparametric** and, more convolutedly, **semiparametric** statistical models. Learn about these in STAT-S 425...

Notation: Random variables and vectors

Random variables will (as usual in statistics) be denoted by capitals.

A vector $\boldsymbol{y}$ consists of elements $(y_1, y_2, \ldots, y_n)$. A vector is a column unless otherwise stated.

Example:

- Suppose I'm going to sample n independent observations from a standard Normal distribution. Represent the results as the random vector $\boldsymbol{Y}$ whose elements $(Y_1, Y_2, \ldots, Y_n)$ are IID standard normal random variables.
  - We write $\boldsymbol{Y} \overset{IID}{\sim} \text{Normal}(0, 1^2)$, where $\sim$ means "is distributed as."
  - Usually we assume things are IID unless otherwise stated, so we can just write $\boldsymbol{Y} \sim \text{Normal}(0, 1^2)$.
- Now suppose I get results $(0.57, -2.66, -1.00\ldots)$ Then $\boldsymbol{y} = (0.57, -2.66, -1.00\ldots)$, and $y_1 = 0.57$, $y_2 = -2.66$, $y_3 = 1.00\ldots$

Notation: Density functions

Recall that with one observation, we can specify the probability distribution with a **probability density function** (PDF).

Example: Let Y be a (continuous) Uniform distribution on (a, b). Its density function is

$$f(y) = \frac{1}{b-a}$$

for $a \leq y \leq b$.
(It's zero otherwise; we don't always bother to write this down.)

Density functions: Some reminders

For a continuous random variable,

- The density function is the derivative of the cumulative distribution function (CDF).
- We can obtain the probability of being in an interval $[x_1, x_2)$ by integrating the density:

$$P(x_1 \leq X < x_2) = \int_{x_1}^{x_2} f(x)\, dx.$$

Since these are convenient properties, we'd like them to hold for all random variables, not just continuous ones. Then "density" could be used in a general sense to cover both discrete and continuous distributions.

Joint density functions

With more than one observation, we specify the probability distribution using a **joint density function**, $f(\boldsymbol{y}; \boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ denotes the true unknown value (or a hypothesized value) of the parameter(s).

Carefully defining the joint density function in general requires multivariate calculus. Fortunately, in the IID case, the joint density is just a product:

$$f(\boldsymbol{y}; \boldsymbol{\theta}_0) = f(y_1; \boldsymbol{\theta}_0) \times f(y_2; \boldsymbol{\theta}_0) \times \cdots \times f(y_n; \boldsymbol{\theta}_0)$$

The likelihood function

The likelihood function is the (joint) density function evaluated at the observed data, and regarded as a function of $\boldsymbol{\theta}$ alone:

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}; \boldsymbol{y}) = f(\boldsymbol{y}; \theta) \quad \text{for } \boldsymbol{\theta} \in \Theta$$

where Θ is the set of possible values of $\boldsymbol{\theta}$.

- As before, it will often be more numerically convenient to deal with log-likelihood or negative log-likelihood in practice.
- Suppose we make one observation, x. Then the likelihood is the PDF considered as a function of the parameters $\boldsymbol{\theta}$.

The maximum likelihood estimator/estimate

Given observations $\boldsymbol{y} = (y_1, \ldots, y_n)$,

any $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_s) \in \Theta$ that maximizes $L(\boldsymbol{\theta})$ over Θ is called a **maximum likelihood estimator/estimate**, or **MLE**, of the unknown true parameter vector $\boldsymbol{\theta}$.

Notes:

- The MLE might not be *unique*: there might be more than one $\hat{\boldsymbol{\theta}} \in \Theta$ that gives the maximum value of $L(\boldsymbol{\theta})$.
- The MLE might not exist: it might never attain its maximum value (e.g. it might only approach it asymptotically.)

Finding the MLE

When there's one parameter, θ, **draw a graph of the likelihood $L(\boldsymbol{\theta})$.**
Look at your graph.

1. If it looks like the maximum is at the edge, ...
2. If it looks like the maximum is not at the edge:
   - 2.1 Find the log-likelihood $l(\boldsymbol{\theta})$ (and plot it).
     - It's usually easier to maximize the log-likelihood rather than the likelihood itself.
     - Any $\boldsymbol{\theta}$ that maximizes the log-likelihood also maximizes the likelihood.
   - 2.2 Find the derivative $l'(\boldsymbol{\theta})$, by differentiating $l(\boldsymbol{\theta})$ with respect to θ.
   - 2.3 Set $l'(\boldsymbol{\theta})$ to zero and solve for $\hat{\theta}$.
   - 2.4 **Check that you have a maximum.**

Continuous models and likelihood

one parameter and one data cases

Example: Uniform distribution

Suppose we have a Uniform random variable where the lower bound is known to be 0 and the upper bound is unknown b. Its PDF is thus

$$f(x) = \begin{cases} \frac{1}{b} & 0 \leq x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

Suppose we make one observation: $x = 5$.

1. Plot the likelihood for $0 < b < 10$.
2. What is the MLE for b?

Application: The time I have to wait for a bus (in minutes) is Uniform$(0, b)$, Suppose I waited five minutes for the bus today: $x = 5$. What does this tell us about b?

Example: Normal distribution with known σ

Suppose we have one observation from a Normal distribution: $x_1 = 5$.

Also suppose (unrealistically) that the SD is known to be 1.

What value of μ maximizes the likelihood?

- We'll remind you what the Normal PDF is in a moment, but for now, recall that it's given by the `dnorm()` function.

Example: Normal distribution with known σ

Instead of maximizing the likelihood, let's minimize the negative log-likelihood:

```
normal.nll <- function(mu){
    lik <- dnorm(5, mu, 1)
    nll <- -log(lik)
    return(nll)
}
optimize(normal.nll, c(-10, 10))
```

Continuous models and likelihood

multivariate case

Example: The Normal model

Recall (or look up) that the Normal PDF is

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right], \quad -\infty < y < \infty.$$

- What does "Y is normally distributed" mean?

Each observation $y_1, \ldots, y_n$ was drawn from a Normal distribution with parameters μ (the mean) and σ^2 (the variance), where μ is a real number and σ^2 is a positive real number.

Assuming IID observations, the Normal model is

$$f(\boldsymbol{y}; \mu, \sigma^2) = \prod_{i=1}^{n} f_{\mu,\sigma^2}(y_i)$$

where $f_{\mu,\sigma^2}(y_i)$ is the PDF of a Normal(μ, σ^2) random variable evaluated at y_i.

Example: The Normal model

Take the PDF and treat the data y as fixed:

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} f_{\mu,\sigma^2}(y_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right] \end{aligned}$$

Now take the log to find the log-likelihood.

Multivariate problems

If you've done multivariate calculus, you can extend this idea to find the maximum with respect to multiple parameters. (It involves taking partial derivatives.)

Occasionally this leads to an elegant, closed-form solution. Linear regression is a notable example of this.

However, most of the time, a numerical solution is necessary, i.e. you need to use a computer. This is fine, because we have computers now.

(If you really want to know how to deal with this theoretically, take Calc 3, then S420.)

Finding the MLEs of μ and σ^2 theoretically

1. Write down the likelihood.
2. Write down the log-likelihood.
3. Take MATH-M 311.
4. Find the partial derivative with respect to μ and verify this is 0 when $\hat{\mu} = \bar{y}$.
5. Find the partial derivative with respect to σ^2 and plug in $\mu = \bar{y}$.
6. Set the result equal to 0 and find the MLE $\hat{\sigma}^2$.
7. Double check $(\hat{\mu}, \hat{\sigma}^2)$ maximizes $L(\mu, \sigma^2)$.

The Normal MLEs (theoretical solutions)

We find the MLE for the Normal is

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

for the mean and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

for the variance.
To compute these MLEs in R:

```
mean(y);  mean((y - mean(y))^2)
```

Normal model

MLEs, unbiasedness, invariance principle

NBA heights

Here are the heights, in inches, of a sample of 19 NBA players:

```
NBAheights <- c(81, 81, 71, 74, 76, 78, 80, 78, 80, 81,
74, 78, 76, 76, 80, 78, 75, 79, 84)
```

What's a good probability model?

How about the good old IID Normal(μ, σ^2):

$$f(\boldsymbol{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right]$$

Checking normality

Draw a normal QQ plot of the data using `qqnorm()`.

- If this produces something close to a straight line, a Normal(μ, σ^2) may be a good probability model.
- If it curves upwards, maybe try models the *logged* data as Normal.
- Otherwise, try something else.

Write a NLL function

The input will be a vector where the first element is μ and the second is σ^2.

```
normal.nll <- function(params){
    mu <- params[1]
    sigma2 <- params[2]
    sigma <- sqrt(sigma2)
    lik <- dnorm(NBAheights, mu, sigma)
    nll <- -sum(log(lik))
    return(nll)
}
```

Use optim()

```
optim(par = c(72, 9), fn = normal.nll, hessian = TRUE)

$par
[1] 77.895389  9.358824
$value
[1] 48.20336
...
$hessian
[,1]           [,2]
[1,]  2.0301696502 -0.0001414531
[2,] -0.0001414531  0.1084285035
```

Was there an easier way?

```
> mean(NBAheights)
[1] 77.89474
> var(NBAheights)
[1] 9.877193
```

The MLE of μ is just the sample mean $\bar{y}$.
But the MLE of σ^2 is not quite the sample variance s^2 that R gives you...

Recall: The Normal MLE (Theoretical solutions)

We find the MLEs for the Normal is

$$\hat{\mu} = \bar{y}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

To find the latter in R:

```
mean((y - mean(y))^2)
```

What happened to the minus one?

Recall that the **sample variance** (i.e. what the `var()` function in R gives you) is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

The reason (some) statistician prefer this to the MLE is it's an **unbiased** estimator. An unbiased estimator is one where the expected value of the estimator equals the true parameter value:

$$\mathrm{E}[S^2] = \sigma^2$$

Simulation to study bias

Write a function to simulate Normal$(0, 1^2)$ data sets of size n, and calculate both $\hat{\sigma}^2$ and s^2:

```
norm.sim <- function(n){
    data <- rnorm(n)
    sigma2 <- mean((data - mean(data))^2)
    s2 <- var(data)
    return(c(sigma2, s2))
}
```

Replicate this a bunch of times, e.g. with $n = 10$:

```
sim.results <- replicate(10000, norm.sim(10))
```

The first row of `sim.results` gives values of $\hat{\sigma}^2$; the second row gives values of s^2. If a statistic is unbiased, the average of its row should be equal to the true variance, which is 1.

Does unbiasedness matter?

In the old days, statisticians worked hard to develop unbiased estimators.

Nowadays statisticians aren't so strict about unbiasedness anymore.

An estimator that's slightly biased may be a bit more accurate overall than an unbiased one.

Of course, the best choice is to take a big sample size, so that it doesn't matter whether you're dividing by n or $n - 1 \ldots$

Summary: Estimating the Normal distribution

We parameterize a Normal distribution via its expected value μ and variance σ^2.
For IID data:

- The MLE of μ is the sample mean, $\bar{X}$.
- The MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

However, for many purposes, we use the **sample variance** (which is unbiased) instead:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Functions of the parameters

It may be the case that instead of wanting to estimate a parameter $\boldsymbol{\theta}$, we wish to estimate a function of that parameter, $g(\boldsymbol{\theta})$.
e.g. We parameterized our Normal distribution in terms of its variance σ^2.

But in addition, or instead, we might be interested in estimating the standard deviation.

Do we have to solve for two separate MLEs? Or is there a shortcut?

Example: Variance and SD

Suppose σ^2 is the variance and σ is the SD. Note that there's a one-to-one relationship between variance and SD (since SD can't be negative.)

Suppose we know that the MLE of the variance is $\hat{\sigma}^2$, giving a likelihood equal to $f(\boldsymbol{y}; \hat{\mu}, \hat{\sigma}^2)$.
If we now re-parameterize in terms of standard deviation, then $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ will also give a likelihood equal to $f(\boldsymbol{y}; \hat{\mu}, \hat{\sigma}^2)$.
It's easy to show by contradiction that it's impossible to choose a value of σ that gives a higher log-likelihood.

So $\hat{\sigma}$ is the MLE of the standard deviation.
(The converse holds as well: if we've found the MLE for the SD, just square it to get the MLE for the variance.)

Transformations

More generally, the MLE is *invariant* to transformations of the parameters.

That means we have substantial freedom to do transformations without messing up the inference. For example:

- We might transform so that a variable, or variables, or the errors, are well-modeled by a Normal distribution.
- Alternative parameterizations may be more convenient in some cases, e.g. SD vs. variance vs. precision (reciprocal of the variance.)

Example: Parameterizing the Normal

We observe a whole bunch of adult women's heights.

Suppose that assuming IID Normal data, we've already found the MLE for the mean was 63.8 inches and for the variance was 8.2 inches squared.

Now, however, our weird friend wants to know:

1. The coefficient of variation σ/μ;
2. The probability a random woman is no taller than 60.0 inches.

Can we find MLEs for these?

Example: Coefficient of variation

We found the MLEs for the mean was $\hat{\mu} = 63.8$ inches, for the variance was $\hat{\sigma}^2 = 8.2$ inches^2, and thus for the SD is $\hat{\sigma} = \sqrt{8.2} \approx 2.864$ inches.

The coefficient of variation is σ/μ.

To find the MLE for the coefficient of variation, we can just do $\hat{\sigma}/\hat{\mu} = 2.864/63.8 \approx 0.045$.

Example: Normal probability

To find the MLE for the probability a random woman is no taller than 60 inches (i.e. the CDF at 60), use `pnorm()`:

```
> pnorm(60, 63.8, sqrt(8.2))
[1] 0.09225144
```

According to the MLE, about 9% of women are 5 feet tall or less.

R Exercise

If the MLEs of the parameters of a Normal distribution are $\hat{\mu} = 63.8$ inches and $\hat{\sigma}^2 = 8.2$ inches squared, what's the MLE of the interquartile range?

Recall that the IQR is the third quartile (75th percentile) minus the first quartile (25th percentile.)

(This is like asking: if those were the correct parameters for the Normal, what would the IQR be? Is there an R function that can help you?)

MLEs: More continuous models

Uniform, bias, variance and MSE

Uniform distributions

Recall that the continuous uniform distribution has PDF

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

We already saw the special case where a is known to be 0. In that case, the MLE of b is $\max\{x_i\}$, i.e. the highest observation.

- In the general case, **the MLE of a is $\min\{x_i\}$ and the MLE of b is $\max\{x_i\}$.**
- What about the MLE(s) of the function of a and/or b?

Is anything actually Uniform?

The Uniform seems too simple to be true very often. Is it ever actually useful?

- Random number generators often aim to generate observations from a Uniform distribution.
- Sometimes things with a hard lower bound and a hard upper bound can be well-approximated by a Uniform. (But if the true PDF isn't flat, there may be better choices, like a Beta distribution, possibly rescaled.)
- Basically any continuous distribution can be approximated as a mixture of Uniforms, for the same reason that any curve can be approximated by a bunch of rectangles.

How accurate is the Uniform MLE? (optional)

Let's suppose we're estimating the maximum b of a Uniform distribution using its MLE $\hat{b} = \max\{X_i\}$, i.e. the maximum of n IID observations. Suppose for simplicity the minimum of the Uniform distribution is known to be $a = 0$.

How accurate is $\hat{b}$? Here are several quantities we might want to know:

- **Bias** $= \mathrm{E}(\hat{b}) - b$.
- **Variance** $= \mathrm{Var}(\hat{b})$. (The standard error is the square root of this.)
- **Mean squared error (MSE)** $= \mathrm{E}[(\hat{b} - b)^2]$. This may be the best measure of "overall" error.

We can estimate all these things using either theory (STAT-S 420/621) or simulation.

Running the simulation (optional)

Write a function to find $\hat{b}$:

```
uniMLE <- function(n, b){
    data <- runif(n, min = 0, max = b)
    b.mle <-  max(data)
    return(b.mle)
}
```

Run it a whole bunch of times:

```
uniMLE.sim <- replicate(10000, uniMLE(n = 5, b = 100))
hist(uniMLE.sim)
```

Bias, variance, MLE (optional)

The bias is estimated as:

```
> mean(uniMLE.sim) - 100
[1] -16.76703
```

The variance:

```
> var(uniMLE.sim)
[1] 196.1467
```

The mean squared error:

```
> mean((uniMLE.sim - 100)^2)
[1] 477.2605
```

Using theory: Distribution of the MLE (optional)

Let $Y = \max(X_1, \ldots, X_n)$, where the X's are IID Uniform$(0, b)$. Let's find the CDF of Y.

Choose some y such that $0 \leq y < b$.

For Y to be less than or equal to y, all n of the X's have to be less than or equal to y.

- The probability that $X_i \leq y$ is y/b.
- The probability that all the X's are $\leq y$ is thus $y/b \times \cdot \times y/b = (y/b)^n$.

That's the CDF:

$$P(Y \leq y) = \prod_{i=1}^{n} P(X_i \leq y) = \frac{1}{b^n} y^n, \quad 0 \leq y < b.$$

Using theory: PDF of the MLE (optional)

We have the CDF:

$$P(Y \leq y) = \frac{1}{b^n} y^n, \quad 0 \leq y < b.$$

Differentiate to find the PDF:

$$f(y) = \begin{cases} \frac{n}{b^n} y^{n-1} & 0 \leq y < b \\ 0 & \text{otherwise.} \end{cases}$$

Using theory: Expected value (optional)

Recall that for a continuous random variable, the expected value is

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

For our estimator, this becomes

$$\begin{aligned} E(Y) &= \int_0^b \frac{n}{b^n} y^n \, dy \\ &= \left[\frac{n}{b^n} \frac{y^{n+1}}{n+1} \right]_0^b \\ &= \frac{n}{b^n} \frac{b^{n+1}}{n+1} \\ &= \frac{n}{n+1} b \end{aligned}$$

Using theory: Bias (optional)

The bias is the expected value minus the true value:

$$\begin{aligned} \text{Bias} &= \frac{n}{n+1} b - b \\ &= \frac{n-(n+1)}{n} b \\ &= -\frac{1}{n+1} b \end{aligned}$$

If you enjoyed this, you can also find the variance and MSE via calculus.
Or (at least in this course) just do the simulation...

MLEs: More continuous models

skewed data: exponential and gamma; MOM

Skewed data

The Normal and Uniform are widely used symmetric distributions. (So is the t-distribution.)

However, most real data sets aren't symmetric.

What should you do if you have skewed data?

- Surprisingly often, a simple transformation like a log makes the data close to symmetric.
- If that doesn't work, there are lots of skewed distributions. We'll introduce a couple here: the **exponential** and the **gamma**.

Why Exponential?

The Exponential is often used to model times, such as lifetimes or "inter-event" times.

- A smoke alarm emits alpha particles at random times. The times between emissions almost exactly follow an Exponential distribution.
- Suppose you have a lightbulb such that every day it has the same chance of breaking (condition on it lasting until that day.) e.g.:
  - It has a 1% chance of breaking on Day 1;
  - If it survives Day 1, it has a 1% chance of breaking on Day 2;
  - If it survives through Day 1000, it has a 1% chance of breaking on Day 1001.

Note: Humans are not like lightbulbs.

The Exponential distribution

The Exponential distribution with "rate" parameter $\lambda > 0$ has CDF

$$F(x) = 1 - e^{-\lambda x}, \quad x > 0$$

and PDF

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0.$$

In R: `pexp()`, `dexp()`.

- $\mathrm{E}(X) = 1/\lambda$ (integration by parts)
- $\mathrm{Var}(X) = 1/\lambda^2$ (more annoying integration by parts)

A quick estimate

For the Exponential, $\mathrm{E}(X) = 1/\lambda$.

Not knowing much else, it would seem we could get a decent guess for λ by substituting $\bar{x}$ for $\mathrm{E}(X)$.

Call this estimate $\hat{\lambda}_{MOM}$:

$$\begin{aligned} \bar{x} &= 1/\hat{\lambda}_{MOM} \\ \hat{\lambda}_{MOM} &= \frac{1}{\bar{x}} \end{aligned}$$

Is $\hat{\lambda}_{MOM}$ the same as the MLE?

Finding the Exponential MLE

Suppose we observe n observations, $x_1, \ldots, x_n$, from an Exponential(λ) distribution. The likelihood (for $\lambda > 0$) is

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

The log-likelihood is

$$\begin{aligned} l(\lambda) &= \sum_{i=1}^{n} [\log \lambda - \lambda x_i] \\ &= n \log \lambda - \lambda \sum_{i=1}^{n} x_i \end{aligned}$$

Finding the Exponential MLE, continued

$$l(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

Take the derivative:

$$l'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

Setting to zero and solving for λ gives

$$\begin{aligned} 1/\hat{\lambda} &= \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} \\ \hat{\lambda} &= 1/\bar{x} \end{aligned}$$

i.e. the MLE for λ is one over the sample mean.
(So $\hat{\lambda}_{MOM}$ is the same as the MLE. This isn't always the case...)

Method of moments

The **method of moments** is often an easy way to get rough parameter estimates.

- With one unknown parameter, find an expression for $\mathrm{E}(X)$.
  Then substitute the sample mean $\bar{x}$ for $\mathrm{E}(X)$ and solve for the parameter.
- With two unknown parameters, find expressions for $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.
  Then substitute the $\bar{x}$ for $\mathrm{E}(X)$ and $\hat{\sigma}^2$ for $\mathrm{Var}(X)$ and solve for the parameters.

Why MOM?

In general, when the method of moments and maximum likelihood give you different answers, the MLE is better.

The MOM can still be useful, however, as a way of getting good initial values for algorithms to find the MLE, as well as for sanity checking numerical MLEs.

MOM example: Negative binomial

Recall we parameterized the Negative Binomial using a size parameter r and a probability p:

$$f(x) = \binom{x+r-1}{x} p^r (1-p)^x, \quad x = 0, 1, \ldots$$

Under this parameterization,

$$\begin{aligned} \mathrm{E}(X) &= \frac{r(1-p)}{p} \\ \mathrm{Var}(X) &= \frac{r(1-p)}{p^2} \end{aligned}$$

What are the MOM estimates for the parameters in terms of $\bar{x}$ and $\hat{\sigma}^2$?

MOM example: Negative binomial

Substitute:

$$\begin{aligned} \bar{x} &= \frac{r(1-p)}{p} \\ \hat{\sigma}^2 &= \frac{r(1-p)}{p^2} \end{aligned}$$

Multiply the second line by p:

$$p\hat{\sigma}^2 = \frac{r(1-p)}{p} = \bar{x}$$

Solve for p:

$$\hat{p}_{MOM} = \frac{\bar{x}}{\hat{\sigma}^2}$$

Rearrange the first equation and solve for r:

$$\hat{r}_{MOM} = \frac{\hat{p}_{MOM}}{1-\hat{p}_{MOM}}\bar{x}$$

The Gamma distribution

The Exponential distribution is often an inaccurate model for times. A more flexible family is the **gamma distribution**.

Let X_1, X_2, X_k be IID Exponential random variables with rate λ.

Then

$$Y = X_1 + \cdots + X_k$$

is a gamma random variable with "shape" parameter k and rate parameter λ.

The Gamma distribution

The PDF of the gamma is

$$f(x) = \frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x}, \quad x > 0$$

where $\Gamma(k)$ is the so-called *Gamma function*.
– When k is a natural number, $\Gamma(k)$ is equal to $(k-1)!$.
– When k is a positive real,

$$\Gamma(k) = \int_0^\infty t^{k-1} e^t \, dt$$

It requires annoying numerical integration, but we can get R to do it...

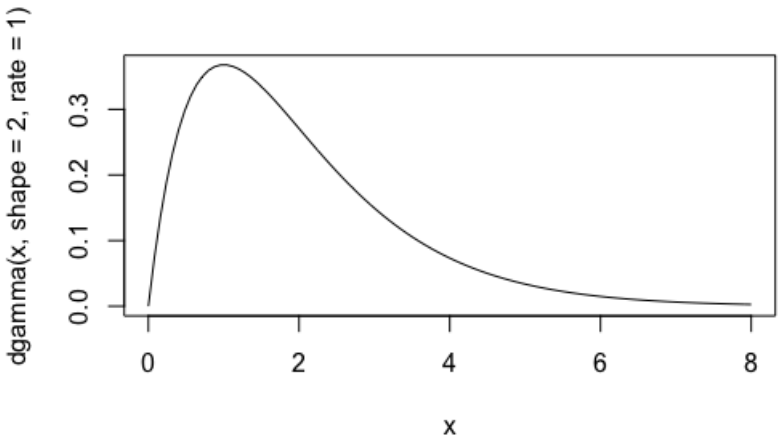
```
> gamma(c(4, 4.5, 5))
[1]  6.00000 11.63173 24.00000
```

You don't want to integrate that, so just get the PDF with R:

```
dgamma(x, shape = ..., rate = ...)
```

Plotting the Gamma distribution

```
curve(dgamma(x, shape = 2, rate = 1), from = 0, to = 8)
```



Expected value and variance

Recall that an Exponential(λ) random variable has expected value $1/\lambda$ and variance $1/\lambda^2$.

The sum of k IID Exponential(λ) random variables has a Gamma(k, λ) distribution.

So the Gamma(k, λ) has:

- Expected value $= 1/\lambda + 1/\lambda + \cdots + 1/\lambda = k/\lambda$
- Variance $= 1/\lambda^2 + 1/\lambda^2 + \cdots + 1/\lambda^2 = k/\lambda^2$

Conveniently, these formulae for expected value and variance hold even when k isn't a whole number!

Finding the MLE

The bad news is that there are no closed form solutions for the MLEs of k and λ.

The good news, as always, is we have R, and can use our usual method:

- Write a negative log-likelihood function;
- Feed it into `optim()`.

But note it can take quite a large sample size to converge to the true parameter values...

Return to MOM

Let's find method of moments estimates for k and λ.

$$\begin{aligned} \bar{x} &= \frac{k}{\lambda} \\ \hat{\sigma}^2 &= \frac{k}{\lambda^2} \end{aligned}$$

Multiply the second line by λ:

$$\lambda\hat{\sigma}^2 = \frac{k}{\lambda} = \bar{x}$$

Solve for λ:

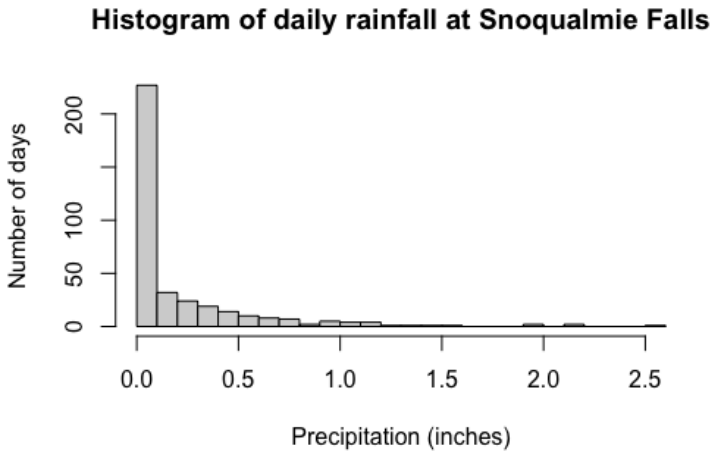
$$\hat{\lambda}_{MOM} = \frac{\bar{x}}{\hat{\sigma}^2}$$

Rearrange the first equation and solve for k:

$$\hat{k}_{MOM} = \hat{\lambda}_{MOM} \cdot \bar{x} = \frac{\bar{x}^2}{\hat{\sigma}^2}$$

Example: Snoqualmie Falls

The file `snoqualmie.txt` contains the recorded precipitation (in inches) on a sample of 365 days (non-consecutive) at Snoqualmie Falls, near Seattle.



Modeling rainfall

How do we model the highly right-skewed distribution of daily precipitation at Snoqualmie Falls?

- Take a log?
- Exponential distribution?
- Gamma distribution?

Fitting an Exponential

With the Exponential, we have an explicit formula for the MLE:

$$\hat{\lambda} = \frac{1}{\bar{x}}$$

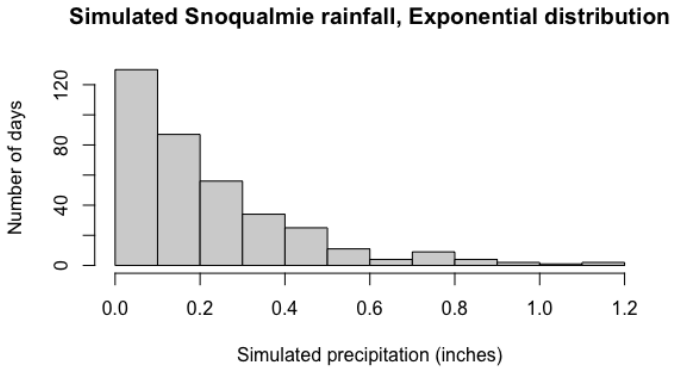
So our guess for λ is

```
> mean(snoq)
[1] 0.205589
```

How well does this fit the data?

Look at simulated data

```
sim <- rexp(365, rate = 1/mean(snoq))
hist(sim)
```



Gamma: Find the MOMs

To fit a gamma distribution, start with the MOMs:

$$\hat{\lambda}_{MOM} = \frac{\bar{x}}{\hat{\sigma}^2}$$

$$\hat{k}_{MOM} = \frac{\bar{x}^2}{\hat{\sigma}^2}$$

Find $\bar{x}$ and $\hat{\sigma}^2$ from the data:

```
x.bar <- mean(snoq)
sigma2.hat <- mean(snoq^2) - mean(snoq)^2 # secret formula
```

Then the MOMs are

```
lambda.mom <- x.bar / sigma2.hat
k.mom <- x.bar^2 / sigma2.hat
```

Write a negative log-likelihood function

```
snoq.nll <- function(pars){
    k <- pars[1]
    lambda <- pars[2]
    lik <- dgamma(snoq, shape = k, rate = lambda)
    loglik <- log(lik)
    nll <- -sum(log(lik))
    return(nll)
}
```

Now try evaluating this at the MOMs:

```
snoq.nll(c(k.mom, lambda.mom))
```

Oops!

What's the problem?

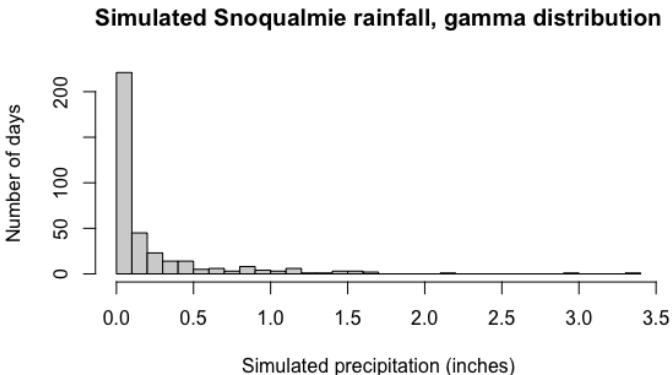
The gamma PDF is undefined at $x = 0$, except for the special case where the shape is 1 (which is just the Exponential distribution.)

Possibilities:

- Just use the Exponential with its MLE—didn't look great
- Use the gamma with its MOM
- Use a **mixture** model:
  - Sometimes the precipitation follows a gamma distribution;
  - Sometimes it's just zero.

Look at simulated gamma data

```
gamma.sim <- rgamma(365, shape = k.mom, rate = lambda.mom)
hist(gamma.sim, breaks = ...)
```



Mixture models and zero-inflation

Snoqualmie rainfall problems

Previously we tried to fit a gamma distribution to Snoqualmie precipitation data.

We found that for a Gamma distribution, the *method of moments* estimates are given by:

- Shape: $k_{MOM} = \frac{\bar{x}^2}{\hat{\sigma}^2}$
- Rate: $\lambda_{MOM} = \frac{\bar{x}}{\hat{\sigma}^2}$

However, we were unable to use maximum likelihood because the data had zeroes, and the Gamma density at zero is either 0 or undefined.

For the purpose of simulation, we could just use the MOM estimates.

However, then we'd never get any true zeroes in our simulation...

Mixture models

Sometimes a population is made up of two or more subpopulations with different distributions.

e.g. American adults are (approximately) composed of men and women. Men's heights are approximately Normal; women's heights are approximately Normal but with different parameters.

- If you have data that includes sex, it makes sense to model men's and women's heights separately.
- If you have data that doesn't include sex, it may still make sense to model the overall population of heights as a *mixture* of two Normal distribution with different parameters, with some probability p of following the first Normal and a probability $1 - p$ of following the second Normal.

This particular problem turns out to be quite challenging to deal with numerically, so we'll punt it until later in the semester and do some easier ones...

Dealing with zeroes

Suppose we have a data set that would be well-modeled by a Gamma distribution, except that it has some zeroes.
A simple solution is to treat an observation X as coming from a mixture of two distributions:

- With probability p, $X = 0$ exactly.
- With probability $1 - p$, X comes from a Gamma(k, λ) distribution.

We want to estimate p, k, and λ. Is there an easy way?

Gamma with zeroes

1. Estimate p by finding the proportion of your data set that's equal to zero.
2. Now throw all the zeroes out of your data set.
3. From the remaining data, get MOM estimates of k and λ.
4. Use these as initial values to find MLEs for k and λ.

(This kind of model is sometimes called a **hurdle** model: if you get over the hurdle, then you can use the gamma distribution.)

Example: Snoqualmie rainfall

Find the proportion of zeroes:

```
snoq <- scan("snoqualmie.txt")
p.hat <- mean(snoq == 0)
p.hat
```

The MLE of p is 0.397. Now take out the zeroes:

```
snoq.no0 <- snoq[snoq != 0]
```

Example: Snoqualmie without zeroes

Find the sample mean and plug-in variance of this reduced data set:

```
x.bar <- mean(snoq.no0)
sigma2.hat <- mean(snoq.no0^2) - mean(snoq.no0)^2
```

Remember from last time the MOMs are $\hat{\lambda}_{MOM} = \bar{x}/\hat{\sigma}^2$ and $\hat{k}_{MOM} = \bar{x}^2/\hat{\sigma}^2$:

```
lambda.mom <- x.bar / sigma2.hat
k.mom <- x.bar^2 / sigma2.hat
```

Fit the gamma

```
snoq.no0.nll <- function(pars){
    k <- pars[1]
    lambda <- pars[2]
    lik <- dgamma(snoq.no0, shape = k, rate = lambda)
    loglik <- log(lik)
    return(-sum(loglik))
}
snoq.no0.mle <- optim(par = c(k.mom, lambda.mom), fn = snoq.no0.nll)
k.mle <- snoq.no0.mle$par[1]
lambda.mle <- snoq.no0.mle$par[2]
```

I get MLEs of $\hat{k} = 0.793$ and $\hat{\lambda} = 2.325$.

Simulate data

```
# How many zeroes?
x <- rbinom(1, 365, p.hat)
# Generate 365 - x obs from the gamma
sim <- rgamma(365 - x, shape = k.mle, rate = lambda.mle)
# Add zeroes
sim <- c(sim, rep(0, x))
summary(sim)
# Draw a histogram
hist(sim)
```

Does this look like the raw data?

Zero-inflated Poisson model

Another common "zero" problem often occurs when trying to use count models like the Poisson.

Very often, the proportion of zeroes in the data is much higher than what a Poisson would predict.

The problem is now slightly trickier, as the Poisson will generate *some* true zeroes. So we can't just estimate the proportion of extra zeroes by taking the sample proportion.

- (There is a closed-form mathematical solution by Lambert, but it's complicated...)

What's the PMF?

Usually the PMF of a Poisson is

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x = 0, 1, 2 \ldots$$

In particular, $f(0) = e^{-\lambda}$.

Now suppose there's a probability $(1 - p)$ of drawing from a Poisson(λ) distribution and a probability p of drawing from a distribution that's all zeroes.

The PMF becomes

$$f(x) = \begin{cases} (1-p)\frac{\lambda^x e^{-\lambda}}{x!} & x = 1, 2, 3 \ldots \\ p + (1-p)e^{-\lambda} & x = 0 \end{cases}$$

Simulate some data

First simulate some Poisson(1) data:

```
data <- rpois(10000, 1)
```

Randomly change some observations to zeroes, with a 40% probability:

```
change <- rbinom(10000, 1, 0.4)
data[change == 1] = 0
```

The true parameters are $p = 0.4$ and $\lambda = 1$. Can we recover these using MLE?

Write an NLL function

```
sim.nll <- function(pars){
    p <- pars[1]
    lambda <- pars[2]
    pois.lik <- dpois(data, lambda)
    real.lik <- rep(NA, length(data))
    which0 <- (data == 0)
    real.lik[which0 == FALSE] = (1 - p) * pois.lik[which0 == FALSE]
    real.lik[which0 == TRUE] = p + (1 - p) * pois.lik[which0 == TRUE]
    loglik <- log(real.lik)
    return(-sum(loglik))
}
```

Try optim()

```
> optim(par = c(0.5, 1), fn = sim.nll)
$par
[1] 0.4037843 0.9696499

$value
[1] 10313.92

...
```

It's pretty close!

Try it out (R Exercise)

The file `fish.csv` on Canvas contains a variable `count` that records the number of fish caught by 250 groups visiting a park. Read it into R using e.g. `read.csv()`.

1. Draw a histogram of the data.
2. Using MLE, fit a zero-inflated Poisson model to the data by finding $\hat{p}$ and $\hat{\lambda}$.
3. Simulate from this zero-inflated Poisson model using the MLE parameters. Draw a histogram. Does it look like the raw data?

Summary: Models for distributions

We've now seen how to fit the most common distributional models:

- *Discrete:* Bernoulli/Binomial, Poisson, Negative Binomial
- *Continuous:* Normal (maybe after a log), Uniform, Exponential, Gamma

If maximum likelihood works, use MLE to find parameters (maybe after using MOM to find initial values.) Otherwise, use MOM.

We've also seen that we can manually make small modifications to our models, e.g. zero-inflation, and still use MLE.