Mathematical Statistics II

Introduction to Point Estimation

Jesse Wheeler

Contents

1 Introduction

Overview

- We will formally introduce the idea of point estimation.
- In addition to an introduction, we will introduce the concept of the empirical distribution, as well as methods of moment estimators.
- The material for this section largely comes from Chapter 8 of Rice (2007).

2 Point Estimation: An introduction

Point estimation

- In the previous lecture(s), we provided an example of Bayesian vs Frequentsist point-estimation via first principles.
- That is, using the various interpretations, we could reason an estimate for the probability p in a binomial experiment.
- We are now interested in studying approaches for more general cases.
- Given a dataset and a chosen model, how can we estimate parameters?
- We will first start with some notation, and motivating examples.
- Term *model* in this class will generally refer to a probability model, and can be based on a discrete or continuous probability measure.

  Normal Model

The Normal (or Gaussian) family of distributions arises often in the real world. Examples include human heights (conditioned on gender), rainfall amounts, and many biological measurements are approximately normal (or log-normal).

Given a set of observations $x_1, x_2, \ldots, x_n$, we may *model* these as iid normal $X_i \sim N(\mu, \sigma^2)$, and our goal being using the data to estimate the values of μ or σ.

Regression

Sometimes the probability model is *implicit*, but present. Consider the regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

We often think of fitting this regression model by minimizing the average squared-error: $(Y_i - \hat{Y}_i)^2$. However, this approach typically corresponds to an implicit probability model for the error terms ε_i, namely a normal distribution with mean 0. In this case, we might want to estimate β_0, β_1, and σ^2, which is $\text{Var}(\varepsilon_i)$.

Poisson Process

Another common example is a Poisson Process model. Many real-world phenomena are well-approximated by a Poisson process, over space or time. Examples include arrival times at a gas station, number of meteors landing in a geographic area, radioactive decay, etc. Here, there is only one parameter we want to estimate using data, namely the rate λ.

Parameter Estimation

- All of the above examples have the common feature that we pick a *model*, and we want to use the model to describe the data-generating process.
- More accurately, however, we pick a candidate *family* of models; (Gaussian family, Poisson Family, Linear Regression family, etc).
- Generally, the exact model needed within a *family* of models is determined by a few parameters.
  - If the family is Gaussian, the model is determined by μ and σ^2.
  - If the family is Poisson, the model is determined by λ.
  - If the family is linear-Gaussian regression, the model is determined by β_0, β_1, and σ^2.

Example: Gamma-Rainfall

- The Gamma distribution depends on two parameters, α and λ:

$$f_X(x; \alpha, \lambda) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x}.$$

- The Gamma distribution is quite flexible, and works as a useful model for various situations.
- One example is modeling rainfall amounts per-storm under two conditions, cloud seeding vs not cloud seeding (simulated data, couldn't find original data).
- A Gamma distribution fits both samples well, but we get different parameters α and λ for the two different samples
- Differences in the respective distributions are reflected in differences in the parameters α and λ.

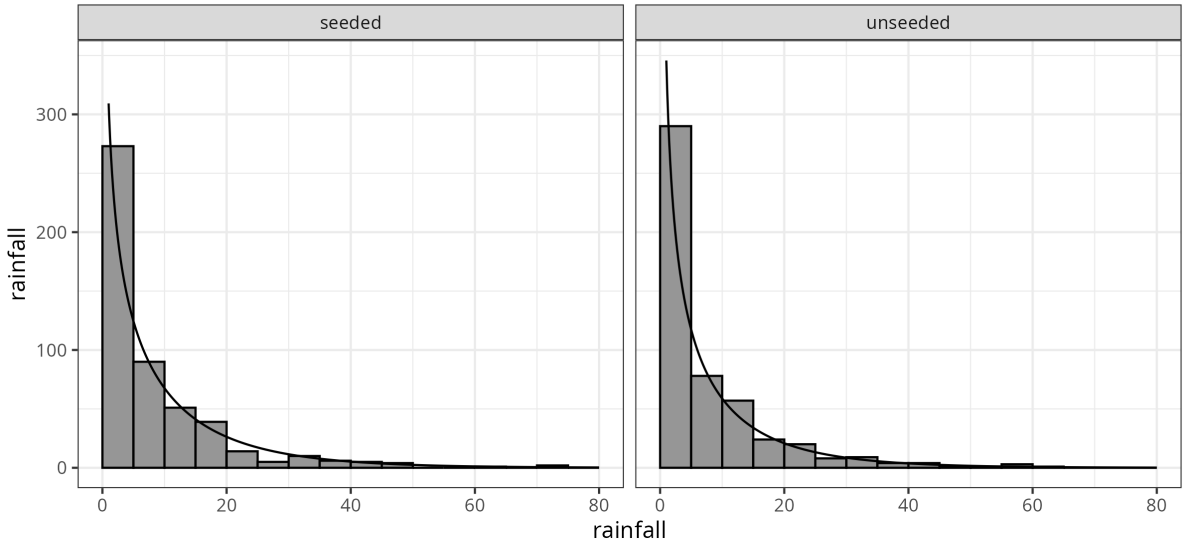


Figure 1: Data and model fit to two different Gamma distributions.

Two-sample Rainfall

Notation and generalizations

- We will generalize by using the following ideas and notations.
- We will denote the *observed data* as $x_1^*, x_2^*, \ldots, x_N^*$, and use the shorthands $x_{1:N}^*$ if we emphasize the entire collection, and x^* if the emphasis is not needed.
- We assume that the data are realizations of random variables $X_1, X_2, \ldots, X_N$, again using the notation $X_{1:N}$ for the collection of N random variables, or X if this is not needed.
- In general, the data x_i^* and random variables X_i can be multivariate, but focus primarily on the univariate case.
- We will be interested in fitting a probabilistic model $f_{X_{1:N}}(x_{1:N}; \theta)$ using the data. The model may correspond to a discrete probability, or a continuous probability. In these cases, f is usually a pmf of pdf, respectively.
- Subscripts will be dropped occasionally if it is not necessary. For instance, $f(x; \theta)$ is taken to mean the model of all data $x = x_{1:N}$, and would formally be expressed as $f_{X_{1:N}}(x_{1:N}; \theta)$.
- This approach is sometimes called "function overload"; it's not my favorite approach, but it is convenient. The meaning of the function is primarily understood by the arguments and context.
- The function $f(x; \theta)$ belongs to a particular *family* of models, indexed by θ, which is generally multivariate.

Normal model example
Suppose we observe the following data: 3.49, 2, 3.38, 1.62, 2.18, and we would like to fit a normal model to the data, assuming the data are iid. Then $x_1^* = 3.49$, $x_2^* = 2$, and so forth, and the model family

depends on $\theta = (\mu, \sigma^2)$, and the model can be expressed as:

$$\begin{aligned} f(x;\theta) &= f_{X_{1:5}}(x_{1:5};\mu,\sigma^2) \\ &= \prod_{i=1}^{5} f_{X_i}(x_i;\mu,\sigma^2) \\ &= \prod_{i=1}^{5} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2} \end{aligned}$$

Our goal is to estimate μ, σ^2 using the observed data $x^*_{1:5}$.

Notation and Generalization (continued)

- Our goal now is to develop general procedures for estimating θ, using observed data x^*, and a proposed family of models $f(x;\theta)$.
- We will develop three main approaches: (1) Method of Moments (2) Maximum Likelihood Estimation, and (3) Bayesian estimation.
- In this section, we will focus only on method of moments estimators.
- Once point estimation techniques are developed, we will provide theory about these estimates and their uncertainty; discussing bias, variance, and optimality of estimates.

3 Brief Introduction to R

R introduction

- Before we start looking at real-data examples, let's introduce some basic R coding principles that will help us calculate moments from the data.
- R is a programming language, but for the sake of this class, we'll just treat it as a statistics calculator.
- For now, we will only focus on the most simple data types and operations: creating objects, vectors, and computing summary statistics.
- First, saving objects in R. We can use `=` (like most languages), or the assignment operator: `<-`

```
x <- 2
x + 2

[1] 4
```

- A vector in R is a collection of objects of the same data type. In this class, we will only need to use numeric data types

```
x <- c(1, 2, 3, 4, 5)
class(x)

[1] "numeric"
```

```
mean(x)

[1] 3

sum(x)

[1] 15
```

- Some fast ways of building vectors include:

```
1:5  # this gives 1, 2, 3, 4, 5

[1] 1 2 3 4 5

seq(1, 10, by = 2)  # Gives 1, 3, 5, 7, 9

[1] 1 3 5 7 9
```

- For generating random numbers, we can use the syntax: `rdist`.

```
rnorm(n = 10, mean = 2, sd = 1)

 [1] 1.7531041 0.7844391 3.5614051 2.4273102 0.7989765 3.0524585
 [7] 0.6949364 1.3073924 2.6026489 1.8022469

rpois(n = 7, lambda = 5)

[1] 2 6 1 5 5 9 6

rbeta(n = 3, shape1 = 0.8, shape2 = 1.3)

[1] 0.51652672 0.10386537 0.05986089
```

- Lastly (and maybe most important), function documentation and help is readily available by appending a question mark: `?rnorm`

```
?mean
?rnorm
?sd
```

4 Method of Moments

Motivation

- The Method of Moments (MoM) estimation technique is a simple idea.
- Pick a family of models $f(x;\theta)$, and observed data x^*.
- The family of models will have theoretical moments, i.e., $\mathrm{E}[X^k]$.
- Generally, these moments can be expressed in terms of the model parameters, θ.
- Thus, we will estimate $\hat{\theta}$ so that the *data moments* match the theoretical moments.

The empirical distribution

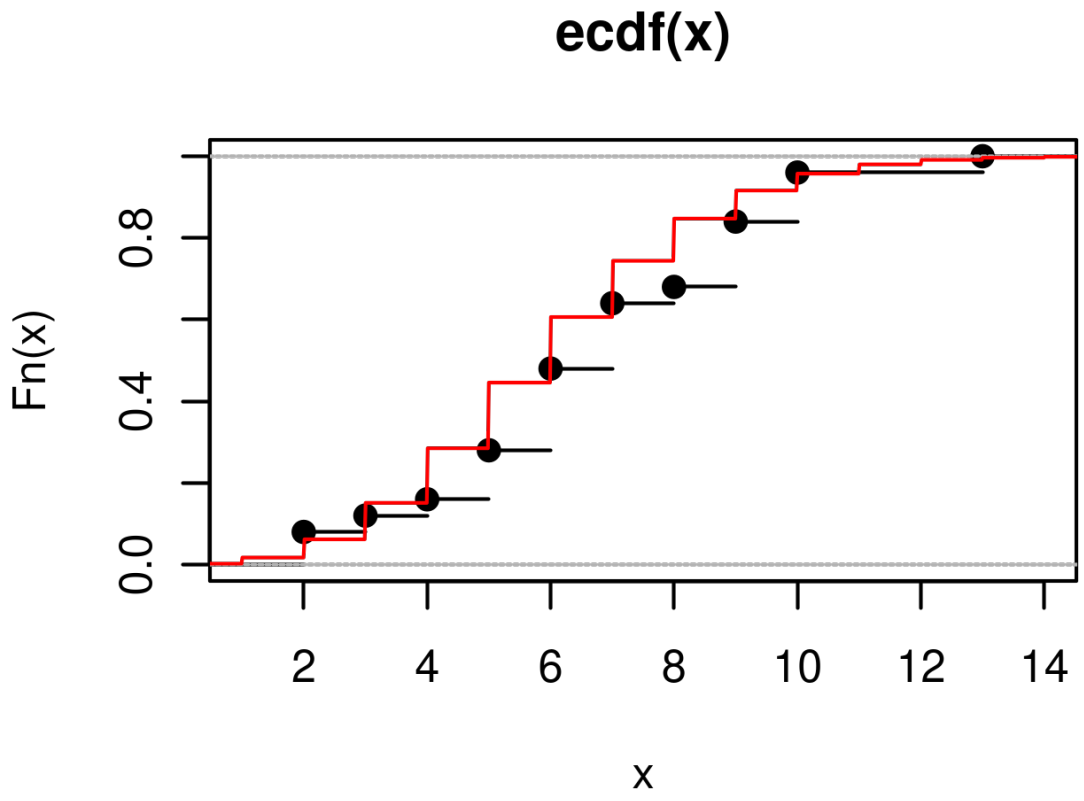
- One justification of this approach considers the *empirical distribution* of observed data.
- Let $X_1, X_2, \ldots, X_N$ be random variables, representing a possible data sample.
- We will assume that X_i are iid, from some distribution F_θ (F_θ is the cdf here).
- We will define the empirical distribution function as:

$$F_n(t) = \frac{1}{N} \sum_{i=1}^{N} I[X_i \leq t].$$

- When we observe a specific dataset x^*, we can plug in these numbers to get a specific distribution that is not random.
- A few things to note is that $F_n(t)$ is a proper CDF.
- By the law of large numbers, $F_n(t) \xrightarrow{a.s.} F_\theta(t)$ for every point t.
- The Glivenko–Cantelli theorem also strengthens this statement by saying that the convergence is uniform, in the sense that $\sup_t |F_n(t) - F_\theta(t)|$ converges to zero.

Example: empirical distribution function for Poisson Data:

```
set.seed(123)  # Reproducible results
x <- rpois(n = 25, lambda = 6)
plot(ecdf(x))
lines(
  x = seq(1e-16, 15, length.out = 1000),
  y = ppois(seq(1e-16, 15, length.out = 1000), 6),
  col = 'red'
)
```



Method of Moments Estimation

- It can be shown that the kth moment of the empirical distribution is

$$\hat{\mu}_k = \frac{1}{N}\sum_{i=1}^{N} X_i^k.$$

- Method of Moments idea:
  - For many commonly used parametric families (e.g., Gaussian, Poisson), the distribution is completely specified by a small set of parameters.
  - These parameters are typically explicit functions of the moments of the distribution (e.g., mean and variance for the Gaussian).
  - Although the moment generating function (MGF) uniquely determines the entire distribution, in many model families, the relevant parameters are uniquely determined by just the first few moments.
  - Therefore, as the empirical moments computed from data converge to the true moments (by the Law of Large Numbers), it is natural to estimate model parameters by equating empirical and theoretical moments—leading to the method of moments estimators.

Method of Moments: generalized version

- To summarize mathematically, let $\mu_k = E[X^k]$ be the theoretical kth moment.
- Let $\hat{\mu}_k = \frac{1}{N}\sum_{i=1}^{N} X_i^k$ be the kth sample moment.
- $\hat{\mu}_k$ is an estimate of μ_k; however, we don't want an estimate of μ, we want an estimate of θ!
- For models with finite parameters, $\theta = (\theta_1, \ldots, \theta_k)$, we can often express θ_i as a function of $(\mu_1, \ldots, \mu_k)$:
$$\theta_i = g_i(\mu_1, \ldots, \mu_k)$$
.
- Thus, our estimate of θ_i would be found by plugging in the empirical moments:
$$\hat{\theta}_i = g_i(\hat{\mu}_1, \ldots, \hat{\mu}_k).$$

Examples

Poisson Distribution
Suppose we observe data $x^*_{1:N}$, and want to fit a Poisson model. Since the Poisson distribution only has one parameter (λ), our goal is to use x^* to estimate λ.
The first moment of the Poisson distribution is $\mu_1 = E[X_i] = \lambda$. Thus, the function $g_1(\mu_1) = \mu_1 = \lambda$, and our estimate should be
$$g_1(\hat{\mu}_1) = \frac{1}{N}\sum_{i=1}^{N} X_i = \hat{\lambda}.$$

Real-data example

Poisson distribution with real data
The National Institute of Science and Technology collected data about asbestos fibers on filters. Asbestos dissolved in water was spread on a filter, and the number of fibers in each of 23 grid squares were counted:

```
 [1] 31 29 19 18 31 28 34 27 34 30 16 18 26 27 27 18 24 22
[19] 28 24 21 17 24
```

- From our previous work, we know that the method of moments estimator for the Poisson Distribution is just
$$\hat{\lambda} = \frac{1}{N}\sum_{i=1}^{N} X_i = \bar{X}_N$$
.
- For this specific dataset, we can calculate that in `R` using the `mean` function.
- I have the data saved in a vector `x`, already, so I get the result: `mean(x) =` 24.91.

The mean can be calculated as:

```
mean(x)

[1] 24.91304
```

- What about the error associated with this estimate?

Sampling Distribution

- As always, we are interested in the the uncertainty related to our estimates.
- In most cases, we cannot directly calculate uncertainty of estimates, and we will have to rely on more advance theory, discussed later.
- Sometimes, however, we can calculate some form of uncertainty based on the form of the estimator, and model assumptions.
- The last (and next) models are such cases.

Sampling Distribution
Most estimates $\hat{\theta}$ of θ are functions of the random variables $X_1, X_2, \ldots, X_N$. Thus, $\hat{\theta}$ is also a random variable. The distribution of $\hat{\theta}$ is called the *sampling distribution*.

- In most cases, the exact distribution of $\hat{\theta}$ is unknowable.
- Instead, we often get approximations to this distribution, and in particular, the variance of the distribution, in order to quantify uncertainty of the estimator.
- For the Poisson model and the method of moments estimator, however, we can calculate this exactly.

Sampling distribution of Poisson MoM estimator
Let $X_1, \ldots, X_N$ be modeled as iid from a Poisson(λ) distribution. If $\hat{\lambda}$ is the method of moments estimator of λ, what is its sampling distribution?

- From our previous work, we found the estimate to be:

$$\hat{\lambda} = \frac{1}{N}\sum_{i=1}^{N} X_i,$$

  or the sample average.

- Last semester, you proved that the sum of n independent Poisson(β) random is Poisson($n\beta$). Using this information, and assuming X_i is Poisson(λ), we have:

$$S = \sum_{i=1}^{N} X_i \sim \text{Poisson}(\lambda n).$$

- Thus, the estimator $\hat{\lambda} = S/n$ is a transformed Poisson random variable, such that:

$$\begin{aligned} P(\hat{\lambda} = k) &= P(S = Nk) \\ &= \frac{(N\lambda)^{Nk} e^{-N\lambda}}{(Nk)!}, \end{aligned}$$

  for all non-negative integers k. This defines the *sampling distribution* of the MoM estimator $\hat{\lambda}$.

- We can also calculated the expected value and variance of $\hat{\lambda}$, using the fact that S is Poisson:

$$E(\hat{\lambda}) = \frac{1}{N}E(S) = \lambda, \qquad \text{Var}(\hat{\lambda}) = \frac{1}{N^2}\text{Var}(S) = \frac{\lambda}{N}.$$

- A few things to note:

  - The estimate is unbiased: $E[\hat{\lambda}] = \lambda$.
  - The variance shrinks at a rate of $1/N$.
  - The CLT says that $\hat{\lambda}$ (which is a sample mean), is approximately normally distributed.
  - The standard deviation of a sampling distribution is called the *standard error* of the estimator.
- We can't know the exact sampling distribution, because it depends on the true value λ.
- However, we can approximate the sampling distribution by substituting $\hat{\lambda}$ for λ.
- This is unbiased, and the mean-square-error is then the sum of squared bias, and variance. The variance also decreases linearly is N, so it should be a good approximation even for moderate N.

For our specific dataset, we can approximate the standard error as:

```
sqrt(mean(x) / length(x)) |> round(2)
```

```
[1] 1.04
```

The "pipe" operator in R `|>` takes the output of one function, and inputs it as the first argument into the next function.

Example: Model Checking

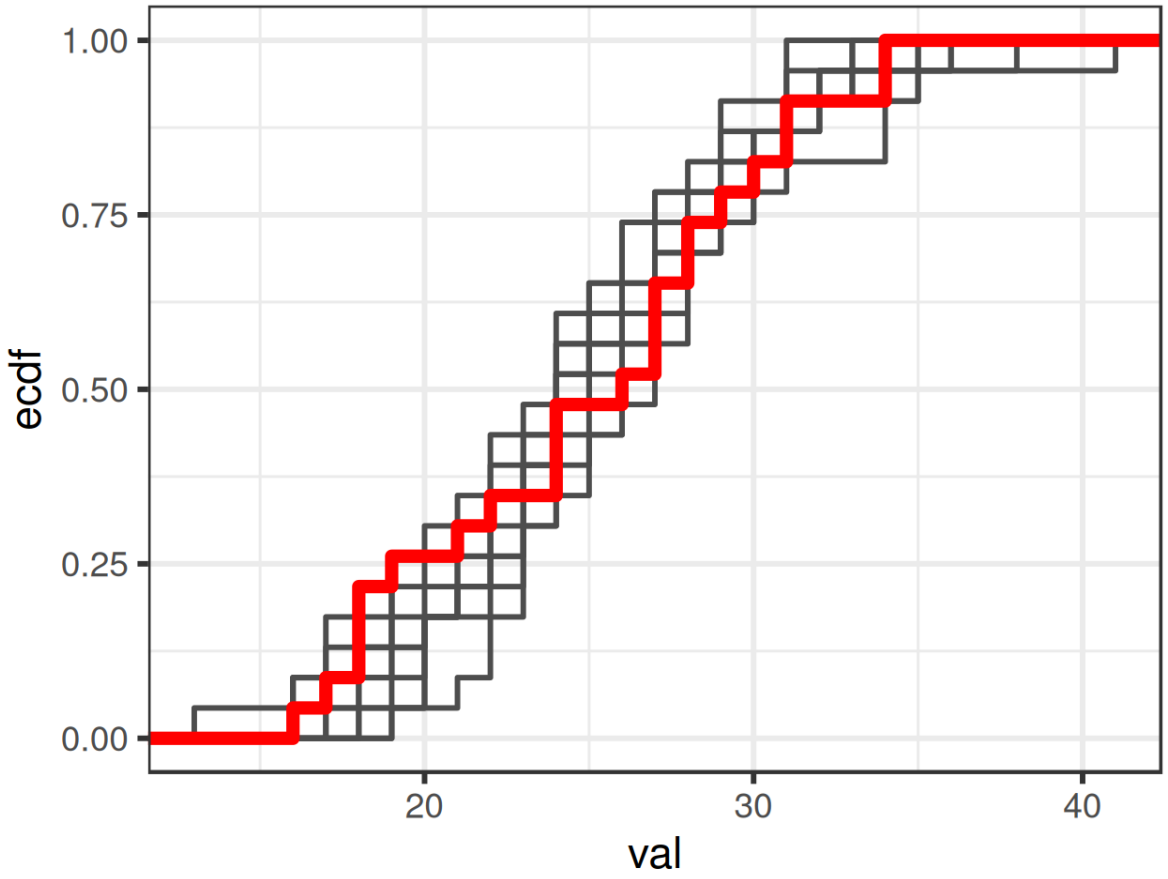
- In addition to estimating the stochastic uncertainty due to sampling distribution, we might want to assess inductive uncertainty due to model selection.
- This is a large topic, one we will revisit in much more detail later.
- For now, I want to present a simple idea that can be done with what we already know.
- This will be based on a *parametric bootstrap*, a technique we will discuss in detail later in the course.
- The idea: we claim that the data come from a Poisson(λ), and we estimated λ.
- If the data really are Poisson(λ), then our estimate $\hat{\lambda} \approx \lambda$.
- We can simulate many different data from Poisson($\hat{\lambda}$), and compare this to our real data.
- If the model is reasonable, then our data shouldn't look too different from the simulations.
- We can compare the empirical distribution functions from the real data, and simulations.

ECDF: R code

```
library(tidyverse)
lambda <- mean(x)
results <- replicate(10, rpois(length(x), lambda = lambda)) |>
  as.data.frame()
colnames(results) <- paste0("sample_", 1:10)
results |>
  pivot_longer(
    cols = everything(), names_to = "replicate",
    names_prefix = "sample_", values_to = "val"
  ) %>%
  ggplot(aes(val, group = replicate)) +
```

```
  stat_ecdf(geom = "step", col = 'grey30') +
  stat_ecdf(
    data = data.frame(val = x, replicate = '0'),
    geom = 'step', col = 'red', linewidth = 1.2
  ) + theme_bw()
```

Plot output



Example: Normal Distribution

Normal Distribution
Suppose we observe N observations, and we want to model them as iid $N(\mu, \sigma^2)$. Find the method of moments estimator $\theta = (\mu, \sigma^2)$.

- In this case, the θ that indexes the model is two-dimensional.
- Under the assumption that the data X_i are iid normal, we have the following theoretical moments:

$$\mu_1 = E[X] = \mu$$
$$\mu_2 = E[X^2] = \mu^2 + \sigma^2$$

- In this case, we need to find functions g_1 and g_2 such that:

$$\theta_1 := \mu = g_1(\mu_1 = \mu, \mu_2 = \mu^2 + \sigma^2)$$
$$\theta_2 := \sigma^2 = g_2(\mu_1 = \mu, \mu_2 = \mu^2 + \sigma^2)$$

- Here, the functions are obvious (indeed, you can basically skip the step above, but I wanted to show how the method fits with the more general notation / approach).

- Specifically, we have:

$$\mu = \mu_1$$
$$\sigma^2 = \mu_2 - \mu_1^2.$$

- Replacing this with sample moments $\hat{\mu_1} = \frac{1}{n}\sum_i X_i$, and $\hat{\mu_2} = \frac{1}{n}\sum_i X_i^2$, we have:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i = \bar{X}_N,$$

and

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i X_i^2 - \big(\frac{1}{N}\sum_{i=1}^{N} X_i\big)^2.$$

A little algebra shows that the second equation can be written as:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} \big(X_i - \bar{X}_N\big)^2.$$

- Note that this is another case where we can get the sampling distribution of the estimators.

- Specifically,

$$\hat{\mu} = \bar{X}_N \sim N(\mu, \sigma^2/N),$$

and

$$n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-1}.$$

- Furthermore, $\hat{\mu}$ and $\hat{\sigma^2}$ are independent.

Example: Gamma Distribution

MoM estimator for Gamma distribution
Consider modeling the data $X_1, \ldots, X_N$ as iid Gamma(α, λ). Find the method of moments estimator for these data.

- The first step is to find the first few moments of a Gamma distribution.

- Last semester, we derived the moment generating function for this distribution; this can be used to calculate the first few moments: $\mu_1 = E[X_i]$, $\mu_2 = E[X_i^2]$.

$$\mu_1 = \frac{\alpha}{\lambda}$$
$$\mu_2 = \frac{\alpha(\alpha+1)}{\lambda^2}.$$

- To apply the method of moments estimation approach, we need to express the parameters α and λ in terms of these moments. Using a substitution method, we can find:

$$\mu_2 = \frac{\alpha(\alpha+1)}{\lambda^2} = \frac{\alpha^2}{\lambda^2} + \frac{\alpha}{\lambda^2}.$$

  Plugging in the expression for μ_1:

$$\mu_2 = \mu_1^2 + \frac{\mu_1}{\lambda}.$$

  Now we want to solve for α and λ, so first solving for λ:

$$\lambda = \frac{\mu_1}{\mu_2 - \mu_1^2}.$$

  Now solving for α, we can use the first moment equation to get:

$$\alpha = \lambda\mu_1 = \frac{\mu_1^2}{\mu_2 - \mu_1^2}.$$

- Now we just need to replace the theoretical moments with the sample moments!

$$\hat{\lambda} = \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2}, \qquad \hat{\alpha} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}.$$

- Now we can simplify a bit. The first data moment $\hat{\mu}_1$ is always just the sample mean, $\bar{X}_N$. As we saw in the last example, we also have:

$$\hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{N}\sum_{i=1}^{N} X_i^2 - \left(\bar{X}_N\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2,$$

  which is the (biased) sample variance. For convenience, we'll denote:

$$\hat{\sigma}^2 := \hat{\mu}_2 - \hat{\mu}_1^2.$$

  Thus, the estimates can be expressed as:

$$\hat{\lambda} = \frac{\bar{X}_N}{\hat{\sigma}^2}, \qquad \hat{\alpha} = \frac{\bar{X}_N^2}{\hat{\sigma}^2}.$$

- Unlike the previous examples, it's not immediately clear how to estimate the sampling distributions in this case. We will discuss approaches to this later in the semester.

Example: Muon Decay

MoM estimator for Muon Decay
In *statistical physics*, the angle ϕ at which electrons are emitted in muon decay is modeled using the following density:

$$f(x;\,\alpha) = \frac{1+\alpha x}{2}, \quad -1 \le x \le 1,$$

and where the parameter α satisfies $-1 \le \alpha \le 1$, and $x = \cos\phi$. Supposing we observe data $X_1, X_2, \ldots, X_N$, what is the method of moments estimator for α?

- As always, we will first need to find the theoretical moments of this distribution, and relate them to the parameters of interest.

- The first moment, denoted $\mu_1 = E[X_i]$ is found via integration:

$$\mu_1 = \int_{-1}^{1} x \frac{1+\alpha x}{2} dx.$$

  Some basic calculus gives:

$$\mu_1 = \alpha/3.$$

- Thus, the parameter of interests is described as a function of the theoretical moments via the equation:

$$\alpha = 3\mu_1.$$

- Replacing theoretical moments with sample moments provides our estimate:

$$\hat{\alpha} = 3\hat{\mu}_1 = 3\bar{X}_N.$$

Acknowledgments

- Compiled on January 15, 2026 using R version 4.5.2.

- We acknowledge students and instructors for previous versions of this course / slides.

References

Rice JA (2007). *Mathematical statistics and data analysis*, volume 371. 3 edition. Thomson/Brooks/Cole Belmont, CA. 1