Data and Statistics, Maximum Likelihood

Math 10A

November 16, 2017

Substitution

Ribet in Providence on AMS business

Today's class conducted by GSI Teddy Zhu

What is statistics?

Statistics is the science of collection, organization, and interpretation of data.

Typical procedure:

- Start with a question
- Collect relevant data
- Analyze the data
- Make inferences based on the data

Statistical inference is the heart of modern statistics.

Much of statistical reasoning is built on the mathematical framework that we've spent the last couple of weeks on – probability theory.

Sampling

Typically we are interested in a population – for example, all Berkeley students, all registered voters in the U.S., all blue whales, all bacteria of a certain species.

It's often unreasonable to collect data for every member of a large population, so we work with a **sample** of the population.

When collecting a sample, it's important to avoid **sampling bias**, which is when the method of sampling systematically fails to represent the population.

Example: Posting a survey on Facebook groups to determine how much sleep Berkeley students get (selection bias, response bias)

Simple random sample (SRS): each member of the population has the same chance of being selected for the sample (gold standard of sampling)

Data

Data can be **quantitative** or **qualitative**.

- Quantitative: height, weight, temperature, age, time, ...
- Qualitative: eye color, ethnicity, gender, college major, whether or not someone smokes, ...

Qualitative data can sometimes be meaningfully converted into something quantitative: for example, in a sample of adults, we could assign a 1 to those who voted in the last election, and a 0 to individuals who did not.

Why is this a meaningful thing to do?

- The average of the 1's and 0's is equal to the proportion of people in the sample who voted.
- Used in advanced linear models (not in this class)

Analyzing quantitative data

Suppose we have some quantitative data from our sample, in the form of a list of values: $x_1, x_2, x_3, \ldots, x_n$.

What are some ways of analyzing this data?

- **sample mean**: $\bar{x} = \dfrac{x_1 + x_2 + \ldots + x_n}{n}$
- **sample variance**: How spread out is the data in the sample? How does this relate to the population variance?
- Does the data appear to belong to a certain **family of probability distributions**? If so, can we determine or estimate the **parameter(s)** of the distribution?

The third bullet point is the topic of the rest of this class meeting.

Some families of probability distributions

- **Uniform distribution**: $f(x|a,b) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$

  Parameters: a, b $(a < b)$

- **Pareto distribution**: $f(x|p) = \begin{cases} \dfrac{p-1}{x^p} & \text{if } x \ge 1 \\ 0 & \text{if } x < 1 \end{cases}$

  Parameter: $p > 1$ ("shape parameter")

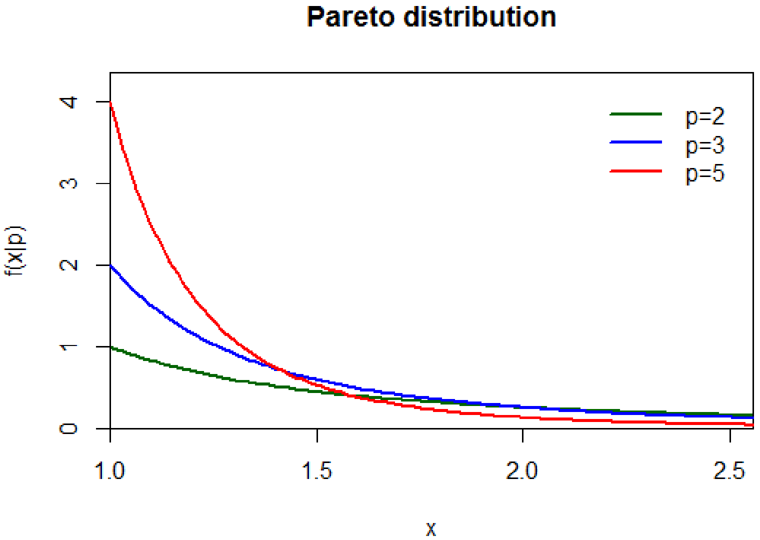
- **Exponential distribution**: $f(x|c) = \begin{cases} ce^{-cx} & \text{if } x \ge 0 \\ 0 & \text{if } x < 0 \end{cases}$

  Parameter: $c > 0$ ("rate parameter")

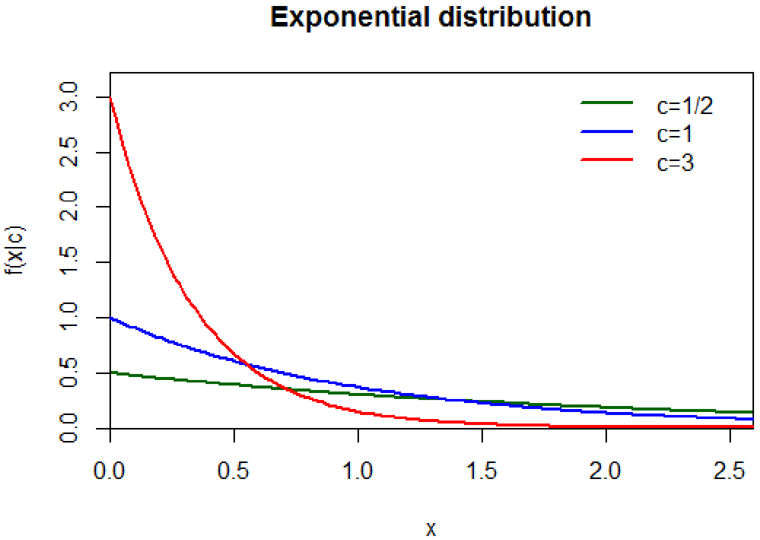
- **Normal distribution**: $f(x|\mu,\sigma) = \dfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

  Parameters: μ (mean), $\sigma > 0$ (standard deviation)

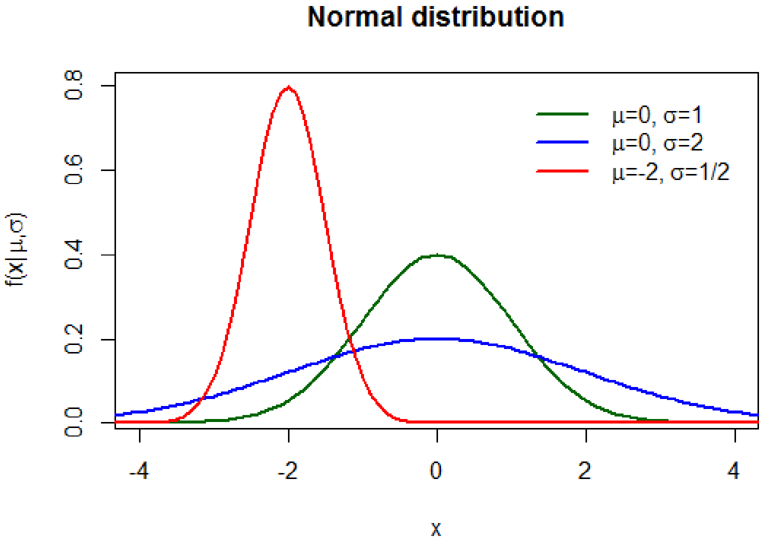
Visual: Pareto distribution



Visual: Exponential distribution

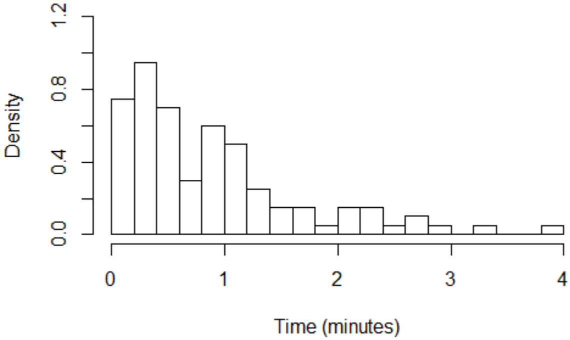


Visual: Normal distribution



Example 1

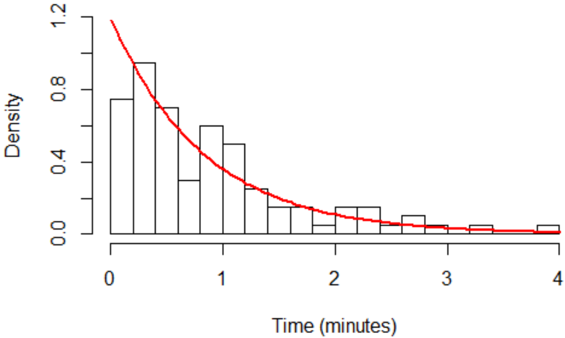
A popular fast food chain wishes to optimize service. To do this, they collect data on the time between customers entering the restaurant (*interarrival times*) during prime time hours. Below is a **histogram** of 100 observations:



In a histogram, the area of the rectangle above an interval $[a, b]$ is the proportion of data lying within the interval $[a, b)$.

Example 1

Below is the same histogram, now with the PDF of an exponential distribution with some parameter drawn in red.



Looks like a good fit! It makes sense to model the interarrivals times of customers as exponentially distributed random variables.

How do we determine *which* exponential distribution?

Likelihood function

Given some observed data $x_1, x_2, \ldots, x_n$ and a family of (continuous) probability distributions with PDFs $\{f(x|\theta)\}$ (θ is the parameter), the **likelihood function** is

$$\boxed{L(\theta) = L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta) = f(x_1|\theta) \cdot f(x_2|\theta) \cdots f(x_n|\theta)}$$

The likelihood function measures the relative likelihood of observing the observed data, assuming that the data comes from the specified probability distribution with parameter θ.

Likelihood function

Remarks:

- The likelihood function does not represent a probability for continuous distributions (It's not always between 0 and 1.)
- The product form of the likelihood function is derived from the underlying assumption that each observation is *independent* and *identically distributed*.
- The likelihood function is a function of θ, not $x_1, x_2, \ldots, x_n$ (common misconception!) The x_i's are observed data, so they're all just constants – we just write them as x_i for the sake of generality.

Maximum Likelihood Estimation

Idea: We can estimate the value of the parameter by finding the value of θ which *maximizes* the likelihood function $L(\theta)$.

This makes sense intuitively: Associated to each possible value of θ is a "likelihood" of observing the observed data *if* the true value of the parameter were equal to θ. The idea is to pick the value of θ which makes this likelihood as large as possible, and use that as our estimate of the population parameter.

Definition: The **maximum likelihood estimator** of θ, denoted $\hat{\theta}_{MLE}$, is the value of θ which maximizes $L(\theta)$.

Back to Example 1

We decided that the exponential distribution is a suitable model for the interarrival times of customers.

Recall: An exponential distribution with parameter $c > 0$ has PDF

$$f(x|c) = \begin{cases} ce^{-cx} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

If our observed values are $x_1, x_2, \dots, x_{100}$, then the likelihood function is

$$L(c) = \prod_{i=1}^{100} f(x_i|c) = \prod_{i=1}^{100} ce^{-cx_i} = c^{100} e^{-c \sum_{i=1}^{100} x_i}.$$

Maximizing the likelihood function

How can we find c which maximizes $L(c) = c^{100}e^{-c\sum_{i=1}^{100} x_i}$?

Remember calculus? – take a derivative, set it equal to 0, and solve for c.

WAIT! In most cases, it's not so easy to take the derivative of the likelihood function, because the likelihood function is a product of a bunch of functions, and derivatives of products aren't so clean. (In the particular case of Example 1, it's actually not too bad.)

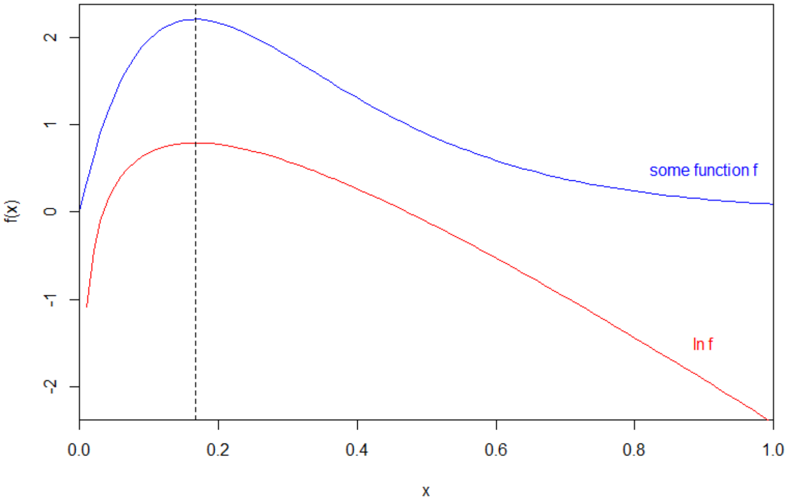
Technique: Maximize the **log likelihood function** instead:

$$\boxed{\ell(\theta) = \log L(\theta)}$$

Fact: θ^* maximizes $L(\theta)$ if and only if θ^* maximizes $\ell(\theta)$.

This is because the function $g(x) = \ln(x)$ is an **increasing** function.

Visual: log of a function



Back to Example 1

The log-likelihood function in Example 1 is

$$\ell(c) = \ln L(c) = \ln(c^{100} e^{-c \sum_{i=1}^{100} x_i}) = 100 \ln c - c \sum_{i=1}^{100} x_i.$$

The derivative is

$$\ell'(c) = \frac{d}{dc}\ell(c) = \frac{100}{c} - \sum_{i=1}^{100} x_i.$$

Setting $\ell'(c)$ equal to 0 and solving for c yields the following formula for the maximum likelihood estimator for c:

$$\boxed{\hat{c}_{MLE} = \frac{100}{\sum_{i=1}^{100} x_i} = \frac{1}{\overline{x}}}$$

Remark

Remark: The mean of an exponential distribution with parameter c is $\frac{1}{c}$. The result of our computations above says that the maximum likelihood estimator of the parameter given the sample data is simply 1 divided by the sample mean.

That makes sense!

Consistency of MLE (not on exam)

The idea of maximum likelihood makes sense intuitively, but is it always a good way to estimate a population parameter?

It is a theorem in statistics that in fact, the maximum likelihood estimator $\hat{\theta}_{MLE}$ *converges* to the true population parameter θ_0 as the size of the sample n tends to ∞.

Any estimator that satisfies the property above is called a **consistent** estimator.

Example 2 (discrete)

Suppose you are given a biased coin which lands heads with some probability p_0 (unknown).

You decide to flip the coin 200 times, and you want to find the maximum likelihood estimator of p_0.

Let $X_i = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ flip is heads} \\ 0 & \text{if } i^{\text{th}} \text{ flip is tails} \end{cases}$

For a coin that lands heads with probability p, $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$.

There's a clever way to write this **probability mass function**:

$$P(X_i = x) = p^x (1 - p)^{1-x}$$

Check!

Likelihood for discrete distributions

Given a discrete distribution (only integer values possible) with probability mass function $f(k|\theta) = P(X = k|\theta)$ depending on a parameter θ, the likelihood function given a set of observed values $x_1, x_2, \ldots, x_n$ is defined to be

$$L(\theta) = L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta) = f(x_1|\theta) \cdot f(x_2|\theta) \cdots f(x_n|\theta).$$

This looks exactly the same as the previous formula for continuous distributions...

The difference is that in the continuous case f is the PDF, and in the discrete case f is the PMF.

In the discrete case, the likelihood function actually is the probability of observing the observed values, assuming that the data comes from the specified probability distribution with parameter θ.

Back to Example 2

You flip the coin 200 times, and your observed values are $x_1, x_2, \ldots, x_{200}$, where $x_i = 1$ if the i^{th} flip was heads, and $x_i = 0$ if the i^{th} flip was tails (the x_i's are observed values, so they are not random.)

The likelihood function is

$$L(p) = \prod_{i=1}^{200} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{200} x_i}(1-p)^{200-\sum_{i=1}^{200} x_i}.$$

Remember, the goal is to find the value of p which maximizes the expression above.

Example 2

As before, we can make our lives easier by instead maximizing the log-likelihood function.

First, let $s = \sum_{i=1}^{200} x_i$. The log-likelihood function is

$$\ell(p) = \ln L(p) = \ln\left(p^s(1-p)^{200-s}\right) = s\ln p + (200-s)\ln(1-p).$$

The derivative is

$$\ell'(p) = \frac{d}{dp}\ell(p) = \frac{s}{p} - \frac{200-s}{1-p}.$$

Setting $\ell'(p)$ equal to 0 and solving for p gives us

$$\boxed{\hat{p}_{MLE} = \frac{s}{200} = \frac{\sum_{i=1}^{200} x_i}{200} = \bar{x}}$$ (= proportion of heads in sample).

Remark

We found that the maximum likelihood estimator for p in this problem is actually just the proportion of heads in the sample – makes sense again.

Since the MLE is a consistent estimator, that means the proportion of heads in the sample converges to p_0, the coin's true chance of heads.

... but we already knew that, by **the law of large numbers**.

Beat Stanfurd

2009 Golden Bears posing with the Axe