# Week 11 Part I: Parameter Estimation

Yilin

DigiPen Institute of Technology

# STATISTICS

# Statistics

Statistics is a very broad field of study that is mostly concerned with the analysis and interpretation of **data**. The goals of this chapter are:

- Study inferential statistics - how to interpret data
- Apply probability theory to reconstruct the underlying probability distribution that governs a given data set.
- Estimate (using probability theory) unknown parameters corresponding to such distributions
- Measure “confidence” that our estimate is correct given the data.
- Learn how to discard “bad” hypotheses when running experiments
- Not fall victim to statistical fallacies (Important!)

# PARAMETER ESTIMATION

# Example 1: Toxins

Suppose a town's water supply is contaminated by a potentially toxic substance.

- Blood tests are administered to 100 residents.
- 22 residents tested positive for the presence of the substance in their blood

**Question**

What is a good way to estimate the proportion, $p$, of people who have the substance present in their blood?

One estimate we can make is the average:

$$\hat{p} = \frac{\text{\# people testing positive}}{\text{\# people tested}} = 0.22$$

**Question**

Can we determine to what extent this estimate is accurate?

# Definitions

First we will introduce some notation.

## Definition (Sample, Population, Parameter, Estimator)

- Measurements are made on a **sample** of a **population**.
- The (unknown) probability $p$ we seek is a **parameter** for a model of the population.
- Any value $\hat{p}$ that we construct based off the data is an **estimator** based on the sample for the true value of $p$.

# Biased vs Unbiased

- In general, we cannot determine the true value of a parameter. Instead we seek a good enough **approximation**.
- If we wish to determine the accuracy of the estimator $\hat{p}$, like in the example, we need to consider **bias.**

## Definition

An estimate $\hat{p}$ for a parameter with true value $p$ is **unbiased** if

$$E(\hat{p}) = p,$$

and is **biased** if $E(\hat{p}) \neq p$. Note that $\hat{p}$ is an r.v. and is endowed with a user-defined probability distribution.

# Some things to keep in mind

- In the example, we computed one value of $\hat{p}$.
- In reality we can only record one outcome of this experiment.
- In theory, the number 0.22 is one outcome from the sample space of the r.v. $\hat{p}$, which is endowed with some probability distribution.
- We do not know this distribution, so it is whatever we make it given the problem.
- In terms of the Law of Large Numbers, if we could repeat this experiment infinitely (not 100) many times we would be able to determine the **true value** of $p$: that is the true average of the r.v. $\hat{p}$.
- Being unbiased tells us that it conforms to this rule, that $E(\hat{p}) = p$.

**Note**

Being biased isn't necessarily "bad" and is often useful when making predictions.

# Measuring Bias

**Question**

How can we determine if an estimator is biased or unbiased?

Back to the example,

- Let $X$ be the r.v. representing the number of individuals in a 100 person sample who test positive for the toxin.
- Let $\hat{p}$ correspond to the estimator determined by the results for a given sample, i.e.

$$\hat{p} = \frac{X}{100}.$$

- Now $\hat{p}$ is more obviously an r.v. than it was before.

# Observations vs R.V.s

What happened earlier was we *observed* the event

$$(X = 22)$$

Remember, that doesn't mean that $X = 22$ always, this just means that from one experiment we measured an outcome of 22. Moreover, we observed the event

$$\hat{p} = \frac{22}{100} = 0.22.$$

Remember, $\hat{p}$ is an **estimator** of the parameter $p$, the true proportion of the **whole** population in which the toxin is present.

# Set-up

Now we can check whether the estimator $\hat{p}$ is biased or not. Let us compute $E(X)$ in terms of the true proportion $p$.

**Note**

If $p$ is the **true proportion** (i.e. true probability that an individual has the toxin), then $X$ is binomially distributed (also assuming each individual was independently selected).

# Using the Binomial Distribution

In this case, $X$ is binomial with $n = 100$, $p = p$ (the true probability) and $k =$ the number of individuals in the sample testing positive, then

$$P(X = k) = \binom{100}{k} p^k (1-p)^{100-k}.$$

We know that

$$E(X) = np = 100p,$$

therefore

$$E(\hat{p}) = E\left(\frac{X}{100}\right) = \frac{E(X)}{100} = \frac{100p}{100} = p.$$

We have shown that, under certain assumptions (independence, mostly) that

$$E(\hat{p}) = p,$$

so we can conclude that this estimator is **unbiased**.

# Summary

- We tested a sample size of 100, yielding $k = 22$ positive results.
- By "common sense" we set the estimator $\hat{p}$ for the true proportion of the population that is positve, to be

$$\hat{p} = \frac{k}{100}.$$

- We then used probability theory (understanding of the binomial distribution) and an assumption of independence to determine that $E(\hat{p}) = p$.
- That is, the estimator $\hat{p}$ is unbiased without knowing an exact value for $p$.

There are more ways than just "common sense" to find estimators..

# MAXIMUM LIKELIHOOD ESTIMATORS

# Maximum Likelihood Estimation

## Idea

- Suppose we have a probabilistic model dependent on an unknown "true" parameter $p$.
- Suppose further that we are given observable data.
- We wish to determine the **likelihood** (probability) of that data occurring as a function of the unknown $p$.

# Definition: Likelihood Function

Define

$$\begin{aligned} L(p) &:= \text{Probability of the observed data occuring given the value of } p. \\ &= \text{Probability of the observed data conditional on the value of } p \\ &= P(\text{Observed data} \mid p). \end{aligned}$$

We call $L(p)$ a **likelihood function**.

**Note**

While $L(p)$ is a probability we use the term **likelihood** to distinguish this setting from that of formal probabilistic models. That is, likelihood is used to try to determine the unknown underlying probability model whereas formal probabilities are determined when the model is known.

# Example 2: Comparing Likelihoods

Back to the blood test example, recall that our sample size was 100 and 22 tested positive. If $p$ is the true probability, then we concluded

$$L(p) = P(X = 22) = \binom{100}{22} p^{22}(1-p)^{78}.$$

If the true value for $p$ is, say, 0.2, then

$$L(p) = L(0.2) = \binom{100}{22} (0.2)^{22}(.8)^{78} \approx 0.0849.$$

If the true value for $p$ is 0.24, then

$$L(p) = L(0.24) = \binom{100}{22} (0.24)^{22}(.76)^{78} \approx 0.0857.$$

If the true value for $p$ is in fact 22, then

$$L(0.22) = \binom{100}{22} (0.22)^{22}(.78)^{78} \approx 0.0959.$$

So $p = 0.22$ is **more likely** than $p = 0.2$ or $p = 0.24$.

# Maximizing Likelihoods

We want to maximize the likelihood of our estimators. Thus..

## Definition

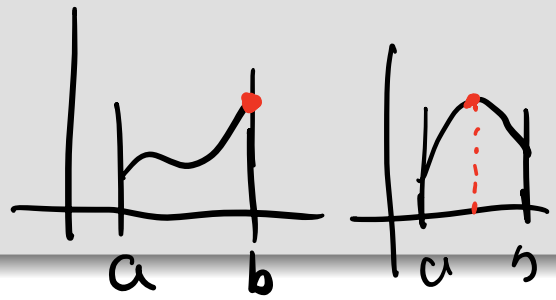
Given a likelihood function $L(p)$, **maximum likelihood estimator** of the parameter $p$ is the value of $p$ for which $L(p)$ takes its **maximum value**.

Recall

## Theorem

*If a differentiable function $f$ defined on $[a, b]$ takes on its maximum value $f(x)$ at a point $x$, then one of the following is true:*

- *$x$ is a critical point, i.e. where $f'(x) = 0$, or*
- *$x$ is an endpoint, either $a$ or $b$.*




To find the point $p$ where $L(p)$ attains its maximum, we find its values at critical points and endpoints.

# Maximizing Likelihood in Example 2

In the previous example,

$$L(p) = \binom{100}{22} p^{22}(1-p)^{78}.$$

Since $L(p)$ is a function defined for $0 \le p \le 1$, if it is *differentiable* on this interval we can use calculus to find its maximum value. We compute

$$\begin{aligned} L'(p) &= \frac{\mathrm{d}}{\mathrm{d}p} L(p) \\ &= \binom{100}{22} \frac{\mathrm{d}}{\mathrm{d}p} (p^{22}(1-p)^{78}) \\ &= \binom{100}{22} (22p^{21}(1-p)^{78} - 78p^{22}(1-p)^{77}) \end{aligned}$$

product rule

# ..Cont'd

$L'(p) = 0$ when

$$0 = (22p^{21}(1-p)^{78} - 78p^{22}(1-p)^{77}) = p^{21}(1-p)^{77}(22(1-p) - 78p).$$

This can only be zero when $p = 0$, $p = 1$ or

$$22(1-p) - 78p = 22 - 100p = 0 \implies p = 0.22.$$

Now we check, which is the greatest: $L(0)$, $L(0.22)$, or $L(1)$? We can check

$$\begin{aligned} L(0) &= \binom{100}{22} 0^{22}(1-0)^{78} = 0 \\ L(1) &= \binom{100}{22} 1^{22}(1-1)^{78} = 0 \\ L(0.22) &= \binom{100}{22} (0.22)^{22}(0.78)^{78} \approx 0.0959. \end{aligned}$$

# Checking the values

So the **maximum likelihood estimator** is

$$\hat{p} = 0.22,$$

as we computed earlier.

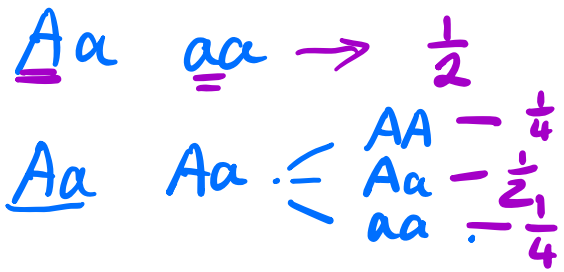
## Note

- We used a more mathematical method to compute $\hat{p}$ than our first guess, which is what we used before.
- While the likelihood is maximized, note that $L(0.22) \approx 0.10$ so the likelihood that the probability of the observed data occurring is only about 10%.

# Example 3: Puppies

Suppose a male and female dog produce 100 (independent) puppies so that

- Each puppy is homozygous of type **aa**
- The mother is **Aa**.
- The father is unknown: there are only two candidates for the father, **Aa** or **aa**.
- Let $F_1$ be the candidate that is of type **aa** and $F_2$ be the candidate of type **Aa**.




**Goal**

We seek the candidate that has the highest likelihood of occurring.

By observation, we might guess that $F_1$ has the highest likelihood. The method we are about to outline is a way to formalize this.

# Likelihood of $F_1$

First, we can compute $L(F_1)$ as

$$\begin{aligned} & L(F_1) \\ = \; & P(\text{all puppies } \mathbf{aa}|F_1) \\ = \; & P(\text{Puppy 1 is } \mathbf{aa}|F_1) \cdot P(\text{Puppy 2 is } \mathbf{aa}|F_1) \cdots P(\text{Puppy 100 is } \mathbf{aa}|F_1) \\ = \; & (0.5) \cdots (0.5) \\ = \; & (0.5)^{100}. \\ \approx \; & 7.89 \times 10^{-31}. \end{aligned}$$

Note that we used

$$\begin{aligned} P(\text{Puppy } i \text{ is } \mathbf{aa}|F_1) &= P(\mathbf{a} \text{ from } \mathbf{Aa} \text{ mother}) \cdot P(\mathbf{a} \text{ from } \mathbf{aa} \text{ father}) \\ &= (0.5)(1) = 0.5. \end{aligned}$$

# Likelihood of $F_2$

On the other hand $L(F_2)$ is

$$\begin{aligned} & L(F2) \\ = \; & P(\text{all puppies } \mathbf{aa}|F_2) \\ = \; & P(\text{Puppy 1 is } \mathbf{aa}|F_2) \cdot P(\text{Puppy 2 is } \mathbf{aa}|F_2) \cdots P(\text{Puppy 100 is } \mathbf{aa}|F_2) \\ = \; & (0.25)^{100} \\ \approx \; & 6.22 \times 10^{-61}. \end{aligned}$$

Note that we used

$$\begin{aligned} P(\text{Puppy } i \text{ is } \mathbf{aa}|F_2) &= P(\mathbf{a} \text{ from } \mathbf{Aa} \text{ mother}) \cdot P(\mathbf{a} \text{ from } \mathbf{Aa} \text{ father}) \\ &= (0.5)(0.5) = 0.25. \end{aligned}$$

# Conclusion

- While both $L(F_1)$ and $L(F_2)$ are practically zero, we see that $F_1$ yields a higher likelihood.
- So $F_1$ is more likely thus the father is more likely to be **aa** than **Aa**.

The above example was nice because we only needed to compare the likelihoods of two parameters.

**Question**

What if we consider continuous parameters, like the rate $\lambda$ from an exponential distribution?

# Example 4: Soda

A soda manufacturer wants to estimate the time it takes for their soda to go flat. They test 10 identical bottles of soda and get the following times (in days) that they go flat:

$$
\begin{aligned}
t_1 &= 0.908 \\
t_2 &= 0.088 \\
t_3 &= 1.764 \\
t_4 &= 0.619 \\
t_5 &= 0.038 \\
t_6 &= 0.575 \\
t_7 &= 0.539 \\
t_8 &= 1.196 \\
t_9 &= 0.791 \\
t_{10} &= 0.311
\end{aligned}
$$

# Constructing a Likelihood Function

The average waiting time is

$$\overline{T} = 0.6833.$$

If the waiting time $T$ is exponentially distributed with parameter $\lambda$, then we wish to find a maximum likelihood estimator for $\lambda$. We compute

$$L(\lambda) = \lambda e^{-\lambda t_1} \cdot \lambda e^{-\lambda t_2} \cdots \lambda e^{-\lambda t_{10}} = \lambda^n e^{-\lambda 10 \cdot \frac{t_1+t_2+\cdots+t_{10}}{10}} = \lambda^n e^{-10\lambda \overline{T}}$$

**Note**

This is a product of the **probability densities** and not an actual probability.

The probability of an individual data point occurring is 0 in a continuous model. What we are doing here is maximizing the probability density rather than the probability itself.

# More Calculus

Proceeding, we use calculus to find the $\lambda$ where $L(\lambda)$ attains its maximum. We compute, noting $\overline{T} = 0.6833$ and $10\overline{T} = 6.833$,,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda}L(\lambda) &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\lambda^{10}e^{-\lambda(6.833)}\right)\\ &= 10\lambda^{9}e^{-\lambda(6.833)} - ((6.833))\lambda^{10}e^{-\lambda(6.833)}\\ &= \lambda^{9}e^{-\lambda(6.833)}\left(10 - (6.833)\lambda\right)\end{aligned}$$

Thus $L'(\lambda) = 0$ if and only if

$$0 = 10 - 6.833 \cdot \lambda \implies \lambda = \frac{10}{6.833} \implies \lambda \approx 1.4634.$$

# Finding $\hat{\lambda}$

Since $0 \leq \lambda < \infty$ we only need to test the critical point 1.4634 and the endpoint 0,

$$L(0) = 0^{10}e^{-0} = 0,$$

and

$$L(1.4634) = (1.4634)^{10}e^{-10(1.4634)(0.6833)} \approx 0.00205.$$

So 1.4634 gives us approximately the maximum likelihood, hence 1.4634 is the "most likely" value of $\lambda$, hence we set

$$\hat{\lambda} = 1.4634.$$

# Example 5: Disconnections

Suppose we have a computer network where

- 8 users experience a disconnection.
- Assume that the number of users experiencing disconnections follows a Poisson distribution with an unknown parameter $\Lambda$ (the average number of disconnections per user).
- We seek a maximum likelihood estimator for $\Lambda$.

If $N$ is the number of disconnections, from the Poisson distribution we have a likelihood function in

$$L(\Lambda) = P(N = 8) = \frac{e^{-\Lambda}\Lambda^8}{8!}.$$

# Solution

To find the maximum likelihood,

$$\begin{aligned} L'(\Lambda) &= \frac{\mathrm{d}}{\mathrm{d}\Lambda}(L(\Lambda)) \\ &= \frac{\mathrm{d}}{\mathrm{d}\Lambda}\left(\frac{e^{-\Lambda}\Lambda^8}{8!}\right) \\ &= \frac{e^{-\Lambda}8\Lambda^7 - e^{-\Lambda}\Lambda^8}{8!} \\ &= \frac{e^{-\Lambda}\Lambda^7}{8!}(8-\Lambda). \end{aligned}$$

Thus the critical points of $L(\Lambda)$ are $\Lambda = 0$ and $\Lambda = 8$. So the maximum likelihood estimator for $\Lambda$ is

$$\hat{\Lambda} = 8,$$

distributed

$$P(N = k) = \frac{e^{-8}8^k}{k!}.$$