50.034 - Introduction to Probability and Statistics

Week 8 – Lecture 13

January–May Term, 2019

SINGAPORE UNIVERSITY OF
TECHNOLOGY AND DESIGN

Lecture 13 Assigned Readings: 7.1, 7.2

Outline of Lecture

- Statistical models
- Parameter space and parameters as random variables
- Statistical inference, statistic
- Bayesian philosophy versus frequentist philosophy
- Prior and posterior distributions
- Likelihood function

Uncertainty in real-world random processes

A lighting company has designed a new light bulb model. They are interested in finding out how long each light bulb is likely to last.

- The lifespan of every light bulb is random!
- To model the lifespan, we need to make some assumptions.

Model Assumptions: The lifespan (in hours) of every light bulb follows an exponential distribution with parameter λ. All light bulbs share the same parameter λ. We do not know the value of λ.

Model Setup: Let $X_1, X_2, \ldots$ be a sequence of iid R.V.'s, each having an exponential distribution with parameter λ.

- We know that $\mathbf{E}[X_i] = \frac{1}{\lambda}$.
- By the law of large numbers, the sample mean $\overline{X}_n \xrightarrow{p} \frac{1}{\lambda}$.
- Hence, if we take a sufficiently large random sample of sample size n, then the sample mean $\overline{X}_n$ is approximately $\frac{1}{\lambda}$, and we can use $\frac{1}{\overline{X}_n}$ as an estimate for the parameter λ.

Uncertainty in real-world random processes

Note: No matter how large our sample size n is, we can never find out the precise value of λ.

- Based on the actual observed values for $X_1, \dots, X_n$, the value of $\frac{1}{\overline{X}_n}$ that we compute is only an approximate value for λ.
- Hypothetically, to find the precise value of λ, we would need to observe the values of the entire infinite sequence $X_1, X_2, \dots$

Thus, we say that the parameter λ is hypothetically observable.

In contrast, each X_k is an observable R.V., since we can carry out an experiment to observe the value of X_k (lifespan of k-th bulb).

A non-observable R.V. that could be inferred from observable R.V.'s (e.g. functions of observable R.V.'s) is called a latent R.V.

- Hypothetically observable R.V.'s are latent R.V.'s.

The model setup (all observable and latent R.V.'s), together with all model assumptions, is usually called a **statistical model**.

Statistical model

Definition: A statistical model consists of the following:

- A collection of R.V.'s $\{X_1, X_2, X_3, \dots\}$ (could be finite or infinite)
  - These R.V.'s could be observable or latent.
- A family of possible joint distributions for observable R.V.'s.
  - e.g. iid with exponential distribution
- Assumptions on the parameters of the joint distributions.
  - e.g. parameter λ is unknown (but hypothetically observable).

Important Idea: Given any unknown parameter λ, we could treat λ as a random variable, and carry out experiments to draw some conclusions about λ.

- Since all R.V.'s have distributions, it then makes sense to consider the distribution of λ.
  - i.e. "distribution of parameter of distribution" makes sense!
- Hence, in any statistical model, it is important to specify whether an unknown parameter is an unknown constant, or a random variable (as well as whether the distribution of the parameter is known or not).

Parameter Space

The parameters of a distribution are numerical attributes whose values determine the distribution completely.

- We have already seen many examples of parameters:
  - Binomial distribution with parameters n and p.
  - Poission distribution with parameter λ.
  - Normal distribution with parameters μ and σ.
  - Bivariate normal distribution with parameters $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho$.
- Each parameter could be treated as a known constant, an unknown constant, a R.V. whose distribution is known, a R.V. whose distribution is unknown, etc.

Given any parameter θ, the set of all possible values for θ is called the parameter space of θ.

- What is considered "possible" depends on the context.
- If μ is the mean of a normal distribution representing the average height (in cm) of a person, then we could take the parameter space of μ to be the interval $[0, 300]$.

Parameters as random variables

Light Bulb Example: $X_1, X_2, \ldots$ is a sequence of iid R.V.'s, each having an exponential distribution with parameter λ.
If we treat λ as a R.V., then the parameter space of λ is the set of all positive real numbers.

- If we assume that the lifespan of every light bulb must be > 1 hour, then since $\frac{1}{\lambda}$ represents the average lifespan, we could restrict the parameter space to just the interval $(0, 1)$.

This new perspective of treating parameters of distributions as random variables opens up **new questions**:

- e.g. what is the conditional probability that $\lambda \leq 0.002$, given the observed values for the random sample $\{X_1, \ldots, X_{100}\}$?
    - Such a question can be interpreted as: What is the probability that the actual average light bulb lifespan is ≥ 500 hours, given the lifespans of 100 randomly selected light bulbs?
- If we had considered λ as an unspecified constant, then the question doesn't quite make sense:
    - Either $\lambda \leq 0.02$ with probability 1, or $\lambda \leq 0.02$ with probability 0, depending on the actual value of λ.

Statistical inference

Given a statistical model, we can make statistical inferences.

Definition: A statistical inference is any procedure that produces a probabilistic statement concerning the statistical model.

- Typically involves making inferences or conclusions based on experimental data.

Examples of different kinds of statistical inferences:

- Estimation: e.g. observe the values of a large random sample, compute the sample mean, and use the computed value to approximate the parameter of the distribution. (more in later lectures..)
- Constructing confidence intervals: e.g. using the observed values of a large random sample, find a suitable interval (a, b) in $\mathbb{R}$, such that the population mean μ is contained in the interval (a, b) with 95% confidence. (more in later lectures..)
- Hypothesis Testing: e.g. given a threshold α, and given a hypothesis that the population mean μ satisfies $\mu > \alpha$, use the observed values of a large random sample to decide whether to accept or reject the hypothesis. (more in later lectures..)

Statistic

Definition: Let $\mathcal{S} = \{X_1, \ldots, X_n\}$ be a set of n observable R.V.'s. A statistic of $\mathcal{S}$ is a function of the R.V.'s in $\mathcal{S}$.

- Note: A statistic is a random variable! Different observed values for $X_1, \ldots, X_n$ give different values for the statistic.
- If $h(x_1, \ldots, x_n)$ is a real-valued function in terms of n variables, then $h(X_1, \ldots, X_n)$ is a statistic.
- More generally, for any algorithm with the observed values for $X_1, \ldots, X_n$ as input, and whose output is a numerical value (i.e. a real number), the output of the algorithm is a statistic.
- Examples: sample mean, $\max(X_1, \ldots, X_n)$, $\min(X_1, \ldots, X_n)$.

Interpretation: A statistic is a descriptive summary of some given set of observable R.V.'s. For example, if our set is a random sample, then a statistic (e.g. sample mean, max value, min value) gives us a good representation of the R.V.'s.

- A statistic is much easier to interpret, as compared to raw data (e.g. a list of all observed values of the R.V.'s).

Light Bulb Example Revisited

Statistical Model:

- $X_1, X_2, \ldots$ is a sequence of iid observable R.V.'s, each having an exponential distribution with parameter λ.
- λ is a R.V. whose parameter space is the interval $(0, 1)$.

To do computations with this statistical model, we first have to specify the distribution of λ. (Otherwise we cannot even start any calculations!)

Question: What if we do not know the distribution of λ?

- Note: λ is hypothetically observerable, not observable!

Answer: We could start with an initial guess, i.e. a "prior distribution".

- maybe λ has the uniform distribution on $(0, 1)$.
- maybe $\lambda = 0.002$ (i.e. distribution given by $\Pr(\lambda = 0.002) = 1$).

As we sequentially observe the values of the observable R.V.'s $X_1, X_2, \ldots$, we get more information about how likely our "prior distribution" describes the actual distribution of λ.

Recall: Prior and posterior probabilities (Lecture 3)

Fair coin versus biased coin: Your friend has two coins, a fair coin, and a biased coin that always gives heads. He randomly selects one of the coins, and asks if the selected coin is fair.

- Let A be the event "selected coin is fair".
- Your *prior* guess: $\Pr(A) = 0.5$. Without more information, you have no reason to favour A (coin is fair) or A^c (coin is biased).
- You toss the coin 10 times and record all 10 outcomes.

Suppose the event $B =$ "all heads for 10 tosses" occurs.

Event B would strongly suggest that the selected coin is NOT fair. How should you update your guess, given that B has occurred?

In other words, what should $\Pr(A|B)$ be?

Gathering experimental evidence to check your prior guess is common practice. In such a scenario, $\Pr(A)$ is called the prior probability, and $\Pr(A|B)$ is called the posterior probability.

Prior and posterior distributions

Consider a statistical model with observable R.V.'s $X_1, \ldots, X_n$. Let θ be a parameter (possibly one of many parameters) of the joint distribution of $X_1, \ldots, X_n$, and treat θ as a random variable.

The prior distribution of θ is the initial distribution specified for θ.

- This is the distribution we specify before observing any data (i.e. before gathering the observed values for $X_1, \ldots, X_n$)
- Sometimes "prior distribution" is simply called "prior".

After we have some observed values, say $X_1 = x_1, \ldots, X_n = x_n$, then the conditional distribution, consisting of all conditional probabilities of the form $\Pr(\theta \in C | X_1 = x_1, \ldots, X_n = x_n)$ (over all possible $C \subseteq \mathbb{R}$), is called the posterior distribution of θ.

Interpretation: The prior distribution of θ is the initial guess for the distribution of θ, while the posterior distribution of θ is the updated guess, after taking into account experimental evidence, i.e. the observed values $X_1 = x_1, \ldots, X_n = x_n$.

Bayesian philosophy

(Lecture 3) The Bayesian philosophy is based on Bayes' theorem. The main idea of this philosophy is that the probability of a random event can be **updated** with new evidence, as follows:

- The event of interest (your hypothesis, e.g. "Medicine A is better than Medicine B") is assigned a prior probability.
- As we gather experimental evidence, we update our guess on how likely the hypothesis is true with the posterior probability.
- If A is the event of interest, and B is the event representing experimental evidence, then $\Pr(A)$ is the prior probability, and $\Pr(A|B)$ is the posterior probability.
- The posterior probability $\Pr(A|B)$ can then be computed using Bayes' theorem.

In other words: As you get new information, you update your belief on how likely a given hypothesis is true.

Bayesian philosophy versus frequentist philosophy

Same probability theory, but different interpretations.

Bayesian philosophy

- Probabilities can be assigned to both data and hypotheses (e.g. hypothesis: "Medicine A is better than Medicine B").
  - e.g. there is 80% probability that Medicine A is better than Medicine B.
- Requires a prior for computing probabilities of hypotheses. Probabilities can be updated with new information.
  - e.g. after some clinical trials, it is concluded that there is 95% probability that Medicine A is better than Medicine B.

Frequentist philosophy

- Probabilities are assigned only to data, not hypotheses.
  - either Medicine A is better than Medicine B, or Medicine A is not better than Medicine B.
- Probabilities represent the limiting relative frequencies of the outcomes of an experiment as you repeat the experiment infinitely many times. Hypotheses are not repeatable.
  - No priors are needed; hypotheses don't have probabilities.

Bayesian philosophy versus frequentist philosophy

While a few statisticians argue over which philosophy is "correct" or "better", most other statisticians just use methods from both. In this course, we shall learn methods from both philosophies.

- Many complicated real-world problems require a mix of both kinds of methods, depending on what data is available, what experiments can be done, and how much computing power is available.

Bayesian methods:

- Popular in the 19th century. Popular again in the 21st century (especially in machine learning, robotics, genetics).
- Tends to be computationally more intensive
  - Parameters of distributions are R.V.'s, not constants.

Frequentist methods:

- Popular in the 20th century (especially in life sciences).
- Tends to be computational less intensive.
  - Parameters of distributions are constants.

Prior distributions: A closer look

Consider a statistical model with observable R.V.'s $X_1, \ldots, X_n$. Suppose θ is a parameter of the joint distribution of $X_1, \ldots, X_n$, where θ is treated as a random variable.

- If θ is discrete, then the pmf of θ is called the prior pmf of θ.
- If θ is continuous, then the pdf of θ is called the prior pdf of θ.

Note on terminology: In the course textbook, the symbol ξ is commonly used to denote the pmf/pdf of a parameter θ.

- The parameter space of θ is sometimes denoted by Ω.
- This is because the sample space of θ (as a R.V.) can be taken to be the parameter space itself.
  - In other words, the parameter space is both the sample space and the set of possible values.
- Hence, the distribution of θ is an assignment of probabilities to all subsets of the parameter space of θ.

Note: The pmf/pdf of θ is usually written as $\xi(\theta)$.

- So θ is also used as the variable of the function $\xi = \xi(\theta)$.

Example 1

Fair coin versus biased coin: Your friend has two coins, a fair coin, and a biased coin that always gives heads. He randomly selects one of the coins, and asks if the selected coin is fair.

- Let A be the hypothesis "selected coin is fair".
- Suppose your initial guess is $\Pr(A) = 0.8$.
- You toss the coin 10 times and record all 10 outcomes.

Statistical model:
Let $X_1, \ldots, X_{10}$ be iid Bernoulli random variables with parameter θ, where $X_i = 1$ if the i-th coin toss is heads, and $X_i = 0$ otherwise. The parameter θ is a discrete R.V. whose **prior pmf** is

$$\xi(\theta) = \begin{cases} 0.8, & \text{if } \theta = 0.5; \\ 0.2, & \text{if } \theta = 1; \\ 0, & \text{otherwise.} \end{cases}$$

Example 2

Light Bulb Example: The lifespan (in hours) of every light bulb follows an exponential distribution with parameter λ.

All light bulbs share the same parameter λ, where λ is a R.V. whose distribution is unknown to us.

- We shall assume the parameter space of λ is $(0, 1)$.
- Initial guess: λ has uniform distribution on $(0, 1)$.
- We shall measure the lifespans of 1000 light bulbs.

Statistical model:
Let $X_1, \ldots, X_{1000}$ be iid exponential R.V.'s with parameter θ, where each X_i represents the lifespan (in hours) of the i-th light bulb.
The parameter θ is a continuous R.V. whose **prior pdf** is

$$\xi(\theta) = \begin{cases} 1, & \text{if } 0 < \theta < 1; \\ 0, & \text{otherwise.} \end{cases}$$

Remark on notation

Example: The exponential R.V. X with parameter θ has pdf

$$f(x) = \begin{cases} \theta e^{-\theta x}, & \text{if } x \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

To indicate that the pdf is conditional on the given value of θ, we write the pdf as $f(x|\theta)$, to remind us that it is a **conditional pdf**. If $X_1, \ldots, X_n$ are iid exponential R.V.'s, each with parameter θ, then the joint conditional pdf given the value of θ is

$$\begin{aligned} f(x_1, \ldots, x_n|\theta) &= f(x_1|\theta) \cdots f(x_n|\theta) \\ &= \begin{cases} \theta^n e^{-\theta(x_1 + \cdots + x_n)}, & \text{if } x_i \geq 0 \text{ for all } i; \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

We could further simplify notation by writing $f_n(\mathbf{x}|\theta)$ or $f(\mathbf{x}|\theta)$, where the boldfaced $\mathbf{x}$ represents $(x_1, \ldots, x_n)$.

- Similarly, we write $p_n(\mathbf{x}|\theta)$ or $p(\mathbf{x}|\theta)$ in the discrete case.

Posterior distributions: A closer look

Consider a statistical model with observable R.V.'s $X_1, \ldots, X_n$. Suppose θ is a parameter of the joint distribution of $X_1, \ldots, X_n$, where θ is a discrete or continuous R.V. with pmf/pdf $\xi(\theta)$.

- If θ is discrete, then the posterior pmf of θ is the conditional pmf of θ given $X_1 = x_1, \ldots, X_n = x_n$.
- If θ is continuous, then the posterior pdf of θ is the conditional pdf of θ given $X_1 = x_1, \ldots, X_n = x_n$.
- In either case, the pmf/pdf is denoted by $\xi(\theta|x_1, \ldots, x_n)$, or more simply, $\xi(\theta|\mathbf{x})$.

$$\boxed{\text{posterior distribution}} = \boxed{\text{conditional distribution of the parameter given the data}}$$

Note: When we write "the posterior pdf of θ is $\xi(\theta|\mathbf{x})$", the two θ's have different meanings.

- The first θ is a R.V.
- The second θ is a variable of a function.

Calculating posterior distributions using Bayes' theorem

Consider a statistical model with observable R.V.'s $X_1, \ldots, X_n$. Suppose θ is a parameter of the joint distribution of $X_1, \ldots, X_n$, where θ is a R.V. with parameter space Ω.

Theorem: (**Bayes' theorem**+**Law of total probability** for R.V.'s)

- If $X_1, \ldots, X_n$ are **discrete** with joint conditional pmf $p_n(\mathbf{x}|\theta)$ and marginal joint pmf $p(\mathbf{x})$, and if θ is **discrete** with prior pmf $\xi(\theta)$, then the posterior pmf of θ is

$$\xi(\theta|\mathbf{x}) = \frac{p_n(\mathbf{x}|\theta)\xi(\theta)}{p(\mathbf{x})} = \frac{p_n(\mathbf{x}|\theta)\xi(\theta)}{\sum_{\theta' \in \Omega} p_n(\mathbf{x}|\theta')\xi(\theta')} \quad (\text{for } \theta \in \Omega).$$

- If $X_1, \ldots, X_n$ are **discrete** with joint conditional pmf $p_n(\mathbf{x}|\theta)$ and marginal joint pmf $p(\mathbf{x})$, and if θ is **continuous** with prior pdf $\xi(\theta)$, then the posterior pdf of θ is

$$\xi(\theta|\mathbf{x}) = \frac{p_n(\mathbf{x}|\theta)\xi(\theta)}{p(\mathbf{x})} = \frac{p_n(\mathbf{x}|\theta)\xi(\theta)}{\int_{\Omega} p_n(\mathbf{x}|\theta')\xi(\theta')\, d\theta'} \quad (\text{for } \theta \in \Omega).$$

Calculating posterior distributions (continued)

Consider a statistical model with observable R.V.'s $X_1, \ldots, X_n$.
Suppose θ is a parameter of the joint distribution of $X_1, \ldots, X_n$, where θ is a R.V. with parameter space Ω.

Theorem: (**Bayes' theorem**+**Law of total probability** for R.V.'s)

- If $X_1, \ldots, X_n$ are **continuous** with joint conditional pdf $f_n(\mathbf{x}|\theta)$ and marginal joint pdf $f(\mathbf{x})$, and if θ is **discrete** with prior pmf $\xi(\theta)$, then the posterior pmf of θ is

$$\xi(\theta|\mathbf{x}) = \frac{f_n(\mathbf{x}|\theta)\xi(\theta)}{f(\mathbf{x})} = \frac{f_n(\mathbf{x}|\theta)\xi(\theta)}{\sum_{\theta' \in \Omega} f_n(\mathbf{x}|\theta')\xi(\theta')} \quad (\text{for } \theta \in \Omega).$$

- If $X_1, \ldots, X_n$ are **continuous** with joint conditional pdf $f_n(\mathbf{x}|\theta)$ and marginal joint pdf $f(\mathbf{x})$, and if θ is **continuous** with prior pdf $\xi(\theta)$, then the posterior pdf of θ is

$$\xi(\theta|\mathbf{x}) = \frac{f_n(\mathbf{x}|\theta)\xi(\theta)}{f(\mathbf{x})} = \frac{f_n(\mathbf{x}|\theta)\xi(\theta)}{\int_{\Omega} f_n(\mathbf{x}|\theta')\xi(\theta')\,d\theta'} \quad (\text{for } \theta \in \Omega).$$

Example 3

Same scenario as in Example 1: Your friend has two coins, a fair coin, and a biased coin that always gives heads. He randomly selects one of the coins, and asks if the selected coin is fair.

- Initial guess: Pr("selected coin is fair") $= 0.8$.
- You toss the coin 10 times and record all 10 outcomes.

Statistical model:
Let $X_1, \ldots, X_{10}$ be iid Bernoulli random variables with parameter θ, where $X_i = 1$ if the i-th coin toss is heads, and $X_i = 0$ otherwise. The parameter θ is a discrete R.V. whose prior pmf is

$$\xi(\theta) = \begin{cases} 0.8, & \text{if } \theta = 0.5; \\ 0.2, & \text{if } \theta = 1; \\ 0, & \text{otherwise.} \end{cases}$$

Suppose all 10 tosses give heads, i.e. $X_1 = \cdots = X_{10} = 1$.

Question: What is the **posterior pmf** of θ?

Example 3 - Solution

Solution: First, note that the conditional pmf of X_i is

$$p(x_i|\theta) = \begin{cases} \theta^{x_i}(1-\theta)^{1-x_i}, & \text{if } x_i = 0 \text{ or } 1; \\ 0, & \text{otherwise;} \end{cases}$$

Using Bayes' theorem (for R.V.'s) and the law of total probability (for R.V.'s), the posterior pmf of θ is

$$\xi(\theta|\mathbf{x}) = \frac{p_n(\mathbf{x}|\theta)\xi(\theta)}{\sum_{\theta' \in \Omega} p_n(\mathbf{x}|\theta')\xi(\theta')} \quad (\text{for } \theta \in \Omega),$$

where $p_n(\mathbf{x}|\theta)$ is the joint conditional pmf of $X_1, \ldots, X_{10}$ given by:

$$\begin{aligned} p_n(\mathbf{x}|\theta) &= p(x_1|\theta) \cdots p(x_{10}|\theta) \\ &= \begin{cases} \theta^{(x_1+\cdots+x_{10})}(1-\theta)^{10-(x_1+\cdots+x_{10})}, & \text{if every } x_i = 0 \text{ or } 1; \\ 0, & \text{otherwise;} \end{cases} \end{aligned}$$

Example 3 - Solution (continued)

We are given that $x_1 = \cdots = x_{10} = 1$, hence

$$p_n(\mathbf{x}|\theta) = p(1, \dots, 1|\theta) = \theta^{10}.$$

Since the possible values for θ are 0.5 and 1, we then get

$$\begin{aligned}\sum_{\theta' \in \Omega} p_n(\mathbf{x}|\theta')\xi(\theta') &= p_n(\mathbf{x}|0.5)\xi(0.5) + p_n(\mathbf{x}|1)\xi(1) \\ &= (\tfrac{1}{2^{10}})(0.8) + 0.2.\end{aligned}$$

Therefore, the posterior pmf of θ (given $X_1 = 1, \dots, X_{10} = 1$) is

$$\xi(\theta|\mathbf{x}) = \frac{p_n(\mathbf{x}|\theta)\xi(\theta)}{\sum_{\theta' \in \Omega} p_n(\mathbf{x}|\theta')\xi(\theta')} = \begin{cases} \frac{(\frac{1}{2^{10}})(0.8)}{(\frac{1}{2^{10}})(0.8)+0.2}, & \text{if } \theta = 0.5; \\ \frac{0.2}{(\frac{1}{2^{10}})(0.8)+0.2}, & \text{if } \theta = 1; \\ 0, & \text{otherwise}; \end{cases}$$

Example 3 - Solution (continued)

After simplifying, we get that the posterior pmf of θ is

$$\xi(\theta|\mathbf{x}) \approx \begin{cases} 0.003891, & \text{if } \theta = 0.5; \\ 0.9961, & \text{if } \theta = 1; \\ 0, & \text{otherwise.} \end{cases}$$

Our updated distribution for θ now looks very different.

- Originally, we guessed that "coin is fair" with 80% probability.
- With our experimental evidence (10 heads for 10 tosses), our updated guess becomes "coin is biased" with 99.6% probability.

Sensitivity Analysis

Question: What if we started with a different prior distribution?

- If we began with a different prior distribution, could the posterior distribution be very different?

Sensitivity analysis refers to a general process of trying out different prior distributions and analyzing how similar or different the resulting posterior distributions are.

- Fortunately, in many statistical models with sufficient data, the posterior distribution would usually be approximately the same, independent of what prior distribution was used.

Coin Toss Example: (10 heads in 10 tosses)

Prior distribution	Posterior distribution
Pr("coin is fair") $= 0.5$	Pr("coin is biased") ≈ 0.9990
Pr("coin is fair") $= 0.8$	Pr("coin is biased") ≈ 0.9961
Pr("coin is fair") $= 0.9$	Pr("coin is biased") ≈ 0.9913
Pr("coin is fair") $= 0.99$	Pr("coin is biased") ≈ 0.9118

Likelihood Function

Recall: When computing the posterior distribution of θ given $X_1 = x_1, \ldots, X_n = x_n$, we used the formula

$$\xi(\theta|\mathbf{x}) = \frac{\boxed{p_n(\mathbf{x}|\theta)}\xi(\theta)}{p(\mathbf{x})} \qquad \text{or} \qquad \xi(\theta|\mathbf{x}) = \frac{\boxed{f_n(\mathbf{x}|\theta)}\xi(\theta)}{p(\mathbf{x})},$$

depending on whether $X_1, \ldots, X_n$ are discrete or continuous.

In either case, if we consider the joint conditional pmf/pdf as a function only in terms of the variable θ, where $\mathbf{x} = (x_1, \ldots, x_n)$ are given fixed values, then this (univariate) function is called the likelihood function of the parameter θ.

- i.e. likelihood functions are functions of parameters of a statistical model, given specific observed values.

Interpretation: The likelihood function of the parameter θ, when substituted with the parameter value θ, is a measure of how likely θ is the actual parameter of the statistical model, given the observed data, i.e. the observed values $x_1, \ldots, x_n$.

Summary

- Statistical models
- Parameter space and parameters as random variables
- Statistical inference, statistic
- Bayesian philosophy versus frequentist philosophy
- Prior and posterior distributions
- Likelihood function

Reminders:

- There is **Mini-quiz 3** (15mins) next week during cohort class.
  - Tested on materials from Lectures 11–13 only.
  - This is Week 8. Today's lecture is Lecture 13.
- The **mid-term exam** is held this Wednesday, 2–4pm, at the MPH. Please come at least 10 minutes early!
- The **class participation assignment** is due this week during cohort class, both report and presentation (max 4 minutes!).