MATH 450: Mathematical statistics

Vu Dinh

Departments of Mathematical Sciences
University of Delaware

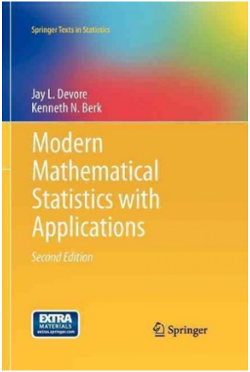
August 27th, 2019

General information

- Classes: Tuesday-Thursday: 9:30am-10:45am.
  Recitation Hall, Room 101
- Office hours: Ewing Hall 312
  - Tuesday 1:30pm-3pm
  - Wednesday 10:30am –12pm
  - By appointments
- Website:

  `http://vucdinh.github.io/m450f19`

Textbook



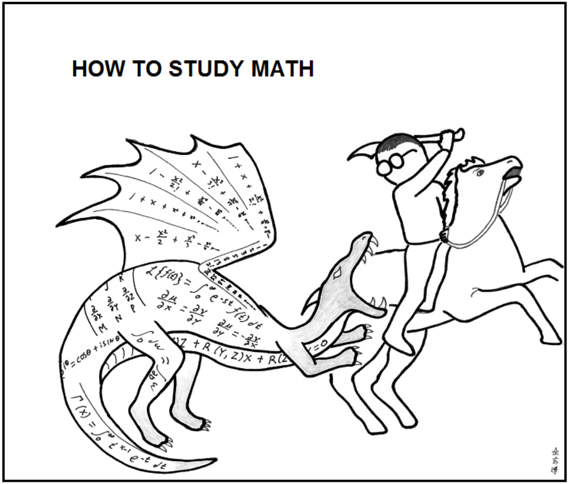
– *Modern mathematical statistics with applications* (Second Edition). Devore and Berk.

– Electronic copies of the book are available (free) at UD Library.

Evaluation

- Overall scores will be computed as follows:
  25% homework, 10% quizzes, 25% midterm, 40% final
- No letter grades will be given for homework, midterm, or final. Your letter grade for the course will be based on your overall score.
- The lowest homework scores and the lowest quiz score will be dropped.
- Letter grades you can achieve according to your overall score.
  - $\geq$ 90%: At least A
  - $\geq$ 75%: At least B
  - $\geq$ 60%: At least C
  - $\geq$ 50%: At least D

Homework



Don't just read it; fight it!

--- Paul R. Halmos

Homework

- Assignments will be posted on the website every other Tuesday (starting from the first week) and will be due on Thursday of *the following week*, *at the beginning of* lecture.
- No late homework will be accepted.
- The lowest homework scores will be dropped in the calculation of your overall homework grade.

Quizzes and exams

- At the end of some chapter, there will be a short quiz during class.
- The quiz dates will be announced at least one class in advance.
- The lowest quiz score will be dropped.

There will be a (tentative) midterm on 10/24 and a final exam during exams week.

Open source statistical system R

`http://cran.r-project.org/`

Tentative schedule

(Tentative) Class schedule:

Week	Chapter	Note
1 (Aug 27-29)	1	
2 (Sep 3-5)	6.1 and 6.2	HW1 (due 09/05)
3 (Sep 10-12)	6.2 and 6.3	
4 (Sep 17-19)	7.1	HW2 (due 09/19)
5 (Sep 24-26)	7.2	
6 (Oct 1-3)	7.3 and 7.4	HW3 (due 10/03)
7 (Oct 8-10)	8.1 and 8.2	
8 (Oct 15-17)	8.3 and 9.1	HW4 (due 10/17)
9 (Oct 22-24)	Review + Midterm exam	Midterm exam (10/24)
10 (Oct 29-31)	9.2 and 9.3	
11 (Nov 5-7)	10.1 and 10.2	HW5 (due 11/07)
12 (Nov 12-14)	10.2 and 10.3	
13 (Nov 19-21)	12	HW6 (due 11/21)
14	—————————————-	Thanksgiving week (no class)
15 (Dec 3-5)	Review	
16	—————————————	Final week

Topics

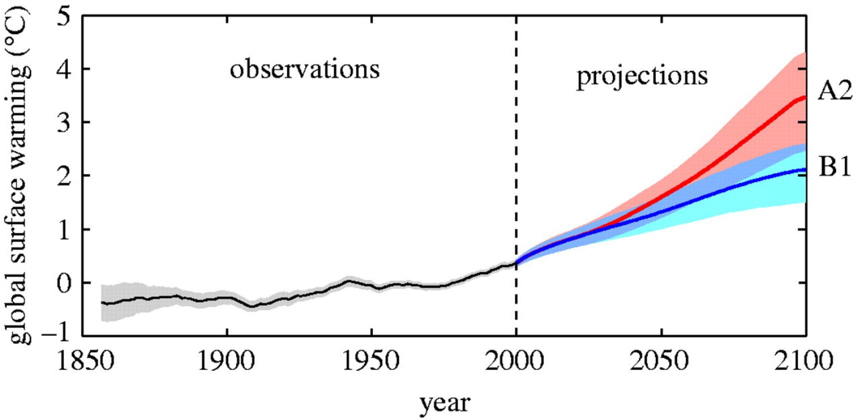
Week 2	*Chapter 6: Statistics and Sampling Distributions*
Week 4	Chapter 7: Point Estimation
Week 6	*Chapter 8: Confidence Intervals*
Week 9	*Chapter 9: Test of Hypothesis*
Week 11	Chapter 10: Two-sample inference
Week 12	Regression

Mathematical statistics

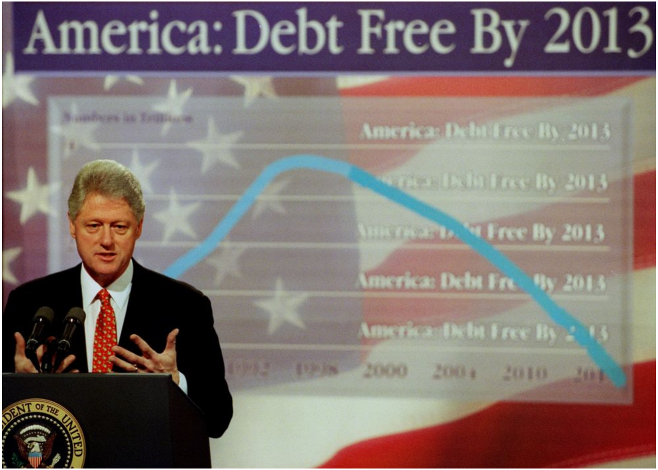
- Statistics ~~is a branch of mathematics that~~ deals with the collection, organization, analysis, interpretation and presentation of data
- "...analysis, interpretation and presentation of data"
  $\rightarrow$ mathematical statistics
  - descriptive statistics: the part of statistics that describes data
  - inferential statistics: the part of statistics that draws conclusions from data

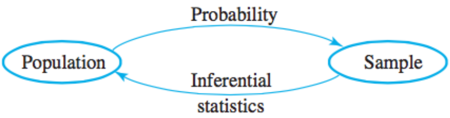
Modelling uncertainties

— Modern statistics is about making prediction in the presence of uncertainties



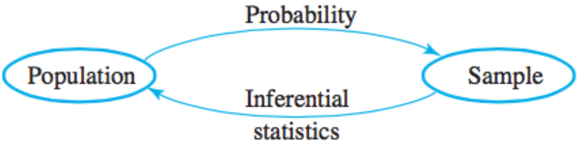
— It is difficult to make predictions, especially about the future.





- *population*: a well-defined collection of objects of interest
- when desired information is available for all objects in the population, we have what is called a *census*
  $\rightarrow$ very expensive
- a *sample*, a subset of the population, is selected

Random sample



Definition

The random variables $X_1, X_2, ..., X_n$ are said to form a (simple) random sample of size n if

1. the X_i's are independent random variables
2. every X_i has the same probability distribution

Week 1: Probability review

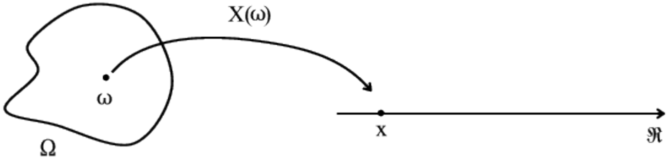
Overview

- Axioms of probability
- Conditional probability and independence
- *Random variables*
- *Special distributions*
- Bivariate and multivariate distributions

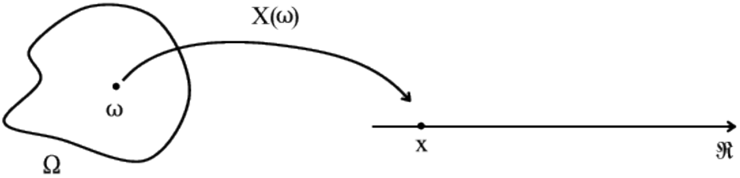
Most important parts

- Expectation and variance of random variables (discrete and continuous)
- Computations with normal distributions
- Bivariate and multivariate distributions

Random variables



- random variables are used to model uncertainties
- Notations:
  - random variables are denoted by uppercase letters (e.g., X);
  - the calculated/observed values of the random variables are denoted by lowercase letters (e.g., x)



Definition

Let S be the sample space of an experiment. A real-valued function $X : S \to \mathbb{R}$ is called a random variable of the experiment.

Discrete random variable

Definition

A random variables X is discrete if the set of all possible values of X

- is finite
- is countably infinite

Note: A set A is countably infinite if its elements can be put in one-to-one correspondence with the set of natural numbers, i.e, we can index the element of A as a sequence

$$A = \{x_1, x_2, \ldots, x_n, \ldots\}$$

A random variable X is described by its *probability mass function*

Definition *The* ***probability mass function*** *p of a random variable X whose set of possible values is $\{x_1, x_2, x_3, \dots\}$ is a function from $\mathbf{R}$ to $\mathbf{R}$ that satisfies the following properties.*

(a) $p(x) = 0$ *if* $x \notin \{x_1, x_2, x_3, \dots\}$.

(b) $p(x_i) = P(X = x_i)$ *and hence* $p(x_i) \geq 0$ $(i = 1, 2, 3, \dots)$.

(c) $\sum_{i=1}^{\infty} p(x_i) = 1$.

Represent the probability mass function

- As a table

x	1	2	3	4	5	6	7
$p(x)$	.01	.03	.13	.25	.39	.17	.02

- As a function:

$$p(x) = \begin{cases} \frac{1}{2}\left(\frac{2}{3}\right)^x & \text{if } x = 1, 2, 3, \dots, \\ 0 & \text{elsewhere} \end{cases}$$

Expectation

Definition *The **expected value** of a discrete random variable X with the set of possible values A and probability mass function p(x) is defined by*

$$E(X) = \sum_{x \in A} x p(x).$$

We say that E(X) exists if this sum converges absolutely.

The expected value of a random variable X is also called the **mean**, or the **mathematical expectation**, or simply the **expectation** of X. It is also occasionally denoted by $E[X]$, $E(X)$, EX, μ_X, or μ.

Exercise

Problem

A random variable X has the following pmf table

X	0	1	2
probability	0.25	0.5	0.25

What is the expected value of X?

Law of the unconscious statistician (LOTUS)

Theorem 4.2 *Let X be a discrete random variable with set of possible values A and probability mass function $p(x)$, and let g be a real-valued function. Then $g(X)$ is a random variable with*

$$E[g(X)] = \sum_{x \in A} g(x)p(x).$$

Exercise

Problem

A random variable X has the following pmf table

X	0	1	2
probability	0.25	0.5	0.25

- *What is $E[X^2 - X]$?*
- *Compute $Var[X]$*

$$Var[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

Continuous random variables

Overview

- Continuous random variables
- Distribution functions
- Working with the standard normal distribution $\mathcal{N}(0, 1)$
- Working with the normal distributions $\mathcal{N}(\mu, \sigma^2)$
- Linear combination of normal random variables

Reading: Sections 4.1, 4.2, 4.3

Continuous random variable

Definition

Let X be a random variable. Suppose that there exists a nonnegative real-valued function $f : \mathbb{R} \to [0, \infty)$ such that for any subset of real numbers A, we have

$$P(X \in A) = \int_A f(x)dx$$

Then X is called **absolutely continuous** or, for simplicity, **continuous**. The function f is called the **probability density function**, or simply the **density function** of X.

Whenever we say that X is continuous, we mean that it is absolutely continuous and hence satisfies the equation above.

Let X be a continuous r.v. with density function f, then

- $f(x) \geq 0$ for all $x \in \mathbb{R}$
- $\int_{-\infty}^{\infty} f(x)\ dx = 1$
- For any fixed constant a, b,

$$P(a \leq X \leq b) = \int_a^b f(x)\ dx$$

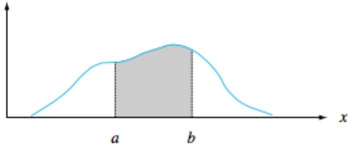


Figure 4.2 $P(a \leq X \leq b)$ = the area under the density curve between a and b

Expectation

Definition *If X is a continuous random variable with probability density function f, the* ***expected value*** *of X is defined by*

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

The expected value of X is also called the **mean**, or **mathematical expectation**, or simply the **expectation** of X, and as in the discrete case, sometimes it is denoted by EX, $E[X]$, μ, or μ_X.

Lotus

Theorem 6.3 *Let X be a continuous random variable with probability density function $f(x)$; then for any function $h: \mathbf{R} \to \mathbf{R}$,*

$$E[h(X)] = \int_{-\infty}^{\infty} h(x) f(x)\, dx.$$

Example

Problem

Let X be a continuous r.v. with density function

$$f(x) = \begin{cases} 2x & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

where c is some unknown constant.

- *Compute $P(X \in [0.25, 0.75])$*
- *Compute $E[X]$ and $Var(X)$.*

Definition

If X is a random variable, then the function F defined on $(-\infty, \infty)$ by

$$F(t) = P(X \leq t)$$

is called the distribution function of X.

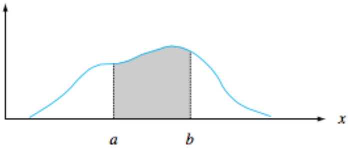


Figure 4.2 $P(a \leq X \leq b)$ = the area under the density curve between a and b

Distribution function

For continuous random variable:

$$F(t) = P(X \leq t) = \int_{(-\infty,t]} f(x)dx$$
$$= \int_{-\infty}^{t} f(x)dx$$

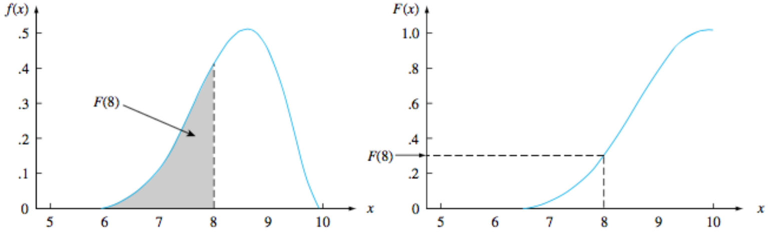


Figure 4.5 A pdf and associated cdf

Distribution function

For continuous random variable:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx = F(b) - F(a)$$

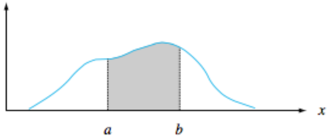


Figure 4.2 $P(a \leq X \leq b)$ = the area under the density curve between a and b

Moreover:

$$f(x) = F'(x)$$

Example

Problem

The distribution function for the duration of a certain soap opera (in tens of hours) is

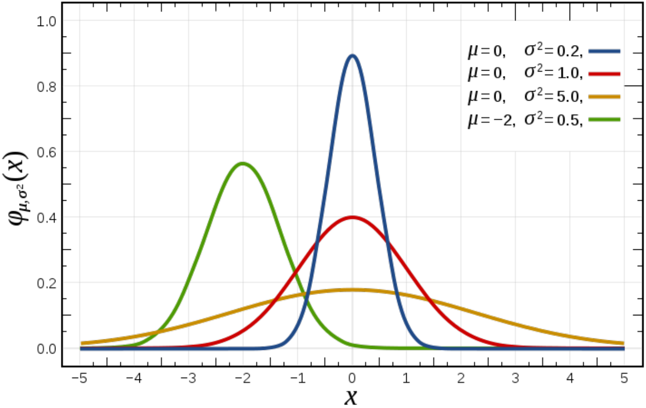
$$F(y) = \begin{cases} 1 - \frac{16}{y^2} & \textit{if } y \geq 4 \\ 0 & \textit{elsewhere} \end{cases}$$

Find $P[4 \leq Y \leq 8]$.

Normal random variables

Reading: 4.3

$\mathcal{N}(\mu, \sigma^2)$



$$E(X) = \mu, Var(X) = \sigma^2$$

$\mathcal{N}(\mu, \sigma^2)$

- $E(X) = \mu, \; Var(X) = \sigma^2$
- Density function

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

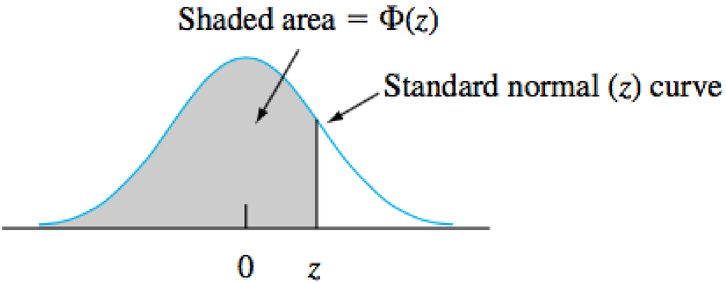
Standard normal distribution $\mathcal{N}(0, 1)$

- If Z is a normal random variable with parameters $\mu = 0$ and $\sigma = 1$, then the pdf of Z is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

  and Z is called the *standard normal distribution*
- $E(Z) = 0, \; Var(Z) = 1$

$\Phi(z)$



$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} f(y)\, dy$$

$\Phi(z)$

Table A.3 Standard Normal Curve Areas (*cont.*) $\Phi(z) = P(Z \leq z)$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9278	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

Exercise 1

Problem

Let Z be a standard normal random variable.
Compute

- $P[Z \leq 0.75]$
- $P[Z \geq 0.82]$
- $P[1 \leq Z \leq 1.96]$
- $P[Z \leq -0.82]$

Note: The density function of Z is symmetric around 0.

Exercise 2

Problem

Let Z be a standard normal random variable. Find a, b such that

$$P[Z \leq a] = 0.95$$

and

$$P[-b \leq Z \leq b] = 0.95$$

- $E(X) = \mu, \; Var(X) = \sigma^2$
- Density function

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Shifting and scaling normal random variables

Problem

Let X be a normal random variable with mean μ and standard deviation σ.
Then

$$Z = \frac{X - \mu}{\sigma}$$

follows the standard normal distribution.

Shifting and scaling normal random variables

If X has a normal distribution with mean μ and standard deviation σ, then

$$Z = \frac{X - \mu}{\sigma}$$

has a standard normal distribution. Thus

$$P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right)$$
$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

$$P(X \le a) = \Phi\left(\frac{a-\mu}{\sigma}\right) \quad P(X \ge b) = 1 - \Phi\left(\frac{b-\mu}{\sigma}\right)$$

Exercise 3

Problem

Let X be a $\mathcal{N}(3, 9)$ random variable. Compute $P[X \leq 5.25]$.