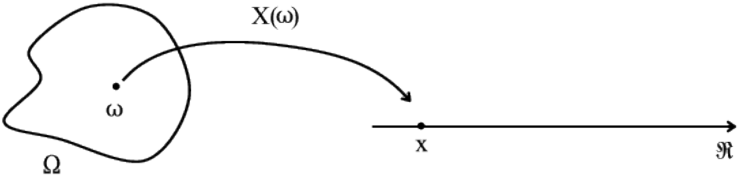
MATH 450: Mathematical statistics

August 29th, 2019

Lecture 2: Working with normal distributions

Week 1: Probability review

Random variable



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 variables

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

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

Discrete random variable

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 $\{x_1, x_2, x_3, \ldots\}$ is a function from $\mathbf{R}$ to $\mathbf{R}$ that satisfies the following properties.*

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

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

(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:

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

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.

Properties

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,

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

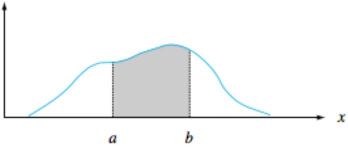


Figure 4.2 $P(a \leq X \leq b)$ = the area under the density curve between a and b

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 $h: \mathbf{R} \to \mathbf{R}$,*

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

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 $Var(X)$.*

Distribution function

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.

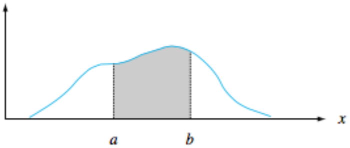


Figure 4.2 $P(a \leq X \leq b)$ = the area under the density curve between a and b

Distribution function

For continuous random variable:

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

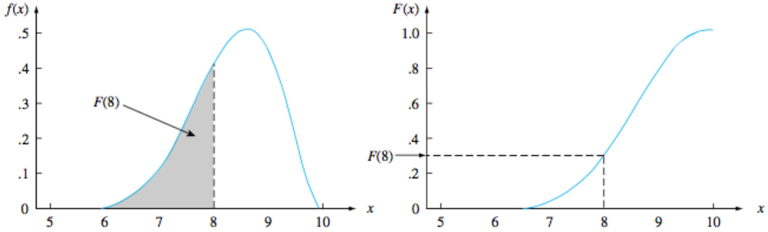


Figure 4.5 A pdf and associated cdf

Distribution function

For continuous random variable:

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

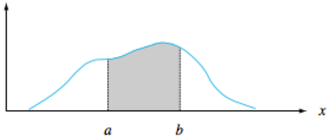


Figure 4.2 $P(a \leq X \leq b)$ = the area under the density curve between a and b

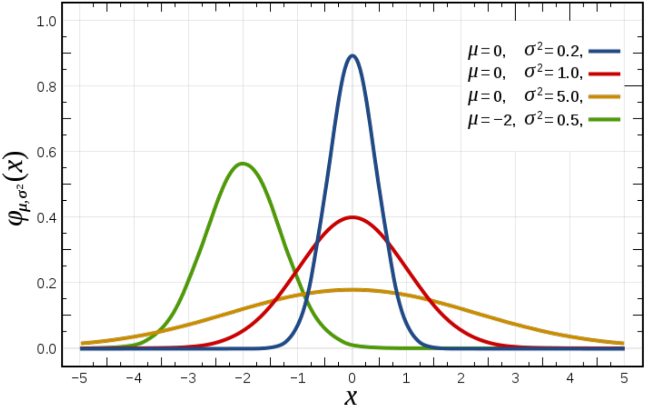
Moreover:

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

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

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

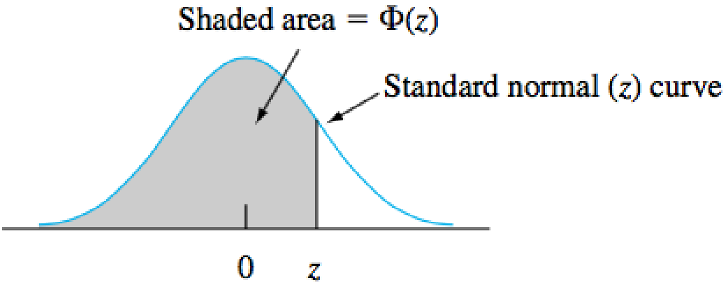
Standard normal distribution $\mathcal{N}(0, 1)$

- 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*
- $E(Z) = 0, \ Var(Z) = 1$

$\Phi(z)$



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

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

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

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

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

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

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 $P[X \leq 5.25]$.

Descriptive statistics

- The Mean
- The Median
- Trimmed Means

Measures of locations: mean

The **sample mean** $\bar{x}$ of observations $x_1, x_2, \ldots, x_n$ is given by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Measures of locations: median

Step 1: ordering the observations from smallest to largest

$$\tilde{x} = \begin{cases} \text{The single middle value if } n \text{ is odd} & = \left(\dfrac{n+1}{2}\right)^{\text{th}} \text{ordered value} \\ \text{The average of the two middle values if } n \text{ is even} & = \text{average of } \left(\dfrac{n}{2}\right)^{\text{th}} \text{ and } \left(\dfrac{n}{2}+1\right)^{\text{th}} \text{ordered values} \end{cases}$$

Median is not affected by outliers

Measures of locations: trimmed mean

- A α% trimmed mean is computed by:
  - eliminating the smallest α% and the largest α% of the sample
  - averaging what remains
- $\alpha = 0 \rightarrow$ the mean
- $\alpha \approx 50 \rightarrow$ the median

Measures of variability: deviations from the mean

The **sample variance**, denoted by s^2, is given by

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{S_{xx}}{n - 1}$$

The **sample standard deviation**, denoted by s, is the (positive) square root of the variance:

$$s = \sqrt{s^2}$$

Working with vectors in R

- manually create a vector a with entry values

$$a = c(1, 2, 6, 8, 5, 3, -1, 2.1, 0)$$

- create a zero vector with length $n = 25$

$$a = rep(0, 25)$$

- $a[i]$ is the i^{th} element of a
- manipulate all entries at the same time using 'for' loop

Boxplots

Order the n observations from smallest to largest and separate the smallest half from the largest half; the median $\tilde{x}$ is included in both halves if n is odd. Then the **lower fourth** is the median of the smallest half and the **upper fourth** is the median of the largest half. A measure of spread that is resistant to outliers is the **fourth spread** f_s, given by

$$f_s = \text{upper fourth} - \text{lower fourth}$$

Boxplots

40 52 55 60 70 75 85 85 90 90 92 94 94 95 98 100 115 125 125

The five-number summary is as follows:

smallest $x_i = 40$ lower fourth $= 72.5$ $\tilde{x} = 90$ upper fourth $= 96.5$
largest $x_i = 125$

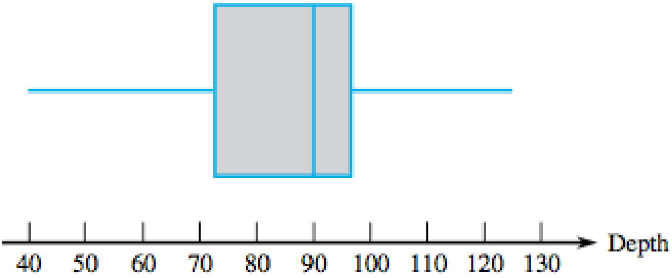
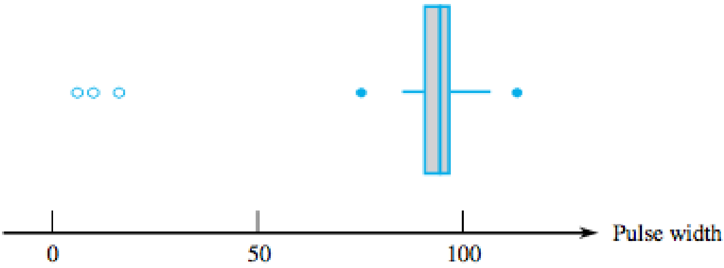


Figure 1.17 A boxplot of the corrosion data

Boxplot with outliers

Any observation farther than $1.5f_s$ from the closest fourth is an **outlier**. An outlier is **extreme** if it is more than $3f_s$ from the nearest fourth, and it is **mild** otherwise.



Comparative boxplots