MATH 450: Mathematical statistics

Vu Dinh

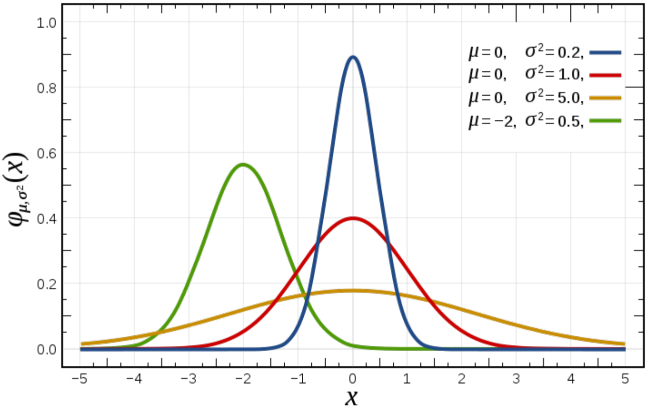
February 15th, 2019

Lecture 3: Simulations with R

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 $\mathcal{N}(0,1)$

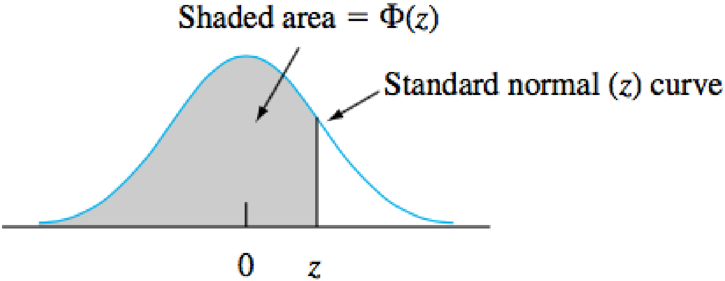
- If Z is a normal random variable with parameters $\mu = 0$ and $\sigma = 1$, then the pdf of Z is

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

  and Z is called the standard normal distribution

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

$\Phi(z)$



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

$\Phi(z)$

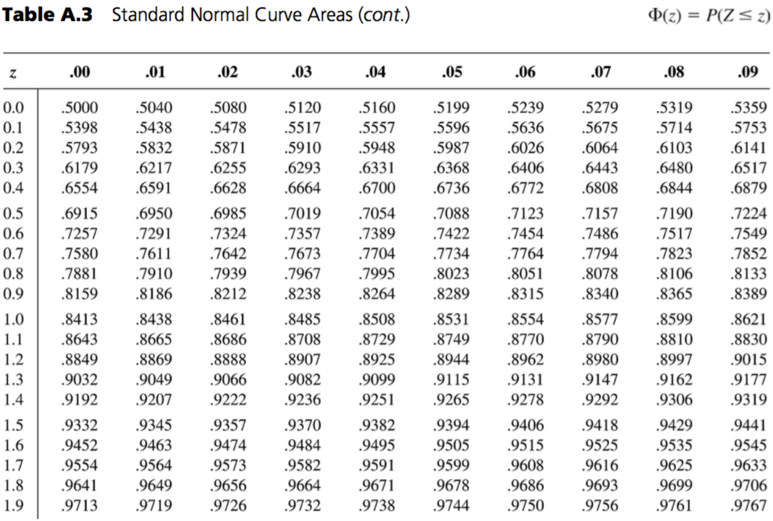


Table A.3 Standard Normal Curve Areas (*cont.*) $\Phi(z) = P(Z \leq z)$

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

Shifting and scaling normal random variables

Problem

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

Then

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

ôŔřŹ 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

$$\begin{aligned} P(a \leq X \leq b) &= P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$

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

Exercise 3

Problem

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

Linear combination of normal random variables

Theorem

Let $X_1, X_2, \ldots, X_n$ be independent normal random variables (with possibly different means and/or variances). Then

$$T = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$$

ôŔřŹ also follows the normal distribution.

Linear combination of normal random variables

Theorem

Let $X_1, X_2, \ldots, X_n$ be independent normal random variables (with possibly different means and/or variances). Then

$$T = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$$

ôŔřŹ also follows the normal distribution.

What are the mean and the standard deviation of T?

- $E(T) = a_1 E(X_1) + a_2 E(X_2) + \ldots + a_n E(X_n)$
- $\sigma_T^2 = a_1^2 \sigma_{X_1}^2 + a_2^2 \sigma_{X_2}^2 + \ldots + a_n^2 \sigma_{X_n}^2$

Example 1

Problem

Assume that

$$X_1 \sim \mathcal{N}(10, 9) \qquad \textit{and} \qquad X_2 \sim \mathcal{N}(30, 16)$$

are independent.

What is the distribution of $X_1 - X_2$?

Example 2

Problem

A concert has three pieces of music to be played before intermission. The time taken to play each piece has a normal distribution.
Assume that the three times are independent of each other. The mean times are 15, 30, and 20 min, respectively, and the standard deviations are 1, 2, and 1.5 min, respectively.

What is the distribution of the length of the concert?

Working with R

Reading: 4.3

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

Pictorial methods

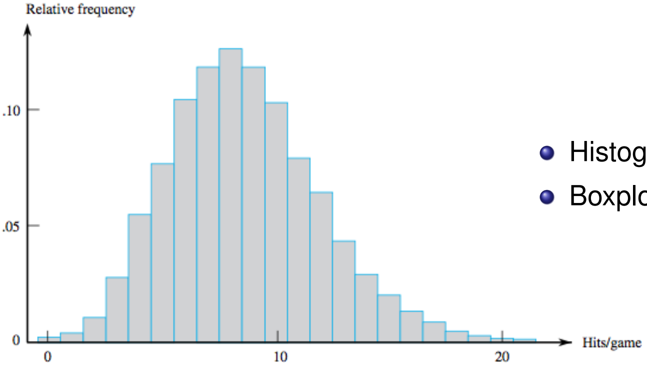


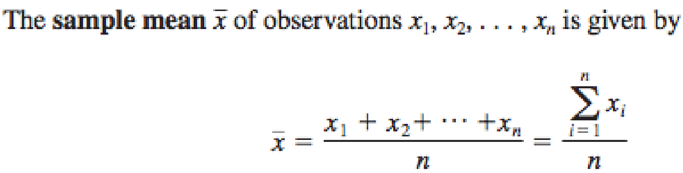
Figure 1.6 Histogram of number of hits per nine-inning game

- Histograms
- Boxplots

1.3: Measures of locations

- 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

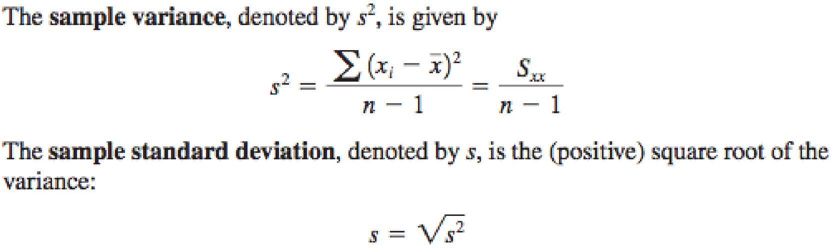
$$\tilde{x} = \begin{cases} \text{The single middle value if } n \text{ is odd} & = \left(\dfrac{n+1}{2}\right)^{\text{th}} \text{ordered value} \\ \text{The average of the two middle values if } n \text{ is even} & = \text{average of } \left(\dfrac{n}{2}\right)^{\text{th}} \quad \text{and} \quad \left(\dfrac{n}{2}+1\right)^{\text{th}}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

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
- $\alpha = 0$ → the mean
- $\alpha \approx 50$ → the median

Measures of variability: deviations from the mean



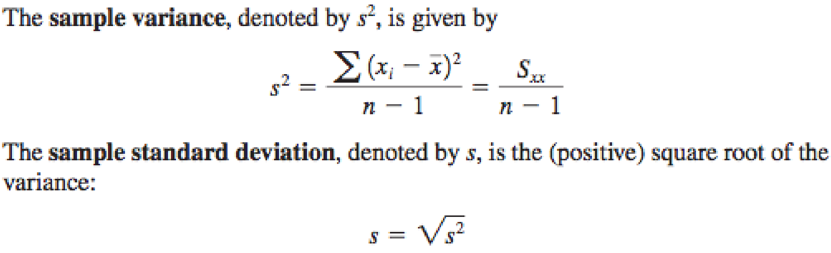
The **sample variance**, denoted by s^2, is given by

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

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

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

Measures of variability: deviations from the mean



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

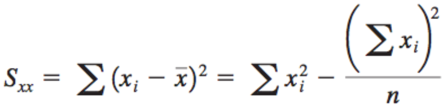
$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{S_{xx}}{n - 1}$$

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

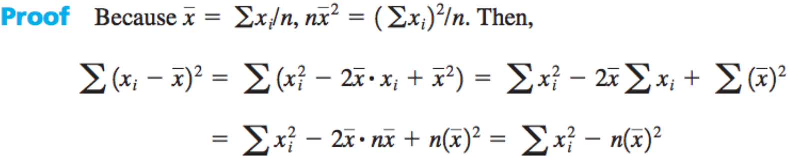
$$s = \sqrt{s^2}$$

- Why squared? Because it is easier to do math with x^2 than $|x|$
- Why $(n-1)$? Because that makes s^2 an unbiased estimator of the population variance (Chapter 7)

Computing formula for s^2



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

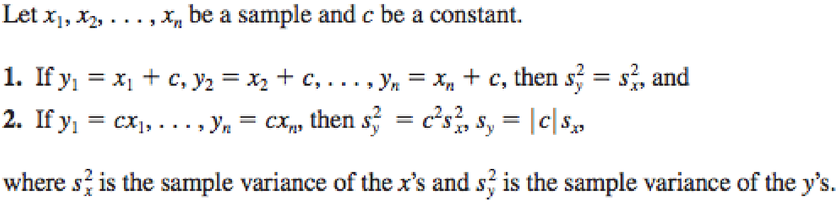


Proof Because $\bar{x} = \sum x_i/n$, $n\bar{x}^2 = (\sum x_i)^2/n$. Then,

$$\sum (x_i - \bar{x})^2 = \sum (x_i^2 - 2\bar{x} \cdot x_i + \bar{x}^2) = \sum x_i^2 - 2\bar{x} \sum x_i + \sum (\bar{x})^2$$

$$= \sum x_i^2 - 2\bar{x} \cdot n\bar{x} + n(\bar{x})^2 = \sum x_i^2 - n(\bar{x})^2$$

Properties of the sample standard deviation

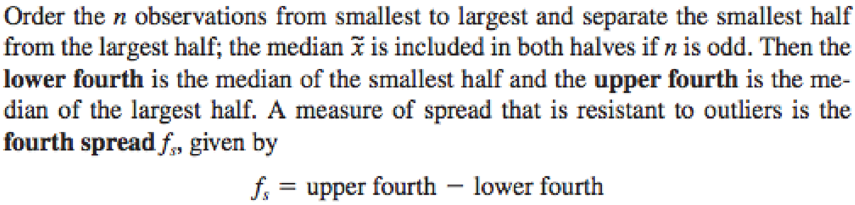


Let $x_1, x_2, \ldots, x_n$ be a sample and c be a constant.

1. If $y_1 = x_1 + c$, $y_2 = x_2 + c, \ldots, y_n = x_n + c$, then $s_y^2 = s_x^2$, and
2. If $y_1 = cx_1, \ldots, y_n = cx_n$, then $s_y^2 = c^2 s_x^2$, $s_y = |c| s_x$,

where s_x^2 is the sample variance of the x's and s_y^2 is the sample variance of the y's.

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

$$\underbrace{40\ \ 52\ \ 55\ \ 60\ \ 70\ \ 75\ \ 85\ \ 85\ \ 90\ \ \rlap{$\overbrace{}$}90}\ \ 92\ \ 94\ \ 94\ \ 95\ \ 98\ \ 100\ \ 115\ \ 125\ \ 125$$

The five-number summary is as follows:

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

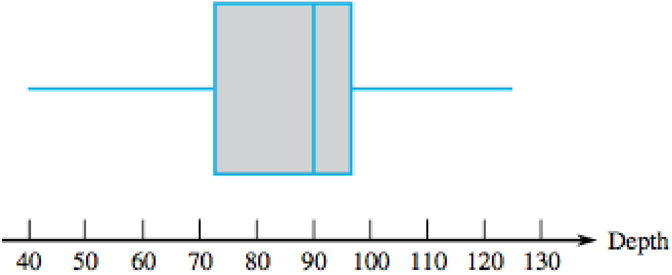
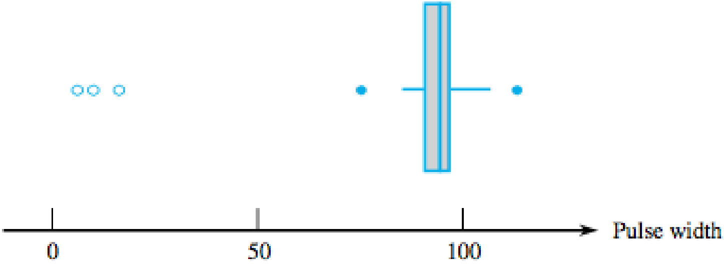


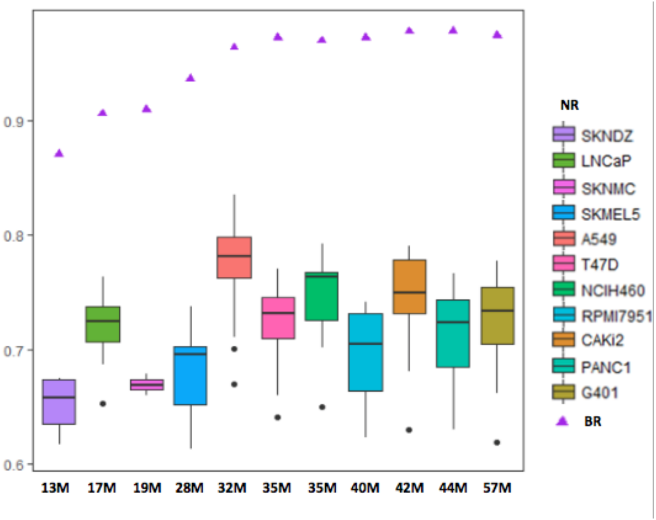
Figure 1.17 A boxplot of the corrosion data

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



Random sample

- Experiment: throw a fair die 2 times
- Before the experiment, denote the random variables that describes the outcome of the first throw and the second throws by X_1 and X_2, respectively
- X_1 and X_2 have the same probability mass function

$$p(x) = 1/6, \quad x = 1, 2, 3, 4, 5, 6$$

- Do the experiment, obtain outcomes x_1 and x_2
- x_1 may be different from x_2

Random sample

Definition

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

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

Each of the X_i's is called an instance, or a realization of the distribution.

Sometimes, people refer to X_i as copies of the same random variable X

Random sample

Given a probability distribution, one want to create a sample of it

1. in real life: statistical sampling
2. using computer: simulation

Simulate uniform distribution

- the uniform distribution on (a,b) has density

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in (a,b) \\ 0 & \text{elsewhere} \end{cases}$$

- To generate the uniform distribution on $(0,1)$, use the function *runif*

$$b = runif(200)$$

Simulate a biased coin

- Assume that we want to simulate a Bernoulli random variable

$$p(x) = \begin{cases} 0.6 & \text{if } x = 0 \\ 0.4 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

- Step 1: generate u from the uniform distribution on $(0, 1)$
- Step 2: If $u < 0.4$, then set $x = 1$; otherwise set $x = 0$

Simulate discrete random variables

- Question: How to simulate samples from the following distribution

$$p(x) = \begin{cases} 0.2 & \text{if } x = 3 \\ 0.3 & \text{if } x = 5 \\ 0.5 & \text{if } x = 7 \\ 0 & \text{otherwise} \end{cases}$$

Simulate continuous random variables

- Question: How to simulate samples from the following distribution

$$f(x) = \begin{cases} 2e^{-2x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

- The distribution function of X is

$$F(x) = \begin{cases} 1 - e^{-2x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

- Step 1: generate u from the uniform distribution on $(0, 1)$
- Step 2: Solve equation $F(x) = u$
- Set x as the solution

$$x = -\frac{1}{2}\ln(1 - u)$$

Why?

Theorem

Let X be a continuous random variable with probability distribution function F. Then $F(X)$ is a uniform random variable over $(0,1)$.

Proof.

Let $Y = F(X)$, then $Y \in [0,1]$ and for all $y \in (0,1)$

$$F_Y(y) = P[Y \leq y] = P[F(X) \leq y] = P[X \leq F^{-1}(y)] = F(F^{-1}(y)) = y$$

thus

$$f_Y(y) = \begin{cases} 1 & \text{if } y \in (0,1) \\ 0 & \text{otherwise} \end{cases}$$

□