STA 313: Topics in Statistics

Al Nosedal.
University of Toronto.

Fall 2015

"essentially, all models are wrong, but some are useful"

George E. P. Box

(one of the great statistical minds of the 20th century).

Basic Concepts of Probability

A random event is an event which has more than one possible outcome. A probability may be associated with each outcome. The outcome of a random event is not predictable, only the possible outcomes and their probabilities are known.

Probability

- Probability is the numerical value of the chance of occurrence of one or more possible results of an unpredictable event.
- The set of all possible exclusive elementary events E_i is denoted by Ω. The probability of occurrence of E_i, $P(E_i)$, should have the following properties:
  1. $P(E_i) > 0$ for all i.
  2. $P(E_i \cup E_j) = P(E_i) + P(E_j)$, E_i and E_j are mutually exclusive.
  3. $P(\Omega) = 1$.

Frequency

Consider the throwing of a die. Suppose that having thrown the die N times, where we obtained "5" k times as a result of the throw. The probability of getting "5" as the result of a throw can be taken to be

$$p = \frac{k}{N},$$

i. e. the number k of "fortunate" events divided by the whole number of events N. This ratio is called relative frequency.

Law of Addition of Probabilities

The events are called exclusive when the occurrence of one of them implies that none of the others occurs. There are two exclusive events A and B. Consider the event C which we suppose to happen if one of the two events A or B happens. The probability of the event C is given by

$$P(C) = P(A) + P(B).$$

The probabilities are given approximately by the relative frequencies: $P(A) \approx \frac{k_A}{n}$ and $P(B) \approx \frac{k_B}{n}$.
The event C happens $k_C = k_A + k_B$ times and the whole number of the events is n.

$$P(C) \approx \frac{k_C}{n} = \frac{k_A + k_B}{n} = \frac{k_A}{n} + \frac{k_B}{n} \approx P(A) + P(B).$$

Law of Multiplication of Probabilities

Suppose that having thrown a die two times. The event A happens, if we get odd number as the result of the first throw. The event B happens, if we get "1" as the result of the second throw. The event C happens if the result of the first throw is A and the result of the second throw is B. The probability of the event C is given by

$$P(C) = P(A)P(B).$$

The probabilities are given approximately by the relative frequencies: $P(A) \approx \frac{k_A}{n}$ and $P(B) \approx \frac{k_B}{n}$.
The event C happens $k_C = k_A k_B$ times and the whole number of the events is nn.

$$P(C) \approx \frac{k_C}{nn} = \frac{k_A k_B}{nn} = \frac{k_A}{n}\frac{k_B}{n} \approx P(A)P(B).$$

Conditional Probability

There are two random events A and B. The probability that B happens when A happens is given by the conditional probability of B given A written $P(B|A)$:

$$P(B|A) \approx \frac{k_{A\cap B}}{k_A} = \frac{k_{A\cap B}/n}{k_A/n} \approx \frac{P(A \cap B)}{P(A)}$$

where the probabilities are approximated with the relative frequencies.
The events A and B are said to be independent if

$$P(B|A) = P(B)$$

which means that the occurrence of A is irrelevant to the occurrence of B and vice versa.

Random Variables

With a random event A one way associate a random variable X, which takes different possible numerical values $x_1, x_2, ...$ corresponding to different outcomes. The corresponding probabilities $P(x_1), P(x_2), ...$ form a probability distribution of the random variable.

Discrete Random Variable

The random variable X and its probability distribution is called discrete, when it can take its value from a finite or infinite set of discrete values $x_1, x_2, ...$ The distribution of the random variable is given by the table:

x_1	x_2	. . .	x_n	. . .
p_1	p_2	. . .	p_n	. . .

where $x_1, x_2, ..., x_n, ...$ are the possible values of the random variable, while $p_1, p_2, ..., p_n, ...$ are their probabilities. The probability that X takes the value x_i is given by $P(X = x_i) = P(x_i) = p_i$.

Discrete Random Variable

The probabilities $p_1, p_2, ..., p_n, ...$ should satisfy the following conditions:

- All should be positive: $p_i > 0$.
- Their sum should be equal to 1: $\sum_i p_i = 1$.

Expected Value

The mathematical expectation value $E(X)$ of the random variable X is given by

$$E(X) = \sum_i x_i p_i.$$

Variance

The variance $V(X)$ of the random variable X is given by

$$V(X) = E[(X - E(X))^2] = E(X^2) - [E(X)]^2.$$

Continuous Random Variable

The random variable X is called continuous if it may have any value from the (a, b) interval. The probability that one observation is from the (c, d) subinterval of the (a, b) interval is given by

$$P(c \leq X \leq d) = \int_c^d f(x)dx,$$

where $f(x)$ the probability density function.

Continuous Random Variable

The probability density function $f(x)$ should satisfy the following conditions:

- It should be positive: $f(x) > 0$.
- Its integral over the whole interval should be 1: $\int_a^b f(x)dx = 1$.

Cumulative distribution function

The cumulative distribution function (cdf) is given by

$$F(x) = \int_a^x f(t)dt.$$

It has the following properties:

- $F(a) = 0$,
- $F(b) = 1$, and
- $0 < F(x) < 1$, if $a < x < b$.

Cumulative distribution function

The function $F(x)$ is a monotone increasing function:

$$F(x_1) < F(x_2), \quad \text{if} \quad x_1 < x_2.$$

The probability that the random variable X is in the (c, d) interval is given by

$$P(c \leq X \leq d) = F(d) - F(c).$$

Expected Value

The mathematical expectation value $E(X)$ of the random variable X is given by

$$E(X) = \int_a^b xf(x)dx.$$

Variance

The variance $V(X)$ of the random variable X is given by

$$V(X) = \int_a^b [x - E(X)]^2 f(x) dx = E(X^2) - [E(X)]^2.$$

Binomial Distribution

A random variable Y is said to have a **binomial distribution** based on n trials with success probability p if and only if

$$p(y) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}, \ y = 0, 1, 2, ..., n \text{ and } 0 \leq p \leq 1.$$

Mean and Variance

Let Y be a binomial random variable based on n trials and success probability p. Then

$$\mu = E(Y) = np \text{ and } \sigma^2 = V(Y) = npq.$$

Geometric Distribution

A random variable Y is said to have a **geometric probability** distribution if and only if

$$p(y) = q^{y-1}p, \ y = 1, 2, 3, ..., \ 0 \leq p \leq 1.$$

Mean and Variance

If Y is a random variable with a geometric distribution,

$$\mu = E(Y) = \frac{1}{p} \qquad \sigma^2 = V(Y) = \frac{1-p}{p^2}.$$

Poisson Distribution

A random variable Y is said to have a **Poisson probability distribution** if and only if

$$p(y) = \frac{\lambda^y}{y!} e^{-\lambda}, \ y = 0, \ 1, \ 2, ..., \ \lambda > 0.$$

Mean and Variance

If Y is a random variable possessing a Poisson distribution with parameter λ, then

$$\mu = E(Y) = \lambda \text{ and } \sigma^2 = V(Y) = \lambda.$$

Uniform Distribution

If $\theta_1 < \theta_2$, a random variable Y is said to have a continuous **uniform probability distribution** on the interval $(\theta_1,\ \theta_2)$ if and only if the density function of Y is

$$f(y) = \begin{cases} \frac{1}{\theta_2 - \theta_1}, & \theta_1 \leq y \leq \theta_2, \\ 0, & \text{elsewhere.} \end{cases}$$

Mean and Variance

If $\theta_1 < \theta_2$ and Y is a random variable uniformly distributed on the interval $(\theta_1,\ \theta_2)$, then

$$\mu = E(Y) = \frac{\theta_1 + \theta_2}{2} \text{ and } \sigma^2 = V(Y) = \frac{(\theta_2 - \theta_1)^2}{12}.$$

Exponential Distribution

A random variable Y is said to have an **exponential distribution** with parameter $\beta > 0$ if and only if the density function of Y is

$$f(y) = \begin{cases} \frac{1}{\beta} e^{-y/\beta}, & 0 \leq y < \infty, \\ 0, & \text{elsewhere.} \end{cases}$$

Mean and Variance

If Y is an exponential random variable with parameter β, then

$$\mu = E(Y) = \beta \quad \text{and} \quad \sigma^2 = V(Y) = \beta^2.$$

Gamma Distribution

A random variable Y is said to have a **Gamma distribution** with parameters $\alpha > 0$ and $\beta > 0$ if and only if the density function of Y is

$$f(y) = \begin{cases} \frac{y^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)} e^{-y/\beta}, & 0 \leq y < \infty, \\ 0, & \text{elsewhere.} \end{cases}$$

where $\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy$.

Gamma Function

The quantity $\Gamma(\alpha)$ is known as the **gamma function**. Direct integration will verify that $\Gamma(1) = 1$. Integration by parts will verify that $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ for any $\alpha > 1$ and that $\Gamma(n) = (n - 1)!$, provided that n is an integer.

Mean and Variance

If Y has a Gamma distribution with parameters α and β, then

$$\mu = E(Y) = \alpha\beta \quad \text{and} \quad \sigma^2 = V(Y) = \alpha\beta^2.$$

Normal Distribution

A random variable Y is said to have a **normal probability distribution** if and only if, for $\sigma > 0$ and $-\infty < \mu < \infty$, the density function of Y is

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right], \quad -\infty < y < \infty.$$

Mean and Variance

A Normal Distribution is described by a Normal density curve. Any particular Normal distribution is completely specified by two numbers, its mean μ and standard deviation σ.
The mean of a Normal distribution is at the center of the symmetric Normal curve. The standard deviation is the distance from the center to the change-of-curvature points on either side.

Homework

This course uses R. R is an open-source computing package which has seen a huge growth in popularity in the last few years. R can be downloaded from https://cran.r-project.org

Please, download R and bring your laptop next time.

Homework Problem 1

In order to verify the accuracy of their financial accounts, companies use auditors on a regular basis to verify accounting entries. The company's employees make erroneous entries 5% of the time. Suppose that an auditor randomly checks three entries.

a. Find the probability distribution for Y, the number of errors detected by the auditor.

b. Construct a graph for $p(y)$.

c. Find the probability that the auditor will detect more than one error.

Homework Problem 2

The probability distribution for a random variable Y is given by

y	p(y)
0	1/8
1	1/4
2	3/8
3	1/4

Find the mean, variance, and standard deviation of Y.

Homework Problem 3

Suppose that Y is a discrete random variable with mean μ and variance σ^2 and let $W = 2Y$.

i) Find $E(W)$.

ii) Find $V(W)$.

Homework Problem 4

A fire-detection device utilizes three temperature-sensitive cells acting independently of each other in such a manner that any one or more may activate the alarm. Each cell possesses a probability of $p = 0.8$ of activating the alarm when the temperature reaches 100 degrees Celsius or more. Let Y equal the number of cells activating the alarm when the temperature reaches 100 degrees.

a. Find the probability distribution for Y.

b. Find the probability that the alarm will function when the temperature reaches 100 degrees.

Homework Problem 5

How many times would you expect to toss a balanced coin in order to obtain the first head?

Homework Problem 6

Suppose that the probability of engine malfunction during any one-hour period is $p = 0.2$. Find the probability that a given engine will survive two hours.
(Suggestion: Try to solve this problem by "hand" and using R.)

Homework Problem 7

Industrial accidents occur according to a Poisson process with an average of three accidents per month. During the last month, six accidents occurred. Does this number seem highly improbable if the mean number of accidents per month, μ, is still equal to 3? Does it indicate an increase in the mean number of accidents per month?
(Suggestion: Try to solve this problem by "hand" and using R.)

Homework Problem 8

Suppose that Y has a Binomial distribution with $n = 2$ and $p = 1/2$. Find $F(y)$, i.e. the cumulative distribution function of Y. (Suggestion: Try to solve this problem by "hand" and using R.)

Homework Problem 9

Let X be a random variable with $p(x)$ given in the table below.

x	1	2	3	4
p(x)	0.4	0.3	0.2	0.1

a) Find an expression for the function $F(x) = P(X \leq x)$.
b) Sketch the function given in part a).

Homework Problem 10

Suppose that Y possesses the density function

$$f(y) = \begin{cases} cy, & 0 \leq y \leq 2, \\ 0 & \text{elsewhere.} \end{cases}$$

a. Find the value of c that makes $f(y)$ a probability density function.
b. Find $F(y)$.
c. Graph $f(y)$ and $F(y)$.
d. Use $F(y)$ to find $P(1 \leq Y \leq 2)$.
e. Use $f(y)$ and geometry to find $P(1 \leq Y \leq 2)$.

Homework Problem 11

The proportion of time per day that all checkout counters in a supermarket are busy is a random variable Y with density function

$$f(y) = \begin{cases} cy^2(1-y)^4, & 0 \leq y \leq 1, \\ 0, & \text{elsewhere.} \end{cases}$$

a. Find the value of c that makes $f(y)$ a probability density function.
b. Find $E(Y)$.

Homework Problem 12

The change in depth of a river from one day to the next, measured (in feet) at a specific location, is a random variable Y with the following density function:

$$f(y) = \begin{cases} k, & -2 \leq y \leq 2, \\ 0, & \text{elsewhere.} \end{cases}$$

a. Find the value of k that makes $f(y)$ a probability density function.
b. Obtain the cumulative distribution function for Y.

Homework Problem 13

If Y has an exponential distribution and $P(Y > 2) = 0.0821$, what is

a. $E(Y)$.

b. $P(Y \leq 1.7)$?