Bayesian Data Analysis - Assignment 1

September 17, 2017

1 Basic probability theory and terms

a)

probability is the measure of the likelihood of a given event's occurrence, which is expressed as a number between 0 and 1.

probability mass is a function that gives the probability that a discrete random variable is exactly equal to some value. ($f_X(x) = P(X = x) = P(\{s \in S : X(s) = x\})$)

probability density is a function of a continuous variable whose integral over a region gives the probability that a random variable falls within the region. ($P(a \leqslant X \leqslant b) = \int_a^b f_X(x)\,dx$)

probability mass function (pmf) is a function that gives the probability that a discrete random variable is exactly equal to some value.

probability density function (pdf) is a function of a continuous variable whose integral over a region gives the probability that a random variable falls within the region.

probability distribution is a function that provides the possibilities of occurrence of all the different possible values (events).

discrete probability distribution is a table (or a formula) listing all possible values that a discrete variable can take on, together with the associated probabilities.

continuous probability distribution describes the probabilities of the possible values of a continuous random variable.

cumulative distribution function (cdf) is a function that gives probability that random variable is less than or equal to a value. ($F_X(x) = P(X \leqslant x)$)

b)

sampling distribution is the probability distribution of sample statistics based on randomly selected samples from the same population.

observation model is a mathematical model (probability distribution) that relates the parameters of the model to the observations.

statistical model is a class of mathematical model (probability distribution) on sample space, which embodies a set of assumptions concerning the generation of some sample data, and similar data from a larger population.

likelihood is a function of the parameters of a statistical model given data, which is equal to the probability (density) assumed for those observed outcomes given those parameter values. ($L(\theta \mid x) = P(x \mid \theta)$)

2 Basic computer skills

The language used is Python. The source code is attached in the appendix.

a)

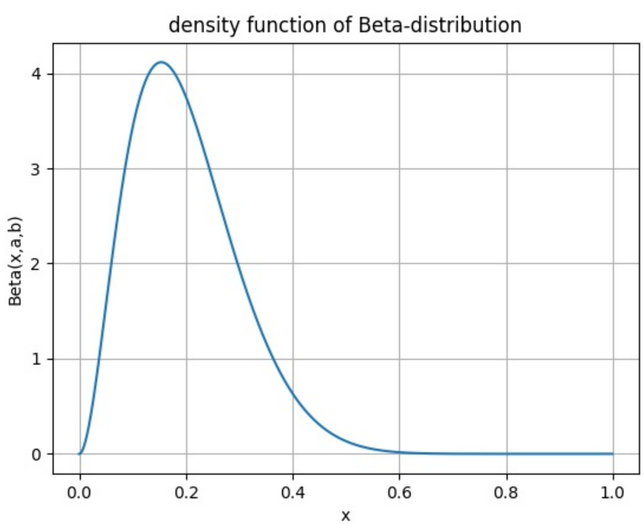

Figure 1: density function

b)

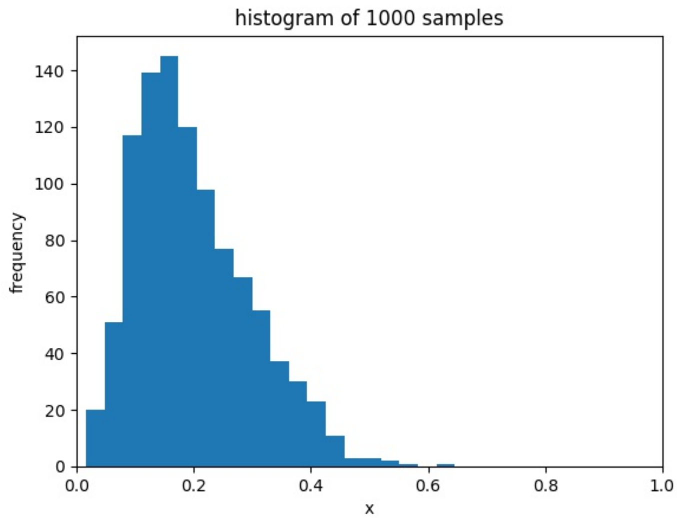

Figure 2: histogram of the samples

We can see that its curve trend is very similar to that of the density function.

c)

	mean	variance
sample	0.199498	0.009546
true	0.200000	0.010000

From the table, we can see that the sample mean and variance match (roughly) to the true mean and variance of the distribution.

d)

The central 95 %-intercal: [0.05141249 0.41952854]

3 Bayes' theorem

Assume that A is the event that having lung cancer, B is the event that test gives a positive result.

$$
\begin{aligned}
P(A) &= 0.001 \\
P(B \mid A) &= 0.98 \\
P(B \mid \bar{A}) = \bar{P}(\bar{B} \mid \bar{A}) &= 0.04
\end{aligned}
$$

$$
\begin{aligned}
P(B) &= P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A}) \\
&= 0.04094
\end{aligned}
$$

If the test gives a positive result, the probability of having lung cancer:

$$
\begin{aligned}
P(A \mid B) &= \frac{P(B \mid A)P(A)}{P(B)} \\
&= \frac{49}{2047} \approx 2.4\%
\end{aligned}
$$

There is only a possibility of 2.4% having lung cancer when the test gives a positive result. So I think it is not a good idea to introduce the test to market. (The claimed 97% successful rate maybe comes from a unreasonable ratio of healthy testees to testees having lung cancer. A relatively high proportion of testees having lung cancer will cause higher succeessful rate. However, we should take into account that lung cancer is rare in general population)

4 Bayes' theorem

The probabilities of selecting box A, B, C:

$$
\begin{aligned}
P(A) &= \frac{2}{5} \\
P(B) &= \frac{1}{10} \\
P(C) &= \frac{1}{2}
\end{aligned}
$$

The probabilities of picking up a red ball from box A, B, C separately:

$$
\begin{aligned}
P(red \mid A) &= \frac{2}{7} \\
P(red \mid B) &= \frac{4}{5} \\
P(red \mid C) &= \frac{1}{4}
\end{aligned}
$$

The probabilities of picking up a red ball:

$$\begin{aligned} P(red) &= P(A)P(red \mid A) + P(B)P(red \mid B) + P(C)P(red \mid C) \\ &= \frac{2}{5} \times \frac{2}{7} + \frac{1}{10} \times \frac{4}{5} + \frac{1}{2} \times \frac{1}{4} \\ &= \frac{447}{1400} \approx 31.9\% \end{aligned}$$

If a red ball is picked up, the probabilities of picking up from box A, B, C separately:

$$\begin{aligned} P(A \mid red) &= \frac{P(red \mid A)P(A)}{P(red)} \\ &= \frac{160}{447} \\ P(B \mid red) &= \frac{P(red \mid B)P(B)}{P(red)} \\ &= \frac{112}{447} \\ P(C \mid red) &= \frac{P(red \mid C)P(C)}{P(red)} \\ &= \frac{175}{447} \end{aligned}$$

So it mostly came from box C.

5 Bayes' theorem

$$\begin{aligned} P(fraternal \quad twins) &= \frac{1}{125} \\ P(identical \quad twins) &= \frac{1}{300} \\ P(males) = P(females) &= \frac{1}{2} \\ P(male \quad twins \mid fraternal \quad twins) &= \frac{1}{4} \\ P(male \quad twins \mid identical \quad twins) &= \frac{1}{2} \end{aligned}$$

The probability of birth of male twins:

$$\begin{aligned} P(male \quad twins) &= P(fraternal \quad twins)P(male \quad twins \mid fraternal \quad twins) \\ &\quad + P(identical \quad twins)P(male \quad twins \mid identical \quad twins) \\ &= \frac{11}{3000} \end{aligned}$$

The probability that the male twins are identical twins:

$$
\begin{aligned}
&P(identical \quad twins \mid male \quad twins) \\
=&\frac{P(male \quad twins \mid identical \quad twins)P(identical \quad twins)}{P(male \quad twins)} \\
=&\frac{5}{11} \approx 45.5\%
\end{aligned}
$$

Appendix

Source code

```
import numpy as np
from scipy.stats import beta
import matplotlib.pyplot as plt

# a)
mean = 0.2
var = 0.01
a = mean*((mean*(1-mean)/var)-1)
b = a*(1-mean)/mean

x = np.arange(0.0, 1.0, 0.001)
y = beta.pdf(x, a, b)

plt.figure()
plt.plot(x, y)
plt.xlabel('x')
plt.ylabel('Beta(x,a,b)')
plt.title("density function of Beta-distribution")
plt.grid(True)


# b)
samples = beta.rvs(a, b, size=1000)
plt.figure()
plt.hist(samples, 20)
plt.xlim([0, 1])
plt.xlabel('x')
plt.ylabel('frequency')
plt.title("histogram of 1000 samples")


# c)
sample_mean = np.mean(samples)
sample_var = np.var(samples, ddof=1)
beta_mean, beta_var = beta.stats(a, b, moments="mv")
# beta_mean = a/(a+b)
# beta_var = (a*b)/(((a+b)**2)*(a+b+1))
print("sample mean = %.6f, "
"sample variance = %.6f"
% (sample_mean, sample_var))
```

```
print("true mean = %.6f, "
"true variance = %.6f"
% (beta_mean, beta_var))

# d)
cp95_interval = np.percentile(samples, [2.5, 97.5])
print("The central 95%-intercal:"
" {}".format(cp95_interval))

plt.show()
```